TopSuite: A meta-suite for protein structure prediction using deep neural networks

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Daniel Mulnaes
aus Gladsaxe

Düsseldorf, den 22.10.2020

Aus dem Institut für Pharmazeutische und Medizinische Chemie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Holger Gohlke

2. Prof. Dr. Gunnar Schröder

Tag der mündlichen Prüfung:

EIDESSTATTLICHE ERKLÄRUNG

Ich, Daniel Mulnaes, versichere an Eides Statt, dass die Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Diese Dissertation wurde in der vorgelegten oder einer ähnlichen Form noch bei keiner anderen Institution eingereicht, und es wurden bisher keine erfolglosen Promotionsversuche von mir unternommen.

Düsseldorf, im Oktober 2020

To my Family and TzuJung

"Chemists seek precise answers to well-defined problems, whereas biologists are content with approximate answers to complex problems."

- Arthur Kornberg

TABLE OF CONTENT

1. LIST OF PUBLICATIONS

This thesis is based on the following publications (Contributions in parenthesis):

I. **Daniel Mulnaes (85%)**, and Holger Gohlke
TopScore: Using deep neural networks and large diverse datasets for accurate protein model quality assessment
Journal of Chemical Theory and Computation, 2018, 14, 6117-6126.
Impact factor reported for 2019: 5.4

II. **Daniel Mulnaes (60%),** Nicola Porta, Rebecca Clemens, Irina Apanasenko, Jens Reiners, Lothar Gremer, Philipp Neudecker, Sander Smits, Holger Gohlke.
TopModel: A deep neural network and model quality driven meta-approach to template-based protein structure prediction
Journal of Chemical Theory and Computation, 2019, Submitted.
Impact factor reported for 2019: 5.4

III. Holger Gohlke, Ulrike Hergert, Tatu Meyer, **Daniel Mulnaes (10%),** Manfred K. Grieshaber, Sander H.J. Smits and Lutz Schmitt.
Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion
Journal of Chemical Information and Modelling, 2013, 53, 2493–2498.
Impact factor reported for 2018: 3.8

IV. Zeli Zhang, Qinyong Gu, Ananda Ayyappan Jaguva Vasudevan, Anika Hain, Björn-Philipp Kloke, Sascha Hasheminasab, **Daniel Mulnaes (5%),** Kei Sato, Klaus Cichutek, Dieter Häussinger, Ignatio G. Bravo, Sander H.J. Smits, Holger Gohlke and Carsten Münk.
Determinants of FIV and HIV Vif sensitivity of feline APOBEC3 restriction factors
Retrovirology, 2016, 13, 46.
Impact factor reported for 2018: 3.4

V. Dalibor Milić, Markus Dick, **Daniel Mulnaes (10%)**, Christopher Pfleger, Anna Kinnen, Holger Gohlke and Georg Groth.
Recognition motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1
Nature Scientific Reports, 2018, 8, 3890.
Impact factor reported for 2018: 4.1

The following publications were additionally co-authored:

VI. Nils Widderich, Marco Pittelkow, Astrid Höppner, **Daniel Mulnaes (10%),** Wolfgang Buckel, Holger Gohlke, Sander H.J. Smits, Erhard Bremer.
Molecular dynamics simulations and structure-guided mutagenesis provide insight into the architecture of the catalytic core of the ectoine hydroxylase.
Journal of Molecular Biology 2014, 426, 586-600.

VII. Sakshi Khosa, Benedikt Frieg, **Daniel Mulnaes (10%)**, Diana Kleinschrodt, Astrid Höppner, Holger Gohlke, Sander H.J. Smits.
Structural basis of lantibiotic recognition by the nisin resistance protein from Streptococcus agalactiae.
Nature Scientific Reports 2016, 6, 18679.

VIII. Prakash Chandra Rathi, **Daniel Mulnaes (40%)** and Holger Gohlke.
VisualCNA: a GUI for interactive Constraint Network Analysis and protein engineering for improving thermostability.
Bioinformatics 2015, 31, 2394–2396.

2. ABBREVIATIONS

CCD	Charge Coupled Device
MAD	Multi-wavelength Anomalous Dispersion
MIR	Multiple Isomorphous Replacement
MR	Molecular Replacement
NMR	Nuclear Magnetic Resonance spectroscopy
2D-HSQC	2D Heteronuclear Single Quantum Correlation
NOE	Nuclear Overhauser Effect
TROSY	Transverse relaxation-optimized spectroscopy
DDC	Dipole-dipole coupling
CSA	Chemical Shift Anisotropy
Cryo-EM	Cryo-Electron Microscopy
SAS	Small Angle Scattering
SANS	Small Angle Neutron Scattering
SAXS	Small Angle X-ray Scattering
FRET	Förster Resonance Energy Transfer
CASP	Critical Assessment of Protein Structure Prediction
MD	Molecular Dynamics
MSA	Multiple Sequence Alignment
PSSM	Position Specific Scoring Matrix
BLAST	Basic Local Alignment Search Tool
PSI-BLAST	Position Specific Iterated BLAST
CSI-BLAST	Context Specific Iterated BLAST
RPS-BLAST	Reverse Position Specific BLAST
DELTA-BLAST	Domain Enhanced Lookup Time Accelerated BLAST

HMM	Hidden Markov Model
MQAP	Model Quality Assessment Program
RMSD	Root Mean Square Deviation
TM-Score	Template Modelling Score
GDT_TS	Global Distance Test Total Score
CAD Score	Contact Area Difference Score
lDDT	Local Distance Difference Test
DCA	Direct Coupling Analysis
MI	Mutual Information
DNN	Deep Neural Network
RF	Random Forest
SVM	Support Vector Machine
PDB	Protein Data Bank
DBN	Deep Belief Network
DCNN	Deep Convolutional Neural Network
DCAE	Deep Convolutional Auto-Encoder
DRNN	Deep Recurrent Neural Network
GAN	Generative Adversarial Networks

3. ABSTRACT

All known life to date depends on proteins. Proteins are essential molecular machines that participate in every process within cells and play vital roles in, for example, cell structure, cell signalling, cell division, motor function, metabolism, and immune responses. Proteins owe their diverse functions to their vast array of different structures. Knowing the three-dimensional structure of a protein is therefore a critical step towards understanding its function. That knowledge can, in turn, help researchers to figure out how to modulate the protein function, leading to new and/or improved drugs or cleaner and more environmentally sustainable industrial processes.

At present, resolving a protein structure experimentally is laborious, time consuming and cost intensive. Therefore, being able to predict a protein structure accurately using computational methods is of high interest in biochemical, biomedical, and biotechnological research. Many methods for computational structure prediction have been developed in the last two decades, but no single method is consistently the best for every protein. Since different methods use different ideas, databases, algorithms and machine learning techniques, they provide different answers to the same types of problems. Consequently, integrating multiple so-called primary methods into a single meta-method harnesses their strengths and counteracts their weaknesses. This results in more robust and accurate structure predictions. However, when the majority of primary methods consent on the wrong prediction, out-numbering those that make the right one, meta-methods that rely on majority consensus make wrong structure predictions.

The goal of this thesis is to provide a toolbox and fully automated protein structure prediction workflow called TopSuite. This workflow consists of multiple meta-methods, each of which solve different tasks for protein structure prediction. Rather than using consensus though, these meta-tools make use of deep neural networks that have been trained on large datasets to learn when, and how much, to trust each primary method. As such, the TopSuite meta-methods are able to go against the majority when needed, and yield predictions that are significantly better than any of their respective primary methods.

Furthermore, the utility of TopSuite, in particular the template-based structure prediction workflow TopModel, is demonstrated through the application to target proteins of high biological, medical, and industrial importance.

4. ZUSAMMENFASSUNG

Proteine sind lebensnotwendig für alle bekannten Organismen. Als „molekulare Maschinen“ sind Proteine an allen Prozessen in der Zelle beteiligt und üben wichtige Funktionen aus, beispielsweise im Bereich der Zellstruktur, Signalweiterleitung, Zellteilung, Motorik, Stoffwechsel und Immunantwort. Proteine verdanken ihre vielfältigen Funktionen einer Vielzahl unterschiedlicher Strukturen. Die Kenntnis der räumlichen Struktur eines Proteins ist daher ein entscheidender Schritt zum Verständnis der Proteinfunktion. Dieses Wissen kann wiederum helfen herauszufinden, wie diese Funktion moduliert werden kann, was zu neuen und/oder verbesserten Arzneimitteln sowie saubereren und umweltverträglicheren industriellen Prozessen führen kann.

Gegenwärtig ist das experimentelle Aufklären einer Proteinstruktur arbeitsaufwendig, langwierig und kostenintensiv. Daher ist es für die biochemische, biomedizinische und biotechnologische Forschung von großem Interesse, Proteinstrukturen mithilfe von computer-gestützten Methoden genau vorhersagen zu können. In den letzten zwei Jahrzehnten wurden viele Methoden zur Vorhersage der Proteinstruktur entwickelt; jedoch ist keine einzelne Methode durchweg die beste für jedes Protein. Da unterschiedliche Methoden auf verschiedenen Ideen, Datenbanken, Algorithmen und Techniken des maschinellen Lernens aufbauen, produzieren sie unterschiedliche Ergebnisse bei der Lösung gleicher Arten von Problemen. Daher birgt die Kombination verschiedener Methoden zu einer Metamethode die Möglichkeit, die Stärken der Primärmethoden auszunutzen und ihren Schwächen entgegenzuwirken. Dies führt zu robusteren und genaueren Strukturvorhersagen. Kommt jedoch die Mehrheit der Primärmethoden zur gleichen falschen Vorhersage, so überstimmt diese die eventuell richtigen Vorhersagen anderer Methoden. Dies würde dazu führen, dass auch die konsensbasierte Metamethode eine falsche Strukturvorhersage trifft.

Das Ziel dieser Arbeit ist es, eine Programmsuite und einen vollautomatisierten Workflow zur Proteinstrukturvorhersage zu erstellen: TopSuite. Die Suite enthält Metamethoden, die jeweils unterschiedliche Aufgaben bei der Vorhersage von Proteinstrukturen erfüllen. Anstelle eines Konsens verwenden diese Metamethoden *Deep Neural Networks*, die auf großen Datenmengen trainiert wurden, um zu lernen, wann und inwieweit jeder Primärmethode zu vertrauen ist. Daher können die TopSuite Metamethoden bei Bedarf gegen die Mehrheit der Primärmethoden entscheiden und Vorhersagen treffen, die deutlich besser sind als die, der jeweiligen primären Methoden.

Des Weiteren wird die Nützlichkeit von TopSuite, insbesondere der templat-basierten Strukturvorhersage, bei der Anwendung auf Proteine von hoher biologischer, medizinischer und industrieller Bedeutung demonstriert.

5. INTRODUCTION

Knowing the three-dimensional structure of a protein is a key component to understand its stability [1], function [2], structural evolution [3], and interactions with ligands [4-5] or other proteins [6]. At present, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the most prominent experimental methods for resolving protein structures. Recent advances in Cryo-Electron microscopy (Cryo-EM), however, [7-8] have resolved structures at near atomic resolution. Despite these advances, experimental structure determination, regardless of the method, is both cost intensive and time consuming, making it impractical for structural genomics and costly for research groups with limited resources.

Consequently, protein structure prediction is an essential part of knowledge-based protein engineering [9], drug-design and -discovery [10], as well as function assignment [11-12]. In the last decades, many approaches to computational structure prediction have been developed. This raises the question of which method to use for a given protein-coding gene of interest, also known as the target sequence. Which biological information can be derived from structure prediction depends on its accuracy: While highly confident models based on homologous templates are generally suitable for computational ligand docking and virtual compound screening, even models of medium confidence can be useful for identification of functionally important sites and disease-associated mutations [13].

Since the inception of the field, protein structure prediction has been considered the Holy Grail of bio-informatics [14], and the majority of computational methods in the field has been developed to aid in the process of predicting protein structure. Consequently, much work has been focused on integrating methods developed by different people and computational groups, to harness their strengths and counteract respective weaknesses. Because these methods are based on different ideas, databases, methods and machine learning algorithms, they all provide different solutions to the same types of problems. These include tasks such as: 1) searching sequence databases for matches that share common ancestry with the target sequence; 2) aligning multiple sequences to one another in order to build a multiple sequence alignment; 3) predicting physical features (*e.g.* secondary structure or solvent accessibility) for the target sequence; 4) identifying suitable template structures by aligning a target sequence to a template; 5) aligning protein 3D structures to one another; 6) constructing 3D structural models from a list of templates and alignments; 7) evaluating model quality of an ensemble of protein 3D structural models.

The most common underlying methodology used for integrating different methods is the concept of consensus [15-23]. At its core, the idea of consensus is to submit the target sequence to a large ensemble of different so-called primary predictors and then merge the outputs of those methods in a clever way. If each of the primary predictor outputs reflects part of the truth, the resulting meta-prediction should be closer to reality than any single primary prediction. Meta-servers that implement different variants of this fundamental concept have shown to be one of the major advances in the field of structure prediction [24-25]. Furthermore, the best performing methods use some variation of majority-based consensus [24-25]. A major drawback of consensus, however, is that sometimes the majority is not correct, and when the majority of methods agree on the wrong prediction, consensus will drive the prediction away from the truth. This can in part due to many methods being highly correlated with one another, as they build on similar methodologies, databases and machine learning concepts. This means that if a particular task is difficult and prone to mistakes, correlated methods are prone to make the same mistakes, driving the majority away from the truth.

In this thesis, I present the development of several meta-methods for protein structure prediction, overcoming the problems of majority voting consensus by using deep neural networks (DNNs) to combine the inputs of different primary predictors. In this way, the DNNs learn which methods to trust most in which situations and allows the meta-methods to perform much better than traditional consensus.

6. BACKGROUND

In this section, I will briefly cover some of the major experimental and computational methodologies that have carried the field of computational structure prediction to its current state. While these developments have happened more or less in tandem with one another, the next section is grouped according to increasing complexity of methods, while attempting to maintain some level of chronological order.

6.1 EXPERIMENTAL STRUCTURE DETERMINATION

6.1.1 X-RAY CRYSTALLOGRAPHY

The currently most accurate experimental method to determine protein structures is X-ray crystallography, in which a highly concentrated solution of the purified protein is gradually condensed until it crystallizes. Upon crystallization, most of the proteins in solution arrange themselves into a highly regular and periodically repeating lattice. The exact structure of this lattice depends on the crystallization conditions and the protein itself. The crystallized sample is cooled, usually to around 100 Kelvin and then exposed to an X-ray beam from multiple angles. Cooling mitigates radiation damage and increases the lifetime of the crystal in the beam about 100-fold [26]. The resulting diffraction pattern, in terms of scattering angles and intensities of the X-rays, can be measured and used to reconstruct the electron density of the protein and infer the mean positions of atoms [27]. However, since light detectors such as Charge Coupled Devices (CCDs) only measure the intensity of the X-rays, the information about the phases is lost. In order to resolve the electron density, this information has to be recovered by shifting the phases and back calculating the phases from the resulting shifts. Resolving the phases, also known as solving the phase-problem, can be achieved in several ways. In Multi-wavelength Anomalous Dispersion (MAD), absorption and re-emission of X-rays at multiple wave-lengths by low orbit electrons leads to a shift in phases [28]. In Multiple Isomorphous Replacement (MIR), heavy metals are incorporated into the structure by soaking [29] or by modifying the protein with unnatural amino-acids with heavier elements such as seleno-cysteine or seleno-methionine [30], which also leads to a shift in phases. Another option, known as Molecular Replacement (MR), requires the reconstruction of the packing in the crystal lattice with a model from a close homologue, allowing for the calculation of the phases from the model density [31]. This however, requires a very high-quality model, since the crystal packing has to be resolved without atomic clashes.

The accuracy and quality of X-ray crystallography can in part be attributed to the rigidification of the protein structure caused by crystallization and cooling. However, the physiological form of the protein is much more flexible at biologically relevant temperatures. Since it is not known *a priori*, which conditions lead to formation of stable crystals, many thousands of different conditions with varying parameters such as protein concentration, salt concentration, buffer composition, temperature, pH, and ionic strength are generally carried out in the hope of finding a set of experimental conditions that lead to crystal formation [32]. This is time and cost intensive, and does not guarantee that a suitable set of conditions is ever found. Difficulties often arise when flexible parts of the protein prevent it from arranging into a stable crystal, making it necessary to cut these pieces out [33-35]. For X-ray crystallography, larger proteins are especially problematic, because all of their atoms have to arrange themselves in the same pattern in order to form a stable crystal. Since large proteins tend to have multiple structural domains, which can differ in their relative spatial orientation, obtaining a crystal for such proteins is particularly rare. Consequently, crystallization becomes less likely with increasing protein size [27]. An outline of an X-ray crystallography workflow for protein structure determination can be seen in Figure 1.

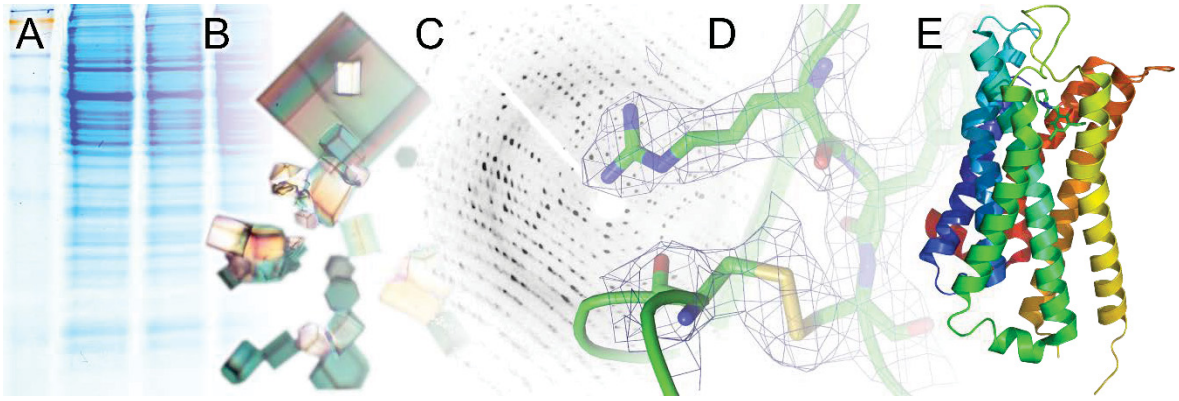


Figure 1. Outline of X-ray crystallography workflow. A) Protein expression and purification. **B)** Crystallization. **C)** Collection of diffraction patterns. **D)** Reconstruction of electron density from diffraction patterns. **E)** All-atom model derived from the electron density.

Because of their abundance in protein structure databases, crystal structures make up the majority of the targets used for training TopScore **(Chapter 9, Publication I)** and TopModel **(Chapter 10, Publication II)**.

6.1.2 NMR SPECTROSCOPY

Protein structure determination is also possible using nuclear magnetic resonance spectroscopy (NMR spectroscopy). The big advantage of NMR spectroscopy in terms of structure determination is that flexible proteins, which do not readily arrange into a crystal

lattice, can be resolved in aqueous solution at room temperature rather than in a frozen state. In NMR spectroscopy a strong external magnetic field is applied to the protein solution, which causes nuclei with a spin to orientate themselves in the direction of, or opposite to, that of the external field. By supplying electromagnetic radiation with so-called pulse frequencies, specific types of spin nuclei absorb the radiation energy and subsequently emit some of it, which can be measured as a so-called NMR spectrum. The exact frequency of the energy transition, termed a chemical shift, depends on the magnetic field around the nucleus, which in turn depends on the chemical environment around the nucleus shielding it from the external magnetic field.

In general, two types of spectroscopy experiments are carried out: Correlation spectroscopy, in which the chemical shifts are centered on a single frequency and correlated resonances are measured, and Nuclear Overhauser Effect (NOE) spectroscopy, in which the relaxation of the resonances is observed. The former is used to unambiguously assign specific signals to specific residues, while the latter allows for assignment of distance restraints between atoms not covalently bound to each other [36].

Which type of nuclei is targeted in NMR, also termed the isotopically labelled nuclei, differs depending on experiment, but common ones include hydrogens and labelling with ^{13}C and/or ^{15}N, the latter of which are fed to the organism producing the protein. By using multiple different labels at once, the coupling between different signals and their coupling constants can be calculated. This allows the signals to be assigned to specific atom pairs. The calculation of such 2D spectra prevents the overlap of different peaks, a typical example of which is 2D heteronuclear single quantum correlation (2D-HSQC) spectra. Using the calculated chemical shifts and coupling constants it is possible to calculate different types of restraints for the protein, most commonly torsion angle restraints and short-range (<5 Å) NOE distance restraints.

Using the calculated torsion angle restraints and NOE distance restraints the structure of proteins of about 20 kilo-Dalton (about 180 amino-acids) can generally be resolved using computational toolboxes such as the Crystallography and NMR System (CNS) [37] to generate protein models that satisfy the experimental restraints [38-39]. Larger protein systems, however, tend to have more complex shapes and intra-molecular interactions, which causes more overlapping peaks even in 2D spectra. Thus, structural elucidation of large proteins using NMR generally require higher dimensionality spectra such as 3D or 4D spectra, and is much more difficult to perform.

Large proteins tend to have shorter transverse correlation times and therefore the NMR signal decay more rapidly, leading to wide lines in the NMR spectra and consequently poor resolution. Transverse relaxation-optimized spectroscopy (TROSY) experiments [40] seek to alleviate this problem. Because of nuclear spin coupling along chemical bonds (also known as J-coupling), peaks in HSQC spectra appear as multiplets. While these peak components correspond to the same signal, they have different relaxation times due to interference between relaxation mechanisms such as dipole-dipole coupling (DDC) and chemical shift anisotropy (CSA). TROSY is performed at a high magnetic field strength (CSA is field strength dependent, DDC is not) to select the peak component for which the DDC and CSA relaxations cancel each other out due to destructive interference. This results in a single sharp peak in the spectrum, which significantly increases NMR resolution.

However, since large proteins are generally less soluble, even with TROSY and multi-dimensional NMR, the determination of large protein structures with NMR is difficult as a soluble sample is required [39, 41]. An outline of an NMR workflow for protein structure determination can be seen in Figure 2.

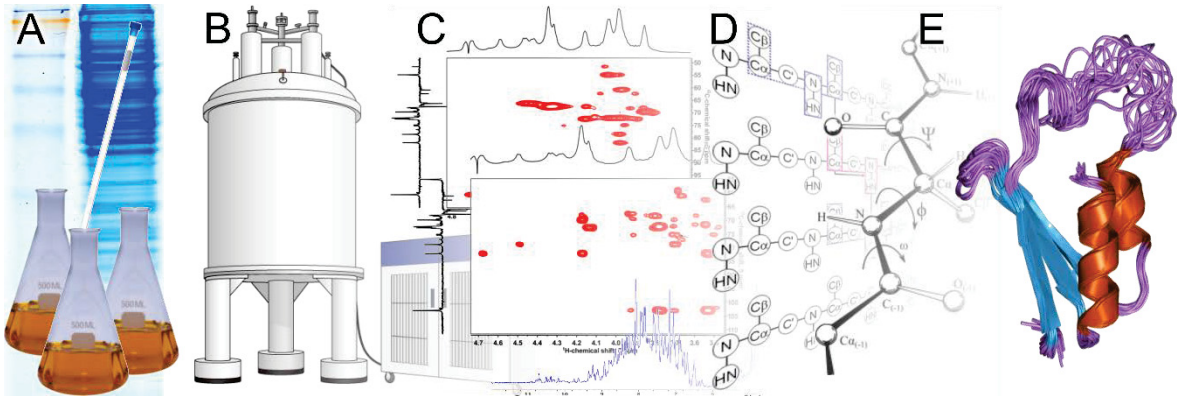


Figure 2. Outline of NMR Spectroscopy workflow. A) Isotope labeling, protein expression and purification. **B)** NMR measurement. **C)** 1D/2D/3D spectra analysis. **D)** Peak assignment and derivation of secondary structure and NOE restraints. **E)** Structure calculation by simulated annealing to fulfill NMR restraints.

In the validation of TopModel **(Chapter 10, Publication II)** on a particularly difficult target protein, we showed that the structure prediction from TopModel agreed well with secondary structure and NOE restraints from NMR experiments.

6.1.3 CRYO-ELECTRON MICROSCOPY

Another experimental method for structure determination, which has seen great advanced in recent years, is Cryo-Electron Microscopy (Cryo-EM) [7-8]. In Cryo-EM, a sample protein in solution is rapidly frozen and fixed in vitreous ice on a carbon film typically reinforced with

a copper grid for structural support. Ideally, the ice is thin enough to accommodate the proteins but not much thicker, preventing molecule overlap when the sheet is photographed. The thin ice sheet is then exposed to an electron beam, which subsequently passes through one or more lenses to magnify the image before it hits a detector, usually a CCD camera. This results in a blurry 2D image. The blurring comes from the fact that the energy from the beam is partially absorbed by the sample resulting in so-called beam damage and movement of the proteins, as well as the limitation of resolution due to the electron wavelength. One of the key advances in Cryo-EM is the development of ways to flash freeze the samples with liquid ethane to mitigate beam damage [42]. Another advancement is to use multiple images of the protein. By stacking these images and processing them with advanced image processing tools, the thermal movement of the sample can be measured and to some degree cancelled out, which greatly improves image quality and structural resolution [43]. By rotating the sample, images can be taken from multiple angles, allowing a 3D image to be computationally reconstructed using programs such as EMAN2 [44] or RELION [45]. Combining all of the above advances has allowed current state-of-the-art Cryo-EM methods to resolve large macromolecule-complexes at medium resolution [46]. Compared to X-ray crystallography and NMR spectroscopy, only a very small sample of just a few µl of a low concentration protein solution is needed, which helps to prevent protein aggregation. The greatest disadvantage of Cryo-EM is the low signal to noise ratio due to movement of the sample, and because of this, it is the method with the lowest resolution - typically 4-6 Å. From the computational perspective, the fact that high-resolution Cryo-EM structure determination is relatively new means that high-resolution Cryo-EM structures are rare in protein structure databases. An outline of a Cryo-EM workflow for protein structure determination is shown in Figure 3.

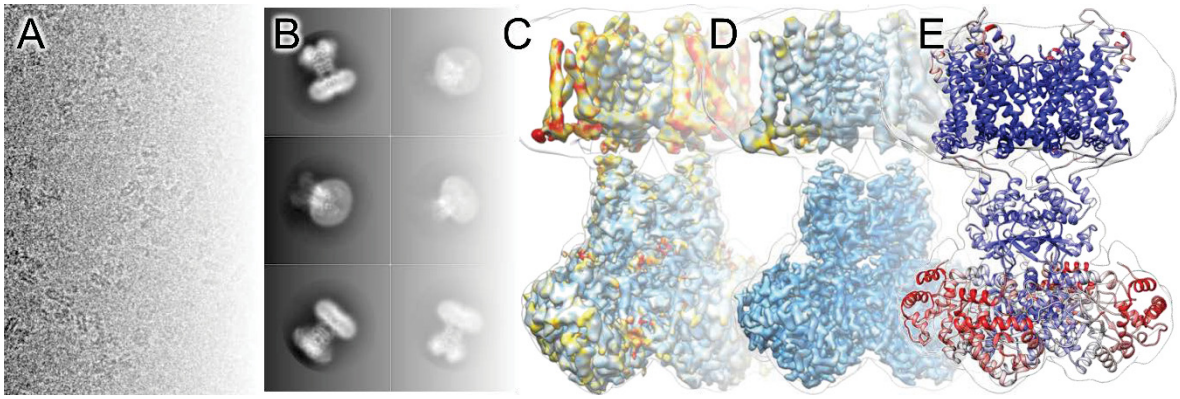


Figure 3. Outline of Cryo-EM workflow adapted from [47]. A) Image of the vitreous ice sheet with the embedded sample. **B)** Subset of the selected particle images showing the protein from different orientations.

C) Initial electron density calculated from the particle images. **D)** Refined density map calculated by image re-focusing on the top and bottom domains respectively. **E)** All-atom model derived from the Cryo-EM density.

6.1.4 LOW-RESOLUTION METHODS

For the reasons mentioned in the previous sections, high-resolution structural information is often not obtainable. Multi-domain proteins, large proteins, disordered proteins and trans-membrane proteins to name a few, each pose their own set of additional technical difficulties, mainly related to molecular flexibility and solubility, for resolving their 3D structure. Therefore, several techniques have been developed to get low-resolution structural information in a faster and cheaper fashion than the previously mentioned methods. In this segment, I will briefly cover a few of the more widely used methods, their advantages and their drawbacks.

Small Angle Scattering (SAS) covers two methods that are highly similar in nature: Small Angle Neutron Scattering (SANS) and Small Angle X-ray Scattering (SAXS). Both methods involve the bombardment of a sample with either neutron rays or X-rays at small angles, typically 0.1-10 degrees. By measuring the scattering patterns, the size and shape of the molecules in the sample can be determined at a resolution of up to 10 Å. The benefit compared to X-ray crystallography is that it is not required to obtain a crystal of the sample. Thus, the measurements can be done much cheaper and faster. Compared to NMR the methods also work for large proteins beyond the practical size limitation of NMR. The low resolution of SAS methods, however, means that, although they can determine the average shape of the protein in solution, they are not accurate enough to provide detailed atomistic information [48]. Therefore these methods are most suitable for large multi-domain proteins or protein-protein complexes, to obtain information about the over-all arrangement of structural units relative to each other [49]. In the validation of TopModel **(Chapter 10, Publication II)** on a particularly difficult target protein, we showed that the structure prediction from TopModel agreed well with the shape and scattering curves calculated from SAXS experiments.

Chemical cross-linking is a low-resolution method, in which the target protein of interest is chemically modified by covalently attaching polymer linkers to surface residues of the protein. These chemical polymer linkers vary from experiment to experiment but generally function by nucleophilic attack of the amino group of surface accessible lysine residues. Some variations of linkers work by activation with UV light [50-51], which in some cases enables the linking of other types of residues. When cross-linking is used for structure

determination, the cross-linked protein is subsequently digested and subjected to mass spectrometry. The resulting cross-linked peptide pair masses can help to determine which peptides, and therefore which residues in the sequence, were cross-linked before the protein digestion. These residue pairs, combined with the linker length, can be used to infer pseudo-distances between the cross-linked surface residues [52]. These distances are not real Euclidian distances, since the linkers traverse the surface of the protein. The actual distance between the cross-linked residue pairs is always significantly shorter than the linker length, due to the curvature and shape of the protein surface. In theory, these pseudo-distances can be used as discriminators for aiding in protein structure determination by imposing upper boundaries to the distance between the residues. However, in practice this has yet to be shown as a feasible and reliable method for protein structure determination, since up to half of the determined distances are generally incompatible with the native structure [53]. In practice, other more accurate sources of information are critical to resolve a structure *ab initio*. There are four main reasons for the low quality of chemical cross-linking data for structure determination: First, due to experimental constraints, there is ambiguity in determining the exact residues that were cross-linked, since the peptides may contain multiple residues that could have been linked. Second, the residues that can be used as covalent attachment points are generally limited, leading to only a limited number of cross links being possible for a given protein. Third, the majority of cross-linked residues are close by in sequence and as such has little to no discriminatory power for modelling and model selection, since the distance between the residues is trivially determined by their sequence distance. Fourth, even if long-range cross-linked residues are found, the fact that only an upper limit on the residue pseudo-distance can be inferred severely limits the discriminatory power of the data in terms of modelling and model selection.

A similar method to chemical cross-linking is Förster Resonance Energy Transfer (FRET) experiments. In a structure determination FRET experiment, fluorophores (also known as dyes) are attached to chemical linkers, which are then covalently bound to surface residues of the protein [54]. These residues are usually cysteine that are introduced into the target sequence by targeted mutagenesis. Different dyes are attached in pairs, where one dye acts as a donor and the other acts as an acceptor. When a protein labelled with two dyes is exposed to light of a specific frequency, matching the excitation frequency of the donor dye, the dye is excited, and the excitation energy is transferred non-radiatively to the acceptor dye by a dipole-dipole resonance interaction. Subsequently, the acceptor dye re-emits the energy as visible light in a frequency not overlapping with the initial light used for excitation.

The proportion between the amount of light used for initial exposure and the re-emitted light is used to calculate the efficiency of energy transfer between the two dyes. This efficiency is inversely proportional to the sixth power of the distance between the dyes, and can be used to calculate upper and lower bounds for the inter-dye distance [54].

There are several key advantages to FRET compared to chemical cross-linking. First, the distance that can be calculated is a Euclidian distance rather than a path the linker takes on the surface of the protein (*i.e.* protein surface curvature is ignored). Second, both upper and a lower distance bounds can be determined, which increases the discriminatory power of the method. Third, because dyes can be attached to any surface residue pair, the number of distance restraints depends mainly on the amount of point mutations that can be expressed and purified. Fourth, by using different dyes, multiple different residue pair distances can be measured from the same set of point mutations, leading to a much higher amount of distance restraints than cross-linking, with a higher information content since residues distant in sequence can be targeted for labelling. FRET does not come without drawbacks, however, since the linker length must be sufficient, such that the two dyes can be considered freely moving, yet longer linkers also give more uncertainty of the position of the dyes. Furthermore, dyes tend to stick to the surface of the protein. This can hinder the distance calculation in which the two dyes are approximated as freely moving. Finally, it is not always trivial to infer protein distance restrains from the inter-dye distance restraints. As such this method is often most suited for large multi-domain proteins or protein complexes, in which case, FRET can be used to calculate the relative position of larger biological units. These cases are particularly difficult for high-resolution methods such as NMR or X-ray crystallography due to the previously mentioned issues of protein flexibility and solubility. The relative position and orientation of biological units is also particularly difficult for computational methods to predict. This is because interactions between these units tend to be sparse, often transient or highly flexible, and are therefore weakly conserved compared to the structure of the units themselves.

6.2. CASP

The Critical Assessment of Structure Prediction (CASP) is a bi-annual competition that tests methods for protein structure prediction in a fully blind manner. The competition has been going on since 1994 [55] and covers a 3 month period, during which sequences for which the structure is about to be experimentally resolved (but may not be due to experimental difficulties, in which case the target is cancelled) either by X-ray crystallography, NMR

spectroscopy, or, more recently, by high-resolution cryo-EM, is released to the modelling community. The registered members of the modelling community then submit models of the sequences before any experimental data is publicly available. These models are automatically assessed by a battery of model quality evaluation algorithms to identify which method produced the model closest to the experimentally determined structure.

The CASP experiments seek to analyze the results of each CASP round to identify which methodological developments led to the biggest breakthroughs, and to identify which areas of research would be most beneficial to pursue. In order to do so, different CASP categories exist to assess the progress and state-of-the-art in different fields. These include template-based modelling [56], template-free modelling [57], model quality assessment [58], protein contact prediction [59], model refinement [60], and ligand binding-site prediction [61]. Additionally, CASP closely collaborates with the Critical Assessment of Predicted Interactions (CAPRI), which seeks to assess the state-of-the-art in predicting protein-protein interactions through protein-protein docking [62].

Due to the high competitiveness in the field, many structural bio-informatics groups keep their developments and findings accessible to the community only as black box online servers, in order to maintain their ranking in the CASP competition. While papers are generally published for most competing methods, these online servers are often very different from the published methods and not available as stand-alone tools. While this has been pointed out repeatedly by the community, it is unlikely that this is going to change in the future. The high competitiveness, however, has also led to gradual development in the field and has served as a good way for fully blind predictions to shine, rather than evaluating method performance only in a retrospective manner.

In the publication of TopModel **(Chapter 10, Publication II)** we showed that on CASP datasets from CASP competitions 10, 11, and 12, TopModel outperforms all primary predictors, and we show targets from CASP13 on which TopModel produced the second best model out of all predictors in the competition.

6.3 TEMPLATE-BASED STRUCTURE PREDICTION

Historically, template-based structure prediction in the form of homology modelling was the first type of structure prediction to be developed. Since protein structure is highly conserved, modelling the 3D structure of a protein structure from a closely homologous template is often trivial. This is because sequence identity between the target sequence and the template structure is generally so high that alignment between the two is unambiguous and the

alignment thus rarely contains any errors. Furthermore, structural conservation is generally so high that model construction often requires little more than copying of atomic backbone coordinates, reconstruction of side-chains and energy minimization. Due to the sparse population of structure databases, however, this type of modelling is generally very restricted in terms of applicability. When sequence identity decreases, the performance of alignment programs drops significantly, which makes alignment errors more prone to occur. Severe alignment errors often start to appear at around 40% sequence identity between the template and target sequence. The twilight zone for template-based structure prediction [63] is at about 30% sequence identity. When a given template has less than 30% sequence identity to the target structure, the structural overlap between model and native structure (the percentage of equivalent residue pairs superimposable to less than 3.5 Å C_α-atom distance) decreases to about 60%, showing a large difference between model and native structure [63].

6.3.1 HOMOLOGY DETECTION

The first challenge one encounters in computational structure prediction is homology detection. Two proteins are homologous, if they share a common ancestor. To learn from evolution, be that in terms of structure or sequence, information from related proteins from different organisms is required. The task is therefore to select from a database of sequences those matches that share a common ancestor with the target protein of interest (which may or may not represent experimentally determined structures). This is done by aligning the target sequence to sequences in a database and the evaluating these alignments in terms of similarity to determine, whether the match and the target share common ancestry or not.

Accurate and sensitive alignment of sequences has been a goal in bioinformatics for decades and has driven the development of the majority of methods in the field to this day. Sequence alignment requires an alignment algorithm and a scoring function. The scoring function is used by the alignment algorithm to calculate the alignment between the target sequence and the match in order to optimize the score. Furthermore, it is used to compare different database matches to one another so as to select those most likely to be homologous to the target sequence [64]. Both scoring function and alignment algorithm determine the speed and accuracy of the homology search. Some search algorithms may use heuristics to speed up calculations or use simple scoring functions to achieve faster database searches at the cost of lower accuracy [65-66], while others favor more accurate alignment algorithms and complex scoring functions to optimize accuracy at the cost of slow calculation speeds [17, 67].

6.3.2 SUBSTITUTION MATRICES

The first scoring functions for alignment algorithms were substitution matrices, in which similar residues are given a better score if aligned to each other, and where gaps are given a penalty. This penalty is usually divided into a penalty for opening a gap and another smaller penalty for extending it. However, these methods are heuristics based on trial and error. The score for aligning two residues is traditionally defined as the logarithm of the observed frequency of substitution of a target residue to a template residue divided by the frequency of the template residue, known as the log-odds ratio. Substitution matrices are defined either from physiochemical similarity or by multiple sequence alignment analysis. The first widely used substitution matrices, the PAM matrices, were developed in the 1970's [68] by calculating mutation rates from multiple alignments of sequences from closely related species and extrapolating these to longer evolutionary timescales using matrix multiplication. This did not perform well when aligning dissimilar sequences, however, leading to the development of the BLOSUM matrices by Henikoff and Henikoff [69]. The BLOSUM matrices were calculated by analyzing blocks of conserved residues across divergent multiple alignments. To remove bias from highly similar sequences, the alignments were clustered at a specific identity cut-off, and each cluster was weighted when calculating the substitution rates. The BLOSUM62 matrix (with a sequence identity cut-off of 62%) is the most widely used substitution matrix to date.

6.3.3 PAIRWISE SEQUENCE ALIGNMENT

Pairwise sequence alignment algorithms align a target sequence of length N with a potential match of length M. Needleman and Wunsch [70] introduced the dynamic programming algorithm, which solves the problem by representing it as finding the least-cost path through a cost matrix. The Needleman-Wunsch algorithm produces the optimal alignment given the scoring function but has a time and space requirement which is proportional to the product of the two sequence lengths $O(N \cdot M)$. An illustration of this algorithm is shown in Figure 4. The Smith-Watermann algorithm [71] reduces the time complexity by using the observation that conserved residues cluster in specific regions of the sequence. It iteratively aligns only the most similar regions, avoiding the issue of searching the entire $N \cdot M$ cost matrix and effectively decomposing it into several smaller matrices located along the diagonal. This is referred to as local alignment, as it effectively changes the alignment problem from a global to a local one. Different variations of these methodologies have been developed [72] but the core principles remain unchanged. Most importantly, these algorithms are highly dependent

on the scoring function, and even the faster variants are generally too slow to search through very large sequence databases, especially when using complex scoring functions. Therefore, heuristics are generally used for sequence matching in database search methods.

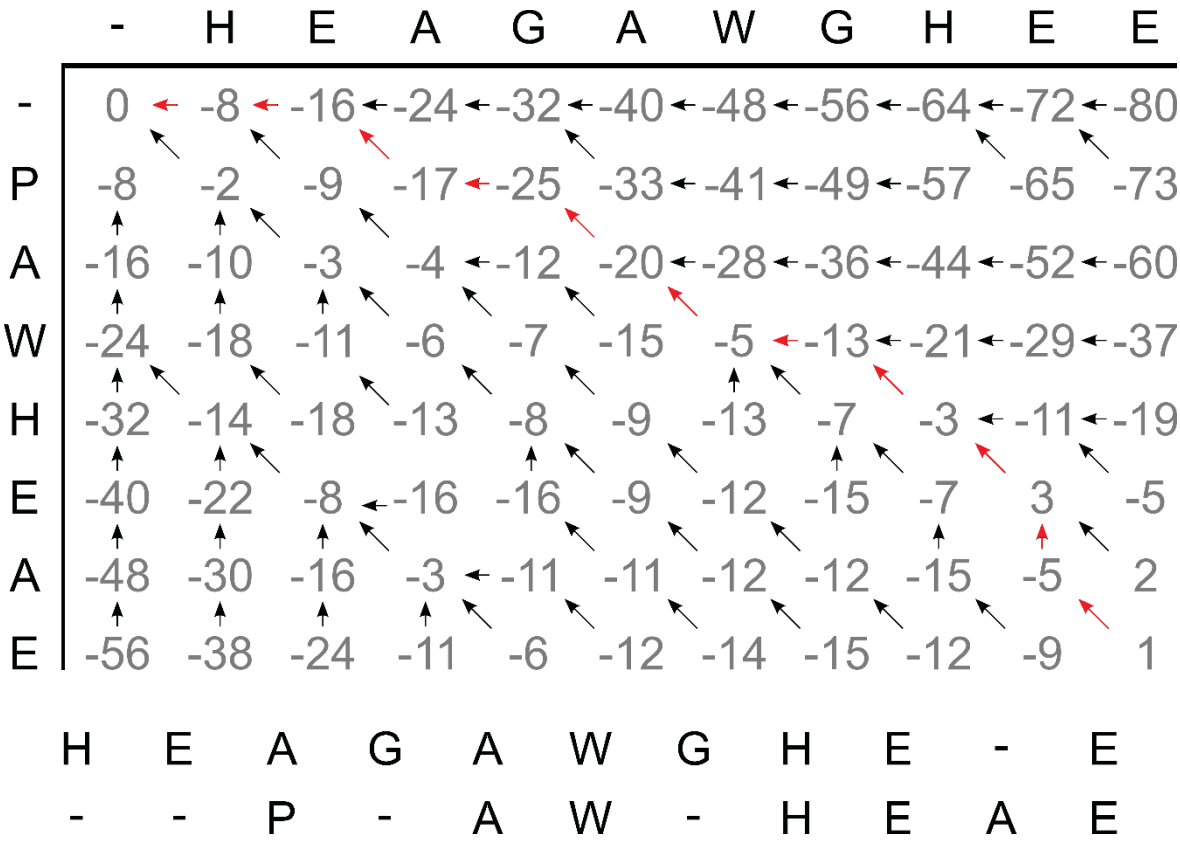


Figure 4. Illustration of the Needleman-Wunsch dynamic programming algorithm for pairwise sequence alignment. First, a cost matrix, *F*, is calculated iteratively based on a scoring matrix, *M*, (in this example the BLOSUM62 matrix) and a gap penalty *d* (in this case -8). For each cell [i,j] in *F*, the score can be computed recursively from the cells [i-1,j-1], [i-1,j] and [i,j-1], using the formula max(F[i-1,j-1]+$M(x_i,y_j)$, F[i-1,j]-d , F[i,j-1]-d). For each cell, the pointer (indicated as arrows) to which of the three cells gave the maximum value is stored. The uppermost row and leftmost column are filled by summing up the gap penalty *d*. Second, the optimal alignment is calculated as the least-cost path through the matrix following the pointers from the lower right and back through the matrix (indicated by red arrows). The resulting alignment is shown underneath.

The most famous heuristic for pairwise alignment used for database searches is the word-search heuristic used in the BLAST [65] and FASTA [66] algorithms. In this heuristic, the target sequence is cut into small fragments called words, for which exact matches are found in a potential match sequence. The relative positions of the matched words in the target are then compared to the positions in the match. Only when the relative positions of the words are comparable between the two, are regular algorithms used, such as the Smith-Watermann or Needleman-Wunch algorithms. This allows these heuristic methods to disregard highly dissimilar sequences with few or no matching words, which increases database search speeds

by orders of magnitude. However, they cannot guarantee that all true hits are found especially for distantly related sequences where only few words are completely conserved.

6.3.4 MULTIPLE SEQUENCE ALIGNMENT

Once a database search has returned a list of related sequences, these can be aligned to form a multiple sequence alignment (MSA) that contains more information about the target sequence than any single match does on its own. The process of calculating an MSA is a computationally expensive NP-complete problem unless heuristics are used, of which three variations are common:

Progressive methods use pairwise alignments (such as the ones generated from a database search) to construct a phylogenetic tree, which is used to first combine similar sequences into a larger alignment, to which more dissimilar sequences are then added. This is computationally efficient, but it can introduce pairwise alignment errors that persist in the final alignment.

Iterative methods are the most common and seek to minimize pairwise alignment errors by iteratively removing subsets of sequences from the alignment and re-aligning them to the alignment from which they were removed, effectively refining the MSA. Examples of methods that use the progressive method with iterative refinement include ClustalW [73], MAFFT [74], and TCOFFEE [75]. However, these methods are not guaranteed to remove pairwise alignment errors.

To remedy the issue of persisting pairwise alignment errors, motif-search methods try to identify highly conserved regions in the initial MSA. The initial MSA may be generated by either a progressive alignment method or a progressive alignment with iterative refinement. These conserved regions are then aligned with global methods (*e.g.* Needleman-Wunch) and the variable regions in between the conserved regions are aligned with local methods (*e.g.* Smith-Watermann). Examples of motif-search methods include FORMATT [76] and 3DCOMB [77].

6.3.5 IMPROVED HOMOLOGY SEARCH

With the ability to generate a large MSA for a target sequence, information such as the amino-acid frequency and gap frequency for each position in the target can be calculated. Such information, called profile information, prompted the development of sequence-profile [78] and profile-profile [79-80] alignment methods, which match sequence profiles rather than sequences themselves. This allows more distantly related homologues to be detected because

more information is available in the sequence profile than in any single member of the MSA itself.

With the rapid growth of sequence databases and the observations that different positions in the protein sequence mutate differently, it became possible to construct position specific scoring matrices (PSSMs) [81-82]. Rather than using a single matrix like BLOSUM62 for the entire sequence, these methods calculate the substitution matrix for every position in the target sequence-based on a MSA of homologous sequences found for example from an initial BLAST search. This encodes more information than the profile, as it includes not only the residue type frequency at each position in the target sequence, but also every binary mutation frequency (mutation of one residue type to another or to a gap).

With the development of PSSMs came the idea of iterated database searching. In this method an initial search is used to construct a PSSM, which is then used to repeat the database search to find more distantly related sequences. The newly identified sequences are then used for updating the PSSM, and the updated PSSM is used for another round of searching. In the first iteration of such a search, the PSSM of each position is generally approximated by the BLOSUM62 matrix. This is the essence of Position Specific Iterated BLAST (PSI-BLAST) [65], which is one of the most widely used bioinformatics tools to date.

PSI-BLAST searches, however, still rely on the initial search to return a significant number of good matches in order to construct a reliable PSSM. This means that the initial standard matrices (usually BLOSUM62) has significant impact on which sequences are detected. To alleviate this issue, context specific matrices were developed based on multiple alignments of proteins with known structure, to map the relationship between physical features (*e.g.* secondary structure and solvent accessibility) and mutation probabilities [83]. By matching a sliding window (±13 residues) of the target sequence to a set of pre-computed context specific substitution matrices, Context Specific Iterated BLAST (CSI-BLAST) [84] is therefore significantly more sensitive than PSI-BLAST.

Reverse Position Specific BLAST (RPS-BLAST) reverses this search methodology by searching through a database of PSSMs to calculate how likely it is for the query sequence to be generated by that PSSM. The latest flavor of BLAST is the development of the Domain Enhanced Lookup Time-Accelerated BLAST (DELTA-BLAST) [64]. DELTA-BLAST first uses RSP-BLAST to find highly scoring pre-calculated PSSMs from the Conserved Domain Database (CDD) [85]. The identified matches are then used to replace the initial PSSM for the matched positions. In this way, the information of the initial PSSM used for searching the

sequence database is more specific to the target, rather than simply being pseudo-counts extrapolated from the BLOSUM matrix. Because the iterated searches depend largely on correct matches being found in the first iteration, this leads to greatly improved specificity and sensitivity of DELTA-BLAST compared to BLAST, PSI-BLAST and CSI-BLAST.

6.3.6 HIDDEN MARKOV MODELS

A major advancement in the field of sequence searching is to extend the methodology of PSSMs and use MSAs to encode probabilistic models known as profile Hidden Markov Models (profile HMMs, or just HMMs for short) [86-90]. An encoded profile HMM encodes the insertion, deletion, and mutational probabilities at each position in the target sequence just like a PSSM. However, HMMs also models the transitional probabilities between each residue and the next residue in the sequence, *i.e.* the probability that a residue is proceeded by a residue of a specific residue type or a gap. HMMs therefore carry more information than PSSMs and, thus, generally perform better for detecting distant homologues.

Once a HMM has been encoded, usually from a MSA, the probability that it would produce a specific sequence can be calculated very quickly, and a HMM can as such be used to rapidly search databases of sequences to identify those likely to match the query, without the need for heuristics such as word-searching. This makes HMM searches more sensitive especially for identifying matches with low sequence identities. The initial HMM is generally initialized from pseudo-counts extrapolated from standard matrices such BLOSUM62. The model is then updated with each search over the database similar to iterated BLAST searches. Because HMMs are probabilistic models, they allow for explicit calculation of the probability, that a sequence is a match, which eliminates the need for cut-offs to the expectation value, which is based on database composition. HMMs also enable very fast comparison between two models, which makes for much more sensitive homology search than comparing a sequence to an HMM. Additionally, once an HMM is matched, all sequences that match it can be matched as well. This allows for more rapid database searches since the HMM's essentially function as clusters of sequences and only if a cluster has been found to match is it required to match each sequence in that cluster [91].

Finally, HMMs can be used to rapidly calculate accurate multiple alignments of many sequences by aligning each sequence to the HMM. This avoids the issue of iterated alignment refinement to some degree and leads to highly accurate MSAs. Thus, both sequence searching and multiple alignment can be done at speeds much faster than with word-search heuristics and often with better accuracy. Since HMMs can be used for scoring,

alignment and search function, methods such as HMMER3 [87], SAM [90], HHBlits [88] and HHSearch [92] are some of the most powerful sequence-based search methods to this day. However, the most sensitive searches, which use HMM-HMM comparisons, requires HMMs to be pre-calculated for each protein domain family in the sequence database, which takes significant computational resources. An illustration of a profile HMM can be seen in Figure 5.

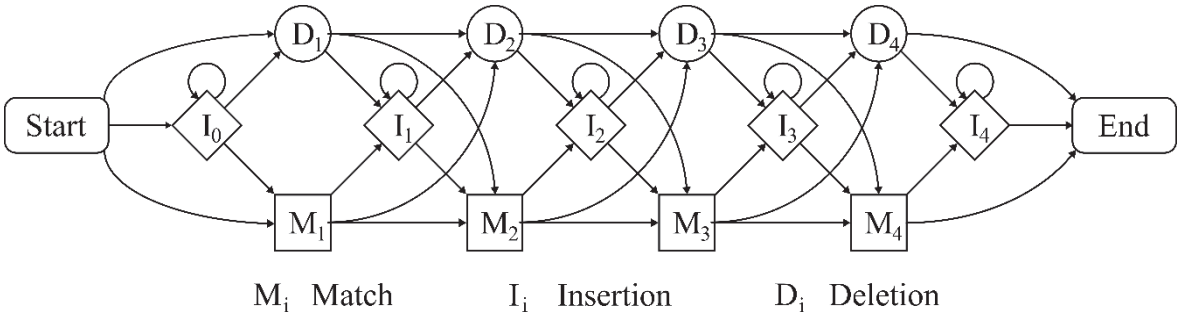


Figure 5: Schematic of a profile HMM. The profile HMM is encoded from a MSA. Match states (M_i) encode the mutational probability of each of the columns in the MSA similarly to the way a PSSM encodes the mutational probability of each position in the target sequence. Insertion states (I_i) encode the probability of insertion of each of the 20 residue types in the MSA and therefore model the highly variable regions of the MSA such as loops. Deletion states (D_i) do not match any residues, however, they make it possible to jump across columns in the MSA and thus model the deletion of residues with position specific probabilities. This is more accurate than, for example, the affine gap penalties commonly used with substitution matrices. Because the HMM is a sequential model, it conveys not only the estimated mutational probabilities (*e.g.*, the probability of observing an Alanine residue at position x mutating to a Tyrosine), but also the transitional probabilities (*e.g.*, the probability of observing the aforementioned mutation, given the mutational probabilities of the previous residue). This makes profile HMMs the most sophisticated probabilistic models of MSAs and, therefore, more efficient tools for detecting distantly related homologous sequences.

6.4 THREADING

So far, the methodologies described focuses on matching a target sequence against a database to retrieve and align homologous sequences. These methods are applicable to both sequence and structure databases, since the structure databases simply contain sequences of protein structures. However, the matching of a sequence against a structure is more difficult, mainly because structure databases such as the PDB [93] are sparsely populated compared to sequence databases such as UniprotKB [94]. This means that, even if a related structure exists, it is far more likely to be distantly related to the target sequence. This in turn makes a false positive match (a template matched by pairwise alignment, which has a different 3D fold than the target protein), much more likely to occur. Threading algorithms seek to circumvent these issues by using a plethora of methods to improve the scoring function used for the alignment between the target sequence and the template. While methodologies such as FASTA [66], DELTA-BLAST [64], SAMT2K [90], HHBlits [88], and FFAS-03 [95] could therefore be considered threading algorithms, in that they perform the same task of identifying potential templates in a structure database, they do so purely by sequence matching.

In the next sections, discussion will focus on methods that are more complex, and use advanced scoring functions for aligning a target sequence to a structure. This is done in three steps: First, the target sequence is used to search a sequence database for matching sequences and generate a MSA using one or more of the previously described methods. Second, the MSA is used to predict physical features of the target sequence, such as secondary structure, solvent accessibility or residue dihedral angles. Finally, these features are used as additional scoring terms for searching through a structural database to find matching structures with similar features. This three-step methodology is shared between all advanced threading algorithms and takes considerably longer than any of the aforementioned methods due to the complex scoring functions.

6.4.1 STRUCTURAL FEATURES

Over the years, the increased availability of sequence data and the maturation of machine learning for analysis and prediction laid the foundation for extending the alignment scoring function beyond simply matching sequences. Since the 3D structure of a protein is far more conserved than its sequence, matching linear features improves the scoring and thus enable the detection of far more distantly related structural homologues. Several such linear structural features have been the target of prediction over the years, most notably secondary structure [96-102], solvent accessibility [96-97, 99, 101-105], residue depth [106], backbone dihedral

angles [101-102, 107], number of contacting residues [17], half-sphere exposure [101-103], and residue disorder [108]. Threading algorithms such as pGenThreader [80], pDomThreader [80], HHSearch [92], LOMETS [17], MUSTER [107], RAPTORX [67] and SPARKX [109] all use one or more of these features to increase the likelihood of selecting good templates and produce high quality pairwise alignments. This comes at the cost of slower search speeds due to the more costly scoring functions.

Additionally, probabilistic modelling [67, 109], depth-dependent alignment of structure fragments [110], multiple template and structure alignment [111], normalized Z-scores [79, 105], and sequence-based solvation potentials [80] have been employed to increase performance of threading alignments by including more information into the scoring function. The most expensive approach to threading [112] uses the construction of crude models of every alignment for every potential template and evaluates the quality based on a 3D energy function or knowledge-based potential, but this is generally far too computationally expensive to be feasible, especially for larger proteins and databases.

In TopModel **(Chapter 10, Publication II)**, FASTA, DELTA-BLAST, HMMER3, HHBlits, HHSearch, FFAS03, SAMT2K, pGenThreader, pDomThreader, LOMETS, MUSTER, RAPTORX, and SPARKX are used as primary threading algorithms. This is aiming to provide multiple diverse threading algorithms for template detection.

6.4.2 META-SERVERS

Meta-approaches have proven to be one of the major advances in template detection and structure prediction [113], as evident by the consistent high ranking of the Zhang meta-server [17] in the blind Critical Assessment of Protein Structure Prediction (CASP) experiments. The meta-server methodology produces structure predictions using information from multiple different primary predictors [17, 25] and either re-ranks or combines their output to produce better predictions than any of the primary predictors.

With the large diversity of methods for threading, it is not surprising that meta-servers, which employ multiple different methods, have been shown to outperform single-method approaches. Meta-servers have several advantages: First, templates that are missed by one method due to differences in database composition, alignment methodology or scoring function, are less likely to be missed by all methods, increasing the chances of a template with the right fold to be represented in the ensemble of templates. Second, the use of multiple threading alignments provides different pairwise alignments for those templates

that were identified by multiple threaders. This allows for calculation of a consensus alignment, which can be constructed to be better than its input alignments.

6.4.3 CONSENSUS

Most meta-servers work based on consensus between different algorithms. The traditional consensus is based on majority voting at either the template selection or model building stage. Majority voting at the template stage could be to select templates based on their similarity to other identified templates, effectively selecting the fold that was found most frequently by most primary predictors. This was implemented in the initial versions of the LOMETS/I-TASSER servers. An example of majority voting at the model building stage could be to build a library of residue contacts from an initial model ensemble and remove contacts that contradict the majority when using those contacts to construct models. This method is used in the most recent versions of the LOMETS/I-TASSER servers [24]. The MULTICOM [25] server generates consensus models during both initial model construction and model refinement. In model construction, multi-template consensus alignments are used, and during refinement, models are clustered and combined with the cluster centroid, either at the global or at the local level. Both LOMETS, I-TASSER and MULTICOM use majority voting, since they converge the results to the fold generated most often across different algorithms. In many cases, especially ones where most methods produce correct folds, majority voting will correct errors and improve the overall modelling result.

However, the problem of majority voting in consensus methods is that different programs have been developed together, building on ideas and methodology from each other, and are as such susceptible to the same pitfalls. In other words, if a difficult target is prone to a particular alignment error, many threading programs are likely to make the same error. Since the erroneous alignment is in the majority, this in turn means that consensus methods based on majority voting may discard a correct alignment and converge on the wrong fold.

In TopModel **(Chapter 10, Publication II)** top-down consensus is used instead of majority voting to filter out false positives based on structural similarity to the best template as identified by a series of deep neural networks. This proved to be very effective, especially for targets where the majority of the identified templates are false positives, since it makes it possible to go against the majority in cases where primary predictors converge on the wrong fold.

6.5 PROTEIN STRUCTURE MODELLING

6.5.1 STRUCTURAL ALIGNMENT

As structural databases of proteins grew, although at a slower pace than sequence databases, it became apparent that protein structure is much more conserved than protein sequences. Thus to determine the relationship between two protein structures, algorithms were designed to align these, and infer residue correspondence based on spatial proximity rather than physiochemical residue similarity. Structural alignments are generally of higher quality than purely sequence-based alignments, and when multiple sequences with known structures are found as matches to a target, aligning these structures can improve the correctness of the alignment [75]. The difficulties in structural alignment arise mainly from the fact that proteins are not rigid but can adopt many conformations. As such, while the fold of two structures might be the same, they may have different conformations that makes structural alignment difficult. Different methods have been employed to overcome these difficulties, ranging from the combination of structural alignments for rigid bodies and sequence alignment for flexible parts [76], the combination of structural alignment algorithms with evolutionary sequence data [114], the alignment of structure fragments [115] or consensus between different structural alignment methods [75]. TopAligner **(Chapter 10, Publication II)** uses all the structure- and sequence-based multiple alignment methods mentioned here to generate an ensemble of different multi-template alignments from which to build models.

6.5.2 MODEL CONSTRUCTION

The construction of a 3D model of the protein is not a trivial problem to solve, as three major challenges has to be overcome: First, the method should construct models, which are close to the native structure. As templates are often used as input, however, model-building software often tends to construct models that are more similar to the template(s) than to the native structure. Second, despite the improvement in threading algorithms, the pairwise alignment between target and template may still contain errors. Many model-building algorithms are not able to correct these errors and these will therefore persist in the model. Finally, the construction of models requires a scoring function to guide and select for the best model. As with threading, however, the more sophisticated the scoring function, the more computationally expensive the modelling becomes. The four most commonly used methods for model construction will be discussed in the following four chapters and illustrated in Figure 6.

6.5.3 RESTRAINT-BASED FOLDING

In restraint-based folding, implemented in popular software such as MODELLER [116], distance restraints are calculated for inter-residue distances in the target sequence. These distances are based on distances between corresponding residues in the template, with the mapping between template and target sequence being given by the input alignment. Starting from a pseudo-random atom positioning based on their original positions in the template; atoms are then moved randomly until the highest number of restraints are fulfilled. There are several advantages to this methodology, most notably the easy inclusion of restraints from different sources. These include geometric restraints based on stereochemistry to guide side-chain arrangement, as well as information from multiple template structures or from predicted features **(Chapter 6.4.1 and 6.7)**. The main disadvantage of this methodology is that it handles modelling of parts of the sequence without template very poorly, especially if these parts adopt secondary structure other than small loops. Furthermore, it is unable to repair alignment errors because the mapping between the target sequence and the template structure (*i.e.* the alignment) is fixed. Restraint-based folding is therefore only generally applicable for template-based structure prediction, and is favored for easy targets with few or no alignment errors. Finally, it tends to construct models that are close to the input templates, and as such performs best for target sequences with highly similar templates. Therefore, the modelling of targets for which the template structures are distant homologues results in lower performance by restraint-based methods. This is mainly a consequence of alignment errors, and the fact that templates are more likely to not fit the native structure. An outline of a restraint-based model construction workflow is shown in Figure 6 A.

6.5.4 FRAGMENT ASSEMBLY

A popular model construction method is fragment assembly, most notably in the form of the ROSETTA software suite [117]. In fragment assembly, the input sequence is first used to generate a library of fragments by cutting the input sequence into overlapping pieces of different sizes, generally 3 and 7 residues long. These pieces are then matched up against a database of residue fragments of the same size. This fragment database is extracted from a large set of protein structures and clustered to obtain a reasonably small set of representative fragments. The scoring function that performs this match considers structural (as described previously) as well as residue similarity when performing the fragment selection. However, since the fragments are generally small (3 and 9 residues usually), it is difficult to obtain a single significant match. In other words, there is a high chance of getting a random match

due to the short fragment size. Therefore, multiple high scoring matches are kept for each fragment of the target sequence, in the hope that one of them has the right conformation.

Once the fragment library is generated, different conformations of the protein are sampled by exchanging the conformation of randomly selected segments with fragments from the library using Monte Carlo sampling, and evaluating if the fragment exchange should be accepted based on an energy function. To speed up this sampling process, two different energy functions are used in ROSETTA, a coarse-grained energy function in which side-chains are represented as a single pseudo-atom, and an all-atom energy function. Initially only backbone conformations are sampled using the coarse-grained energy function to generate a large number of diverse initial structures known as decoys. Then, a subset of high scoring decoys according to the coarse-grained energy function are reconstructed in atomic detail and re-sampled using the all-atom energy function.

The main advantage of fragment assembly is that it is a highly flexible method. The input fragments can, for example, come from detected homologues. This speeds up the convergence since the fragments match homologous structures. Additionally, because no full-structure fragments are used, the conformations of fragment assembly models can often end up closer to the native structure than the input templates. This also means that fragment assembly can be used both for structures with detected templates and for structures without known templates. Because *ab-initio* fragments (with no global homology to the native structure) are used, loops are generally of better quality than from restraint-based methods such as MODELLER. Furthermore, the scoring function used to select fragments and to score decoys can be modified to favor agreement with predicted features **(Chapter 6.4.1 and 6.7)**. The disadvantage of fragment assembly is that the extensive Monte Carlo sampling is extremely computationally demanding, and since the fragments are short, long-range interactions between residues far apart in sequence but close in the structure cannot be captured by the assembly method to narrow down the search space. In other words, even though long-range information can change the energy landscape of folding, the amount of sampling required is still very large. Convergence can therefore take a very long time. This is especially true for large structures, since tens of thousands to hundreds of thousands of models have to be generated. An outline of a fragment assembly model construction workflow is shown in Figure 6 B.

6.5.5 CONTINUOUS ASSEMBLY

Contrary to fragment assembly, which performs global moves where small fragments are replaced across the entire sequence, continuous assembly programs such as I-TASSER assemble large fragments identified by threading. I-TASSER (Iterated Threading ASSEmbly Refinement) [24, 118] uses replica-exchange Monte Carlo sampling to assemble larger continuous pieces of the target sequence. These pieces are built from restraint-based models from templates identified by threading alignments. This is done by sampling regions without template on a lattice and allowing the rigid template-based pieces to move off-lattice. In the first iteration of I-TASSER, one simulation is carried out for each template using consensus restraints extracted from all alignments of all templates to generate initial models. The initial models from this simulation are then clustered, and new restraints are extracted from the largest cluster centroid as well as from templates that structurally align to the centroid. The new restraints are added to the initial restraints and used for a second round of structure re-assembly, starting from the cluster centroid. This allows for alignment errors to be rectified and for the fold to be refined, after which full atomic-detail structures are constructed and energy-minimized.

The advantages of continuous assembly is that it is much faster than fragment assembly, since the fragments identified by threading are both much larger and much more likely to have the right conformation, which makes the sampling faster and the accuracy high. However, it comes with the drawback that sampling outside of the conformational space defined by the threading results is limited compared to fragment assembly, and it is, therefore, limited in terms of flexibility of which threading results can be used. To compensate for this drawback, the most recent version of the I-TASSER server also generates models from a more traditional fragment assembly method QUARK [118] to improve performance for *ab initio* modelling. Another disadvantage that is remedied by the addition of QUARK is that traditionally I-TASSER was unable to correct mistakes in threading if these mistakes were made by the majority of threaders, since the conformational variability is highly biased by the initial template threading results. Inclusion of *ab initio* models from QUARK partially remedies this template-based bias. An outline of a continuous assembly model construction workflow is shown in Figure 6 C.

6.5.6 CONTACT-BASED FOLDING

Contact-based folding is fundamentally different from the previous methods in that it disregards template structures completely. Unlike the previous methods, which obtain

structural information either in the form of highly specific distance restraints from templates (MODELLER), specific structure pieces from threading results (I-TASSER), or less specific structural information from small fragments (ROSETTA), contact-based folding uses no structural information at all.

The main idea behind contact-based folding is that information about residue-residue contacts can be obtained directly from a large MSA, and that given such information the protein can be folded. How such information is obtained from the MSA will be discussed in chapter 6.7. A classic example of contact-based folding is the CONFOLD method [119]. Since the information from contact predictions is historically prone to high false positive rates, contact-based folding is performed in two steps. In the first step, a fully extended conformation of the protein backbone is moved in order to fulfill residue-residue contacts in a manner similar to MODELLER. MODELLER, however, is centered on distance restraints and starts from an input template conformation, whereas CONFOLD is centered on contact restraints and starts from an extended conformation. Therefore, CNS [37] is used as the folding engine, as it starts from an extended conformation and is built around contacts initially developed for resolving structures from short-range NOE restraints from NMR experiments. In the second step, each contact is re-weighted according to how often it was fulfilled, and used for a second round of folding. Additionally, different weights are placed on secondary structure restraints and distance restraints to provide models that are more diverse. This methodology allows for false positive restraints that disagree with the majority of restraints to be down-weighted and prevents them from distorting the final fold. An outline of a contact-based model construction workflow is shown in Figure 6 D.

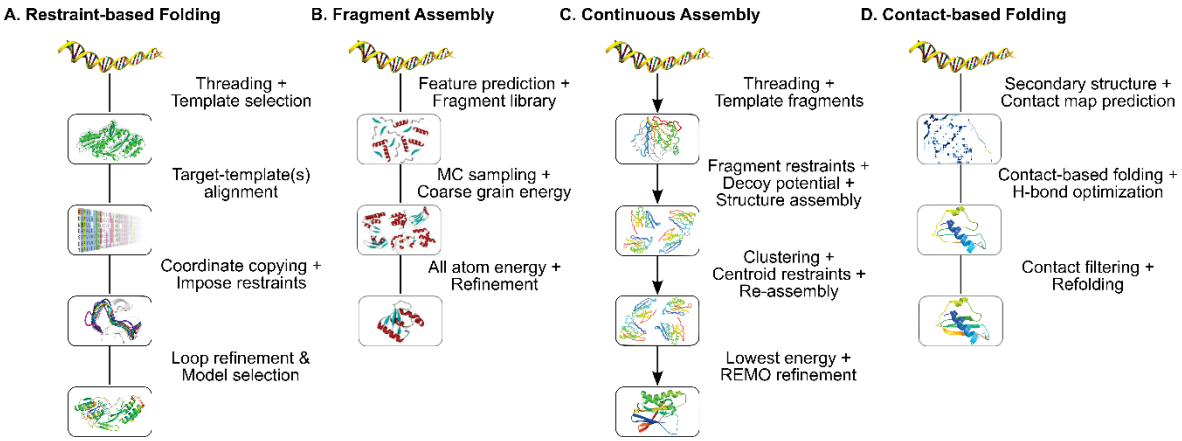


Figure 6. Outlines of different model construction workflows. The detailed description of each type of folding is described in the previous sections. **A. Restraint-based folding** (*e.g.* with Modeller) starts with the identification and selection of templates via threading, followed by the alignment between the target sequence and the selected templates. After alignment, the coordinates of the template(s) are copied as starting points and moved randomly until the distance restraints extracted from the templates using the alignment are fulfilled. Finally, loops with no structure are refined and the final model is selected from the resulting model ensemble. **B. Fragment Assembly** (*e.g.* with ROSETTA) starts with predicting structure features and generating a fragment library using profile-profile alignment between the target sequence and the fragment library, while matching predicted structure features **(Chapter 6.4.1 and 6.7)**. After generating the fragment library, Monte Carlo (MC) sampling is used with a coarse-grain energy function to replace dihedral angels in the target sequence with those of the fragments from the library. The lowest energy decoy is then refined by adding side-chains and using MC with an all-atom energy function. **C. Continuous assembly** (*e.g.* with iTASSER) starts with threading and extraction of high-scoring template fragments identified by the threading, which are used to generate C_α-atom traces of parts of the input sequence. Distance restraints from the fragments are used with a decoy potential to assemble the fragments on a lattice to generate initial decoy structures. The initial decoys are clustered and restraints from the largest cluster are combined with restraints from templates that align well to the centroid of the largest cluster to generate refined decoys by re-assembly of the fragments. Finally, the lowest energy decoy is selected and the full atom model constructed and refined using Replica Exchange Monte Carlo simulations (REMO). **D Contact-based folding** (*e.g.* with CONFOLD) starts with the prediction of secondary structure and a residue contact map. Then, a fully extended peptide is folded using simulated annealing with CNS, to fulfil secondary structure and residue contacts. After initial folding, contacts are filtered to remove those that disagree with the majority, and the folding is repeated using the reduced set of contacts.

6.6 MODEL QUALITY AND REFINEMENT

Once an ensemble of models, also known as decoys, has been produced for the query protein, a common approach is to predict which models are most likely to be correct in order to select these for model refinement [120-121]. The types of errors that typically appear in protein models span a wide range. At the one end of the spectrum are template selection errors, in which the selected templates do not share the same fold as the target sequence. In these cases, the entire model may have a wrong fold or topology. Even if the models do share the same fold as the

native structure, alignment errors may also cause misfolding of local regions. Misalignment of β-sheets and wrong rotations of α-helices are the more severe types of errors while flexible loops generally suffer less from misalignment. At the other end of the spectrum, high-quality models with little or no alignment errors may still suffer from errors in terms of atomic clashes, wrong loop or side-chain conformations, and poor hydrogen bonding.

6.6.1 MODEL QUALITY ASSESSMENT

The prediction of model quality is undertaken by so-called Model Quality Assessment Programs (MQAPs). Because the potential errors span a wide range, different MQAPs tend to focus on different types of errors. Template selection errors are for example often captured by evaluating the agreement between predicted features **(Chapter 6.4.1 and 6.7)**, and alignment errors are generally identified by poor energetics from knowledge- or physics-based potentials. Errors in loop and side-chain conformations or hydrogen bonding are often captured best by methods that evaluate the stereochemistry of the protein backbone and side-chains as well as atom clashes [122].

MQAPs generally tend to focus on the global quality of the protein, assigning a single score for each model in an ensemble of models in order to select one with the least amount of errors. Some methods, however, also predict the local model quality, aiming to identify both how much error is in a model and where in the model these errors occur. Prediction of local model quality is useful especially if multiple models with errors in different structural regions are to be combined.

A key difference between MQAPs is their target value. The target value is a measure of protein error or quality that can be measured when comparing the model to the native structure, but has to be predicted when the native structure is unknown. The different types of model quality scores fall into two over-all categories: Superposition-dependent scores and superposition-independent scores. Superposition-dependent scores are calculated by aligning the model to the native structure and evaluating a score depending on the distance between corresponding residues after alignment. Scores that fall into this category include the LG-Score [123], S-Score [124], TM-Score [125], GDT-TS Score [126], or MaxSub-Score [127]. Superposition-independent scores, instead, measure the consistency of intra-molecular distances and evaluates the structural similarity based on internal coordinates. This has the advantage that no structural alignment is required. This makes these scores less susceptible to being distorted by the structural alignment process, which for example produces artificially high errors for multi-domain proteins even if the domains themselves are

correctly folded, simply because the relative orientation of the domains differ between model and native structure. Scores that fall into this category include the Q-Score [128], lDDT score [129] and CAD score [130].

As with threading, meta-methods that use multiple primary MQAPs to predict errors in protein structures have been shown to be one of the major advances in model quality estimation and model selection for refinement [122]. This is because the focus on multiple different error types and target scores provides both a higher accuracy and a better consistency for different model quality ranges.

In the development of TopScore **(Chapter 9, Publication I)** I used an ensemble of 15 different primary MQAPs and combined their outputs using a two-stage deep neural network. By training the method on six diverse training datasets totaling over 1.5×10^5 models and 2.3×10^7 residues I obtained a much more accurate and consistent performance than any of the primary predictors.

6.6.2 MODEL REFINEMENT

Model Refinement has the goal of driving the best model or ensemble of models towards the global energy minima of the protein, essentially seeking to obtain a model more similar to a crystal structure. Model refinement generally falls into one of two classes, the first is rooted in molecular dynamics simulations (MD-based refinement), and the second is based on model fragmenting and/or averaging (Fragment-based refinement).

MD-based refinement [131-132] has seen marginal success for medium quality starting structures due to the inability to re-fold the starting structure. This is because MD-based refinement has to balance sampling and energy minimization in order to both be able to explore the energetic landscape, find the global energy minimum, and be able to stay in the global energy minimum once found. In other words, if the starting model is of very high model quality, MD-based refinement tends to over-sample the conformations and drives the model into nearby local energy minima, deteriorating the model quality. On the other hand, for starting models with very poor model quality, the energy landscape is too rugged and the sampling too weak to overcome the energy barriers involved in local refolding [133].

Fragment-based model refinement has had somewhat more success, especially in terms of refining models of poor quality [134]. The main reason for this is that fragment-based model refinement can re-fold parts of the structure and, therefore, break bonds that lock the protein into incorrect conformations. This is done using Monte Carlo sampling in the

ROSETTA refinement protocol, for example [134]. In this method, fragments of the structure are randomly replaced with new fragments sampled from an ensemble of decoy structures generated by sampling conformations of the input structure. This model ensemble is then subjected to a genetic algorithm, which iteratively refines the ensemble. In this algorithm, improved model ensembles are generated by combining low energy models (cross breeding), replacing fragments in a low energy model (mutation), or keeping the best low energy models from the previous ensemble (elitism). By consecutively applying this genetic algorithm until convergence, the initial model can be significantly improved towards models of low energy, corresponding to models close to the native structure.

In TopModel, refinement is performed by TopRefiner **(Chapter 10, Publication II)**, in which an ensemble of models is scored with TopScore **(Chapter 9, Publication I)** to identify regions with errors. These regions are then removed and the remaining pieces are used to construct a refined model. Repeating this process proved to significantly improve model quality.

6.7 CONTACT PREDICTION

In previous sections it was described how the prediction of protein features **(Chapter 6.4.1)** can improve the results of threading and model quality assessment. Residue-residue contacts, or just contacts for short, have had such a big impact on structure prediction that it is worthwhile to discuss it in a chapter of its own [59]. Unlike linear 1D features such as secondary structure, residue contacts are 2D, and thus, every residue pair in the target sequence has a value that needs to be predicted. This increase in dimensionality makes contacts more informative, for example, when used as scoring terms for threading [135], but also makes them more expensive to predict.

Accurate *ab initio* prediction of residue-residue contacts is one of the major breakthroughs in the field of *ab initio* protein structure prediction. In *ab initio* prediction the structure of the protein is determined without the use of any structural information [59]. The fundamental basis for the prediction of contacts is the concept of residue co-evolution. Co-evolution is the process, in which the mutation of a residue in a protein leads to a strong bias in which types of mutations proximal residues can adapt if the function and/or stability of the protein are to be maintained. This mutation bias can be detected in large sequence alignments using statistical methods such as direct coupling analysis (DCA) or mutual information (MI) analysis [136-139], often coupled with advanced machine learning techniques [15, 140-142], to provide information about the spatial proximity of residue pairs. Such

information can subsequently be used in down-stream *ab initio* folding simulations with programs such as ROSETTA [117] or CONFOLD [119] to reduce the conformational search space and drive the folding process towards the native state. Other uses of contact prediction include scoring terms for protein model quality assessment [143], contact-based template selection [144], and protein threading [135].

The benefits of residue-residue contact prediction have resulted in a large number of methods being developed in recent years [15, 136-142, 145-151]. One of the most promising advances in contact prediction is the use of deep neural networks, traditionally developed for image recognition [140]. Two slightly different approaches have been employed, in which all contacts in the contact map are either predicted at once [140, 142, 148, 151], or in which a receptive field (*i.e.* a 2D sliding window) scans across the contact map predicting each residue according to the local information in the map around it [150]. The latter has also been used for contact prediction with deep random forests [149]. Predicting all contacts at once is fast and memory efficient during training and evaluation, and allows for modelling of large and complex contact patterns potentially spanning the entire protein. However, it requires vast amounts of training data, as each protein is considered one sample. This can limit its ability to generalize to contact patterns not seen during training, especially for very sparse contact maps generated from small alignments. On the other hand, when using a receptive field, each residue pair is one sample, making it slow, memory demanding, and limited in its ability to explicitly model contact patterns larger than the receptive field size. This limitation however, can improve the models ability to generalize to contact patterns not seen during training, and directly prevents over-training, since no whole-protein pattern is seen by the network.

The accurate prediction of residue-residue contacts and residue-residue distances using deep neural networks has led to a revolution in protein *ab initio* folding [59]. This revolution is founded on the ability to explicitly predict long-range interactions, *i.e.*, those between residues far apart in the target sequence. This improves conformational sampling and allows contact-based methods such as CONFOLD to compete with traditional fragment-based assembly methods such as ROSETTA at a fraction of the computational cost.

6.8. MACHINE LEARNING

Machine learning refers to a number of different techniques which, once trained on a dataset, can convert a set of inputs (features) to an output (prediction) either linearly, *e.g.* classical curve fitting, or non-linearly. It can be thought of as a general-purpose fitting methodology,

which can be used to make prospective predictions. Over all, while machine learning is a very broad topic that cannot be explored in depth in this chapter, it can be broadly categorized into three main areas of research: Random Forests (RF), Support Vector Machines (SVMs) and Deep Neural Networks (DNNs).

SVMs are non-linear classifiers that map the input feature vector into a high-dimensional space and use a hyperplane as a separator for classification by maximizing the distance of data points with different labels to the hyperplane. The data is mapped using a kernel function to keep computational load manageable, and to ensure that the dot product between two vectors can be computed easily. The hyperplane is then calculated as the set of orthogonal vectors that define the plane.

RFs are a generalization of decision trees and are therefore greedy classification algorithms. They decide binary split points for the input features to generate decision trees in which the predicted labels are on each of the leaf nodes. Multiple trees are made (hence the forest), and bootstrapping and random feature subsets are used for each tree to prevent the overfitting on the training data. The benefits of RFs are that they are invariant to normalization and type of input data and as such are applicable to many types of problems, especially ones in which the input features are highly heterogeneous.

Both RFs [152-155] and SVMs [146, 156-158] have seen widespread use in bio-informatics. They do however, come with the drawback that for highly complex tasks they do not scale well. Both types of models scale with the amount of training data and the complexity of the task. For SVMs, the dimensionality of the hyperplane scales with the square of the number of training data points and can become so high that the evaluation on new data points becomes exceedingly slow. Similarly, for RFs, the trees become very large, and the forest size therefore has to be increased to prevent over-fitting, which in turn increases the model size further. Although the evaluation of the RF model is generally fast, the amount of memory required for storing the forest becomes a limiting factor. The practical limits of memory and runtime therefore means that for complex tasks with large amounts of data, such as for example protein contact prediction, RFs and SVMs have poor performance compared to methods such as DNNs, in which the model size remains fixed for a given task, irrespective of the complexity or amount of training data [142].

Recently, DNNs have been the most popular machine learning technique due to their flexibility, high performance and fast compact models. Furthermore the increase in chip

speed has enabled more complex models to be used [159]. For these reasons, DNNs are used extensively throughout TopSuite.

6.8.1 DEEP NEURAL NETWORKS

Deep Neural Networks (DNNs), as the name suggests, are a class of algorithms that were inspired by the connection of neurons in the brain. Each neuron in a DNN takes a set of values as an input (analogous to a brain neuron receiving a signal from multiple other neurons), processes that input by calculating a weighted sum, passes that sum through a so-called activation function, and sends the new transformed signal to one or more other neurons. During training, the connections between the neurons are initially given random weights, and these weights are progressively updated as the network is trained on more and more data. This training is done using the backpropagation algorithm, which calculates the signal gradient with respect to a loss function. The loss function is defined according to the target value. A simple loss function could for example be the difference between the predicted and true value. The backpropagation algorithm adjusts the weights of the network such, that the network with the adjusted weights give a prediction with a lower loss than before the weight adjustment.

The neurons in DNNs are generally arranged in layers, and the signal is propagated from the neurons in one layer to the neurons in the next. Depending on the number of layers and the connections between the neurons in each layer, very complex patterns can be fitted by training the network. The activation function of the neurons, the pattern of connections between the neurons, the number of neurons in each layer, and the number of layers can all be varied to obtain different types of models that can solve different types of problems.

While the number of different types DNNs have increased dramatically in the last decade, most networks fall into one of five different categories: Deep Belief Networks (DBNs), Deep Convolutional Neural Networks (DCNNs), Deep Convolutional Auto-encoders (DCAEs), Deep Recurrent Neural Networks (DRNNs), and Generative Adversarial Networks (GANs) [160]. In this section, each type of network will be briefly discussed with a focus on the applicability of these types of networks to bioinformatics problems.

DBNs were one of the first types of neural nets to be developed. They are generally used for predictions, in which the input vector has a fixed size, and where the input features share no spatial relationship. Examples of these types of problems could be classification of medical samples or linking symptoms to diseases [161]. TopThreader **(Chapter 10, Publication II)** uses a series of DBNs to predict the structural similarity between a putative template structure and the native structure and uses this information to remove false positive

templates and rank the templates according to suitability for modeling. TopScore **(Chapter 9, Publication II)** uses two stages of DBNs to predict the structural quality of an ensemble of models both at the residue-wise and whole-protein level.

DCNNs are particularly useful for problems, in which the input feature vector may vary in size and where the input features have a spacial connection. An example of these types of problems is for example image classification [162], where the images may have different sizes and where the meaning of each pixel is highly correlated to the pixels around it. These types of neural networks have been extensively used for prediction of protein-protein contacts[59, 140, 142, 163] and protein features such as secondary structure and solvent accessibility[98]. TopContact and TopDomain use DCNNs to predict protein structural features such as domain boundaries, secondary structure, transmembrane topology, dihedral angles, solvent accessibility and residue-residue contacts and distances. This is because residue proximity in sequence provides the spacial connection between the input pixels that is required for convolution.

DCAEs are generally used for feature reduction problems, in which a high-dimensional input needs to be reduced to a more manageable size with minimal loss of information. Examples of these types of problems include image compression [164] and image clustering [165].

DRNNs are typically used for problems, in which the input varies in length and share a sequential relationship. Common examples of the use of DRNNs include speech recognition [166], text data mining [167] and genomic sequence analysis [168]. DRNNs have been instrumental to predict linear features of proteins **(Chapter 6.4.1)** as their sequential relationship to each other is captured nicely by this type of network [102, 169-170].

GANs are used often for signal processing, as it learns to generate new samples that follow the same statistical distribution as its training data. This enables it to for example generate images [171], fill in missing parts of an image or improve image resolution [172].

A key issue with standard deep neural network learning is the vanishing gradient problem, which limits how deep (*i.e.* how many layers) a network can be before accuracy stagnates or even declines [173]. Residual convolutional neural networks, originally developed for image recognition [174] bypass this problem by passing along the input signal together with the transformed signal after each neural transformation. This allows for very deep models to be built, and has shown great results for bio-informatics applications such as secondary structure [175] and residue contact prediction [140].

An issue with deep learning is that the random initialization of neurons can lead to performance differences for different DNNs with the same network architecture and training data. These effects have can be minimized by training multiple models with different random seeds and averaging the output [140, 151]. However, with the development of dropout [176], this problem can be solved directly by randomly "switching off" neurons during model training. This forces the network to learn using new neural connections, and effectively learn different models simultaneously, which improves the ability of the neural network to generalize to data not seen during training. The use of dropout has therefore become the standard in image recognition and protein contact prediction [140].

Furthermore, because DNNs have thousands of different fitting parameters (the weights of network neurons) they are prone to over-fitting. Over-fitting happens when the DNN memorizes the training data in order to obtain a perfect performance, which in turn makes it unable to perform reliably for new data. There are several options for avoiding over-fitting, of which two are so common, that they have become standard in the field: Early Stopping and Regularization. In Early Stopping, the networks performance is evaluated on the fly on an independent set of test data and the training is stopped early (hence the name) when performance on the test data starts to deteriorate. Regularization is a mathematical trick, which augments the loss function in order to penalize complexity, by assuming that simpler models are better than complex ones. In practice, this is done by adding a term to the loss function, which slowly pushes the weights of the neuron connections towards zero as the model is trained. In this way, only connections that are critical to the performance of the network (where the cost of switching off the signal is too great) end up being used, while in practice connections that are not important are turned off.

7. SCOPE OF THE THESIS

Since the inception of the field of bioinformatics, two decades of method development has left researchers and scientists, who wish to predict the structure of a protein, with one major question: "*Which method should I use to predict the structure of my protein?*" The large variety of different method for structure prediction led to the answer: "*Use a consensus of different methods*". Method performance, however, depends on both algorithm design, training data, input features and target value, and many methods produce highly correlated results. This correlation between methods can lead majority-based consensus, which assumes each method has a fixed chance of being right, to converge on wrong predictions. Machine learning can go one step further than consensus and learn not only how well a method performs on average, but also in which context it performs well. This enables methods based on, for example, deep neural networks to perform much better than traditional consensus. This is the dominant idea behind the development of TopSuite.

The philosophy of TopSuite is to collect and integrate many diverse primary predictors for a given task and make it easy for a user to provide input data to them all. Then, rather than using majority-based consensus, the output of the different predictors is used as input for deep neural networks, which are trained on large diverse databases to produce high quality meta-predictions. These predictions are then presented to the user in a format that is easily transferrable from one task to the other, in order to link the different programs together seamlessly. The aim of this thesis is to:

1. Develop a deep neural network-based meta-method for determining the quality of a protein structure prediction and identify which parts of a model contain errors **(TopScore, Chapter 9, Publication I)**.
2. Develop a fully automated deep neural network-based meta-method for template-based protein structure prediction using TopScore as the core scoring function **(TopModel, Chapter 10, Publication II)**.
3. Illustrate the gradual development of TopSuite and demonstrate the usefulness of the developed methods by applying them to target proteins of high biological (**Publication III**), medical (**Publication IV**), and industrial (**Publication V**) interest.

8. TOPSUITE

TopSuite is a suite of programs, which has been and is being developed in order to make automated high-quality protein structure prediction easy and accessible to non-expert users and the scientific community as a whole.

TopSuite consists of several different modules, some of which are described in the papers of this thesis, and some of which are still in development and therefore not yet published. In some cases, preliminary versions of the programs have been used in projects such as the prediction of the dimeric state of the GAF domain of ETR1 **(Chapter 21, Publication V)**. The modules of TopSuite can be classified into three major categories: Protein Feature Prediction, Protein Structure Prediction, and Protein Interaction Prediction. Within these three categories, the different modules of TopSuite are as follows (.odules still in development are marked with a *):

1. **Protein Feature Prediction**
    I. **TopDomain*** predicts the location of domain boundaries in the input sequence using a combination of *ab initio*-, co-evolution- and template-based primary predictors and uses a two-stage DNN approach to perform high quality predictions that approximate expert human domain annotations as closely as possible.
    II. **TopContact*** predicts residue-residue contacts and residue-residue distances as well as secondary structure (α-helix, β-strand and coil), relative solvent accessibility and backbone dihedral angles in concert. This is done based on 25 different primary predictors and 3 stages of DNNs to combine the outputs.
2. **Protein Structure Prediction (Chapter 9 and 10, Publication I and II)**
    I. **TopThreader** predicts templates and alignments between the templates and the target sequence. This is done using 12 different primary predictors and predicting the template similarity to the native structure using multi-stage DNNs.
    II. **TopAligner** calculates alignments between provided input structures and a target sequence using different sequence and structure based multiple sequence alignment programs.
    III. **TopBuilder** constructs 3D models of the input alignments using MODELLER and ROSETTA. TopBuilder also provides an easy interface for side-chain refinement and MD-based refinement.

IV. **TopScore** predicts the global and local structural similarity to the native structure using 12 different primary predictors and combines the outputs using a two-stage deep neural network. TopScore consists of two scoring functions: TopScore and TopScoreSingle, the latter of which uses no clustering or ensemble information. TopScore and TopScoreSingle allows for selection of high-quality models as well as identification of errors in specific regions of the models.

V. **TopRefiner** combines and refines an input ensemble of models selected from single-template models and multi-template models calculated using TopThreader, TopAligner and TopBuilder. It does so by effectively identifying poorly modelled regions using TopScore and TopScoreSingle, removing these regions and replacing them with better modelled regions from other models in the ensemble.

VI. **TopModel** predicts protein structure by applying TopThreader, TopAligner, TopBuilder, TopScore and TopRefiner to produce template-based protein structure predictions in a fully automated manner.

3. Protein Interaction Prediction

I. **TopInterface*** predicts protein-protein interactions between two input structures using a combination of conservation-based, co-evolution-based and template-based primary predictors and a three-stage deep neural network for combining the input features into probabilities of residue-residue contacts.

II. **TopDock*** predicts protein-protein complexes using predicted contacts from TopInterface. TopDock uses a deep neural network to predict the best docking solution from fulfillment of predicted contacts from TopInterface, docking energy and model clustering.

III. **TopLigand*** predicts protein-ligand interactions by predicting the binding site and ligand pharmacophore features given a model from TopModel as an input. In doing so, TopLigand opens closed binding pockets and optimizes side-chain conformations to facilitate ligand binding. The binding site and pharmacophore prediction is done using a 3D Deep Convolutional Neural Network.

TopModel is the core workflow of TopSuite as it integrates most of the modules seamlessly and therefore allows for fully automated structure prediction with a single

command. The interaction between the different modules in the TopSuite workflow is shown in Figure 7.

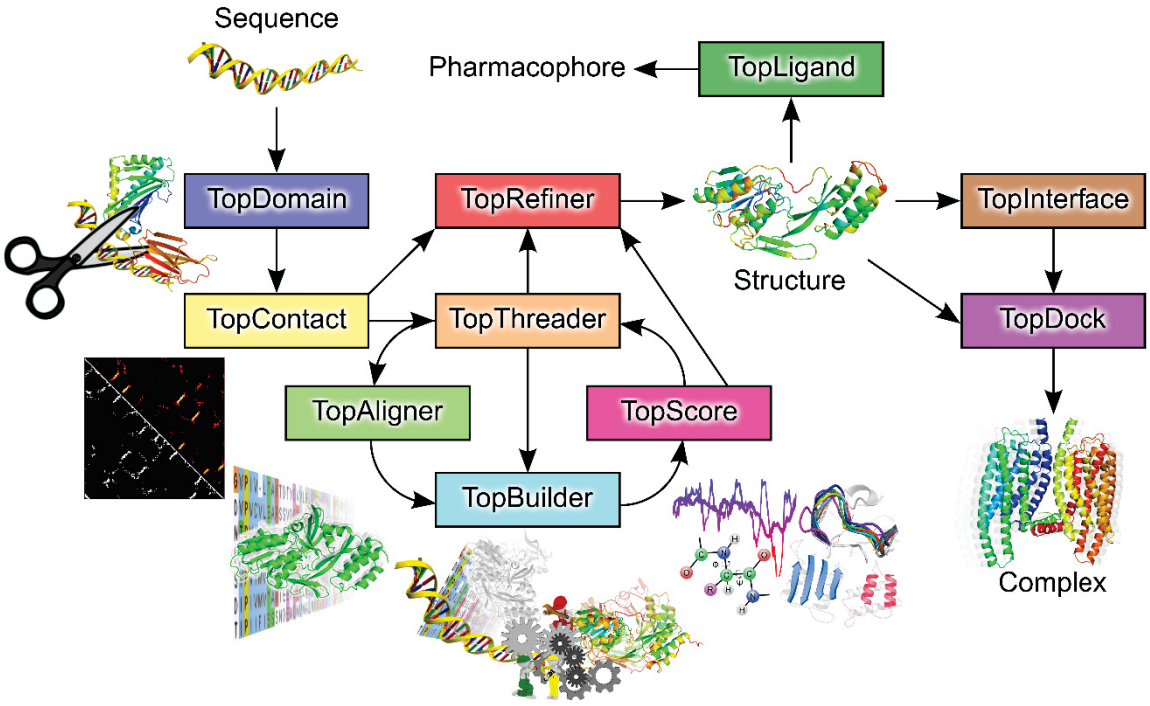


Figure 7. Simplified interaction between TopSuite modules. The target sequence is given as input to TopDomain and the sequence is separated into domains. Each domain is then given as input for TopContact to predict secondary structure, dihedral angles, residue contacts and residue distances. The sequence and the predicted features are then given as input to TopThreader, which searches for templates using different primary threaders. TopThreader uses TopBuilder to build models from the primary threader alignments, template structures and target sequence, which are scored with TopScore, and used by TopThreader together with primary threader scores to rank and cluster templates and remove false positives. TopThreader then uses TopAligner to align templates and construct consensus alignments, which are built with TopBuilder, scored with TopScore, and used together with primary threader scores in TopThreader to rank templates by predicted similarity to the native structure. After template selection, TopAligner is used to generate a large ensemble of pairwise and multi-template alignments from which models are built with TopBuilder and scored with TopScore. Models are selected from the *ab initio* predictions from TopContact, the multi-template ensemble, and the single-template models by TopRefiner, which combines and refines the models to produce a final structure. Predicted structures can then be used as input for TopInterface to predict protein-protein contacts, and the predicted contacts and structures can be used as input for TopDock to produce a protein-protein complex. The predicted structure can also be used as input for TopLigand to predict ligand binding sites and pharmacophore models, which can in turn be used for virtual compound screening.

TopSuite was used in the following publications. Publications described in this thesis are marked with an *:

* Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion. Holger Gohlke, Ulrike Hergert, Tatu Meyer, **Daniel Mulnaes (10%),** Manfred K. Grieshaber, Sander H.J. Smits and Lutz Schmitt. *Journal of Chemical Information and Modelling*. 2013, 53, 2493–2498.

* Determinants of FIV and HIV Vif sensitivity of feline APOBEC3 restriction factors. Zeli Zhang, Qinyong Gu, Ananda Ayyappan Jaguva Vasudevan, Anika Hain, Björn-Philipp Kloke, Sascha Hasheminasab, **Daniel Mulnaes (5%),** Kei Sato, Klaus Cichutek, Dieter Häussinger, Ignatio G. Bravo, Sander H.J. Smits, Holger Gohlke and Carsten Münk. *Retrovirology*; 2016, 13, 46.

* Recognition motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1. Dalibor Milić, Markus Dick, **Daniel Mulnaes (10%)**, Christopher Pfleger, Anna Kinnen, Holger Gohlke and Georg Groth. *Scientific Reports* 2018, 8, 3890.

Molecular dynamics simulations and structure-guided mutagenesis provide insight into the architecture of the catalytic core of the ectoine hydroxylase. Nils Widderich, Marco Pittelkow, Astrid Höppner, **Daniel Mulnaes (10%),** Wolfgang Buckel, Holger Gohlke, Sander H.J. Smits, Erhard Bremer. *Journal of Molecular Biology* 2014, 426, 586-600

Structural basis of lantibiotic recognition by the nisin resistance protein from Streptococcus agalactiae. Sakshi Khosa, Benedikt Frieg, **Daniel Mulnaes (10%)**, Diana Kleinschrodt, Astrid Höppner, Holger Gohlke, Sander H.J. Smits. *Scientific Reports* 2016, 6, 18679.

9. TOPSCORE: USING DEEP NEURAL NETWORKS AND LARGE DIVERSE DATA SETS FOR ACCURATE PROTEIN MODEL QUALITY ASSESSMENT

Daniel Mulnaes (85%), and Holger Gohlke

Journal of Chemical Theory and Computation; 2018, 14, 6117-6126.

9.1 BACKGROUND

In computational structure prediction, it is of vital importance to determine, how close to the real protein structure the predicted model can be expected to be. High-quality models built from closely homologous protein structures are often suitable for investigation of small molecule binding and can therefore serve as starting points for drug-discovery [13]. However, models built from distantly related proteins or without any template may contain errors that limit their ability to answer such detailed biological questions. Models with a medium degree of error can still be useful to answer several biological questions though, *e.g.* understanding effects of disease-associated mutations, functional annotation, or to aid the experimental elucidation of the structure, but are generally not suitable, if fine-grained atomistic information is required.

Errors in protein structure models can range from small differences in side-chain conformations or flexible loop orientations, to frame-shift errors in which misalignment causes residues to be located in wrong secondary structure elements or on the wrong side of β-sheets. On the largest scale, template selection errors, in which the model is based on a wrong template, can cause most or the entire model to be wrongly folded. **(Chapter 6.6.1 Model Quality Assessment)** Drawing conclusions based on such a model can lead researchers to completely wrong conclusions.

Because the errors in protein models span a wide range, different types of errors are detected by different types of Model Quality Assessment Programs (MQAPs). Minor errors in side-chain orientation can be identified by examining bond lengths, bond angles, and steric clashes, while frame-shift errors can be detected by examining energetic interactions between residues using knowledge-based potentials, since these errors generally lead to less favorable interactions between residues. At the fold level, errors can be detected by examining self-consistency between features of the model that can be predicted from the primary sequence. Such features include secondary structure, solvent accessibility, contact

density and residue-residue contacts, and those same features measured in the structural model itself. The reasoning behind this methodology is that if the model is consistent with the independently predicted features, it is more likely to be correct. Finally, errors can also be estimated by examining multiple independent structural models of the same protein. The models could for example stem from different folding simulations or be based on different templates or alignments. Regions of the protein that adopt different folds in an ensemble can be considered less confident than those that adopt the same fold in many models.

An issue facing MQAPs, other than the detection of errors, is the conversion of measured error features into a geometric measure of error that is useful and intuitive to understand. These features include atomic clashes, wrong stereochemistry, unfavorable energetics, disagreement with predicted features and structural inconsistency between independent models, as described in the previous section. Several geometric quality measures have been used in the past. These can be divided into two main groups: Superposition-dependent quality measures, such as the TM-Score and the GDT_TS Score, calculate differences in atomic location after superimposing a model to the known native structure using an algorithm that optimize these measures. Superposition-independent measures on the other hand, such as the lDDT, and CAD scores, evaluates intra-molecular distances and interactions and therefore calculate the consistency between the model and the native structure using internal coordinates. The benefit of superposition-dependent measures is that they reward correct spacial placement of secondary structure elements and domains. However, they over-penalize multi-domain structures, since the super positioning of one domain often leads to very large distance differences for other domains even when these are correctly folded. Superposition-independent measures, which focus on internal coordinate consistency do not suffer from these issues, making them ideal for estimating errors and correct folding.

The goal of TopScore is to identify many different types of errors in predicted protein models with a single program that uses different primary predictors (primary MQAPs), which focus on different types of errors and uses different definitions of model quality.

9.2 RESULTS

In this work we developed two meta-Model Quality Assessment Programs (meta-MQAPs) called TopScore and TopScoreSingle. Meta-methods combine scores from multiple different primary predictors to produce more consistent and accurate predictions than any single method. The output of the different predictors was combined using a two-stage DNN

approach to predict both the global error of the protein model as well as the local error of individual residues. The predicted target score was chosen to be 1-lDDT score. The lDDT score calculates intra-protein all-atom interatomic distance conservation using four different distance cut-offs making it a highly sensitive superposition independent score. We chose 1-lDDT to have low scores correspond to low amount of error in the protein. To ensure robust performance across many different types of models we constructed a composite dataset of model ensembles from many different sources. These include previous CASP experiments, *ab initio* folding trajectories from I-TASSER, model decoy datasets, high-quality homology models, homology models based on distantly related templates, and artificially misfolded decoys from the 3DRobot dataset. Our results show that different primary MQAPs perform very differently, depending on which dataset they are tested on. By optimally combining the outputs of the different methods using the DNNs, we obtained a much more consistent performance across different datasets. Furthermore, we obtained a performance that is significantly better than any of the investigated primary predictors. An excerpt of the performance of TopScore and TopScoreSingle compared to some of the best performing primary predictors on the different datasets in terms of different quality measures is shown in Figure 8 for whole-protein scores and Figure 9 for residue-wise scores. The correlation between the whole-protein TopScore and the true value on the combined dataset as well as an example of the residue-wise performance can be found in Figure 10.

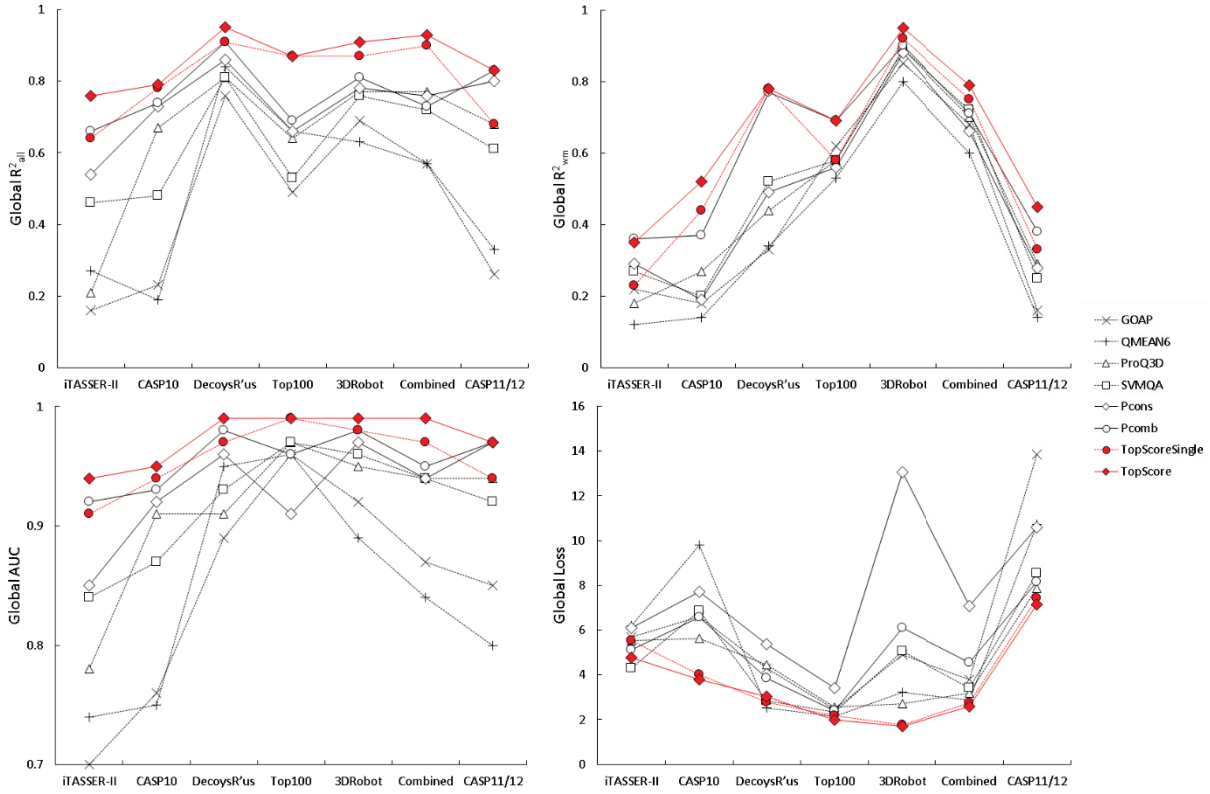


Figure 8. TopScore global performance. TopScore (red circles) and TopScoreSingle (red dashes) global performance compared to a subset of primary predictors (black). Dashed lines represent single-model methods and full lines methods that use clustering information. The 95% confidence intervals were calculated using the Fischer r-to-z transformation. The widest confidence interval for any R_{all}^2 or R_{wm}^2 was 0.01 and 0.12, respectively. Statistical significance was determined by the two-sided Steiger test [177]. Accordingly, the R_{all}^2 and R_{wm}^2 of TopScore and TopScoreSingle are significantly different from any primary MQAP for the combined dataset ($p < 0.05$). In terms of R_{all}^2, for the CASP11/12 dataset, TopScoreSingle is not significantly different from ProQ3D, and neither is TopScore when compared to Pcomb.

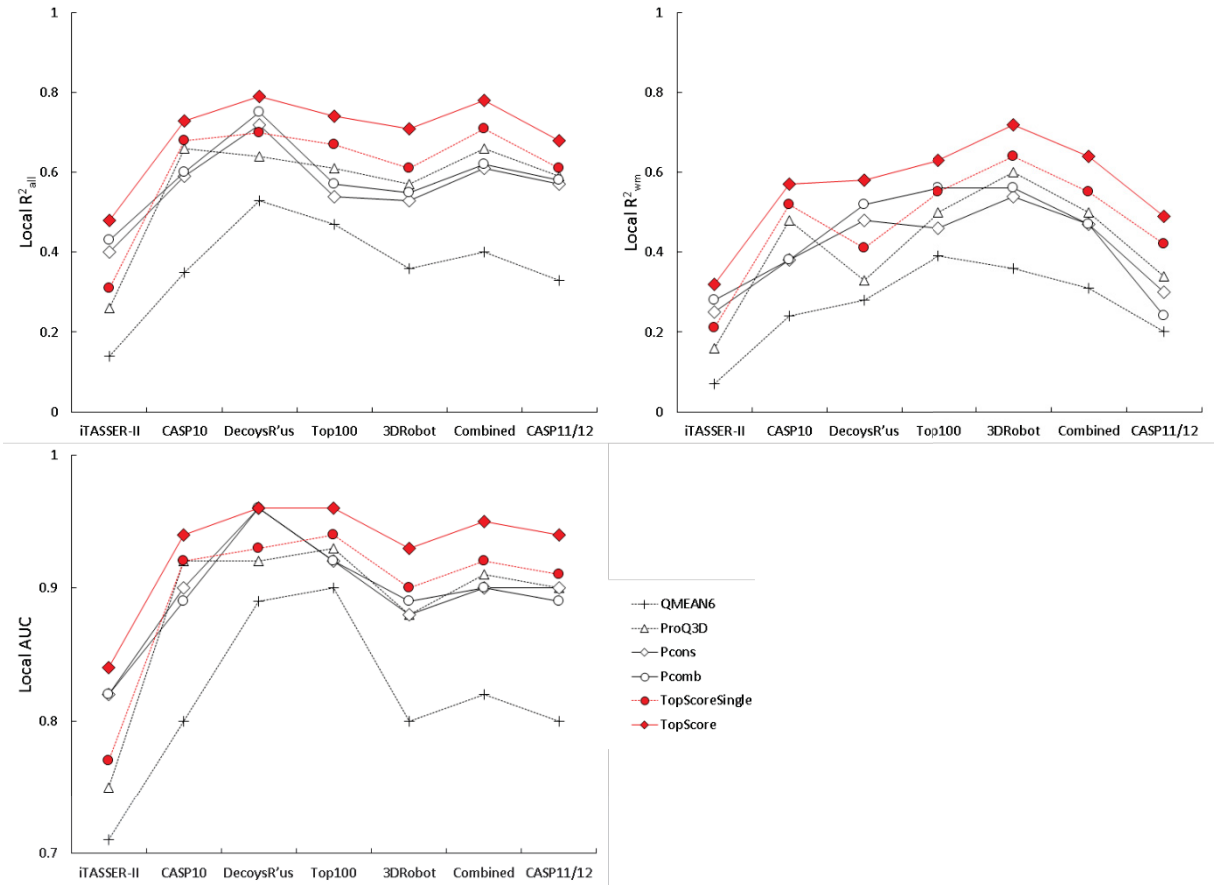


Figure 9. TopScore local performance. TopScore (red circles) and TopScoreSingle (red dashes) local performance compared to a subset of primary predictors (black). Dashed lines represent single-model methods and full lines methods that use clustering information. The 95% confidence intervals and statistical significances are calculated in the same way as for Figure 3. The widest confidence interval for any R_{all}^2 or R_{wm}^2 was 0.001 and 0.17, respectively. The R_{all}^2 and R_{wm}^2 of TopScore and TopScoreSingle are significantly different from any primary MQAP ($p < 0.05$).

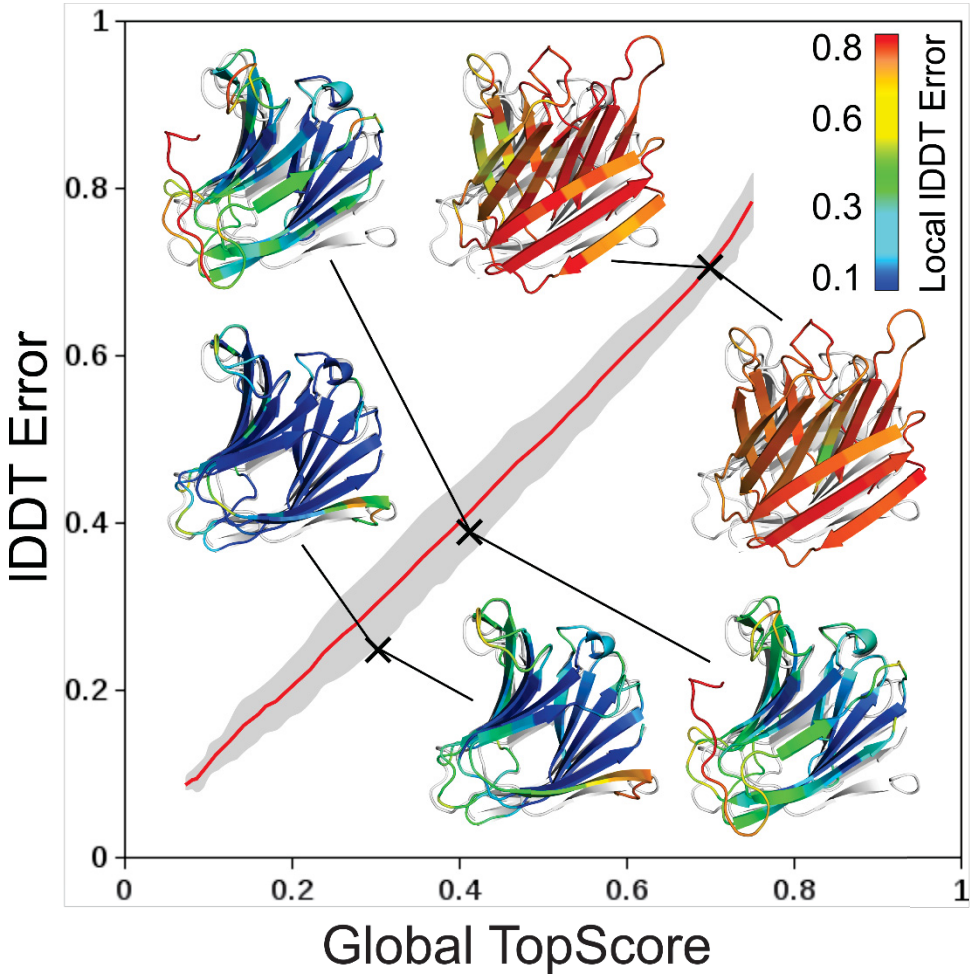


Figure 10. TopScore performance. The global TopScore predictions plotted against the lDDT error of the models for the combined dataset. Three randomly selected example models of PDB ID 4BMB from the 3DRobot dataset are shown colored according to local TopScore error prediction (lower triangle) and true local lDDT error (upper triangle).

9.3 CONCLUSIONS AND SIGNIFICANCE

The development of TopScore and TopScoreSingle is a key part of TopModel. These scoring functions are essential to solve four important steps of the structure prediction workflow:

1. In template selection, TopScore and TopScoreSingle help to discard templates that produce wrongly folded models, thus improving the template selection especially for difficult targets.

2. In template-target alignment, TopScore and TopScoreSingle helps to both identify and rectify parts of alignments that contain errors and produce badly scoring models. This helps produce consensus alignments that favor good scoring models.

3. In model selection, TopScore and TopScoreSingle help to select high-quality models for refinement and model combination, which is required in order to perform high-quality model refinement.

4. In model refinement, TopScore and TopScoreSingle help to identify parts of the input models that contain errors, enabling these regions to be removed such that new refined models can be constructed from the remaining parts.

TopScore and TopScoreSingle were both shown to perform significantly better than any of their primary predictors (see Figures 8 and 9), and to be more consistent in their performance across many different model datasets. This robustness is due to the large dataset used for training (~1.5×10^5 models and ~2.3×10^7 residues) the methods as well as the large diversity of different primary predictors. This makes TopScore and TopScoreSingle ideal scoring functions when predicting protein structures using TopModel.

10. TOPMODEL: TEMPLATE-BASED PROTEIN STRUCTURE PREDICTION AT LOW SEQUENCE IDENTITY USING TOP-DOWN CONSENSUS AND DEEP NEURAL NETWORKS

Daniel Mulnaes **(60%)**, Nicola Porta, Rebecca Clemens, Irina Apanasenko, Jens Reiners, Lothar Gremer, Philipp Neudecker, Sander Smits, Holger Gohlke.

Journal of Chemical Theory and Computation; 2019, Submitted

10.1 BACKGROUND

Protein structure prediction is a core part of bioinformatics that has been in development since the initial conception of the field. This in turn has led to an abundance of different algorithms for solving the different challenges commonly faced in structure prediction, most notably template identification, sequence alignment, model construction, and model refinement. However, no single method consistently outperforms all other methods for every given protein target. In other words, different methods produce the best results for different proteins. It is therefore not surprising that meta-methods, which utilize multiple different algorithms, such as the MULTICOM and Zhang Servers, have shown some of the best over all performances in every CASP competition since their conception.

These methods, however, mainly function as black box online servers **(Chapters 6.2 CASP and 6.4.2 META-SERVERS)**. This is due to the high competition in the field, which discourages the sharing of workflows and methods. This in turn means that users, who do not wish to send their data to remote servers or need large-scale calculations for many proteins, can still be at a disadvantage. Furthermore, meta-methods such as the Zhang and MULTICOM servers generally operate using consensus information from their different primary predictors, which is based on the assumption that the majority is more likely to be correct. In practice, this means that templates that are identified more often by different methods are more likely to be used. While this can often be beneficial, it means that if the correct fold is in the minority, then the consensus drives the model away from the true fold.

These shortcomings incentivized the development of TopModel, which is the core part of TopSuite (Figure 7) and contains the workflow needed for fully automated template-based structure prediction using a top-down consensus methodology. The top-down

consensus methodology aims at identifying the best template and fold using deep neural networks, and then selects templates or models based on their agreement with the top ranked one. Therefore, TopModel can go against the majority and improve models beyond the initial model ensemble, if a good estimate of the template quality and model quality can be calculated.

10.2 RESULTS

In this work, we developed a meta-method for automated template-based protein structure prediction called TopModel. TopModel is available as a stand-alone toolbox for the scientific community and utilizes most of the available stand-alone algorithms for template identification, sequence and structure alignment and model construction. TopModel makes using them easy and intuitive, requiring only a single command-line call for complete structure prediction, while at the same time allowing the user to use each module individually.

TopModel provides much better template selection than its constituent primary methods due to the sophisticated threading module TopThreader, which uses both model quality assessment with TopScore and TopScoreSingle as well as deep neural networks for estimating template quality. Furthermore, due to the use of top-down consensus TopThreader has very few false positives. This can be seen when comparing the ability to select the best template between TopModel and its constituent primary predictors (Figure 11).

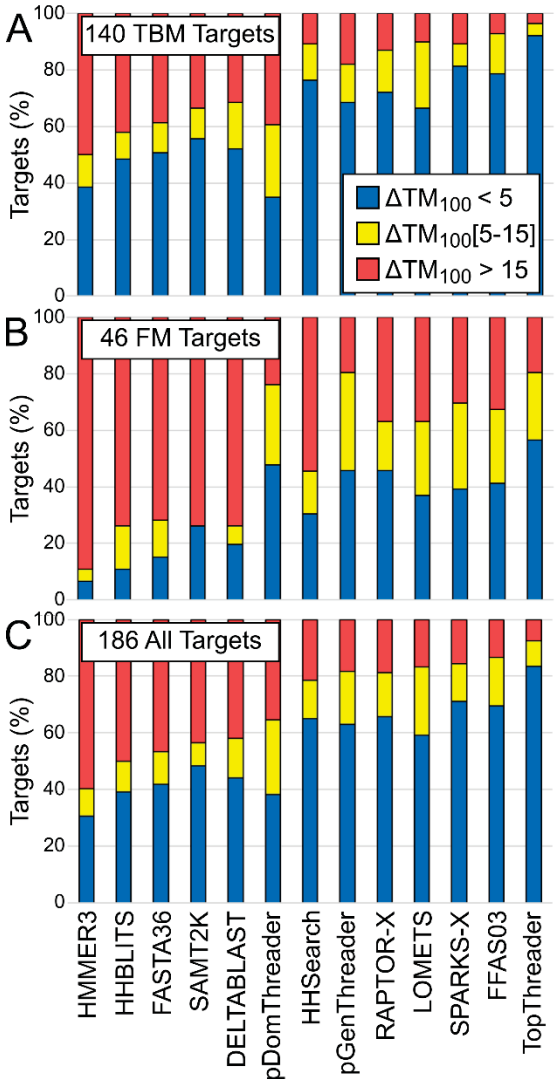


Figure 11. Template enrichment by TopThreader compared to primary threaders. Comparison of template selection performance on the CASP dataset. Performance is evaluated based on the ΔTM_{100} score, which evaluates the difference between the best of the top five ranked templates of a given threader, and the best template found by any threader. For each target, three categories are selected: (I) the best template is found ($\Delta TM_{100} < 5$), (II) an adequate template is found (ΔTM_{100} [5-15]), and (III) no adequate template is found ($\Delta TM_{100} > 15$). The values represent percentages of targets in the CASP dataset for TBM (A), FM (B), and all (C) targets, respectively. Differences between TopThreader and the best primary threader for each subset are highly significant ($p < 0.01$) according to the Ghent implementation of the Freeman-Halton exact test for 3x3 contingency tables [178].

After template identification with TopThreader, TopModel constructs an ensemble of different multi-template alignments using the TopAligner module. In doing so, model quality can be improved since the use of multiple templates not only has the potential to increase the coverage of the target sequence, but also has the ability to improve the target-template alignment. After generating the alignment ensemble, the alignments are used to

generate models with TopBuilder, which works as an interface to ROSETTA and MODELLER, and the resulting models are scored with TopScore and TopScoreSingle.

Finally, TopModel uses an iterative refinement protocol called TopRefiner in which the best scoring single-template models generated by TopThreader as well as the best scoring multi-template models generated by TopAligner and TopBuilder are selected according to TopScore and TopScoreSingle rankings. From these models, regions predicted to contain errors by TopScore and TopScoreSingle are deleted, and the remaining pieces are used to construct meta-models with fewer errors. This process is done iteratively to refine the models, after which fragment-guided MD refinement with ModRefiner is performed to provide a single refined model to the user. The effect of generating multi-template models as well as refining the best models using TopRefiner can be seen in Figure 12.

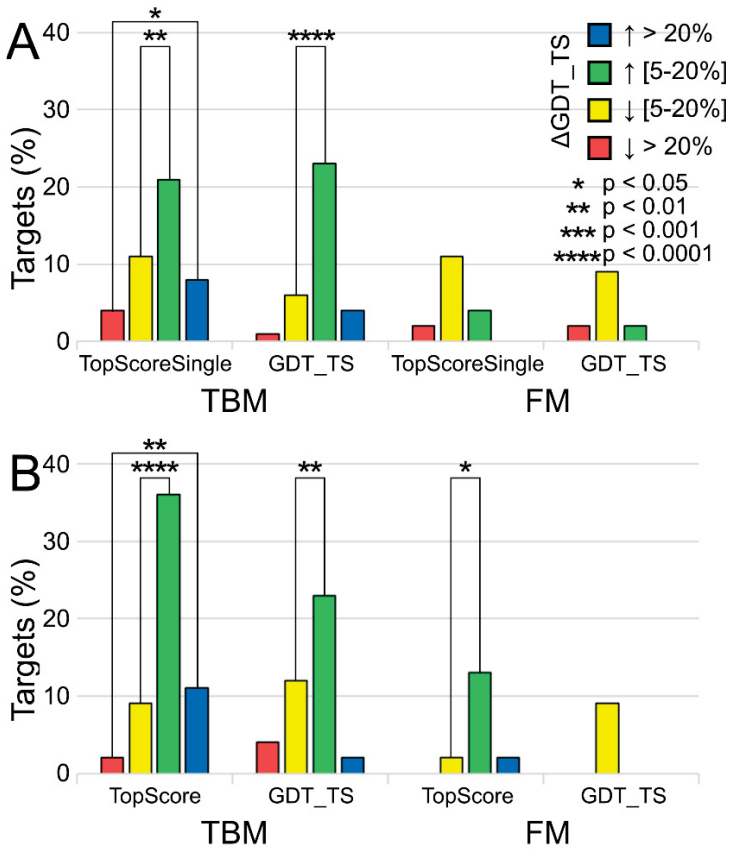


Figure 12. Impact of using TopAligner and TopRefiner on model quality. The relative change in GDT_TS score (ΔGDT_TS) is calculated by comparing a model selected before and after running TopAligner (A) or TopRefiner (B), respectively. **A.** Difference in model quality when selected from a multi/single-template model ensemble from TopAligner/TopThreader compared to selection from a single-template pairwise primary threader model ensemble. **B.** Difference in model quality when selected from the first stage of TopRefiner (before refinement) compared to selection from the last stage of TopRefiner (after refinement). The models are selected either by true GDT_TS or by TopScoreSingle (A) or TopScore (B). Five categories are defined based

on the ΔGDT_TS: No change (ΔGDT_TS < 5%), small increase/decrease (ΔGDT_TS ↑/↓ [5%-20%]; green/yellow), large increase/decrease (ΔGDT_TS ↑/↓ > 20%; blue/red). The "No change" category is the most abundant and is not shown as it reflects no significant change in model quality. Significance is calculated using a one-tailed *t*-test between corresponding increase/decrease categories (blue-red and green-yellow, respectively). The null hypothesis is that the probability of model quality increase of a given amount (5-20% or >20% ΔGDT_TS) is the same as the probability of quality decrease by the same amount. Pairwise comparisons where this hypothesis can be rejected are indicated with brackets and corresponding p-values (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$). The number of samples used is the number of CASP targets in the TBM (140) and FM (46) categories, respectively.

TopModel was validated on the CASP10-12 datasets and showed good performance compared to its primary methods. However, since TopModel is a template-only method, *ab initio* targets for which no good template structures could be identified showed worse models from TopModel compared to servers, which use *ab initio* methods for protein contact and distance prediction as well as domain parsing for large multi-domain structures. The results of the validation on the CASP10-12 dataset can be seen in Figure 13.

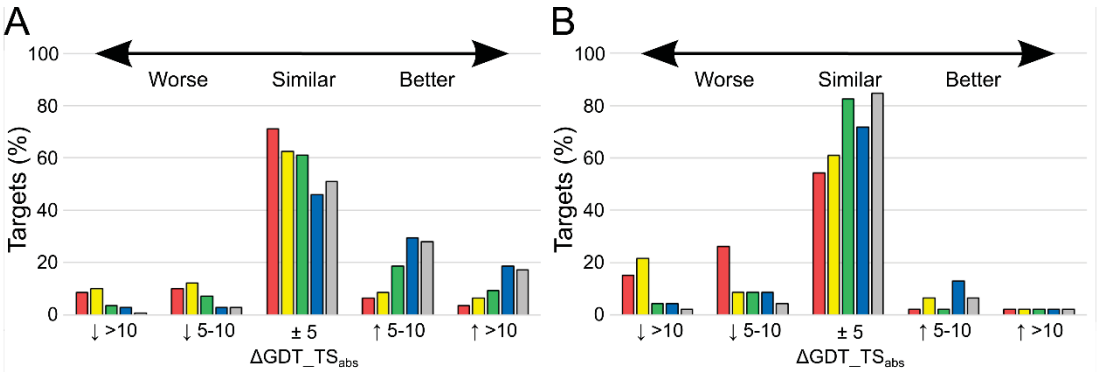


Figure 13. GDT_TS comparisons between TopModel and CASP servers. The bars represent comparison between TopModel and one of four established CASP servers (the Zhang Server (red), the Baker Server (yellow), the HHPred server (green), the Zhou Server (blue)) as well as the average of the top 200 server submissions for each target (gray). The Zhang server and Baker server both make use of *ab initio* folding and domain parsing, putting them at an advantage over TopModel. **A.** ΔGDT_TS_{abs} for CASP TBM targets indicates for how many of CASP TBM targets TopModel shows similar, worse, or better model quality than other established servers. **B.** ΔGDT_TS_{abs} for CASP FM targets indicates for how many of CASP FM targets TopModel shows similar, worse, or better model quality than other established servers.

To demonstrate the utility of TopModel, the workflow was experimentally validated on two *de novo* protein systems showing good agreement with experimental data in terms of crystal structures, NMR spectroscopy experiments, and SAXS experiments. These proteins were the NSR protein from *S. agalactiae* and LipoP from *C. difficile* and showed far better agreement with experimental data than predictions from any of its constituent primary predictors. The results from the NSR protein are shown in Figure 14 and illustrate how

TopModel go against the majority (the center of the distribution) resulting in a model of far greater quality than any of its constituent primary predictors. The results for LipoP from *C. difficile* is shown in Figure 15. They show how, after a short refinement using molecular dynamics, the model of the LipoP show good agreement with both NOE and secondary structure restraints from NMR as well as scattering profile and volumetric shape from SAXS.

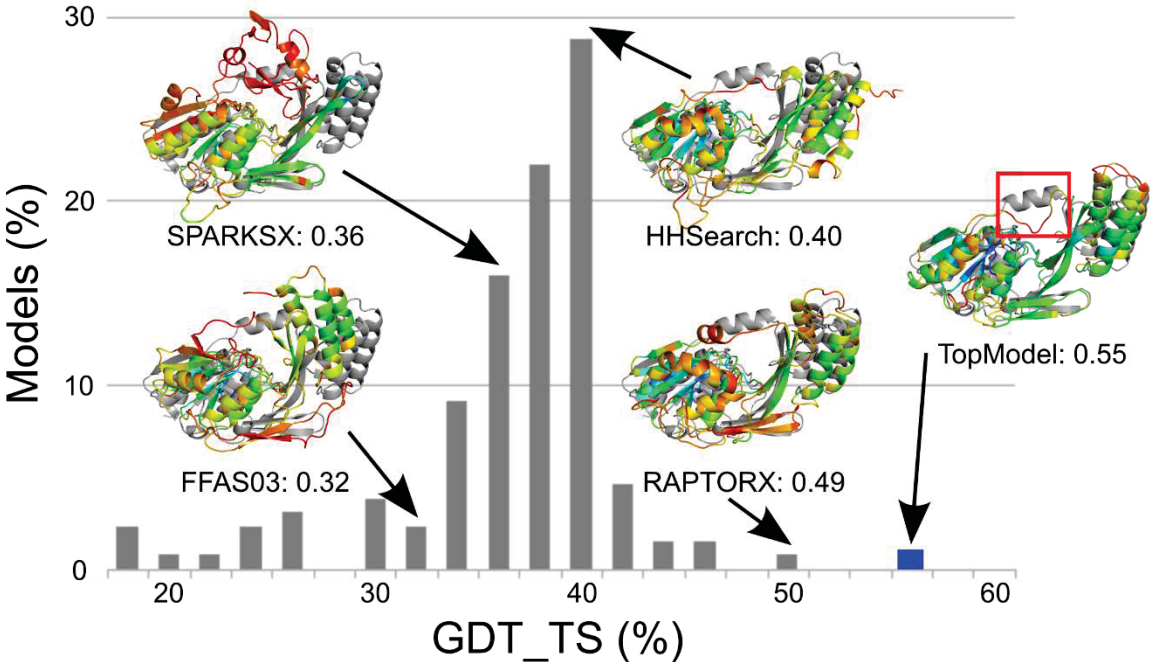


Figure 14. Prospective modelling of the *NSR* protein from *S. Agalactiae*. The model quality distribution (in terms of GDT_TS score) of primary threader models for the *NSR* protein from *S. agalactiae* for prospective modelling before the release of the native structure (gray) to the PDB. The vast majority (82%) of models show an incorrectly threaded N-terminal domain (see SPARKSX model). A minority of models (18%) show a correctly threaded helical domain (HHSearch, RAPTORX, and FFAS03) on a few templates, often with large errors elsewhere in the model (such as β-sheets shown in red). Because TopModel does not use majority voting, the model produced (blue box) is of far better quality (GDT_TS = 55) than those produced by primary threaders (median GDT_TS = 38), while majority voting consensus would produce a model in the middle of the distribution at a GDT_TS of ~38. Model examples from the different bins are colored according to residue-wise lDDT score [129] to the native structure, with red showing incorrectly modelled regions and blue showing perfect agreement with the crystal. The largest error in the TopModel model is the fact that the residues linking the helical bundle with the catalytic core of the protein do not fold into an α-helix (red box). This is because no model from any of the primary predictors correctly fold these residues into a helix, and as such, TopRefiner has no fragment it can select during model fragmenting and refinement, which would produce a helix for these residues. Secondary structure prediction by PSIPRED [99] also fails to identify these residues as helical.

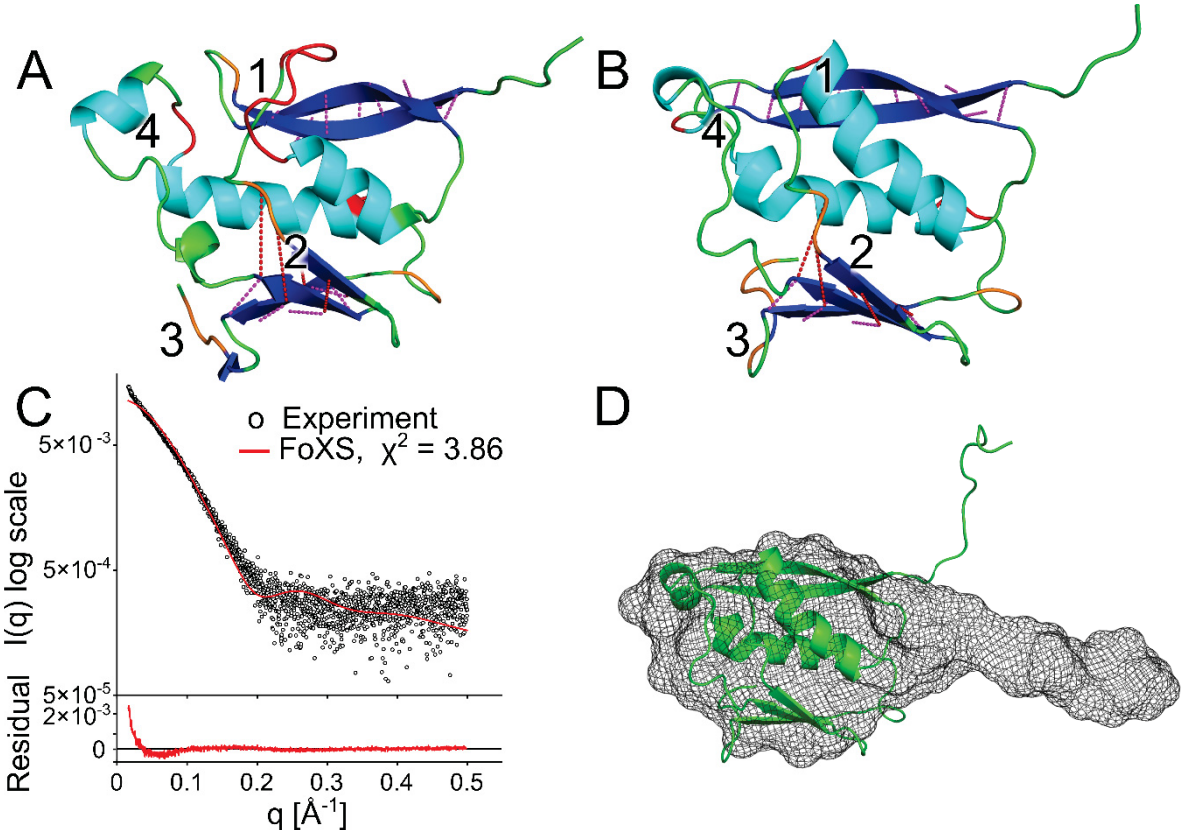


Figure 15. Model of LipoP from *C. Difficile* after MD refinement and selection according to agreement with sparse experimental structural data. A. Agreement of the TopModel model with secondary structure assignments and NOE restraints from NMR. The numbers indicate the location of errors. **Blue:** β-sheet residues showing agreement between model and NMR data. **Orange:** Residues identified to be in a β-strand in NMR but not found so in the model. **Cyan:** α-helical residues showing agreement between model and NMR data. **Red:** Residues identified to be α-helical in NMR but not found so in the model. **Magenta lines:** Experimental β-sheet NOE restraints showing agreement with the model. **Red Lines:** Experimental β-sheet NOE restraints showing a shift of two residue positions of β-strand 3. **B.** Agreement between the model after MD refinement, selection according to agreement with experimental NMR and SAXS data, and model combination with TopBuilder Colors are following the same scheme as in panel A. The extension of α-helix 1 is seen. **C.** Agreement between the experimental scattering data from SAXS (black) and simulated scattering curve of the MD model (red); FoXS [179-180] was used for simulating the scattering curve. The fit plots depict log-intensity *versus* q ($Å^{-1}$), the residuals plot shows the difference between experimental and computed intensity *versus* q ($Å^{-1}$). **D.** The volumetric envelope of LipoP, as calculated from the scattering data using GASBOR [181], is shown in gray mesh. The MD model of LipoP (green) was docked into the volumetric envelope using SUPCOMB [182]. Disagreement with SAXS is found mainly for the disordered tail of LipoP.

10.3 CONCLUSIONS AND SIGNIFICANCE

The idea behind the TopModel methodology is that since no method can be expected to be the best for every protein target of interest, different predictions from different primary predictors are integrated using deep neural networks to select the best candidate template or model. This is, to our knowledge, the first time deep neural networks have been applied to estimate template similarity to the native structure for use in pure template-based structure prediction. Then, using top-down consensus, predictions that agree with the best candidate are selected and used for multi-template modelling. This is, to our knowledge, the first use of top down consensus, rather than majority voting consensus, for template selection and protein structure prediction. During refinement, rather than averaging the models, regions predicted by TopScore and TopScoreSingle to contain errors are removed and replaced by better regions from different models based on different templates or alignments. This is, to our knowledge, one of the first times model refinement has been driven not by energy minimization, but by minimizing the output score of a deep neural network **(TopScore, Chapter 9, Publication I)**. These developments enable TopModel to make structure predictions that go against the majority of its primary predictors and produce models that are significantly better than any of the predictions from any of its constituent primary predictors.

TopModel is the core of TopSuite as it provides the tools and the workflow required for high quality template-based protein structure prediction. TopModel was used in all the application projects mentioned in this thesis, and the structures predicted by TopModel provided valuable insights and good starting points for further study of important biological systems.

11. BINDING REGION OF ALANOPINE DEHYDROGENASE PREDICTED BY UNBIASED MOLECULAR DYNAMICS SIMULATIONS OF LIGAND DIFFUSION

Holger Gohlke, Hergert, U., Meyer, T., **Daniel Mulnaes (5%),**

Grieshaber, M.K., Sander H.J. Smits and Lutz Schmitt.

Journal of Chemical Information and Modelling. 2013, 53, 2493–2498.

11.1 BACKGROUND

Lack of oxygen can be caused either by an environmental change or by an increased oxygen consumption by the organism itself, *i.e.* increased oxygen consumption by muscles during movement. To maintain a continuous flux of energy under conditions of intense physiological activity in which oxygen supply becomes a limiting factor, organisms therefore switch to full or partial anaerobic metabolism. This anaerobic metabolism can follow four main pathways initialized from Phosphoenolpyruvate: (1) The glucose-succinate pathway in which the final product is succinate, (2) the aspartate-succinate pathway that also results in succinate, (3) the glucose-lactose pathway in which the final product is lactate, and (4) the glucose-opine pathway in which the final products are various opines. While the first two are more energy efficient pathways, they are slower than the latter two, and thus serve complimentary roles depending on the duration of hypoxia. In the opine pathway, opine dehydrogenases ensure a constant supply of ATP by maintaining the NADH/NAD+ balance [183].

In this work, we investigated the binding of L-alanine to the Alanopine Dehydrogenase of *Arenicola Marina* (AlaDH*Am*), a member of the opine dehydrogenase family. Although much is known biochemically about this enzyme class, the substrate specificity of different Alanopine Dehydrogenases towards different amino acids, as well as the substrate inhibition has yet to be explained at the molecular level.

Due to recent advances in molecular dynamics (MD) simulation algorithms and hardware, the simulation of unbiased ligand binding and unbinding to their target protein has recently become possible. In addition to the ability to identify the binding region, these simulations can reveal binding and unbinding pathways as well as metastable binding states, and give quantitative estimates of both binding affinity and on/off rates [184-188]. In this study, we go beyond these measurements to examine determinants of substrate specificity starting

not from a crystal structure but from a predicted 3D structural model made by a preliminary version of our structure prediction workflow TopModel.

11.2 RESULTS

Biochemical characterization of AlaDH*Am* revealed a high substrate specificity for L-alanine, and showed that, with about 3- to 4-fold reduction in activity, glycine could also be used as a substrate. For other small amino acids tested, such as L-serine, L-threonine, L-cysteine, or L-valine, none or only negligible activity was found. This indicates a high substrate specificity and shows that the binding site of AlaDH*Am* has evolved to bind L-alanine specifically. In contrast, the AlaDH from *M. Sanguinea* has a much broader substrate specificity allowing also other small amino-acids to form the corresponding opine [189].

Structure prediction with TopModel revealed three structures representing two proteins: Octopine Dehydrogenase (OcDH) from *P. maximus* with either L-arginine (PDB ID: 3C7C) or agmatine (PDB ID: 3IQD) in the binding site, and *M*-(1-D-Carboxylethyl)-L-Norvaline Dehydrogenase (CENDH) from *Arthrobacter Sp.* (PDB ID: 1BG6). Both structures are opine dehydrogenases with sequence identities of 46% and 20% respectively. I constructed a sequence alignment using the structural information of the available structures, and analyzed the sequence conservation of the OcDH binding site compared to AlaDH*Am*, which revealed a high degree of residue conservation. For example, binding-site residues E141 and W279 cannot mediate substrate specificity since they are conserved between OcDH and AlaDH*Am*. The only two sequence differences between the binding sites of OcDH and AlaDH*Am* are residues V208 and N209 located in the kink of a helix-kink-helix structural motif in the binding site. In OcDH, V208 is instead a tyrosine, and there is no residue insertion at position 209, unlike in CENDH, where residue 209 is also inserted. Interestingly, the N209 insertion occupies the same volume that in OcDH is occupied by the bound L-arginine, and thus prevents the binding of large amino acids to AlaDH*Am*. The structure-based alignment can be seen in Figure 16.

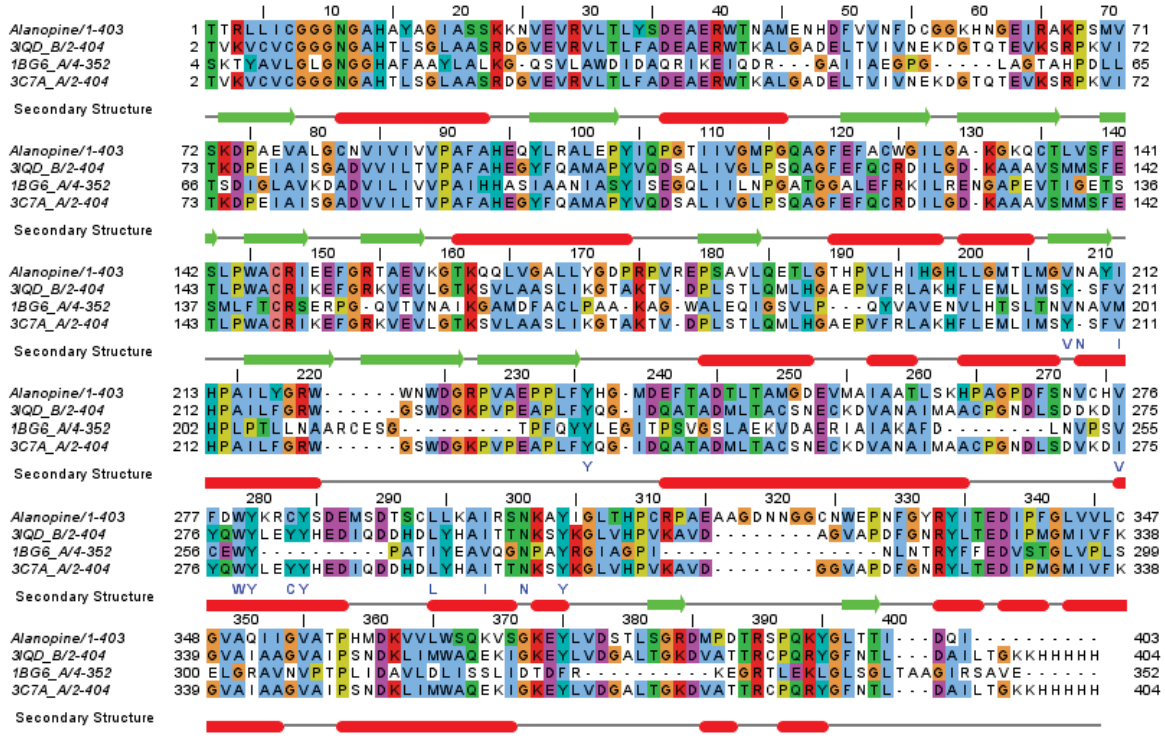


Figure 16: Alignment of sequences of AlaDHA*Am* to the three templates. Red bars and green arrows indicate α-helices and β-strands respectively, of the AlaDH*Am* model as determined by DSSP. Numbers provided on the left and right refer to positions in the respective sequences; numbers provided on top refer to the positions in the AlaDH*Am* sequence. The amino acids are colored according to the ClustalW criteria in Jalview (orange: G; Yellow: P, cyan H and Y; blue hydrophobic amino-acids (A, I, L, M, F, W, V,C); green: polar amino-acids (N, Q, S, T); red: positively charged amino-acids (K,R); magenta: negatively charged amino-acids (D,E)) if the amino-acid profile of the alignment at that position meets a minimum criterion specific for the residue type.

To produce a binding model of L-alanine to AlaDH*Am* the coordinates of the NADH cofactor could be copied from OcDH without steric clashes after structural superposition. However, superposition of L-alanine to the backbone part of the OcDH bound arginine required a geometry optimization leading to a shift in position of 3Å. After constructing the initial model of AlaDHAm, we subjected it to three independent MD simulations of 200 ns each. The simulations showed an overall moderate deviation from the starting structure with a root mean square deviation (RMSD) of C_α atoms ranging from 2.5 to 4 Å. This is comparable to a control simulation of 100 ns of OcDH (PDB ID 3C7C) showing a 1.5-3.5 Å RMSD. While NADH stayed in the binding pocket for the full duration of all three simulations, the L-alanine showed both binding and unbinding from the binding pocket in all three simulations. These remarkable results of binding and unbinding of substrate into bulk solution and back into the binding pocket is therefore one of the few examples of *ab initio* unbiased MD

simulations to date [184-188] that show unbinding and re-binding of the substrate. The results of the MD simulations are shown in Figure 17.

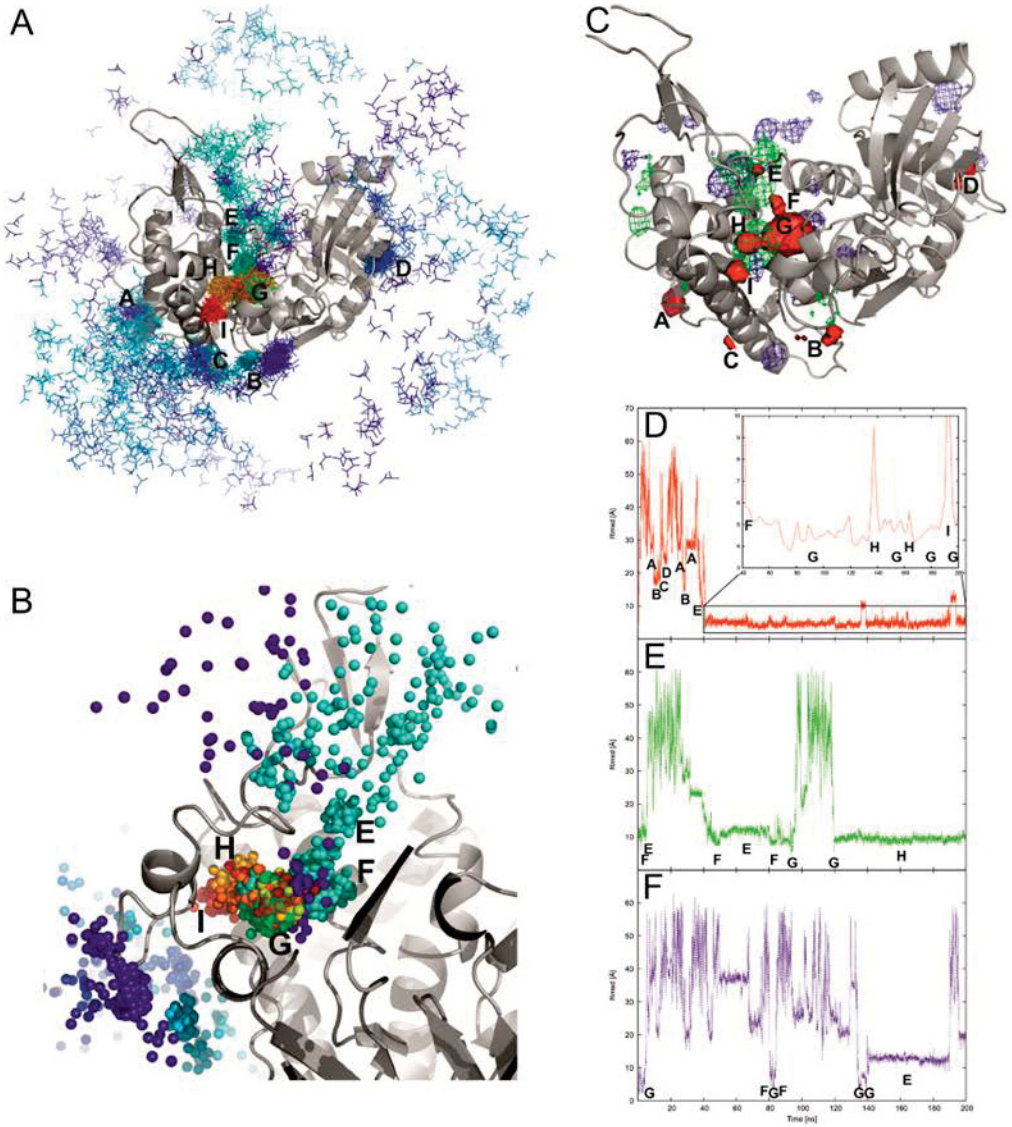


Figure 17. Unbiased MD simulations of L-alanine diffusion in the TopModel model of AlaDH*Am*. (**A-F**) Black letters indicate regions of high density of L-alanine during the MD simulation 1 as identified in panel C. Region G is the predicted binding region. (**A**) Traces of L-alanine extracted from MD trajectory 1generated by MD simulation of 200 ns length of the AlaDH*Am*/NADH/L-alanine system in explicit water; L-alanine reaches the predicted binding region after ~40 ns (see panel D). The time evolution of the MD simulation is color coded from blue (0 ns) to red (200 ns). For clarity, only a conformation closest to the average conformation of AlaDH*Am* is shown (gray cartoon). (**B**) Close-up view of the predicted binding region shown in panel A with the trace of C_α atoms of L-alanine extracted from trajectory 1 shown as spheres. See panel A regarding the color-coding. (**C**) Overlay of density maps extracted from trajectory 1 (red surface), 2 (green mesh) and 3 (blue mesh) showing the frequency of interaction between L-alanine on the surface of AlaDH*Am*; the contour level is 3 sigma. Regions of high density in trajectory 1 are labelled with black letters. The protein conformation is as in panel A. (**D-F**) Root mean square deviations (RMSDs) of the L-alanine atoms during the simulations 1-3, respectively, with respect to the AlaDHAm starting model from TopModel after super-positioning to the starting structure-based on Cα atoms.

In order to estimate quantitative thermodynamic binding properties, substantially more binding and unbinding events would be required. Still, the simulations provide suggestions for energetically favorable binding locations on the surface of AlaDH*Am* as shown in Figure 12. These simulations reveal a potential binding pathway in which alanine successively binds to interaction "hot spots" on the way towards the active site. This was also corroborated by energetics calculations with molecular mechanics generalized born surface area (MM-GBSA) calculations. The identification of these binding regions provides an explanation for the substrate inhibition of AlaDHAm, as the occupation of these spots by alanine would hinder alanopine egress from the binding site, provided that it follows the same successive binding pathway (only in reverse). This assumption is highly likely given the gorge-like shape of the binding funnel.

11.3 CONCLUSIONS AND SIGNIFICANCE

In summary, we presented a biochemical characterization of AlaDH*Am*, which catalyzes the reductive condensation of L-alanine with pyruvate to alanopine. AlaDH*Am* displays a high catalytic efficiency and substrate specificity, and is prone to substrate inhibition. As the 3D structure of AlaDH*Am* is unknown, I predicted the structure with TopModel and we used the substrate-binding model from the homologue OcDH from *P. maximus* to infer the cofactor-binding pose and initial L-alanine binding modes. Unbiased MD simulations of the system captured the binding of L-alanine diffusing from solvent to the putative binding region, located at the helix-kink-helix motif, as observed for binding of L-arginine to OcDH. At the same time, the observed binding of L-alanine provides for the first time a molecular explanation for the role of amino acids 208 and 209 in substrate specificity, the only amino acids within the binding region that differ between OpDHs with different substrates. Finally, the presence of energetically favorable non-native ligand binding states near the binding region provides an explanation for the substrate inhibition of AlaDH*Am*.

Historically, the modelling of AlaDH*Am* was the first to be done with a preliminary version of TopModel **(Chapter 10, Publication II)**. As TopModel was in its infancy, the threading module TopThreader included only a few search tools for template identification such as BLAST, and TopAligner included only SAlign. For Model quality estimation PROCHECK, DOPE, and ANOLEA was used. The availability of high-quality templates and the close homology between the target sequence and the input structure resulted in a high-quality model despite the preliminary status of the modelling workflow and showed the potential of a fully automated pipeline for structure prediction.

12. DETERMINANTS OF FIV AND HIV VIF SENSITIVITY OF FELINE APOBEC3 RESTRICTION FACTORS

Zeli Zhang, Qinyong Gu, Ananda Ayyappan Jaguva Vasudevan, Anika Hain, Björn-Philipp Kloke, Sascha Hasheminasab, **Daniel Mulnaes (10%),** Kei Sato, Klaus Cichutek, Dieter Häussinger, Ignatio G. Bravo, Sander H.J. Smits, Holger Gohlke and Carsten Münk.

Retrovirology; 2016, 13, 46.

12.1 BACKGROUND

The feline immunodeficiency virus (FIV) is a lentivirus with the potential to cause an immunodeficiency disease in domestic cats, which is similar to human immunodeficiency virus type 1 (HIV-1) induced AIDS [190]. Additionally, under experimental conditions, FIV infection in cats has a mortality rate of up to 60 % [191-193]. This makes FIV infection of cats a valuable animal model system for the study of HIV-1 and AIDS [194-196].

APOBEC3 (A3) proteins are anti-viral cytidine deaminase restriction factors found in placental mammals, which counteract lentiviruses such as HIV, FIV, and Simian immunodeficiency virus (SIV) [197-200]. Primates have seven different variants of A3 proteins while felines have four. Some retroviruses counteract the anti-viral A3 proteins by expressing proteins themselves, such as Vif from lentiviruses (HIV, FIV, and SIV) [201-206]. Surprisingly, feline A3 proteins also inhibits HIV and SIV, and HIV-2 and SIV Vif proteins can counteract some feline A3 proteins such as A3Z2Z3.

The A3 proteins target viruses and genetic elements that depend on reverse transcription, but also show some activity against unrelated viruses [207]. The viral protein Vif from lentiviruses works by inhibiting encapsidation of A3 into the virus particles, thereby preventing the deamination of the virus single-stranded DNA cytidines. If Vif is not present, A3 enters the nascent viral particles and introduces G-to-A mutations in the viral genes during reverse transcription, which inhibits the function and stability of the transcribed viral proteins. Furthermore, some A3 proteins inhibit viral replication by reducing reverse transcription and integration [208-213]. In cats, the feline A3 protein A3Z2Z3 is expressed following a read-through transcription and alternative splicing, which introduces a previously untranslated exon in-frame, which in turn encodes for a domain insertion termed the A3 linker domain.

HIV-1 Vif cannot counteract feline A3s, and HIV-1 is therefore inhibited by all feline A3s, with A3Z2Z3 displaying the strongest inhibition [214-217]. The mechanism behind the inability of HIV Vif to degrade feline A3 is unclear, especially since feline A3Z2Z3 and HIV-1 Vif are recovered together using co-immunoprecipitation assays, indicating that they do in fact bind to each other. In contrast, the Vif of SIV from macaques (SIVmac) can degrade feline A3s [218].

To assess the feasibility of generating an animal model for the human system based on FIV, we and others cloned FIV Vif into HIV-1 and proved that in feline cell lines the A3 proteins are the dominant restriction factors against HIV-1 [214, 216]. In order to understand the FIV Vif interaction with feline A3 proteins, we identified in this study important A3 residues and used a homology model of feline A3Z2Z3 generated by our structure prediction pipeline TopModel to describe the structure-function relationship of these potential FIV Vif binding amino acids.

12.2 RESULTS

In this study, we aimed to identify which residues in feline A3s are recognized by Vifs and required for A3 degradation. To identify these residues, chimeric human-feline A3s were tested, and to locate these interaction regions the first structural model of feline A3 was predicted using TopModel.

For modelling the human APOBEC structure, TopModel identified templates 4J4J_A (35 % Identity), 2KBO_A (37 % Identity), and 2RPZ_A (30 % Identity) resulting in a model with 84 % accuracy according to TopScore **(Chapter 9, Publication I)** (TopScore of 0.16). For modelling the feline A3Z2b, TopModel **(Chapter 10, Publication II)** identified templates 3VM8_A (42 % Identity), 2KBO_A (39 % Identity), and 1M65_A (10 % Identity) resulting in a model with 88 % accuracy according to TopScore (TopScore of 0.12). For modelling the feline A3Z3, TopModel identified templates 4J4J_A (31 % Identity), 2KBO_A (36 % Identity), and 2RPZ_A (24% Identity) resulting in a model with 84 % accuracy according to TopScore (TopScore of 0.16). For the linker domain, TopModel identified templates 2XS9_A, 2MMB_A, 2DA4_A, 2LFB_A, and 1FTZ_A with sequence identities ranging from 9-19 %.

In the Z3 domain, we identified residues involved in binding of FIV Vif, and upon mutation, the Vif-induced A3Z3 degradation was blocked. Furthermore, we found additional essential residues for FIV Vif interaction in the A3Z2 domain, which allowed us to construct

FIV Vif resistant A3Z2Z3 mutants. These mutants also showed resistance to the Vif of a lion-specific FIV, indicating an evolutionary conserved Vif-A3 binding. These results can be seen in Figure 18.

The predicted structure of feline A3Z2Z3 from TopModel suggests that the residues interacting with FIV Vif have a unique location at the domain interface of Z2 and Z3, unlike Vif-interacting residues in human A3s. Furthermore, it showed that the linker domain between the Z2 and Z3 domains forms a homeobox-like domain protruding from the Z2Z3 core. HIV-2 and SIV Vifs efficiently degrade feline A3Z2Z3, possibly by targeting this linker domain.

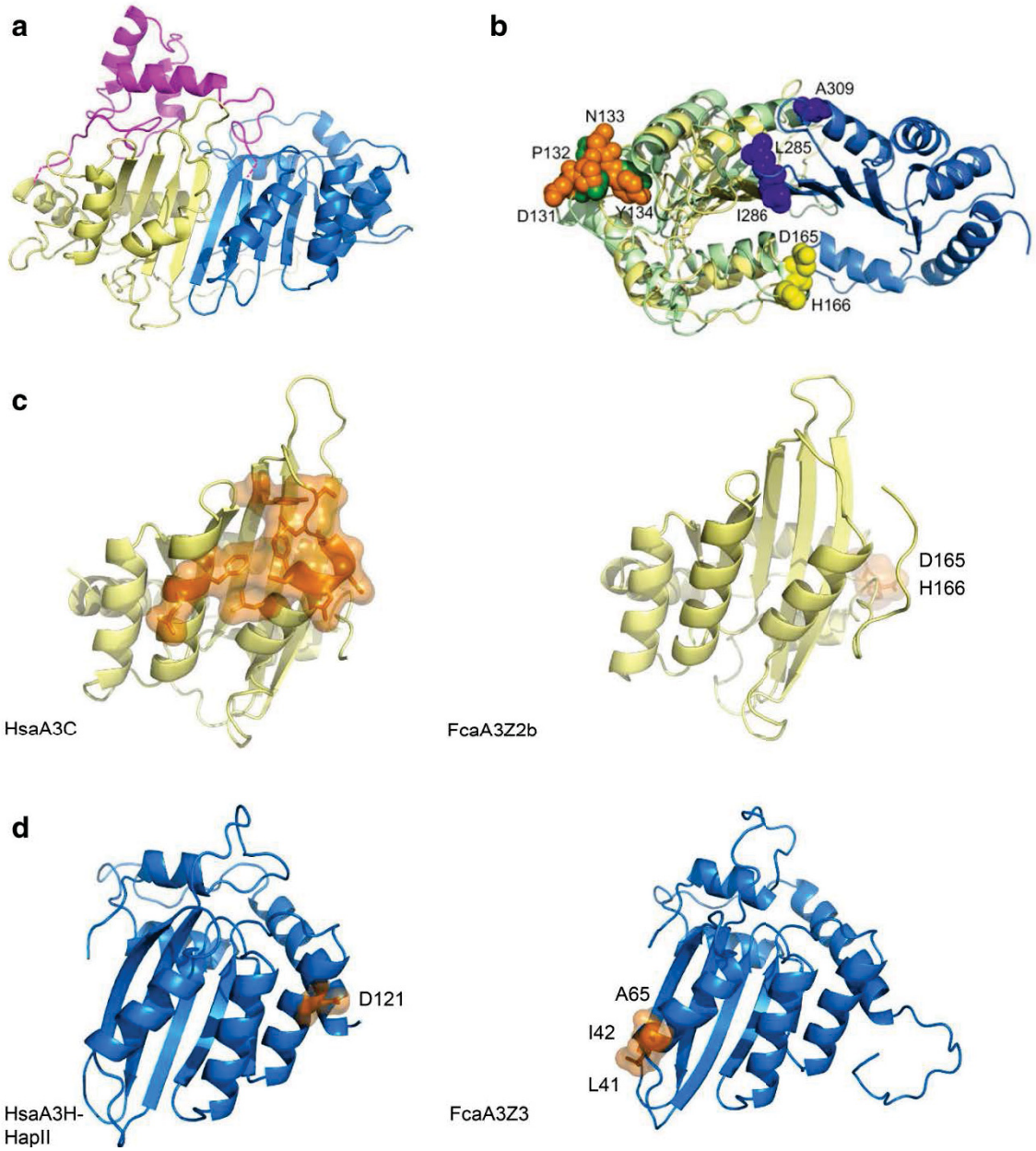


Figure 18. Composite model of feline APOBEC3 predicted by TopModel and the locations of residues mediating the Vif binding. a) Structural model of feline A3Z2Z3 including the homeobox-like linker domain (pink) above the Z2 (yellow) and Z3(green) domains as predicted by TopModel. **b)** The structural model in **a** rotated by 90°. The linker domain and parts of the N-terminus without template were omitted for clarity. Residues in sphere representation in yellow (D165/H166), blue (L285/I286/A309), and orange (D131-Y134) mediate binding of Vif. **c)** The human A3C crystal structure (3VOW) and a structural model of feline A3Z2b built by TopModel depicting the positions of respective HIV-1 Vif and FIV Vif binding sites. The structures are oriented as the Z2 domain in **a**. A structural model of human A3H-HapII and feline A3Z3 built by TopModel depicting the positions of respective HIV-1 Vif and FIV Vif binding sites. The domains are oriented as the Z3 domain in **a**. Key residues involved in Vif binding are labelled (except human A3C), represented in sticks and highlighted with its surface in orange.

12.3 CONCLUSIONS AND SIGNIFICANCE

In this work computational structure prediction with TopModel, biological assays, and sequence analysis were employed to identify residues in feline A3s important for binding of FIV Vif. Our results show that HIV Vif binds human A3s differently than FIV Vif bind feline A3s, and structure prediction with TopModel revealed a linker domain unique to feline A3s. The linker insertion is predicted to form a homeobox-like domain, which is unique to A3s of cats and related species, and not found in human and mouse A3s. Together, these findings indicate a specific and different A3 evolution in cats compared to humans, which is important to consider when using the domestic cat as a model organism for the study of HIV and AIDS.

The modelling of APOBEC3 was more challenging for TopModel than it was to model Alanopine Dehydrogenase **(Chapter 11, Publication III)**, due in part to the more distantly related templates and in part to its multi-domain nature. The inclusion of more primary predictors in TopThreader, TopAligner, and TopScore at this point in the development of TopModel, however, led to high-quality models in spite of these added difficulties. Furthermore, the modelling of the linker domain was the first time an early version of the TopRefiner protocol was used. The results from TopModel proved to be critical in the identification of key residues important for the interaction between FIV Vif and APOBEC3. They also revealed key differences between feline and human A3s, which could have important implications for the development of anti-viral therapies using the domestic cat as a model animal.

13. RECOGNITION MOTIF AND MECHANISM OF RIPENING INHIBITORY PEPTIDES IN PLANT HORMONERECEPTOR ETR1

Dalibor Milić, Markus Dick, **Daniel Mulnaes (10%)**,

Christopher Pfleger, Anna Kinnen, Holger Gohlke and Georg Groth.

Nature Scientific Reports 2018, 8, 3890.

13.1 BACKGROUND

Fruit ripening of crops, such as apples, bananas and tomatoes, is induced by the plant hormone ethylene. To minimize fruit damage and spoilage during transportation due to over-ripening, some industries therefore interfere with ethylene biosynthesis or signaling, by storing and transporting the crops in an unripe state and inducing ripening by ethylene exposure at the final destination. Synthetic peptides derived from Ethylene-Insensitive Protein 2 (EIN2), a central regulator of the ethylene signaling pathway, were recently shown to delay fruit ripening. In particular, the inhibitory peptide NOP-1 derived from EIN2 was shown to delay ripening by interacting with the ETR1 protein, the prototype of the plant ethylene receptor family. ETR1 is a large multi-domain receptor protein with a trans-membrane domain and four cytosolic domains, which forms a dimer in-vivo. Upon ethylene binding, ETR1 starts an intracellular signaling cascade, which ultimately results in altered gene expression and the induction of ripening.

However, despite knowing that upon ethylene binding ETR1 induces fruit ripening, and knowing that NOP-1 inhibits this signal, the molecular mechanism of these interactions is still unknown. Understanding how the binding of ethylene impacts ETR1 to induce intracellular signaling and how this signal is inhibited by NOP-1 is key to understanding fruit ripening at the molecular level as well as figuring out how to best modulate this process to prevent food spoilage during transport.

In this study, we show that the inhibitory peptide NOP-1 derived from EIN2 binds to the GAF domain of ETR1. Furthermore, by combining site-directed mutagenesis, computational structure prediction with TopModel and TopDock, molecular dynamics simulations, and rigidity analysis we reveal the peptide interaction site and a plausible molecular mechanism for the ripening inhibition. This in turn may aid in the future

optimization of peptide inhibitors of fruit ripening such as NOP-1 and decrease spoilage during crop transport.

13.2 RESULTS

To understand the structural basis of interactions between ethylene receptors and inhibitory peptides, heterologous expression was used to produce truncated constructs of ETR1 from the model organism *A. Thaliana*, which successively lack protein domains starting from the C-terminus. The goal was to identify ETR1 domain(s) crucial for the protein-peptide interaction between the inhibitory octapeptide NOP-1 derived from EIN2 $^{219\text{-}221}$, under the assumption that domain truncation has minimal impact on protein stability and dimerization. Microscale thermophoresis was used to characterize the binding and revealed that only once all cytosolic domains of ETR1 had been removed, was the binding of NOP-1 to the receptor abolished. This indicates that the last cytosolic domain that was removed, the GAF domain, binds NOP-1. To investigate the binding of NOP-1 further, a construct containing the receiver domain, the catalytic ATP-binding domain, and the dimerization histidine-phosphotransfer domain was expressed and tested for binding to NOP-1 using microscale thermophoresis. Surprisingly, this construct also showed no binding to NOP-1, which disproved the initial hypothesis that NOP-1 binds to a canonical phosphorylation site in the receiver domain. By process of elimination, the GAF domain was therefore pinpointed as the only binding partner of NOP-1, since the constructs lacking this domain showed no binding of NOP-1. A schematic representation of ETR1 can be seen in Figure 19.

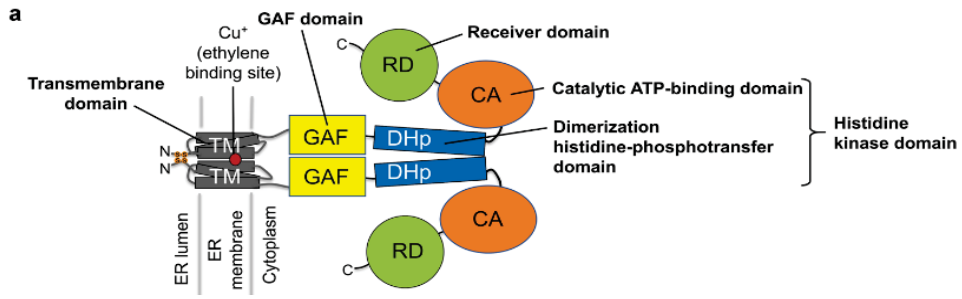


Figure 19. Schematic representation of the full ETR1 protein in its dimeric form. The GAF domain (yellow) was found to mediate the dimer interaction and to be the domain, which interacts with the NOP-1 inhibitory peptide. By successively removing first the Receiver domain, then also the Catalytic ATP-binding domain, then the Dimerization histidine-phosphotransfer domain, and finally the GAF domain, and only observing abolished NOP-1 binding at the last step, the GAF domain could be identified as a NOP-1 binding domain. By constructing a protein with all other domain than the GAF domain and observing no NOP-1 binding, it was shown that only the GAF domain binds NOP-1.

To explore the binding of NOP-1 to the GAF domain of ETR1, I predicted a model of the GAF domain using TopModel, since no experimental structure exists. The identified templates by TopModel all share the same fold, with the top five ranked templates being: 3P01_A (18 % Sequence Identity), 3TRC_A (15 % Sequence Identity), 3CI6_A (13% Sequence Identity), 3W2Z_A (12 % Sequence Identity), and 1YKD_B (15 % Sequence Identity). The final model built by TopModel **(Chapter 10, Publication II)** was assessed with TopScore **(Chapter 9, Publication I)** to be 71% correct, with the majority of inaccuracies being located in the flexible loop regions (residues 228–247 and 257–272: 47 % and 52 % inaccuracies, respectively).

Previous findings suggest that ethylene receptors form a dimer in their simplest functional state, which is also mediated by their GAF domains [222]. I therefore built a dimer model of the GAF domain using a preliminary version of the protein-protein docking software TopDock and a preliminary version of the protein-protein interface prediction software TopInterface, which at the time was integrated into TopDock. TopInterface predicts protein-protein contacts based on a structure-based homology search that is independent of sequence. It does so by using the Phyrestorm [223] clustering tree to rapidly search the PDB database for structures similar to the input, using a 0.5 TM-Score cut-off to select true positives. When the same PDB ID is found as a structural homologue for two queries, their interface is inferred from the interface in the structural homologue. TopInterface identified five different homologous interfaces (PDB ID and chain identifiers given: 3G6O_AB, 3IBJ_AB, 3K2N_AB, 3P01_AB, and 3TRC_AB) all of which indicate that the dimer interface of the GAF domain consists of the N- and C-terminal helices.

I used the residue-residue contacts from each homologous interface for restrained docking of the GAF domains using TopDock, which uses the docking engine HADDOCK [224]. The docking solutions were pooled and clustered by TopDock, and ranked according to HADDOCK energy, cluster size, distance to cluster centroid, and fulfillment of predicted contacts to select a docking solution. Each monomeric subunit of our final model contains a central, antiparallel, seven-strand β-sheet, flanked by one short α-helix (amino-acids 213–220) and three, parallel-oriented α-helices that cover the N- and C-terminal regions (amino-acids 118–173 and 290–305). The N-terminal α-helices of the two monomers together form the dimeric interface resulting in a six-helix bundle in the homo-dimeric structure. The final monomeric and dimeric structures predicted by TopModel and TopDock can be seen in Figure 20.

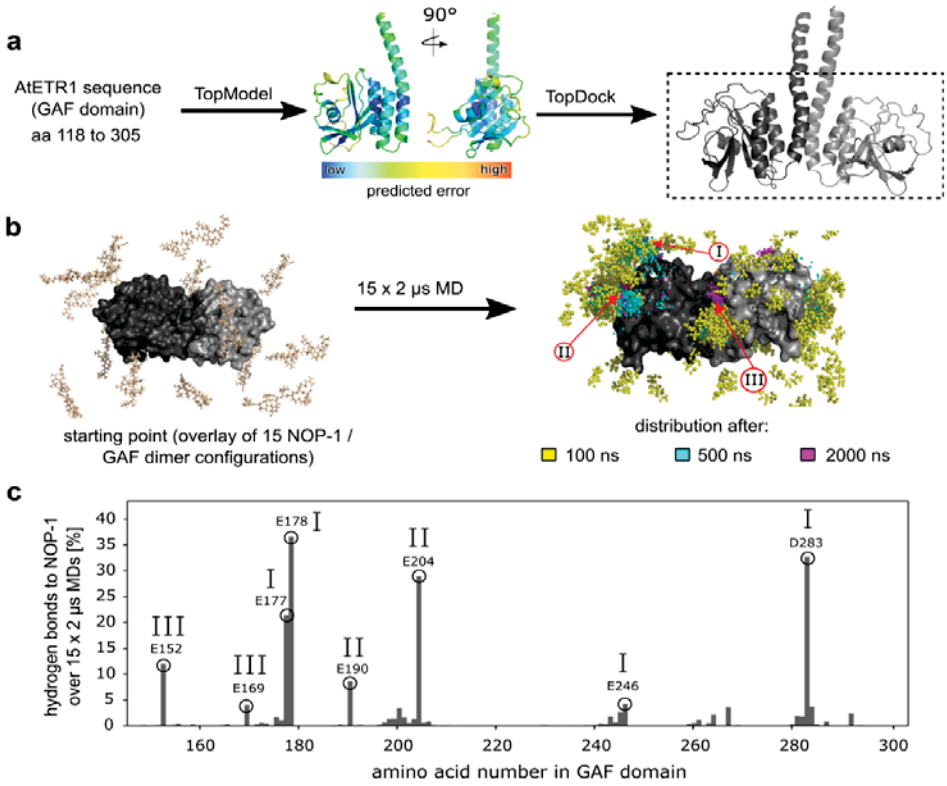


Figure 20. The model of the GAF domain. a) The model is colored according to predicted residue-wise error according to TopScore and docked to form a dimeric model using TopDock. **b)** The truncated dimeric model was solvated and 15 NOP-1 peptides were placed randomly in the solvent. After MD simulations of 15x2µs, the residues interacting with NOP-1 were identified. **c)** Three hot spots were identified according to hydrogen bonding between NOP-1 and the GAF domain, and were analyzed experimentally to identify the most likely binding site for NOP-1.

MD simulations of the protein of 500 ns length in the absence of any peptide ligand revealed overall moderate structural variations within both monomers. Subsequently, 15 MD simulations of 2 µs each with different randomly placed NOP-1 peptides were performed to identify putative binding sites. Three such sites were identified by analyzing hydrogen bonding between the peptide and the GAF domains. To identify which of these potential binding sites are more likely to be the true peptide-binding site, alanine mutations of the binding site residues were combined with intrinsic tryptophan fluorescence quenching experiments. These experiments confirmed that the most likely binding site is located in a negatively charged patch (binding site III) close to the interface between the two monomers, where the positively charged peptides bind. Using a combination of rigidity analysis [225-226] and analysis of the stability of the GAF domains in the MD simulations showed a stabilizing effect of NOP-1 binding. This stabilization may hamper the transmission of ethylene binding

signals (such as a conformational change induced by ethylene binding to the TM domain) from reaching the rest of the transporter, and thus inhibit fruit ripening.

13.3 CONCLUSIONS AND SIGNIFICANCE

The application of TopModel and TopDock to the modelling of the GAF domain of ETR1 enabled an understanding of the inhibitory effect of the NOP1 peptide on fruit ripening at the molecular level. This understanding is rooted in the accurate modelling of the GAF domain with TopModel and dimer construction with TopDock. These accurate predictions enabled the identification of putative binding sites using free ligand diffusion MD simulations of the dimeric model as well as experimental validation of the binding sites. Furthermore, a mode of action was proposed by performing rigidity analyzes with CNA and flexibility analyses of the MD trajectory. This new knowledge could in turn be used to design new improved inhibitors of fruit ripening by targeting this binding site with either small molecules or peptide inhibitors.

The modelling of the GAF domain showed the power of automated structure prediction with TopModel **(Chapter 10, Publication II)**. The automated structure prediction is especially useful for target proteins such as the GAF domain, where only distantly related templates with low sequence identity were found. All of the templates identified for the GAF domain had less than 20% sequence identity, but the final model still had a high quality when modelled with TopModel. Furthermore, it showed the potential of predicting protein-protein interactions with the preliminary version of TopInterface and using those predictions to guide protein-protein docking with TopDock.

14. SUMMARY AND PERSPECTIVES

In this thesis, I have described the ongoing development of TopSuite and its application to different biological systems of interest, resulting in key biological insights and providing a basis for further research.

I have developed a meta-tool for protein model quality estimation with two scoring functions, TopScore and TopScoreSingle. These scoring functions use deep neural networks to combine predictions from many diverse model quality estimation programs. They were trained on a large dataset of models from different sources representing both homology models from closely related and distantly related homologous templates, *ab initio* models from folding simulations, artificially misfolded decoys, and models from previous CASP competitions. TopScore and TopScoreSingle showed a significantly better and more stable performance across the different datasets compared to all state-of-the-art primary predictors (**Chapter 9, Publication I**).

Building on the ability to estimate protein model quality accurately with TopScore and TopScoreSingle, I developed a fully automated template-based protein structure prediction workflow called TopModel. TopModel differs from traditional structure prediction pipelines in two main ways: First, template selection is performed based on predicted template similarity to the native structure using deep neural networks. Then, top-down consensus is used to discard templates, that are structurally different from the best template, rather than selecting the fold most often found using regular consensus. Second, instead of using regular consensus between initial models during model refinement, TopModel predicts the residue-wise error using TopScore and TopScoreSingle and uses the predicted error to locate and correct erroneous regions. Compared to its primary predictors, TopModel has a much better template selection and, compared to other template-based structure prediction workflows, it shows a significant improvement in model quality. However, TopModel is still at a disadvantage compared to methods, which use *ab initio* folding, state-of-the-art contact prediction and protein domain prediction (**Chapter 10, Publication II**).

To demonstrate the usefulness and the power of fully automated protein structure prediction with TopModel, I applied the workflow to several projects in which the structure of the target protein of interest was unknown. Three such examples are detailed in this thesis, namely: (1) Structure prediction of Alanopine Dehydrogenase and the identification of

structural determinants of ligand specificity by comparison to related structures. Furthermore, the substrate binding pathway and substrate inhibition was predicted, using the predicted structure as a starting point molecular dynamics simulations and free energy calculations (**Chapter 11, Publication III**). (2) Prediction of feline and human APOBEC3 protein structures to determine residues important for binding viral Vif proteins from HIV and FIV viruses, respectively. These predictions enabled the identification of key differences in host-pathogen interaction patterns in humans and domestic cats, which are critical to consider, when using cats as model animals for the study of HIV (**Chapter 12, Publication IV**). (3) Prediction of the dimeric structure of the GAF domain of the plant Ethylene Receptor 1 (ETR1), which induces fruit ripening upon binding of the plant hormone ethylene. This prediction enabled the identification of binding sites for the NOP-1 peptide derived from the Ethylene-Insensitive Protein 2 (EIN2), which inhibits the ripening process. These binding sites were identified by combining free ligand diffusion molecular dynamics simulations of the structures predicted by TopModel with experimental validation. The predicted binding sites also provided insights into the inhibition mechanism of NOP-1 on fruit ripening and thus provide a basis for industrial application and improvement of peptide inhibitors of fruit ripening (**Chapter 13, Publication V**).

In all, the results presented in this thesis show that integrating different primary predictors for protein model quality estimation (**Publication I**), and template-based structure prediction (**Publication II**), and combining their outputs using deep neural networks is highly effective. Furthermore, the usefulness of the methods developed in this thesis was demonstrated by applying them to target proteins of high biological (**Publication III**), medical (**Publication IV**) and industrial (**Publication V**) interest.

However, there is still a need for methods that can handle large proteins with multiple domains better, as well as a need for methods that can accurately predict protein structures for which no templates are available. Therefore my ongoing work is focused on developing methods that can identify domains in the target sequence (TopDomain), predict properties of these domains which can be used for *ab initio* folding (TopContact), and predict the interactions between domains or proteins (TopInterface), in order to construct large multi-domain proteins and protein-protein complexes by docking together individual proteins and/or protein domains (TopDock).

15. ACKNOWLEDGEMENTS

When, moving to Düsseldorf, I started a chapter of my life filled with challenges and new experiences, but also a lot of friends and great memories that I will never forget. I want to express my gratitude to some of the people who have made my time here in Germany unforgettable.

Thanks to Prof. Dr. Holger Gohlke, for giving me the possibility of doing a PhD at the Heinrich-Heine University and for giving me the freedom and encouragement to pursue my interests and ideas, as well as his continuous enthusiastic and knowledgeable support and guidance through the years.

Thanks to all my colleagues and friends of the CPC lab. In particular, I would like to thank Stephan Schott-Verdugo, Lukas Wäschenbach, Dr. Christoph G.W. Gertzen and Dr. Prakash Chandra Rathi for their valuable insights, fruitful collaborations and good discussions in the office. I also want to thank Dr. Christoph G. W. Gertzen, Dr. Benoit David, Stephan Schott-Verdugo, Nicola Porta and David Bickel for corrections and feedback on this thesis.

I would like to further thank Prof. Dr. Carsten Münk, Prof. Dr. Dieter Häussinger, Prof. Dr. Georg Groth, Prof. Dr. Lutz Schmitt, Prof. Dr. Klaus Cichutek, Prof. Dr. Erhard Bremer, Prof. Dr. Wolfgang Buckel, Prof. Dr. Manfred K. Grieshaber, Dr. Sander H.J. Smits, Dr. Markus Dick, Dr. Christopher Pfleger, Dr. Prakash Chandra Rathi, Dr. Dalibor Milić, Dr. Zeli Zhang, Dr. Qinyong Gu, Dr. Ananda Ayyappan Jaguva Vasudevan, Dr. Ignatio G. Bravo, Dr. Björn-Philipp Kloke, Dr. Philipp Neudecker, Dr. Sakshi Khosa, Dr. Benedikt Frieg, Dr. Rebecca Clemens, Dr. Astrid Höppner, Dr. Diana Kleinschrodt, Anika Hain, Sascha Hasheminasab, Kei Sato, Ulrike Hergert, Tatu Meyer, Anna Kinnen, Nils Widderich, Marco Pittelkow, and Nicola Porta for fruitful collaborations.

Finally I would like to thank (In alphabetical order) Anuseema Bhadauriya, Dr. Alexander Metz, Dr. Bartholomäus Daniel Ciupka, Dr. Dennis M. Krüger, Bastian Schepers, Birte Schmitz, Dr. Christian Hanke, Christina Nutschel, Daniel Becker, Dr. Denis Schmidt, Dr. Emanuele Ciglia, Filip König, Dr. German Erlenkamp, Dr. Giulia Pagani, Dr. Ido Ben-Shalom, Dr. Jagmohan Saini, Jonas Dittrich, Lukas Wäschenbach, Dr. Markus Dick, Mykola Dimura, Dr. Neha Verma, Pegah Golchin, Dr. Prakash Chandra Rathi, Sabahuddin Ahmad, Susanne Hermans and Dr. Tobias Kröger, for the times we shared together and the memories we will never forget.

A big thanks to my parents for their constant love and support in all moments of my life. Special thanks to TzuJung: I am so glad that I was here in Düsseldorf long enough to meet you, and I am looking forward to filling our future with great and loving memories.

16. REPRINT PERMISSIONS

Publication I

Reprinted (adapted) with permission from:

"TopScore: Using deep neural networks and large diverse datasets for accurate protein model quality assessment"

Daniel Mulnaes and Holger Gohlke

Journal of Chemical Theory and Computation; 2018, 14, 6117-6126.

Publication II

Reprinted (adapted) with permission from:

"TopModel: A deep neural network and model quality driven meta-approach to template-based protein structure prediction"

Daniel Mulnaes, Philipp Neudecker, Nicola Porta, Rebecca Clemens, Sander Smits and Holger Gohlke

Journal of Chemical Theory and Computation; 2019, Submitted.

Publication III

Reprinted (adapted) with permission from:

"Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion"

Holger Gohlke, Ulrike Hergert, Tatu Meyer, Daniel Mulnaes, Manfred K. Grieshaber, Sander H.J. Smits and Lutz Schmitt

Journal of Chemical Information and Modelling. 2013, 53, 2493–2498.

Publication IV

Reprinted (adapted) with permission from:

"Determinants of FIV and HIV Vif sensitivity of feline APOBEC3 restriction factors"

Zeli Zhang, Qinyong Gu, Ananda Ayyappan Jaguva Vasudevan, Anika Hain, Björn-Philipp Kloke, Sascha Hasheminasab, Daniel Mulnaes, Kei Sato, Klaus Cichutek, Dieter Häussinger, Ignatio G. Bravo, Sander H.J. Smits, Holger Gohlke and Carsten Münk

Retrovirology; 2016, 13, 46.

Publication V

Reprinted (adapted) with permission from:

"Recognition motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1"

Dalibor Milić, Markus Dick, Daniel Mulnaes, Christopher Pfleger, Anna Kinnen, Holger Gohlke and Georg Groth

Nature Scientific Reports 2018, 8, 3890.

17. PUBLICATION I

TopScore: Using deep neural networks and large diverse datasets for accurate protein model quality assessment

*Daniel Mulnaes[1] and Holger Gohlke[1,2]**

[1]Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine University, Düsseldorf, Germany

[2]John von Neumann Institute for Computing (NIC), Jülich Supercomputing Center (JSC) & Institute for Complex Systems - Structural Biochemistry (ICS 6), Forschungszentrum Jülich GmbH, Jülich, Germany

JCTC Journal of Chemical Theory and Computation
Article
Cite This: J. Chem. Theory Comput. 2018, 14, 6117−6126
pubs.acs.org/JCTC

TopScore: Using Deep Neural Networks and Large Diverse Data Sets for Accurate Protein Model Quality Assessment

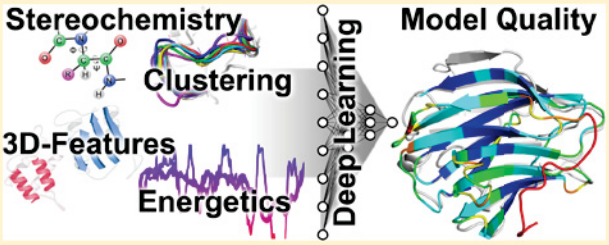
Daniel Mulnaes[†] and Holger Gohlke*[,†,‡]

[†]Department of Mathematics and Natural Sciences, Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Germany

[‡]John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC) & Institute for Complex Systems - Structural Biochemistry (ICS 6), Forschungszentrum Jülich GmbH, Jülich, Germany

Supporting Information

ABSTRACT: The value of protein models obtained with automated protein structure prediction depends primarily on their accuracy. Protein model quality assessment is thus critical to select the model that can best answer biologically relevant questions from an ensemble of predictions. However, despite many advances in the field, different methods capture different types of errors, begging the question of which method to use. We introduce TopScore, a meta Model Quality Assessment Program (meta-MQAP) that uses deep neural networks to combine scores from 15 different primary predictors to predict accurate residue-wise and whole-protein error estimates. The predictions on six large independent data sets are highly correlated to superposition-independent errors in the model, achieving a Pearson's R_{all}^2 of 0.93 and 0.78 for whole-protein and residue-wise error predictions, respectively. This is a significant improvement over any of the investigated primary MQAPs, demonstrating that much can be gained by optimally combining different methods and using different and very large data sets.



1. INTRODUCTION

Protein structure prediction is an established field of structural bioinformatics, but computational models still contain errors that limit their utility to answer biologically relevant questions.[1] Thus, it is paramount to establish which model in an ensemble of predicted structures is the most correct (global scoring problem), and which parts of that model are most reliable (local scoring problem). Model quality assessment programs (MQAPs) are therefore a critical part of automated protein structure prediction.[1] Many MQAPs have been developed in the last few decades. These have been continuously evaluated in the biannual Critical Assessment of Structure Prediction (CASP) blind experiments and led to great improvements in model selection.[2] Current state-of-the-art MQAPs can be broadly divided into four categories:

I. **Single model methods.** These methods are generally fast and memory-efficient and evaluate model quality based on features from a single model. They fall into three groups: 1. Physics- or knowledge-based potentials, such as contact, angle, or distance potentials.[3−8] 2. Methods measuring geometric properties such as bond and dihedral angles, atom volume, packing, or steric clashes.[9−12] 3. Methods that evaluate agreement between features in the model and features predicted from the primary sequence, such as secondary structure and solvent accessibility.[13−18]

II. **Clustering methods.** Clustering methods compare multiple models of the same sequence. They either assume that the native structure is near a cluster center and has structural fragments that are more abundant in the model ensemble or that the models represent sampling of potentially folded or partially folded states.[19−23] Clustering methods are effective and often more accurate than single model methods especially for difficult systems.[2] However, they are computationally expensive, and their performance suffers if the correct fold is underrepresented. Furthermore, they become increasingly slow and memory-dependent as the size of the model ensemble increases, making their use unfeasible for large model ensembles.

III. **Quasi-single methods.** The main idea of quasi-single methods is to use a small independent set of good models as a reference for a cluster-based scoring. Two main approaches have been used in which the reference ensemble is either predicted from the primary sequence[24,25] or selected with single model methods from the input ensemble.[26] The first approach is computationally expensive, however, and does not ensure that the reference ensemble is more correct than the input ensemble. The second approach, on the other hand, only improves ranking of models that are not ranked at the top by single model methods, which would not be selected in any case.

Received: July 6, 2018
Published: September 25, 2018

DOI: 10.1021/acs.jctc.8b00690
J. Chem. Theory Comput. 2018, 14, 6117−6126

IV. **Meta-MQAPs.** Meta methods seek to combine multiple methods to improve predictions. While they are slower than using any individual MQAP method on its own, their accuracy is generally superior, since they find different types of errors using different methods.[26−28]

Despite these remarkable developments, however, the current state-of-the art MQAPs have often been trained with different methods on different data sets with different models or ensemble compositions and with different goals or target functions in mind (see Table S1). This has led to a high degree of MQAP diversity with a focus on different types of errors, which begs the question of which method to use in a general case. Yet, it also makes the MQAPs ideal components for designing a general meta-MQAP. In this study, we present the development and evaluation of a meta-MQAP called TopScore. TopScore was developed and validated on six very large and diverse data sets totaling 910 targets, ~1.7×10^5 models, and ~2.9×10^7 residues employing deep neural networks (DNNs) to combine predictions from 15 different primary MQAPs. TopScore aims to achieve four goals important to model quality assessment:

I. Distinction between wrongly and correctly folded models
II. Assessment of the whole-model error (global score)
III. Ranking models to select the one closest to the native structure
IV. Assessment of the residue-wise error of the model (local score)

TopScore shows a consistent and robust performance across multiple quality measures and data sets of very different compositions regarding the methodology for generating models and the model's error distribution. As such, TopScore proved superior to any of the 15 evaluated state-of-the-art primary MQAPs.

2. METHODS

Target Function. To evaluate the quality of TopScore, a definition of correctness of the scored models with respect to the native structures is needed as a target function. We are interested in definitions of error at the local and global level that are bounded from zero (native) to one (wrong). Many previous MQAPs have used superposition-dependent target scores such as the LG-Score,[29] S-Score,[30] TM-Score,[31] GDT-TS Score,[32] or MaxSub-Score[33] and superposition-independent scores such as the Q-Score[34] and the lDDT score.[35] Superposition-dependent scores face the challenge that calculating the target score depends on the structural alignment of the model and the true structure, a task that comes with its own difficulties and sources of error: The structural alignment of a model to the native structure can, e.g., be difficult if only parts of the model are correct, or if the model is in a different conformation than the native structure.[35]

The lDDT score[35] is an all-atom measure of structure similarity, which is independent of structural superposition. It compares intraprotein rather than interprotein atomic distances given certain distance cutoffs. This makes it insensitive to domain orientation and large conformational changes, while still being a highly accurate measure of structural similarity.[35] It is bounded between 0 and 1 both for local and global scores.[35] To provide an intuitive relation between low values for the correctness assessment and good models (reminiscent, e.g., of a small structural deviation), the target function of TopScore is chosen as 1 − lDDT score, which is referred to as the "lDDT error".

Primary MQAPs. The primary MQAPs are the input programs used for local and global error estimation and upon which the TopScore consensus is calculated. These MQAPs (Table 1) cover stereochemical analysis methods, knowledge-based potentials, clustering methods, methods that compare measured features (e.g., secondary structure and solvent accessibility) with ones predicted from primary sequence, and composite scoring methods that use a combination of the above.

Table 1. Local and Global Primary MQAPs in TopScore[a,c]

MQAP	local score	global score
PROCHECK[9]	deviation score[b]	local score average
MolProbity[10]	average score[b]	MolProbity score
ANOLEA[5,36,37]	ANOLEA energy[b]	local score average
ProSA2003[6]	ProSA2003 combined energy[b]	ProSA2003 energy normalized by length
DOPE[3]	local DOPE energy[b]	global DOPE energy normalized by length
GOAP[4]	not available	GOAP energy normalized by length
ProQ2[14,30]	residue-wise S-Score	predicted LG-Score
ProQ2D[17]	predicted local lDDT score	predicted lDDT score
ProQ3D[17]	predicted local lDDT score	predicted lDDT score
SVMQA[18,38]	not available	average of predicted GDT_TS and TM-Score
QMEAN6[39]	predicted local lDDT score	predicted lDDT score
SELECTpro[13]	not available	SELECTpro energy normalized by length
ModFOLDClust2[19]	*predicted C_α atom distance to native*	*local score average*
SPICKER[20]	*not available*	*relative cluster size*
Pcons[30]	*residue-wise S-Score*	*global S-Score normalized by length*

[a]Clustering-based scores are indicated in italics. [b]Scores smoothed over a five-residue window using triangular smoothing. [c]See Table S1 for a listing of primary MQAP target scores, training set sizes, and reported correlations from the literature.

For PROCHECK, no local or global score exists as of yet. A simple local score was therefore devised as follows: The maximum deviation of bond lengths, bond angles, and dihedral angles within a residue was calculated. This value was smoothed within a five-residue window using triangular smoothing. The global score is calculated as the average across all residues. In MolProbity, four local scores exist: A rotamer-score, a Ramachandran-score, and the maximum deviation from optimal values of the bond angles and bond lengths, respectively. These were normalized to represent probabilities of the value occurring, averaged, and smoothed within a five-residue window using triangular smoothing. For SPICKER, the relative cluster size was used as a quality score, based on the assumption that larger clusters are more likely to contain correct models.

Since the primary MQAPs have output scores in different units, deep neural networks (DNNs) were trained for each MQAP on a large data set of diverse models (see next section) to predict the lDDT error from the output score. This ensures that each MQAP is treated the same afterward and allows for

DOI: 10.1021/acs.jctc.8b00690
J. Chem. Theory Comput. 2018, 14, 6117−6126

evaluation of the primary MQAPs ability to predict the lDDT error on their own. Each primary MQAP is run with default parameters to gain optimal performance, i.e., for the ProQ suite of predictors, side chain repacking is enabled. The performance of each primary MQAP is then compared to TopScore. Due to the need of running multiple different primary predictors, the average runtime of TopScore is about 1.5 min per model in the ensemble or about 5 h for an ensemble of 200 models.

Data Sets. Because we aim to build a scoring function that should be broadly applicable to score models from *ab initio* structure prediction, homology modeling, and CASP ensembles, it is important that the training data reflects as many different types of models and ensembles as possible. This is relevant both in terms of the methods used to generate the models and the composition of the model ensembles for each target. Studies have shown[26,40] that, in particular, two cases are difficult to score correctly: Cases when the large majority of models are of very poor quality and cases with centralized distributions of model scores, i.e., all models score very similarly. Six different data sets are therefore used in order to train the DNNs on different types of model ensembles.

The first data set is generated using our in-house structure prediction meta-tool TopModel,[41−43] which uses multiple state-of-the-art threading and sequence/structure alignment tools to generate a large ensemble of models from different pairwise and multiple alignments of the top three highest ranked template structures. We chose the Top100 protein data set[44] as target structures, due to its diversity and limited size. To simulate differences in modeling difficulty, we performed a screening during which each target structure is predicted from its primary sequence, with cutoffs imposed to template-target sequence identity. The cutoffs were chosen as 90%, 60%, and 30% identity, respectively, emulating trivial, easy, and difficult targets. The Top100 data set consists of 300 targets comprising $\sim 3.7 \times 10^4$ models with in total $\sim 7.5 \times 10^6$ residues.

Furthermore, we selected five independent data sets generated by different methods: First, the 3DRobot data set[40] consists of 200 targets comprising $\sim 6.0 \times 10^4$ models with a uniform error distribution with in total $\sim 8.0 \times 10^6$ residues. Second, from the DecoysR'us data set,[45] we selected all targets with less than 1000 models from the multiple decoy set. This yielded 109 targets comprising $\sim 1.4 \times 10^4$ models with in total $\sim 1.5 \times 10^6$ residues. Third, we selected the iTASSER-II *ab initio* modeling decoy set,[46] which contains $\sim 2.4 \times 10^4$ models of 56 targets with in total $\sim 1.9 \times 10^6$ residues. Fourth and fifth, we selected the CASP10 stage1 and stage2 data sets,[2] which contain 1.6×10^4 models of 116 targets with in total $\sim 4.3 \times 10^6$ residues. We combined the data from all data sets, totaling 781 targets, $\sim 1.5 \times 10^5$ models, and $\sim 2.3 \times 10^7$ residues, into a combined set that was used for training and validating TopScore.

To evaluate TopScore on a completely independent data set, the CASP11 and CASP12 stage1 and stage2 data sets[47] were combined and used. This set contains 129 targets, $\sim 2.1 \times 10^4$ models, and $\sim 5.8 \times 10^6$ residues. From this data set, no models were used during TopScore training; rather, they were only used for evaluation of the trained methods.

To ensure consistency of CASP ensembles, the reference sequence was determined as the longest sequence shared by the most models of an ensemble. Models with missing residues were repaired using Modeller.[48] Missing residues in the native structure were repaired in the same way. For each data set, all primary MQAP scores and target scores were calculated using local versions of the respective programs. The lDDT score was calculated with default settings (GDT_HA) and radius cutoff (R_0 = 15 Å).

Deep Neural Networks. The deep neural networks (DNNs) of TopScore were trained using the Python package SciKit-learn version 1.8.1.[49] The training of all DNNs was done in an identical manner. The data was first divided randomly into training and evaluation sets, leaving 80% of the data for training and 20% for final evaluation. For global scoring, this was done on the target level, while for local scoring all residues were considered independent samples during training, i.e., no information of which model a residue belongs to is retained. The evaluation data was left out of the entire training procedure and was only used to evaluate the final DNNs. The IsolationForest method in SciKit-learn was used to remove the 1% most severe outliers from the training data to ensure a fit of the most representative data.

The DNNs were trained using the MLPRegressor method with the ADAM stochastic gradient descent algorithm[50] and default weight decay settings for regularization to prevent overfitting. To estimate the meta-parameters of the DNNs, the training data was randomly subdivided into training and test sets using the k-fold method in SciKit-learn to perform 5-fold cross-validation. By using a grid-search, the DNN architecture and the neuron type were varied. Architectures ranged from a single-hidden-layer perceptron to a three-hidden-layer perceptron with 10, 20, 40, 80, and 160 neurons in the first layer and subsequent layers having half the neurons of the previous layer. The tested transition functions were logistic, hyperbolic tangent, and rectified linear unit function. For each 5-fold cross-validation split, the DNN was trained on 50% of the data and evaluated on the rest. To prevent overtraining, the training was stopped early if the correlation between predicted and true lDDT errors on the test half decreased. After selecting an optimal architecture and transition function, a DNN was trained on all the training data except for the outliers, again setting aside 50% of the data for testing and applying early stopping to prevent overtraining. The final DNN performance was evaluated on the 20% of the data left out as evaluation data of the initial set at the beginning of the training, which was thus not considered during training at all. In all cases, we observed a difference of less than 1% correlation on evaluation data compared to training data. An identical approach was used for the local scores for each method with three main differences: 1. Primary predictors GOAP, SPICKER, SELECTpro, and SVMQA were left out due to the lack of a local score; 2. Outlier filtering was omitted due to memory constraints arising from a random forest with 2.7×10^7 data points; 3. Based on the outcome of several grid searches, the DNN architecture was fixed to 160, 80, and 40 neurons in consecutive hidden layers to speed up the local DNN training.

After each primary MQAP value was normalized using the MQAP-specific DNN, with specific architectures and transition functions for each primary MQAP, the normalized values were used as inputs for another DNN to calculate the TopScore prediction. This DNN aims to combine the normalized primary MQAP scores in an optimal way, adjusting for cross-correlations and performance differences, to achieve a prediction that is on average better than any single primary MQAP. The DNN was trained in a manner identical to the training of the primary MQAP DNNs for selecting meta-parameters. For the local score predictor, a DNN is trained in

DOI: 10.1021/acs.jctc.8b00690
J. Chem. Theory Comput. 2018, 14, 6117−6126

an identical manner as for the global one. The architecture of the entire TopScore workflow is outlined in Figure 1. The DNN architectures and transition functions of the individual networks as a result of the grid-search optimization can be found in Table S2.

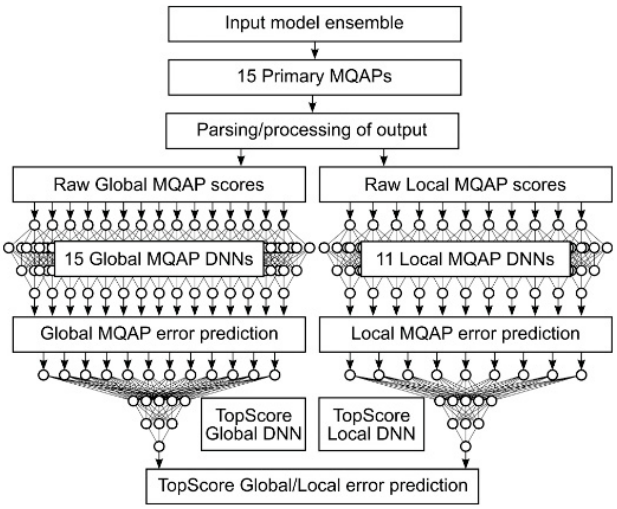


Figure 1. TopScore architecture. TopScore runs 15 primary MQAPs, parses and processes their output, and feeds the raw scores as input for individual primary MQAP DNNs to predict global or local errors. These predictions are used as input for a global or local consensus DNN that predicts the global or local lDDT error. The architecture for TopScoreSingle is the same as shown above except that clustering primary MQAPs are omitted. See Figure S1 for plots of the individual DNN functions and their fit to the true lDDT error. See Table S2 for a description of the individual DNN architectures (layers, neurons, and transition functions).

Performance Measures. MQAP performance can be evaluated in different ways and varies depending on the measure used.[26] We evaluated MQAP performance with four quality measures: The area under the curve (AUC), Pearson's product-moment coefficient of determination (R_{all}^2), the weighted mean Pearson's coefficient of determination (R_{wm}^2), and the "Loss". These were computed as in Dong et al.[26] The AUC is calculated based on the receiver operator characteristic curves for each method using a 0.5 cutoff in the lDDT error. This cutoff was selected because the lDDT error scale spans from 0 to 1 and because for most of the data sets the median lDDT is close to this value (Table S3). It determines how well a MQAP separates good (lDDT error <0.5) from bad (lDDT error >0.5) models (Goal I). R_{all}^2 is the squared correlation coefficient calculated on all the combined data from all ensembles and determines the accuracy with which a MQAP predicts the global lDDT error of the models (Goal II). R_{wm}^2 is calculated using Fisher's r-to-z transformation as the weighted average Pearson's R^2 across all model ensembles and shows the ability of a MQAP to rank the individual model ensembles (Goal III). The loss is the difference in the lDDT error between the highest ranked model and the best model in the ensemble, averaged across ensembles. It indicates the lDDT error that could be avoided if the best model was always selected instead of the top ranked one and, thus, shows the ability for a MQAP to select the best model (Goal IV).

3. RESULTS

DNN Training. To visualize the training of the primary MQAP DNNs, the raw primary MQAP values are binned, and the lDDT error distributions for each bin are compared to predicted values from the DNNs. These can be found in Figure S1 and show the value of outlier filtering and the robustness of the DNN training with regards to overfitting.

TopScore and TopScoreSingle. Previous CASP rounds[2,47,51] have shown that clustering-based methods such as ModFOLDClust2 and Pcons are superior to single-model methods in terms of distinguishing between good and bad models and for local quality assessment. This is a result of the increased amount of information available to the former methods from different models in the ensemble. Single-model methods, on the other hand,[17,18] have shown superior ability to select the best model, especially when the ensemble is highly heterogeneous and the best fold is underrepresented and not part of a cluster. To make sure that TopScore does not gain all of its performance from the clustering methods, as well as to obtain a method that can select the best model if the ensemble is heterogeneous, it is interesting to see the performance when clustering methods are omitted as primary MQAPs. A predictor called TopScoreSingle is therefore trained on all primary MQAPs except SPICKER, Pcons, and ModFOLDclust2, in a manner otherwise identical to that of TopScore (Figure 1). Additionally, we compare TopScore to Pcomb,[52] which is a linear combination of single-model and clustering scores.

The global and local performance for a subset of primary MQAPs compared to TopScore and TopScoreSingle is shown in Figures 2 and 3, respectively. It is clear that prediction of local error is a harder task than prediction of global error, as seen by lower R_{all}^2 and R_{wm}^2 values, which is likely due to the limited information available for a single residue and its environment compared to that of the entire protein.

TopScore Performance. A detailed description of the four quality measures AUC, R_{all}^2, R_{wm}^2, and Loss used to evaluate TopScore performance is given in the Methods section. These measures determine the ability of a MQAP to distinguish between good and bad models, the accuracy with which it predicts the global lDDT error of the models, its ability to rank the individual model ensembles, and its ability to select the best model, respectively. Our analyses (Figures 2 and 3; see Tables S4, S5, and S6 for numerical values) show that, overall, TopScore and TopScoreSingle outperform all other primary MQAPs significantly ($p < 0.05$) with respect to all quality measures on the combined data set, which includes the iTASSER-II, Top100, 3DRobot, DecoysR'us, and CASP10 stage1/2 data sets. Furthermore, on the CASP11/12 test set, we see significantly improved performance of both TopScore and TopScoreSingle compared to primary MQAPs for both global and local scoring for all quality measures, with the exception of global R_{all}^2, according to which the performance of TopScoreSingle is comparable to ProQ3D, and TopScore is comparable to Pcomb. TopScoreSingle shows a decrease in performance on the CASP11/12 data set compared to the combined data set, which is likely due to the data set's difficulty as reflected in a median lDDT error of 0.61 (Table S3). This drop is seen less for TopScore because TopScore also considers clustering MQAPs, which are good at separating good from bad models. This overall remarkable performance of TopScore and TopScoreSingle is due to the DNN's ability to

DOI: 10.1021/acs.jctc.8b00690
J. Chem. Theory Comput. 2018, 14, 6117−6126

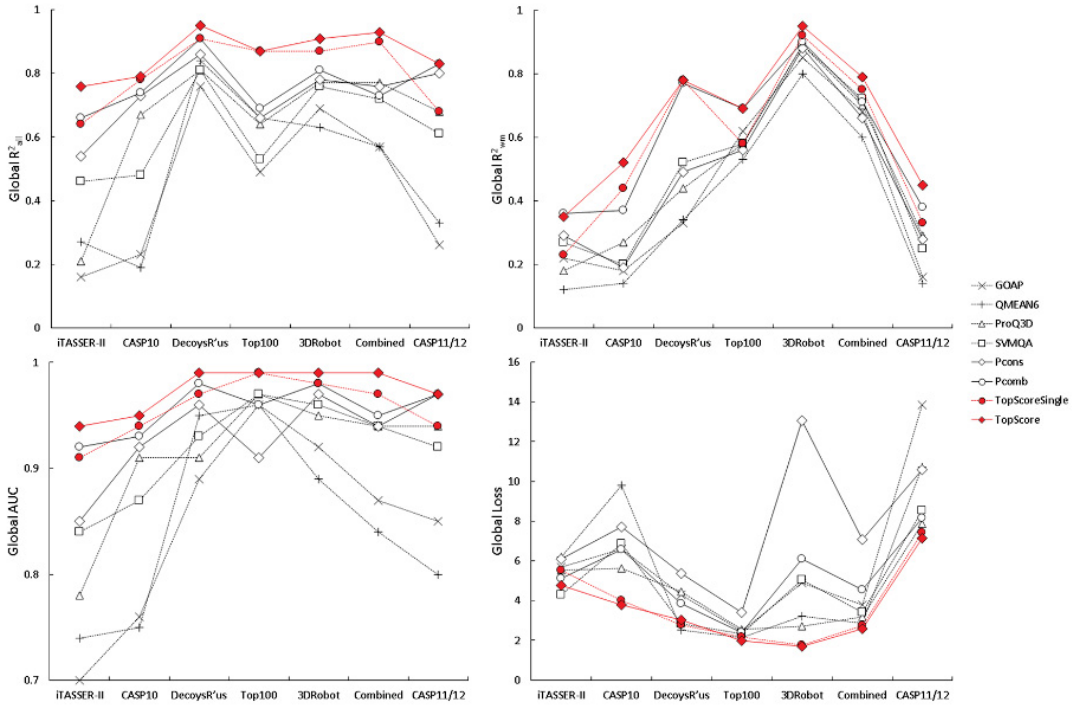


Figure 2. TopScore global performance. TopScore (red circles) and TopScoreSingle (red dashes) global performance compared to a subset of primary predictors (black). Dashed lines represent single-model methods, and full lines represent methods that use clustering information. The 95% confidence intervals were calculated using the Fischer r-to-z transformation. The widest confidence interval for any R_{all}^2 or R_{wm}^2 was 0.01 and 0.12, respectively. Statistical significance was determined by the two-sided Steiger test.[53] Accordingly, the R_{all}^2 and R_{wm}^2 of TopScore and TopScoreSingle are significantly different from any primary MQAP for the combined data set ($p < 0.05$). In terms of R_{all}^2, for the CASP11/12 data set, TopScoreSingle is not significantly different from ProQ3D, and neither is TopScore when compared to Pcomb. See Tables S4 and S5 for numerical values of all investigated MQAPs. See Table S3 for statistics of lDDT distributions of individual data sets.

combine the performance of the primary MQAPs as well as the very large and diverse data sets used for model training.

For the global scores (Figure 2, Tables S4 and S5), individual methods occasionally perform better on a single quality measure and specific data set, which stresses the importance of evaluating MQAPs across multiple quality measures and data sets. One example is SVMQA outperforming TopScoreSingle and TopScore on the iTASSER-II data set in terms of Loss and TopScoreSingle also in terms of R_{wm}^2. Another example is QMEAN6 outperforming TopScore on the DecoysR'us data set in terms of Loss. However, when all measures and data sets are considered, the improvement obtained from TopScore and TopScoreSingle is unequivocal.

For the local scores (Figure 3, Table S6), TopScoreSingle even performs better than clustering-based primary MQAPs for several data sets. This finding indicates that there is still much to gain from combining single model MQAPs even when no clustering information is included and shows that clustering-based methods are no longer exclusively the best for evaluating residue-wise quality. Furthermore, our findings show that MQAP performances vary significantly between different data sets, which also holds for TopScore and TopScoreSingle regardless of whether the data sets were used for their training or not. This finding suggests that the data sets used for the development of each primary MQAP have a large impact on their performance. This finding does not necessarily reflect an overtraining of the MQAPs; rather, it results from different data set compositions and a limit to the degree of generalization possible from limited training data.

Figures 2 and 3 give a good indication of the overall performance of each predictor but do not show if the uncertainty of the predictions is uniformly distributed or if certain ranges of the lDDT error are more difficult to predict. Figure 4 shows that it is easier to predict the global error of very bad models and very good models (lDDT error <0.2 or lDDT error >0.50; standard deviation (SD) = 0.01−0.04) rather than of intermediate ones (lDDT error = 0.2−0.50, SD = 0.04−0.055). This trend is also seen for TopScoreSingle and for the local predictors.

Primary MQAP Performance. When analyzing the performance of primary MQAPs, we see considerable differences between performances on different data sets as well as with respect to different quality measures. When considering the combined data set, some primary MQAPs show a good overall estimate of the lDDT error of a model relative to all models (indicated by R_{all}^2) but are considerably worse at ranking models within an ensemble (indicated by R_{wm}^2). This trend is seen for PROCHECK, MolProbity, and SPICKER. Using SPICKER's cluster size as an example, this is not surprising: Since models belonging to the same cluster have the same (or for other MQAPs a highly similar) score, the ranking of an ensemble will be worse with increasing cluster size. However, an ensemble where the largest cluster is small is less likely to contain good models than one where most models belong to the same cluster. Methods such as GOAP, DOPE, ANOLEA, and SELECTpro show the opposite trend, with a higher R_{wm}^2 than R_{all}^2. These MQAPs fail to correctly estimate the error of some ensembles, leading to a decrease in R_{all}^2

DOI: 10.1021/acs.jctc.8b00690
J. Chem. Theory Comput. 2018, 14, 6117−6126

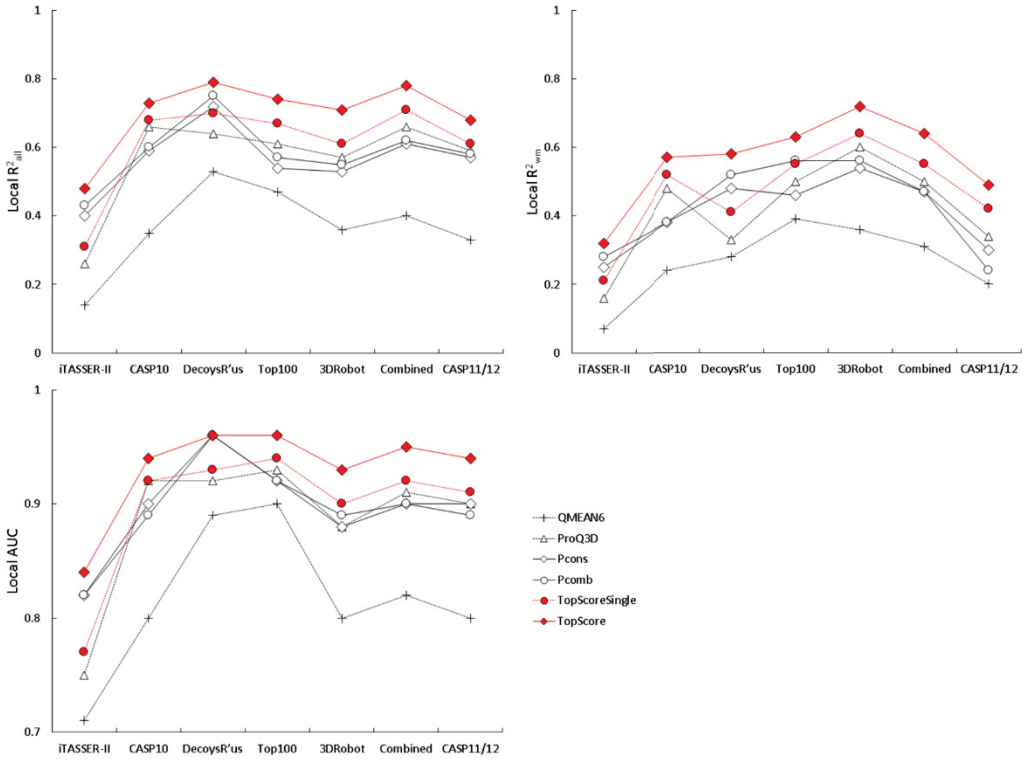


Figure 3. TopScore local performance. TopScore (red circles) and TopScoreSingle (red dashes) local performance compared to a subset of primary predictors (black). Dashed lines represent single-model methods, and full lines represent methods that use clustering information. The 95% confidence intervals and statistical significances are calculated in the same way as for Figure 2. The widest confidence interval for any R_{all}^2 or R_{wm}^2 was 0.001 and 0.17, respectively. The R_{all}^2 and R_{wm}^2 of TopScore and TopScoreSingle are significantly different from any primary MQAP ($p < 0.05$). See Table S6 for numerical values for all investigated MQAPs.

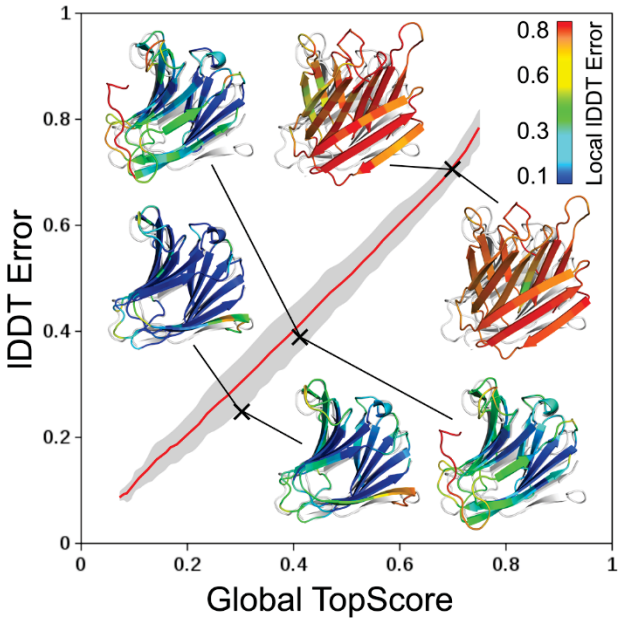


Figure 4. TopScore performance. The global TopScore predictions plotted against the lDDT error of the models for the combined data set. Three randomly selected example models of PDB ID 4BMB from the 3DRobot data set are shown colored according to local TopScore error prediction (lower triangle) and true local lDDT error (upper triangle).

compared to R_{wm}^2. This could be because these types of ensembles were not seen during the training of these MQAPs.

The top performing primary MQAPs in terms of R_{all}^2 and R_{wm}^2 are, as expected, the clustering-based methods ModFOLDClust2 and Pcons, as well as the composite method Pcomb. They show similar global performances, with Pcons performing better than ModFOLDClust2 in most cases, although both show a higher Loss on some data sets than some single model methods. This shows that much of their performance is gained from correctly ranking the majority of the models, but that this does not necessarily lead to the best model being ranked at the top. Since the majority of the models is not selected for further analysis, correctly ranking them has a limited effect on improving the quality of the top ranked model. For local error predictions, Pcons shows significant advantage over ModFOLDClust2, which is likely due to differences in target score (Table S1). Pcomb shows an overall higher performance than most primary MQAPs on most data sets due to its composite nature.

Figures 2 and 3 show that different primary MQAPs can have very different performances for the different data sets. PROCHECK and ANOLEA, for example, have a R_{all}^2 of 0.32 and 0.44 on the Top100 as well as 0.46 and 0.73 on the DecoysR'us data set, respectively, but less than 0.2 for the 3DRobot and iTASSER-II data sets. This is expected as the former data sets contain mostly models close to the native structure, while the latter contain mostly models representing misfolded or unfolded proteins. When most models are close to the native structure, many errors can be found from stereochemistry or loop energetics, while at the other end of

DOI: 10.1021/acs.jctc.8b00690
J. Chem. Theory Comput. 2018, 14, 6117−6126

Table 2. MQAP Cross Correlation: R_{all}^2 between Primary MQAPs and TopScore MQAPs[a]

TopScore & TopScoreSingle Cross-correlations	PROCHECK	MolProbity	ANOLEA	ProSA2003	DOPE	GOAP	QMEAN6	ProQ2	ProQ2D	ProQ3D	SVMQA	SELECTpro	ModFOLDClust2	Pcons	SPICKER	TopScore	TopScoreSingle
PROCHECK	-	0.64	0.01	0.05	0.02	-	0.15	0.14	0.14	0.28	-	-	0.28	0.30	-	0.32	0.32
MolProbity	0.30	-	0.00	0.04	0.06	-	0.12	0.13	0.19	0.26	-	-	0.30	0.32	-	0.31	0.31
ANOLEA	0.01	0.08	-	0.04	0.08	-	0.21	0.16	0.08	0.10	-	-	0.06	0.06	-	0.12	0.14
ProSA2003	0.21	0.14	0.06	-	0.09	-	0.20	0.15	0.15	0.16	-	-	0.09	0.10	-	0.16	0.18
DOPE	0.12	0.19	0.19	0.50	-	-	0.13	0.05	0.05	0.06	-	-	0.02	0.02	-	0.09	0.13
GOAP	0.19	0.23	0.20	0.55	0.64	-	-	-	-	-	-	-	-	-	-	-	-
QMEAN6	0.29	0.42	0.21	0.46	0.57	0.65	-	0.44	0.41	0.48	-	-	0.25	0.28	-	0.51	0.56
ProQ2	0.22	0.21	0.14	0.57	0.47	0.67	0.57	-	0.73	0.7	-	-	0.45	0.51	-	0.65	0.70
ProQ2D	0.26	0.19	0.06	0.55	0.38	0.62	0.51	0.83	-	0.82	-	-	0.48	0.54	-	0.70	0.77
ProQ3D	0.37	0.24	0.08	0.55	0.36	0.62	0.57	0.81	0.91	-	-	-	0.54	0.61	-	0.85	0.93
SVMQA	0.19	0.18	0.15	0.52	0.38	0.69	0.49	0.76	0.74	0.74	-	-	-	-	-	-	-
SELECTpro	0.03	0.02	0.00	0.29	0.34	0.26	0.16	0.23	0.24	0.19	0.18	-	-	-	-	-	-
ModFOLDClust2	0.28	0.13	0.05	0.41	0.24	0.41	0.32	0.57	0.58	0.64	0.61	0.13	-	0.87	-	0.70	-
Pcons	0.32	0.14	0.03	0.38	0.18	0.38	0.29	0.54	0.57	0.64	0.59	0.09	0.91	-	-	0.79	-
SPICKER	0.17	0.06	0.04	0.30	0.19	0.27	0.20	0.38	0.38	0.40	0.40	0.10	0.64	0.56	-	-	-
TopScore	0.40	0.30	0.12	0.53	0.36	0.59	0.61	0.73	0.72	0.82	0.75	0.11	0.81	0.81	0.51	-	0.91
TopScoreSingle	0.41	0.28	0.13	0.54	0.36	0.60	0.61	0.75	0.76	0.85	0.77	0.11	-	-	-	0.9	-

[a]The upper triangle contains R_{all}^2 between local scores and the lower triangle R_{all}^2 between global scores. Values of 0−0.2 have a white background, values of 0.2−0.4 have a light gray background, values of 0.4−0.6 have a medium gray background, and values of 0.6−1.0 have a dark gray background.

the spectrum, folding energy and cluster density are better error estimators.

The ProQ class of predictors (ProQ2, ProQ2D, and ProQ3D) show incremental improvement with each iteration of predictors. ProQ3D performs better than the linear method QMEAN6 on most measures and data sets except for R_{all}^2 on the DecoysR'us and Top100 data sets. SVMQA shows comparable performance to ProQ3D for most data sets but has a significantly higher performance on the iTASSER-II data set, which is in line with ProQ3D being described as having a lower performance for *ab initio* targets.[17] Interestingly, for residue-wise error estimation, ProQ3D outperforms all other primary MQAPs and even shows better performance than clustering-based ones on several data sets.

To analyze the similarity between the individual methods, a cross correlation test was performed to estimate to which degree the predictions of individual primary MQAPs correlate to each other and to predictions from TopScore and TopScoreSingle. The result of this analysis is shown in Table 2. It is not surprising that TopScore and TopScoreSingle have the highest correlation with the best performing primary MQAPs. It is however interesting to see that simple methods such as PROCHECK and MolProbity show higher correlation to TopScore and TopScoreSingle for local scores than knowledge-based potentials such as ProSA2003 and DOPE, though the latter show higher correlation for the global scores. This indicates that while stereochemical violations do not necessarily reflect global correctness they do indicate locations of local error.

DOI: 10.1021/acs.jctc.8b00690
J. Chem. Theory Comput. 2018, 14, 6117−6126

4. CONCLUSION

Model quality assessment is critical for automated protein structure prediction. In this study, we developed a scoring function with consistently high performance for a wide range of data sets and model ensembles resulting from *ab initio* folding, homology modeling, and CASP competitions. We used six large diverse data sets of models to examine the relationship between MQAP scores and the global and local lDDT error using four quality measures. We trained a DNN for each primary MQAP to predict the lDDT error from its raw score, using outlier filtering, weight decay, extensive cross-validation, and early stopping to prevent overfitting. We then combined the individual primary MQAP predictions using DNNs to predict global and local lDDT errors more accurately than any primary MQAP. Our methods TopScore and TopScoreSingle show a significant improvement in prediction of both local and global lDDT errors compared to the tested publicly available MQAPs when all quality measures and data sets are considered. This increase in performance indicates that despite the primary MQAPs predictions being weakly correlated, with the median correlation between them being 0.3 for global scores and 0.15 for local scores, there is still much to be gained from combining knowledge from different MQAPs into a single unified score. Notably, we observed that MQAP performance varies significantly depending on the data set they are evaluated on. This is likely due to varying degrees of difficulty of the data sets as well as the methods' limited abilities to generalize far beyond the data set they were trained on, which clearly stresses the importance of using large and diverse data sets for method training. TopScore was trained on more than 100,000 diverse models from almost 1000 targets spanning close to 30 million residues, which led to a robust performance across all data sets.

As a summary, a comparison of the best (when all different quality measures are considered) single-model, clustering, and hybrid MQAPs with TopScore and TopScoreSingle performance for local and global prediction on the combined data set and the CASP11/12 data set is shown in Table 3.

Table 3. Comparison of the Best Single-Model, Clustering, and Hybrid MQAPs to TopScore and TopScoreSingle

	combined data set (R_{all}^2/R_{wm}^2)		CASP11/12 data set (R_{all}^2/R_{wm}^2)	
MQAP	global	local	global	local
ProQ3D	0.77/0.70	0.66/0.50	0.68/0.29	0.59/0.34
Pcons	0.76/0.66	0.61/0.47	0.80/0.28	0.57/0.30
Pcomb	0.73/0.71	0.62/0.47	0.83/0.38	0.58/0.24
TopScoreSingle	0.90/0.75	0.71/0.55	0.68/0.33	0.61/0.42
TopScore	0.93/0.79	0.78/0.64	0.83/0.45	0.68/0.49

A comparison of all MQAPs can be found in Tables S4 and S6. TopScore as a stand-alone package and the data sets generated for it are available at http://cpclab.uni-duesseldorf.de/software.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00690.

Summary of literature values reported for primary MQAPs as well as numerical tables of the performance of all primary MQAPs for all quality measures and data sets used for both local and global scores, calculated statistics describing the difficulty of each data set, architecture and activation functions of all neural networks, and plots of the activation function and distribution of raw primary scores for the full range of the primary scores and model correctness (PDF)

■ AUTHOR INFORMATION

Corresponding Author

* Phone: (+49) 211 81 13662. Fax: (+49) 211 81 13847. E-mail: gohlke@uni-duesseldorf.de.

ORCID

Holger Gohlke: 0000-0001-8613-1447

Funding

D.M. was funded in part by a scholarship from the CLIB^{2021} Graduate Cluster "Industrial Biotechnology". This work was supported in part by the German Research Foundation (DFG) within the Collaborative Research Center SFB 1208 "Identity and Dynamics of Membrane Systems – From Molecules to Cellular Functions" (TP A03 to H.G.). We are grateful for computational support and infrastructure provided by the "Zentrum für Informations- und Medientechnologie" (ZIM) at the Heinrich Heine University Düsseldorf and the computing time provided by the John von Neumann Institute for Computing (NIC) to H.G. on the supercomputer JURECA at Jülich Supercomputing Centre (JSC) (user ID: HKF7).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to the developers of all primary MQAPs used in this work for making their methods available as standalone methods to the scientific community. In particular, we are thankful to the developers of QMEAN6 and ANOLEA for providing their software upon request.

■ REFERENCES

(1) Uziela, K.; Shu, N.; Wallner, B.; Elofsson, A. Proq3: Improved Model Quality Assessments Using Rosetta Energy Terms. *Sci. Rep.* **2016**, *6*, 33509.

(2) Kryshtafovych, A.; Barbato, A.; Fidelis, K.; Monastyrskyy, B.; Schwede, T.; Tramontano, A. Assessment of the Assessment: Evaluation of the Model Quality Estimates in Casp10. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 112−126.

(3) Shen, M. y.; Sali, A. Statistical Potential for Assessment and Prediction of Protein Structures. *Protein Sci.* **2006**, *15*, 2507−2524.

(4) Zhou, H.; Skolnick, J. Goap: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophys. J.* **2011**, *101*, 2043−2052.

(5) Melo, F.; Feytmans, E. Novel Knowledge-Based Mean Force Potential at Atomic Level. *J. Mol. Biol.* **1997**, *267*, 207−222.

(6) Sippl, M. J. Recognition of Errors in Three-Dimensional Structures of Proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 355−362.

(7) Tosatto, S. C. The Victor/Frst Function for Model Quality Estimation. *J. Comput. Biol.* **2005**, *12*, 1316−1327.

(8) Zhou, H.; Zhou, Y. Distance-Scaled, Finite Ideal-Gas Reference State Improves Structure-Derived Potentials of Mean Force for Structure Selection and Stability Prediction. *Protein Sci.* **2002**, *11*, 2714−2726.

(9) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. Procheck: A Program to Check the Stereochemical Quality of Protein Structures. *J. Appl. Crystallogr.* **1993**, *26*, 283−291.

DOI: 10.1021/acs.jctc.8b00690
J. Chem. Theory Comput. 2018, 14, 6117−6126

(10) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. Molprobity: All-Atom Structure Validation for Macromolecular Crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 12−21.
(11) Pettitt, C. S.; McGuffin, L. J.; Jones, D. T. Improving Sequence-Based Fold Recognition by Using 3d Model Quality Assessment. *Bioinformatics* **2005**, *21*, 3509−3515.
(12) Pontius, J.; Richelle, J.; Wodak, S. J. Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *J. Mol. Biol.* **1996**, *264*, 121−136.
(13) Randall, A.; Baldi, P. Selectpro: Effective Protein Model Selection Using a Structure-Based Energy Function Resistant to Blunders. *BMC Struct. Biol.* **2008**, *8*, 52.
(14) Wallner, B.; Elofsson, A. Can Correct Protein Models Be Identified? *Protein Sci.* **2003**, *12*, 1073−1086.
(15) Wang, Z.; Tegge, A. N.; Cheng, J. Evaluating the Absolute Quality of a Single Protein Model Using Structural Features and Support Vector Machines. *Proteins: Struct., Funct., Genet.* **2009**, *75*, 638−647.
(16) Zhao, F.; Xu, J. A Position-Specific Distance-Dependent Statistical Potential for Protein Structure and Functional Study. *Structure* **2012**, *20*, 1118−1126.
(17) Uziela, K.; Menéndez Hurtado, D.; Shu, N.; Wallner, B.; Elofsson, A. Proq3d: Improved Model Quality Assessments Using Deep Learning. *Bioinformatics* **2017**, *33*, 1578−1580.
(18) Manavalan, B.; Lee, J. Svmqa: Support−Vector-Machine-Based Protein Single-Model Quality Assessment. *Bioinformatics* **2017**, *33*, 2496−2503.
(19) McGuffin, L. J.; Roche, D. B. Rapid Model Quality Assessment for Protein Structure Predictions Using the Comparison of Multiple Models without Structural Alignments. *Bioinformatics* **2010**, *26*, 182−188.
(20) Zhang, Y.; Skolnick, J. Spicker: A Clustering Approach to Identify near-Native Protein Folds. *J. Comput. Chem.* **2004**, *25*, 865−871.
(21) Lundström, J.; Rychlewski, L.; Bujnicki, J.; Elofsson, A. Pcons: A Neural-Network−Based Consensus Predictor That Improves Fold Recognition. *Protein Sci.* **2001**, *10*, 2354−2362.
(22) Ginalski, K.; Elofsson, A.; Fischer, D.; Rychlewski, L. 3d-Jury: A Simple Approach to Improve Protein Structure Predictions. *Bioinformatics* **2003**, *19*, 1015−1018.
(23) Benkert, P.; Schwede, T.; Tosatto, S. C. Qmeanclust: Estimation of Protein Model Quality by Combining a Composite Scoring Function with Structural Density Information. *BMC Struct. Biol.* **2009**, *9*, 35.
(24) Pawlowski, M.; Kozlowski, L.; Kloczkowski, A. Mqapsingle: A Quasi Single-Model Approach for Estimation of the Quality of Individual Protein Structure Models. *Proteins: Struct., Funct., Genet.* **2016**, *84* (8), 1021−1028.
(25) Maghrabi, A. H.; McGuffin, L. J. Modfold6: An Accurate Web Server for the Global and Local Quality Estimation of 3d Protein Models. *Nucleic Acids Res.* **2017**, *45* (W1), W416−W421.
(26) Jing, X.; Dong, Q. Mqaprank: Improved Global Protein Model Quality Assessment by Learning-to-Rank. *BMC Bioinf.* **2017**, *18*, 275.
(27) Pawlowski, M.; Gajda, M. J.; Matlak, R.; Bujnicki, J. M. Metamqap: A Meta-Server for the Quality Assessment of Protein Models. *BMC Bioinf.* **2008**, *9*, 403.
(28) Benkert, P.; Tosatto, S. C.; Schomburg, D. Qmean: A Comprehensive Scoring Function for Model Quality Assessment. *Proteins: Struct., Funct., Genet.* **2008**, *71*, 261−277.
(29) Brown, P. J.; Fuller, W. A. Statistical Analysis of Measurement Error Models and Applications. *Proceedings of the Ams-Ims-Siam Joint Summer Research Conference Held June 10−16, 1989, with Support from the National Science Foundation and the US Army Research Office*. American Mathematical Soc.: 1990; Vol. *112*.
(30) Wallner, B.; Elofsson, A. Identification of Correct Regions in Protein Models Using Structural, Alignment, and Consensus Information. *Protein Sci.* **2006**, *15*, 900−913.
(31) Zhang, Y.; Skolnick, J. Tm-Align: A Protein Structure Alignment Algorithm Based on the Tm-Score. *Nucleic acids Res.* **2005**, *33*, 2302−2309.
(32) Zemla, A. Lga: A Method for Finding 3d Similarities in Protein Structures. *Nucleic acids Res.* **2003**, *31*, 3370−3374.
(33) Siew, N.; Elofsson, A.; Rychlewski, L.; Fischer, D. Maxsub: An Automated Measure for the Assessment of Protein Structure Prediction Quality. *Bioinformatics* **2000**, *16*, 776−785.
(34) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. Optimal Protein-Folding Codes from Spin-Glass Theory. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 4918−4922.
(35) Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. Lddt: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests. *Bioinformatics* **2013**, *29*, 2722−2728.
(36) Melo, F.; Feytmans, E. Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.* **1998**, *277*, 1141−1152.
(37) Melo, F.; Devos, D.; Depiereux, E.; Feytmans, E. ANOLEA: a www server to assess protein structures. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1997**, *5*, 187−190.
(38) Joo, K.; Lee, S. J.; Lee, J. Sann: solvent accessibility prediction of proteins by nearest neighbor method. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 1791−1797.
(39) Benkert, P.; Biasini, M.; Schwede, T. Toward the Estimation of the Absolute Quality of Individual Protein Structure Models. *Bioinformatics* **2011**, *27*, 343−350.
(40) Deng, H.; Jia, Y.; Zhang, Y. 3drobot: Automated Generation of Diverse and Well-Packed Protein Structure Decoys. *Bioinformatics* **2016**, *32*, 378−387.
(41) Zhang, Z.; Gu, Q.; Vasudevan, A. A. J.; Hain, A.; Kloke, B.-P.; Hasheminasab, S.; Mulnaes, D.; Sato, K.; Cichutek, K.; Häussinger, D. Determinants of Fiv and Hiv Vif Sensitivity of Feline Apobec3 Restriction Factors. *Retrovirology* **2016**, *13*, 46.
(42) Khosa, S.; Frieg, B.; Mulnaes, D.; Kleinschrodt, D.; Hoeppner, A.; Gohlke, H.; Smits, S. H. Structural Basis of Lantibiotic Recognition by the Nisin Resistance Protein from Streptococcus Agalactiae. *Sci. Rep.* **2016**, *6*, 18679.
(43) Widderich, N.; Pittelkow, M.; Höppner, A.; Mulnaes, D.; Buckel, W.; Gohlke, H.; Smits, S. H.; Bremer, E. Molecular Dynamics Simulations and Structure-Guided Mutagenesis Provide Insight into the Architecture of the Catalytic Core of the Ectoine Hydroxylase. *J. Mol. Biol.* **2014**, *426*, 586−600.
(44) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. Visualizing and Quantifying Molecular Goodness-of-Fit: Small-Probe Contact Dots with Explicit Hydrogen Atoms. *J. Mol. Biol.* **1999**, *285*, 1711−1733.
(45) Samudrala, R.; Levitt, M. Decoys 'R'us: A Database of Incorrect Conformations to Improve Protein Structure Prediction. *Protein Sci.* **2000**, *9*, 1399−1401.
(46) Zhang, J.; Zhang, Y. A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. *PLoS One* **2010**, *5* (10), e15386.
(47) Kryshtafovych, A.; Barbato, A.; Monastyrskyy, B.; Fidelis, K.; Schwede, T.; Tramontano, A. Methods of Model Accuracy Estimation Can Help Selecting the Best Models from Decoy Sets: Assessment of Model Accuracy Estimations in Casp11. *Proteins: Struct., Funct., Genet.* **2016**, *84*, 349−369.
(48) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using Modeller. *Curr. Protoc. Bioinformatics* **2014**, *47* (1), 5.6.1−5.6.32.
(49) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(50) Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. 2014, arXiv preprint, arXiv:1412.6980. https://arxiv.org/abs/1412.6980 (accessed Oct 2, 2018).

DOI: 10.1021/acs.jctc.8b00690
J. Chem. Theory Comput. 2018, 14, 6117−6126

(51) Kryshtafovych, A.; Fidelis, K.; Tramontano, A. Evaluation of Model Quality Predictions in Casp9. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 91−106.
(52) Larsson, P.; Skwark, M. J.; Wallner, B.; Elofsson, A. Assessment of Global and Local Model Quality in Casp8 Using Pcons and Proq. *Proteins: Struct., Funct., Genet.* **2009**, *77*, 167−172.
(53) Steiger, J. H. Tests for Comparing Elements of a Correlation Matrix. *Psychol. Bull.* **1980**, *87*, 245.

DOI: 10.1021/acs.jctc.8b00690
J. Chem. Theory Comput. 2018, 14, 6117−6126

Supplementary Information

TopScore: Using deep neural networks and large diverse datasets for accurate protein model quality assessment

Daniel Mulnaes[1] and Holger Gohlke[1,2]*

[1]Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine-University, Düsseldorf, Germany

[2]John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC) & Institute for Complex Systems - Structural Biochemistry (ICS 6), Forschungszentrum Jülich GmbH, Jülich, Germany

*Address: Universitätsstr. 1, 40225 Düsseldorf, Germany.
Phone: (+49) 211 81 13662; Fax: (+49) 211 81 13847
E-mail: gohlke@uni-duesseldorf.de

Supplementary Tables

Table S1. Literature values for primary MQAP scores and their target score. See also Table 1.

MQAP	Target Score / Protein systems	Global Score R^2 / R
ANOLEA	C_α RMSD / 23	0.69 / 0.83
ProSA2003	N/A / 167	N/A
DOPE	C_α RMSD / 20	0.76 / 0.87
GOAP	TM-Score / 390	0.39 / -0.63
ModFoldClust2	GDT-TS / 120	0.90 / 0.95
Pcons	LG-Score / 96	0.90 / 0.95
ProQ2	S-Score / 123	0.50 / 0.71
ProQ2D	lDDT-Score / 67	0.72 / 0.85
ProQ3D	lDDT-Score / 67	0.81 / 0.90
SVMQA	SVMQA-Score / 385	0.83 / 0.91
QMEAN6	GDT-TS / 122	0.59 / 0.77

Table S2. TopModel DNN architectures and activation functions (global / local). See also Figure 1.

MQAP	First layer size	Second layer size	Third layer size	Activation function
PROCHECK	160 / 160	80 / 80	0 / 40	relu / relu
MolProbity	160 / 160	80 / 80	40 / 40	relu / relu
ANOLEA	40 / 160	0 / 80	0 / 40	relu / relu
ProSa2003	20 / 160	10 / 80	40 / 0	relu / relu
DOPE	160 / 160	80 / 80	0 / 40	relu / relu
GOAP	10 / 0	5 / 0	02 / 0	tanh / -
QMEAN6	160 / 160	80 / 80	40 / 40	relu / relu
ProQ2	40 / 160	20 / 80	10 / 40	relu / relu
ProQ2D	40 / 160	0 / 80	0 / 40	relu / relu
ProQ3D	80 / 160	40 / 80	20 / 40	relu / relu
SVMQA	160 / 0	80 / 0	40 / 0	relu / -
SELECTpro	20 / 0	10 / 0	0 / 0	relu / -
ModFoldClust2	160 / 160	80 / 80	40 / 40	relu / relu
PCONS	80 / 160	40 / 80	20 / 40	relu / relu
SPICKER	10 / 0	5 / 0	2 / 0	tanh / -
TopScore	160 / 160	80 / 80	40 / 40	relu / relu
TopScoreSingle	160 / 160	80 / 80	40 / 40	relu / relu

Table S2. Single model primary MQAPs are shown in the first section. Clustering primary MQAPs are shown in the second section. TopScore and TopScoreSingle (this study) are shown in the bottom section. The number of neurons and the activation function for each of the potential three layers of the deep neural networks are shown for each primary MQAP as well as for TopScore and TopScoreSingle. For activation functions, "relu" refers to the rectified linear unit function and "thanh" refers to the hyperbolic tangent function. Values before the / correspond to the global score networks and values after the / correspond to the local score networks.

Table S3. Quality and spread of different datasets used for TopScore. See also Figure 2.

Dataset	MOM[a]	MOMAD[b]	MD[c]	MAD[d]
Top100	0.29	0.02	0.29	0.09
3DRobot	0.54	0.09	0.53	0.11
DecoysR'us	0.28	0.02	0.46	0.11
iTASSER-II	0.52	0.02	0.51	0.07
CASP10	0.47	0.03	0.45	0.10
Combined	0.43	0.02	0.46	0.13
CASP11/12	0.60	0.04	0.61	0.14

Table S3. For each model ensemble in a given dataset the lDDT score to the native structure is calculated. For each model ensemble, the median and median absolute deviation of the lDDT scores are then calculated. For each dataset, the median of medians (MOM[a]) and median of median absolute deviations (MOMAD[b]) are calculated across all model ensembles in that dataset. Additionally, the median (MD[c]) and median absolute deviation (MAD[d]) are calculated when all models from all ensembles are pooled together.

Table S4. $R_{all}^2\uparrow$ and $R_{wm}^2\uparrow$ of global primary MQAPs and TopScore. See also Figure 2.

MQAP	iTASSER-II	CASP10	DecoysR'us	Top100	3DRobot	Combined	CASP11/12
PROCHECK	0.00 / 0.00	0.06 / 0.04	0.46 / 0.05	0.32 / 0.19	0.12 / 0.27	0.38 / 0.15	0.04 / 0.05
MolProbity	0.05 / 0.02	0.00 / 0.04	0.46 / 0.02	0.15 / 0.07	0.24 / 0.29	0.30 / 0.13	0.01 / 0.03
ANOLEA	0.14 / 0.06	0.19 / 0.10	0.73 / 0.14	0.44 / 0.28	0.15 / 0.51	0.11 / 0.32	0.14 / 0.11
ProSa2003	0.13 / 0.12	0.21 / 0.13	0.28 / 0.28	0.33 / 0.43	0.54 / 0.75	0.50 / 0.53	0.19 / 0.13
DOPE	0.27 / 0.22	0.19 / 0.11	0.59 / 0.35	0.53 / 0.55	0.47 / 0.75	0.34 / 0.59	0.22 / 0.09
GOAP	0.16 / 0.22	0.23 / 0.18	0.76 / 0.33	0.49 / 0.62	0.69 / 0.85	0.57 / 0.68	0.26 / 0.16
QMEAN6	0.27 / 0.12	0.19 / 0.14	**0.84** / 0.34	**0.66** / 0.53	0.63 / 0.80	0.57 / 0.60	0.33 / 0.14
ProQ2	0.36 / 0.19	0.42 / 0.18	0.77 / 0.42	0.55 / 0.49	0.70 / 0.85	0.69 / 0.65	0.58 / 0.18
ProQ2D	0.27 / 0.20	0.63 / 0.23	0.70 / 0.44	0.51 / 0.49	0.71 / 0.86	0.68 / 0.66	0.63 / 0.25
ProQ3D	0.21 / 0.18	**0.67** / **0.27**	0.81 / 0.44	0.64 / **0.58**	**0.77** / **0.89**	**0.77** / 0.70	**0.68** / **0.29**
SVMQA	**0.46** / **0.27**	0.48 / 0.20	0.81 / **0.52**	0.53 / **0.58**	0.76 / **0.89**	0.72 / **0.72**	0.61 / 0.25
SELECTpro	0.07 / 0.18	0.08 / 0.18	0.03 / 0.33	0.18 / 0.41	0.17 / 0.70	0.11 / 0.51	0.10 / 0.14
ModFoldClust2	0.49 / 0.27	0.65 / 0.04	**0.87** / 0.44	0.61 / 0.49	**0.82** / 0.86	0.75 / 0.61	0.74 / 0.07
Pcons	**0.54** / **0.29**	**0.73** / **0.19**	0.86 / **0.49**	**0.66** / **0.56**	0.78 / **0.88**	**0.76** / **0.66**	**0.80** / **0.28**
SPICKER*	0.44 / 0.22	0.29 / 0.06	0.58 / 0.13	0.41 / 0.17	0.39 / 0.44	0.46 / 0.27	0.45 / 0.10
Pcomb	**0.66** / **0.36**	**0.74** / **0.37**	**0.91** / **0.77**	**0.69** / **0.69**	0.81 / **0.90**	0.73 / 0.71	**0.83** / **0.38**
TopScoreSingle	**0.64** / 0.23	**0.78** / **0.44**	**0.91** / **0.78**	**0.87** / **0.58**	**0.87** / **0.92**	**0.90** / **0.75**	**0.68** / **0.33**
TopScore	**0.76** / 0.35	**0.79** / **0.52**	**0.95** / **0.78**	**0.87** / **0.69**	**0.91** / **0.95**	**0.93** / **0.79**	**0.83** / **0.45**

Table S4. Single model primary MQAPs are shown in the first section. Clustering primary MQAPs are shown in the second section. Pcomb is shown in the third section, and was not included as a primary predictor. TopScore and TopScoreSingle (this study) are shown in the bottom section. In the first and second MQAP sections, the best score is highlighted in **bold**. Pcomb is highlighted in bold if it outperforms all other primary MQAPs. In the fourth section, TopScore is highlighted if it outperforms all other MQAPs, and TopScoreSingle is highlighted if it outperforms all other single primary MQAPs. The 95% confidence intervals were calculated using the Fischer *r*-to-*z* transformation. The widest confidence interval for any R_{all}^2 or R_{wm}^2 was 0.007 and 0.07, respectively. Statistical significance was determined by the two-sided Steiger test[47]. Accordingly, on the combined dataset the R_{all}^2 and R_{wm}^2 of TopScore and TopScoreSingle are significantly different from any primary MQAP ($p < 0.05$). The arrows "↑" and "↓" indicate if a score gets better with increasing or decreasing value respectively.

Table S5. AUC↑ and Loss↓ of global primary MQAPs and TopScore. See also Figure 2.

MQAP	iTASSER-II	CASP10	DecoysR'us	Top100	3DRobot	Combined	CASP11/12
PROCHECK	0.46 / 9.05	0.61 / 9.69	0.90 / 11.46	0.82 / 4.12	0.64 / 10.57	0.75 / 7.60	0.59 / 12.95
MolProbity	0.62 / 6.27	0.50 / 10.99	0.82 / 13.49	0.72 / 3.48	0.72 / 14.74	0.73 / 8.77	0.55 / 12.90
ANOLEA	0.72 / 6.29	0.69 / 9.49	0.86 / 4.09	0.88 / 3.35	0.71 / 11.57	0.63 / 6.20	0.71 / 13.04
ProSa2003	0.66 / 7.78	0.75 / 10.72	0.69 / 4.72	0.86 / 3.94	0.88 / 11.04	0.84 / 6.57	0.75 / 13.38
DOPE	0.77 / 5.09	0.74 / 10.23	0.84 / 3.45	0.96 / 2.81	0.85 / 4.42	0.78 / 3.59	0.77 / 11.92
GOAP	0.70 / 5.68	0.76 / 6.60	0.89 / **2.28**	0.96 / 2.47	0.92 / 4.90	0.87 / 3.77	0.85 / 13.85
QMEAN6	0.74 / 6.20	0.75 / 9.79	**0.95** / 2.53	0.96 / **2.15**	0.89 / 3.22	0.84 / 2.87	0.80 / 10.72
ProQ2	0.82 / 5.96	0.85 / 7.54	0.91 / 4.35	0.96 / 2.81	0.94 / 8.14	0.92 / 4.95	0.91 / 9.51
ProQ2D	0.77 / 5.47	0.90 / 6.65	0.88 / 4.80	**0.97** / 3.01	0.94 / 7.43	0.92 / 4.87	0.92 / 8.48
ProQ3D	0.78 / 5.52	**0.91** / **5.62**	0.91 / 4.46	**0.97** / 2.56	0.95 / **2.71**	**0.94** / **3.18**	**0.94** / **7.86**
SVMQA	**0.84** / **4.28**	0.87 / 6.87	0.93 / 2.83	**0.97** / 2.33	**0.96** / 5.04	**0.94** / 3.41	0.92 / 8.52
SELECTpro	0.60 / 6.23	0.64 / 11.47	0.33 / 3.92	0.85 / 3.91	0.70 / 11.15	0.66 / 6.34	0.69 / 13.43
ModFoldClust2	0.84 / 7.31	**0.92** / 9.97	**0.97** / 6.61	**0.91** / 4.03	**0.98** / **12.82**	**0.94** / 7.95	0.97 / 12.11
Pcons	**0.85** / **6.09**	**0.92** / **7.72**	0.96 / **5.37**	**0.91** / **3.41**	0.97 / 13.07	**0.94** / **7.09**	**0.97** / **10.60**
SPICKER	0.86 / -	0.83 / -	0.88 / -	0.85 / -	0.79 / -	0.83 / -	0.86 / -
Pcomb	**0.92** / 5.11	**0.93** / 6.56	**0.98** / 3.85	0.96 / 2.38	0.98 / 6.09	**0.95** / 4.54	**0.97** / 8.16
TopScoreSingle	**0.91** / 5.54	**0.94** / **4.01**	**0.97** / 2.76	**0.99** / 2.17	**0.98** / **1.76**	**0.97** / **2.75**	**0.94** / **7.43**
TopScore	**0.94** / 4.76	**0.95** / **3.77**	**0.99** / 3.02	**0.99** / **1.97**	**0.99** / **1.69**	**0.99** / **2.59**	**0.97** / **7.15**

Table S5. Sections, highlighting, and symbols are made in an identical manner as for Table S4. For SPICKER no Loss is calculated (Indicated by "-"), since models are ranked by cluster size and multiple models therefore have the same score.

Table S6. R_{all}^2↑, R_{wm}^2↑ and AUC↑ of local primary MQAPs and TopScore. See also Figure 3.

MQAP	Top100	3DRobot	DecoysR'us	iTASSER-II	CASP10	Combined	CASP11/12
PROCHECK	0.15/0.12/0.75	0.07/0.07/0.62	0.24/0.01/0.79	0.00/0.00/0.52	0.05/0.03/0.61	0.22/0.06/0.72	0.04/0.04/0.61
MolProbity	0.18/0.15/0.76	0.01/0.01/0.58	0.19/0.05/0.75	0.00/0.00/0.56	0.04/0.03/0.60	0.21/0.05/0.71	0.04/0.04/0.62
ANOLEA	0.18/0.12/0.73	0.07/0.08/0.63	0.34/0.11/0.81	0.07/0.04/0.64	0.13/0.08/0.67	0.10/0.09/0.64	0.11/0.07/0.66
ProSa2003	0.12/0.08/0.74	0.12/0.12/0.68	0.07/0.08/0.67	0.05/0.05/0.59	0.11/0.09/0.67	0.13/0.09/0.69	0.10/0.07/0.67
DOPE	0.12/0.12/0.71	0.14/0.14/0.69	0.02/0.10/0.53	0.06/0.09/0.61	0.15/0.14/0.69	0.09/0.13/0.64	0.11/0.14/0.66
QMEAN6	0.47/0.39/0.90	0.36/0.36/0.80	0.53/0.28/0.89	0.14/0.07/0.71	0.35/0.24/0.80	0.40/0.31/0.82	0.33/0.20/0.80
PROQ2	0.46/0.34/0.89	0.42/0.45/0.83	0.47/0.16/0.88	0.21/0.10/0.73	0.47/0.29/0.86	0.51/0.34/0.86	0.47/0.22/0.86
PROQ2D	0.44/0.33/0.89	0.47/0.51/0.86	0.47/0.19/0.88	0.23/0.13/0.73	0.64/0.44/0.91	0.55/0.39/0.89	0.50/0.26/0.87
PROQ3D	**0.61/0.50/0.93**	**0.57/0.60/0.88**	**0.64/0.33/0.92**	**0.26/0.16/0.75**	**0.66/0.48/0.92**	**0.66/0.50/0.91**	**0.59/0.34/0.90**
ModFoldClust2	0.49/0.38/0.89	0.36/0.36/0.82	0.61/0.32/0.92	0.32/0.16/0.79	0.56/0.33/0.88	0.53/0.34/0.87	0.54/0.24/0.89
PCONS	**0.54/0.46/0.92**	**0.53/0.54/0.88**	**0.72/0.48/0.96**	**0.40/0.25/0.82**	**0.59/0.38/0.90**	**0.61/0.47/0.90**	**0.57/0.30/0.90**
PCOMB	0.57/**0.56**/0.92	0.55/0.56/0.89	**0.75/0.52/0.96**	**0.43/0.28/0.82**	0.60/0.38/0.89	0.62/0.47/0.90	0.58/0.24/0.89
TopScore	**0.74/0.63/0.96**	**0.71/0.72/0.93**	**0.79/0.58/0.96**	**0.48/0.32/0.84**	**0.73/0.57/0.94**	**0.78/0.64/0.95**	**0.68/0.49/0.94**
TopScoreSingle	**0.67/0.55/0.94**	**0.61/0.64/0.90**	**0.70/0.41/0.93**	**0.31/0.21/0.77**	**0.68/0.52/0.92**	**0.71/0.55/0.92**	**0.61/0.42/0.91**

Table S6. Sections, highlighting, and symbols are made in an identical manner as for Table S4. The 99% confidence intervals and statistical significances are calculated in the same way as for Supplementary Table 5. The widest confidence interval for any R_{all}^2or R_{wm}^2 was 0.002 and 0.08, respectively. The R_{all}^2 and R_{wm}^2 of TopScore and TopScoreSingle are significantly different from primary MQAPs (Steiger test $p < 0.05$).

Supplementary Figures

Figure S1. Training of local (a) and global (b) primary MQAP DNNs.

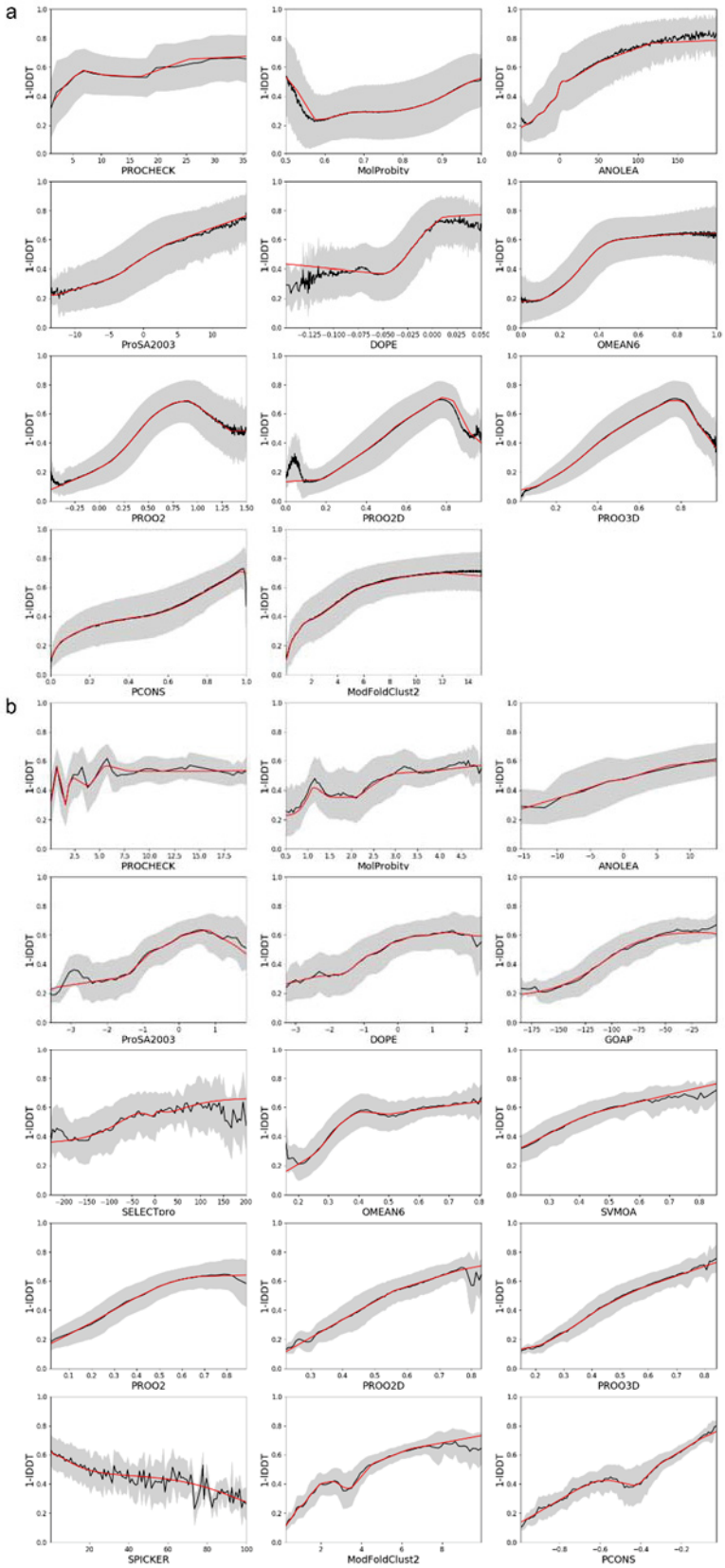


Figure S1. Training of local (a) and global (b) primary MQAP DNNs. The raw MQAP scores are binned into 100 bins. Black lines indicate bin means and gray areas bin standard deviations in terms of lDDT error. The red lines indicate the output from the DNNs on the raw data given the bin MQAP bin means as input.

Supplementary References

1 Deng, H.; Jia, Y.; Zhang, Y. 3DRobot: automated generation of diverse and well-packed protein structure decoys. *Bioinformatics* **2015**, *32*, 378-387.

2 Samudrala, R.; Levitt, M. Decoys 'R'Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci.* **2000**, *9*, 1399-1401.

3 Zhang, J; Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS one* **2010**, *5*, e15386.

4 Kryshtafovych, A.; Barbato, A.; Monastyrskyy, B.; Fidelis, K.; Schwede, T.; Tramontano, A. Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP 11. *Proteins: Struct., Funct., Bioinf.*, **2016**, *84*, 349-369.

5 Boratyn, G.M.; Schäffer, A.A.; Agarwala, R.; Altschul, S.F.; Lipman, D.J.; Madden, T.L. Domain enhanced lookup time accelerated BLAST. *Biol. direct*, **2012**, *7(1)*, 12.

6 O'Sullivan, O.; Suhre, K.; Abergel, C.; Higgins, D. G.; Notredame, C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. M. Biol.* **2004**, *340*, 385-395.

7 Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283-291.

8 Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195.

9 Katoh, K.; Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772-780.

10 Chen, V.B.; Arendall, W.B.; Headd, J.J.; Keedy, D.A.; Immormino, R.M.; Kapral, G.J.; Murray, L.W.; Richardson, J.S.; Richardson, D.C., MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Cryst. D,* **2010**, *66(1)*, 12-21.

11 Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods,* **2012**, *9*, 173-175.

12 Collingridge, P. W; Kelly, S. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC Bioinform.* **2012**, *13*, 117.

13 Melo, F.; Feytmans, E. Novel knowledge-based mean force potential at atomic level. *J. M. Biol.* **1997**, *267*, 207-222.

14 Söding, J., Biegert, A.; Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids Res.* **2005**, *33*, W244-W248.

15 Madhusudhan, M.; Webb, B. M.; Marti-Renom, M. A.; Eswar, N.; Sali, A. Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des. Sel.* **2009**, *22*, 569-574.

16 Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins: Struct., Funct., and Gen.* **1993**, *17*, 355-362.

17 Rychlewski, L.; Li, W., Jaroszewski, L.; Godzik, A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **2000**, *9*, 232-241.

18 Pei, J.; Kim, B.-H.; Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids Res.* **2008**, *36*, 2295-2300.

19 Shen, M.Y.; Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006**, *15*, 2507-2524.

20 Yang, Y.; Faraggi, E.; Zhao, II.; Zhou, Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native

properties of templates. *Bioinformatics* **2011**, *27*, 2076-2082.
21 Daniels, N. M.; Nadimpalli, S.; Cowen, L. J. Formatt: Correcting protein multiple structural alignments by incorporating sequence alignment. *BMC Bioinform.* **2012**, *13*, 259.
22 Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **2002**, *11*, 2714-2726.
23 Peng, J.; Xu, J. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Struct., Funct. Bioinf.* **2011**, *79*, 161-171.
24 Konagurthu, A. S.; Whisstock, J. C.; Stuckey, P. J.; Lesk, A. M. MUSTANG: a multiple structural alignment algorithm. *Proteins: Struct., Funct., Bioinf.* **2006**, *64*, 559-574.
25 McGuffin, L. J.; Roche, D. B. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*, **2010**, *26*, 182-188.
26 Wu, S.; Zhang, Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids Res.* **2007**, *35*, 3375-3382.
27 Wang, S., Peng, J.; Xu, J. Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. *Bioinformatics*, **2011**, *27*, 2537-2545.
28 Wallner, B.; Elofsson, A. Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci.* **2006**, *15*, 900-913.
29 Lobley, A.; Sadowski, M. I.; Jones, D. T. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* **2009**, *25*, 1761-1767.
30 Zhang, Y.; Skolnick, J. SPICKER: A clustering approach to identify near-native protein folds. *J. Comput. Chem.* **2004,** *25*, 865-871.
31 Benkert, P.; Schwede, T.; Tosatto, S. C. QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. *BMC Struct. Biol.* **2009**, *9*, 35.
32 Pearson, W. R. Finding protein and nucleotide similarities with FASTA. *Curr. Protoc. Bioinformatics*, **2016**, 53(1), 3-9.
33 Ray, A.; Lindahl, E.; Wallner, B. Improved model quality assessment using ProQ2. *BMC Bioinf.* **2012**, *13*, 224.
34 Karplus, K. SAM-T08, HMM-based protein structure prediction. *Nucleic acids Res.*, **2009**, *37(suppl_2)*,W492-W497.
35 Uziela, K.; Menéndez Hurtado, D.; Shu, N.; Wallner, B.; Elofsson, A. ProQ3D: improved model quality assessments using deep learning. *Bioinformatics* **2017**, *33*, 1578-1580.
36 Manavalan, B.; Lee, J. SVMQA: support–vector-machine-based protein single-model quality assessment. *Bioinformatics* **2017**, *33*, 2496-2503.
37 Randall, A.; Baldi, P. SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *BMC Struct Biol.* **2008**, *8*, 52.
38 Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids Res.* **1994**, *22*, 4673-4680.
39 Lee, C., Grasso, C.; Sharlow, M. F. Multiple sequence alignment using partial order graphs. *Bioinformatics* **2002**, *18*, 452-464.
40 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids Res.* **2004**, *32*, 1792-1797.
41 Sierk, M. L.; Smoot, M. E.; Bass, E. J.; Pearson, W. R. Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. *BMC Bioinf.* **2010**,

11, 146.

42 Pei, J.; Sadreyev, R.; Grishin, N. V. PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics* **2003**, *19*, 427-428.

43 Do, C. B.; Mahabhashyam, M. S.; Brudno, M.; Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **2005**, *15*, 330-340.

44 Al Ait, L.; Yamak, Z.; Morgenstern, B. DIALIGN at GOBICS - multiple sequence alignment using various sources of external information. *Nucleic acids Res.* **2013**, *41*, W3-W7.

45 Taylor, W. R. Protein structure comparison using iterated double dynamic programming. *Protein Sci.* **1999**, *8*, 654-665.

46 Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids Res.* **2005**, *33*, 2302-2309.

47 Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **1980**, *87*, 245.

18. PUBLICATION II

TopModel: Template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks

Daniel Mulnaes[1], Nicola Porta[1], Rebecca Clemens[2], Irina Apanasenko[3,4], Jens Reiners[2,5], Lothar Gremer[3,4], Philipp Neudecker[3,4], Sander Smits[2,5], Holger Gohlke[1,2,6]*

[1]Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[2]Institute für Biochemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany.

[3]Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[4]Institute for Complex Systems - Structural Biochemistry (ICS-6) & JuStruct, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[5]Center for Structural Studies Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[6]John von Neumann Institute for Computing (NIC) & Jülich Supercomputing Center (JSC), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

TopModel: Template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks

Daniel Mulnaes[1], Nicola Porta[1], Rebecca Clemens[2], Irina Apanasenko[3,4], Jens Reiners[2,5], Lothar Gremer[3,4], Philipp Neudecker[3,4], Sander Smits[2,5], Holger Gohlke[1,4,6]*

[1]Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[2]Institute für Biochemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany.

[3]Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[4]Institute for Complex Systems - Structural Biochemistry (ICS-6) & Jülich Center for Structural Biology (JuStruct), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[5]Center for Structural Studies Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[6]John von Neumann Institute for Computing (NIC) & Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

Author ORCID
Daniel Mulnaes: 0000-0003-2162-5918
Nicola Porta: 0000-0002-6005-4372
Lothar Gremer: 0000-0001-7065-5027
Philipp Neudecker: 0000-0002-0557-966X
Sander Smits: 0000-0003-0780-9251
Holger Gohlke: 0000-0001-8613-1447

*Address: Universitätsstr. 1, 40225 Düsseldorf, Germany.

Phone: (+49) 211 81 13662; Fax: (+49) 211 81 13847

E-mail: gohlke@uni-duesseldorf.de

Abstract

Knowledge of protein structures is essential to understand the proteins' functions, evolution, dynamics, stabilities, interactions, and for data-driven protein- or drug-design. Yet, experimental structure determination rates are far exceeded by that of next-generation sequencing. Computational structure prediction seeks to alleviate this problem, and the Critical Assessment of protein Structure Prediction (CASP) has shown the value of consensus- and meta-methods that utilize complementary algorithms. However, traditionally, such methods employ majority voting during template selection and model averaging during refinement, which can drive the model away from the native fold if it is underrepresented in the ensemble. Here, we present TopModel, a fully automated meta-method for protein structure prediction. In contrast to traditional consensus- and meta-methods, TopModel uses top-down consensus and deep neural networks to select templates and identify and correct wrongly modeled regions. TopModel combines a broad range of state-of-the-art methods for threading, alignment and model quality estimation and provides a versatile work-flow and toolbox for template-based structure prediction. TopModel shows a superior template selection, alignment accuracy, and model quality for template-based structure prediction on the CASP10-12 datasets. TopModel was validated by prospective predictions of the nisin resistance protein NSR protein from *S. agalactiae* and LipoP from *C. difficile*, showing far better agreement with experimental data than any of its constituent primary predictors. These results, in general, demonstrate the utility of TopModel for protein structure prediction and, in particular, show how combining computational structure prediction with sparse or low-resolution experimental data can improve the final model.

Introduction

Knowing the 3D structure of a protein is important to understand its stability [1], dynamics, function [2], structural evolution [3], and interactions with ligands [4, 5] or other proteins [6]. Consequently, protein structure prediction is an essential part of knowledge-based protein engineering [7], drug-design and -discovery [8], and function assignment [9, 10]. At present, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the dominating experimental methods for structure determination, but both are too time consuming to keep up with current high-throughput genome sequencing information. Computational structure prediction has sought to alleviate this problem, and in the last decades, many approaches have been developed, raising the question of which method to use for a given sequence of interest. The biological information that can be derived from a structure prediction depends on its accuracy: High-confidence models based on homologous templates are generally suitable for computational ligand docking and virtual compound screening, while models with medium confidence can be useful for identification of functionally important sites and disease-associated mutations [11].

The field of computational structure prediction has driven many advances in structural bioinformatics, the most important being the development of threading algorithms that seek to identify a template structure most similar to the native structure of a target sequence of interest. These developments include fast and sensitive alignment methods such as: Iterated search methods [12], position-specific-scoring matrices (PSSM's) [13, 14], sequence-profile alignment [15], profile-profile alignment [16, 17], and hidden Markov models (HMM's) [18-22]. The accuracy of threading algorithms has been further improved by adding structural features such as predicted secondary structure, residue contacts, solvent accessibility [23], residue-depth [24], and backbone dihedral angles [25] to the alignment and scoring functions. Additionally, probabilistic modeling [26, 27], depth-dependent alignment of structure fragments [28], multiple template and structure alignment [29], normalized Z-scores [16, 23], and sequence-based solvation potentials [17] have been employed to increase performance. Advances in multiple structure/sequence alignment methods, model building, clustering, and quality estimation have also had a large impact in the field [30, 31]. Meta-approaches have proven to be one of the major advances [32], as evident by the consistent high ranking of the Zhang meta-server [33] in the blind critical assessment of protein structure prediction (CASP) experiments. The meta-server methodology is to produce structure predictions using information from multiple different algorithms [33, 34] and either re-rank or combine their output to produce better predictions than any of their component predictors. Considering the diversity of

optimization procedures, training sets and quality measures, it is not surprising that meta-methods provide more robust results with a higher over-all quality.

Here, we present a meta-approach to template-based structure prediction, which uses a top-down consensus approach rather than traditional majority-voting consensus termed TopModel. The development of TopModel was inspired by the success of meta-approaches in CASP experiments [35]. The CASP experiments, however, are undertaken on a working group rather than an algorithmic level, and competing groups use different algorithms not only for threading, but also for alignment, model construction, model refinement, model evaluation, and model selection. It is therefore difficult to assign the differences in model quality from different groups to improvement of a specific step of the structure prediction workflow.

The aim of TopModel is therefore to individually optimize four steps of the structure prediction pipeline: Template selection, target-template(s) alignment, model selection, and model combination and refinement. By focusing on each step individually, we aim to improve the final quality of models produced by TopModel. TopModel aims to provide a versatile and accurate toolbox for template-based protein structure prediction, expand applicability of existing algorithms for threading, alignment, model quality estimation and refinement via an automated integration of all methods, and yield high quality structure predictions even for low sequence identities that are in agreement with experimental data.

Ab initio folding methods have in recent years seen a large increase in model accuracy due to a revolution in using image recognition deep neural networks for predicting residue-residue contacts and distances [36]. The aim of TopModel, however, is to establish an automated workflow for template-based modeling in order to explore how deep learning can improve template selection and how well the use of structural information from multiple templates and alternate alignments can improve model quality compared to single-template based modeling. In parallel to the development of TopModel, we are working on an *ab initio* folding pipeline that builds on the recent advances in prediction of residue-residue distances, which we aim to combine with the template-based folding in TopModel for improved performance.

Methods and implementation

TopModel. TopModel is a protein structure prediction work-flow with five modules that are executed consecutively or can be used individually. A simplified depiction of the interaction between the different TopModel modules can be seen in Figure 1; a detailed description of each module is given below.

1. **TopThreader.** TopThreader identifies template structures from a target sequence based on predictions from twelve different primary threading programs using a top-down consensus approach instead of traditional majority voting.
2. **TopAligner.** TopAligner makes an ensemble of alignments between the target sequence and the provided templates based on template-template alignments from eight different primary alignment programs and template-target alignments from TopThreader.
3. **TopBuilder.** TopBuilder makes models of the target sequence based on alignments from TopAligner or TopThreader and templates from TopThreader, using Modeller9 [37] and Rosetta [38].
4. **TopScore.** TopScore and TopScoreSingle [39] predicts the global and local error of models based on predictions from fifteen primary model quality assessment programs. TopScoreSingle is similar to TopScore but does not include clustering information and is therefore suitable when the best model is not part of a cluster.
5. **TopRefiner.** TopRefiner selects, combines, and refines models made by TopBuilder based on predicted global and local errors from TopScore and TopScoreSingle.

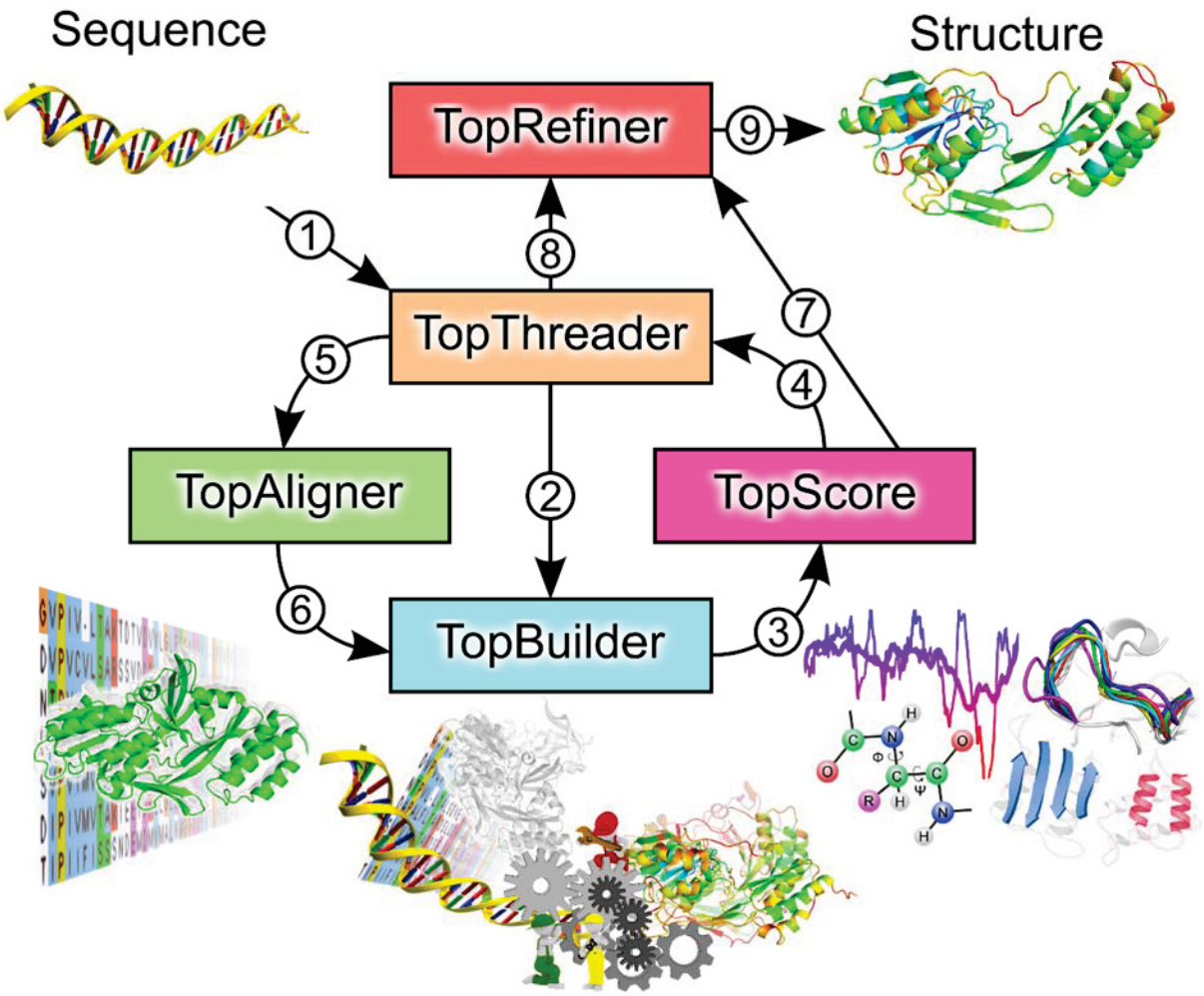


Figure 1. Simplified interaction between TopModel modules. The target sequence is given as input to TopThreader (1), which searches for templates using different primary threaders. TopThreader uses TopBuilder (2) to build models from the primary threader alignments, template structures and target sequence, which are scored with TopScore (3), and used by TopThreader (4) together with primary threader scores to rank and cluster templates and remove false positives. TopThreader then uses TopAligner (5) to align templates and construct consensus alignments which are built with TopBuilder (6), scored with TopScore (3), and used together with primary threader scores in TopThreader (4) to rank templates by predicted similarity to the native structure. After template selection, TopAligner (5) is used to generate a large ensemble of pairwise and multi-template alignments from which models are built with TopBuilder (6) and scored with TopScore (3). Models are selected from the multi-template ensemble (7) and the single-template models (8) by TopRefiner, which combines and refines the models to produce a final model (9).

TopThreader. The threading process is the first and most critical step of template-based protein structure prediction [33]. It has three main goals: (1) Identification of correct template structures for a target sequence, also known as fold recognition or threading, (2) target-template alignment, and (3) ranking of templates according to their similarity to the native structure. The TopModel threading module TopThreader uses a combination of deep neural networks (DNNs),

model quality prediction by TopScore and TopScoreSingle [39], and sequence/structure alignments to predict the TM-Score [40] between each template and the native structure, remove false positive templates, calculate consensus alignments, and rank templates by their predicted TM-Score. The TM-Score is a robust measure of structural similarity between two proteins which is independent of the protein sizes.

Prediction of template quality is similar to protein model quality assessment but not identical. First, template similarity to the native structure differs from model quality because of different possible target-template alignments, which is one of the main determinants of template-based model quality. In other words, a template may be similar to the native structure, but if the target-template alignment is wrong, the resulting model can have a low quality. Consequently, while a template has just one TM-Score to the native structure, models built from different alignments between the target and the template may have different model qualities, which can obscure the detection of the best template. Second, template similarity to the native structure is based on comparison between structures with different sequences and sizes, while model quality is based on comparison between structures of the same size and sequence as the native structure. Thus, while a small partially matching template may have the right fold for a given part of the target sequence, a model based on such a template alone could have a poor quality due to low coverage. These differences are important especially for hard cases, in which threaders may prefer a wrong template with a large coverage over a short template with a correct fold but poor coverage. As such, the prediction of template similarity to the native structure is a challenging task.

TopThreader has eight steps outlined here. In the Supporting Information a detailed description of the TopThreader workflow (Text T1, Figure S1), the DNN training (Text T1, Figure S2), and the primary threading programs (Text T2) can be found.

1. **Primary Threaders.** TopThreader uses twenty primary threading algorithms from twelve primary threaders and selects the top five templates from each threader (Table S1). All threaders are run with default settings following the provided instructions by their respective authors.
2. **Pre-filtering.** Pre-filtering allows the user to discard templates according to cutoffs with respect to, e.g., sequence identity, coverage, experimental method, or submission date. By default, templates with less than 30% coverage and artificially designed proteins are removed.
3. **Alignment Fitting.** TopThreader fits all pairwise threading alignments to the template structures and target sequence to ensure that residues match exactly.

4. **Score templates using DNNs.** TopThreader initially predicts a target-template TM-Score (*Initial Score*) using DNNs. The DNNs' input features include primary threader scores and values calculated from threading alignments such as sequence identity and target coverage.
5. **Redundancy clustering.** TopThreader clusters templates at 90% sequence identity and pairwise TM-Score of 0.9, selecting the cluster centroid with the highest *Initial Score*. Alignments from other threaders/templates in the cluster are transferred to the centroid by superimposing their target-template alignments to the (nearly identical) centroid while minimizing changes to the alignment.
6. **False positive removal.** Removal of false positives is critical to ensure correct fold recognition. TopThreader first clusters templates structurally to remove bias towards folds with many templates. For each cluster, DNNs are used to predict the centroid TM-Score (*Filtering Score*). Templates are then structurally aligned to the best centroid based on *Filtering Score* and TopScoreSingle of a model built from the template. Using a top-down consensus approach, models are discarded if they are dissimilar (TM-Score < 0.4) to the best centroid.
7. **Consensus.** TopThreader uses local and global quality scores of models from different pairwise threading alignments combined with a structural alignment of all templates to calculate consensus alignments for each template.
8. **Ranking.** The final template ranking is based on predicted TM-Score from a DNN with input features from all previous steps. This score, the *TopThreader Score*, has a Pearson's R^2 of 0.77 with the true TM-Score of the template.

A key difference between TopThreader and consensus methods such as the MULTICOM[41] or Zhang servers [42] is that consensus in TopThreader is calculated based on DNN-predicted template similarity (TM-Score) to the native structure and top-down structural comparison to the highest scoring template. This contrasts with traditional consensus approaches like the ones mentioned above, in which the frequency with which a fold is identified is the driving factor of the consensus decision. TopThreader therefore has the advantage that, even if the majority of identified templates or alignments are wrong, it can find true templates and good alignments if the highest scoring template is correct. This selection scheme is a key advantage in cases where correlated threading results produce a bias towards the same false positive templates or wrong alignments, as seen for CASP target T0742 as well as for prospective modeling of the nisin resistance protein NSR from *Streptococcus agalactiae* (*Sa*NSR; see Experimental Validation section). An analogous situation is found in protein model quality assessment, in which

clustering methods (which determine quality based on consensus between models) perform worse at selecting the best model if this model does not belong to a cluster, a task single-model and quasi-single-model methods handle better [43]. In turn, the top-down approach is at a disadvantage if the highest scoring template does not have the correct fold, in which case a potentially correct fold could be discarded when being compared to the highest scoring template.

TopAligner. The use of information from multiple templates can improve model quality by increasing total target coverage or improving pairwise alignments between templates and target by matching structural elements of different templates [26]. This improvement depends heavily on the quality of the templates and their similarity to each other, however. If the quality difference between the best template and other identified templates is large, including sub-par information from bad templates may decrease model quality or distort multiple alignments. Therefore, the TopAligner module calculates an ensemble of pairwise and multiple alignments using every possible combination of the top five compatible (pairwise TM-Score > 0.5) templates. TopAligner uses eight different state-of-the-art programs for template-template alignment (Table S1) and all primary threader and consensus alignments from TopThreader for template-target alignment. Each pairwise template-target alignment is weighted both globally and locally according to the weights calculated by TopThreader from model quality assessment with TopScore, residue-wise lDDT to the best scoring pairwise-alignment model, and residue-wise sequence similarity between target and template. A detailed description of TopAligner and its primary alignment programs can be found in the Supporting Information Text T3 and T4.

TopBuilder. All alignments from TopAligner are modeled using the TopBuilder module, which is also used at the initial modeling stages of TopThreader. TopBuilder uses Modeller9 [37] and the partial thread function of Rosetta [38] to construct models based on alignments and template structures. It includes algorithms for knot detection and elimination, multiple types of loop refinement selected automatically based on loop size, and four methods for model refinement [44-47]. A detailed description of TopBuilder can be found in the Supporting Information Text T5. By default, model refinement is done by side-chain repacking with RASP [45].

TopScore. The ensemble of models generated by TopBuilder is evaluated using TopScore and TopScoreSingle [39]. Since TopAligner produces more alignments based on multiple templates, model selection with TopScore is, due to the use of clustering information, biased towards selecting a multi-template model. As mentioned (see TopAligner section), this bias can in some cases lead to worse models due to inclusion of information from worse templates. Therefore it is key to consider both TopScore and TopScoreSingle when selecting models for refinement

and model combination (see TopRefiner section).

TopRefiner. Previous work [34, 41, 48] has shown that combining different templates or models can improve the accuracy of the final model. Previous work has focused on combining pairwise alignments [41], extracting consensus restraints from templates [42], or averaging models [49]. The TopRefiner module refines models using model quality assessment, model fragmenting, fragment recombination, template/model hybridization and fragment-guided MD refinement in order to remove regions with predicted errors and combine good fragments into full-length models. Models are first selected from the TopAligner (top ranked model for each template combination) and TopThreader (top five primary threader models and top five consensus models according to TopScore and TopScoreSingle) model ensembles. From these models, regions predicted to contain errors by TopScore or TopScoreSingle are removed, and the resulting fragments are re-combined into improved models. After fragment recombination, the models are used to construct new structural alignments to all identified templates, from which hybrid models are built using Rosetta [38]. Finally the best models from each of the previous steps of the refinement are selected and refined with Modrefiner [46] followed by a second round of model fragmenting and recombination. The final model is selected as the highest ranked model in the largest cluster according to TopScore. A detailed description of TopRefiner can be found in the Supporting Information Text T6 and Figure S3.

Data sets

Screening. To train the DNNs of TopThreader on a set of diverse structures and difficulties (with respect to low sequence identity), a screening protocol is used, in which a set of known structures are re-predicted while removing templates with a sequence identity above a given cut-off. The sequence identity cut-offs were chosen as 90%, 60%, and 30%, respectively, to simulate trivial, easy, and difficult modeling situations. A detailed description of the screening can be found in the Supporting Information.

CASP dataset. To evaluate how TopModel performs when compared to other automated methods in the field, the conditions of the CASP10, CASP11, and CASP12 experiments were approximated. By turning on the PDB submission date filter in TopThreader, templates submitted on the day of or after the submission of a CASP target are removed, a procedure similar in nature to the CAMEO experiments [50]. A CASP target was kept if it fulfills three criteria: (1) The target native structure must be submitted to the PDB as of writing this manuscript, to allow for comparison between model and native structure, (2) the target must not have been cancelled during the CASP competition by the organizers, (3) the sequence

identity between the sequence released for prediction and the resolved native structure must be at least 50%. Applying these filtering criterial leaves 140 template-based targets and 46 free modelling targets (Supporting Information Table S2). It is important to note that this approximation will not yield the exact same results as if TopModel was run at the time of each CASP competition. Since threader and sequence databases have been updated since the respective competitions, quality scores (such as e-values and Z-scores) calculated by primary threaders, as well as primary feature predictions (such as secondary structure), will differ from what they would have been at the time of the competition. This can lead to hits that would have been identified with scores above significance cut-offs at the time of CASP competition, but now have scores below the cutoffs for the updated databases. This effect is compounded by database clustering, in particular for threaders that only return a fixed number of hits, of which a significant portion may be released too recently and thus removed by the filter. However, despite these approximations it can serve as a useful indicator of structure prediction performance. None of the CASP targets were considered for the training of the TopThreader DNNs. This dataset will be referred to as the CASP dataset and is used as external evaluation of TopModel performance.

Experimental validation. To evaluate the performance of TopModel on two de novo cases, we modeled the *Sa*NSR protein from the nisin operon of *S. agalactiae* (Uniprot ID A0A140UHB6) [51] before its release to the PDB, and the LipoP from *Clostridium difficile* (Uniprot ID Q18BL3). These structures were then experimentally validated by crystallization [51] or by agreement with SAXS and NMR data (see Experimental Validation).

Results and discussion

Evaluation of TopThreader

The aim of template selection with TopThreader is to retrieve a set of templates ranked according to their similarity (according to TM-Score) to the native structure. To evaluate how well this goal is achieved, we calculate the following: For each target in the CASP dataset (Supporting Information Table S2), the highest TM-Score between the native structure and any template identified in the top five templates of any primary threader is calculated, to find the best obtainable TM-Score given the primary threader results if template selection by TM-Score is perfect. Then, for TopThreader and each primary threader, the highest TM score of the top 5 ranked templates is compared to this best obtainable score. From this comparison we calculate $\Delta TM_{100} = 100 \cdot (\max[TM_{\text{all templates}}] - \max[TM_{\text{top5 templates}}])$. Based on this ΔTM_{100} score for a given target, we define three categories: (I) the best template is found ($\Delta TM_{100} < 5$), (II) an adequate template is found (ΔTM_{100} [5-15]), and (III) no adequate template is found ($\Delta TM_{100} > 15$). We count the frequency of each category for each primary threader and for TopThreader for three subsets of the CASP dataset: (1) Cases assigned by CASP organizers as template-based modeling (TBM) targets, (2) cases assigned as free modeling (FM) targets, and (3) all (TBM+FM) targets. The results are presented in Figure 2 (see Table S3 for numerical values).

The Ghent implementation of the Freeman-Halton exact test for 3x3 contingency tables [52] was used to determine significance between the categorization of TopThreader and each primary threader in terms of the three categories (I, II, III) described above (see Table S4 for summarized normalized tables). Accordingly, all differences are highly significant ($p < 0.01$), for all cases showing a large and significant benefit to selecting templates with TopThreader over any of the tested stand-alone primary threaders.

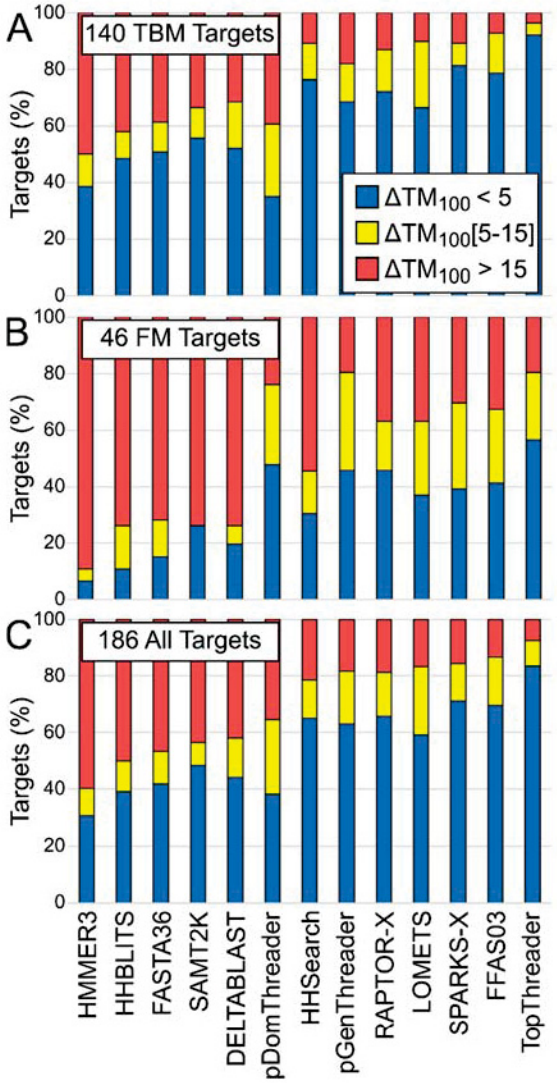


Figure 2. Template enrichment by TopThreader compared to primary threaders. Comparison of template selection performance on the CASP dataset. Performance is evaluated based on the ΔTM_{100} score, which evaluates the difference between the best of the top five ranked templates of a given threader, and the best template found by any threader. For each target, three categories are selected: (I) the best template is found ($\Delta TM_{100} < 5$), (II) an adequate template is found (ΔTM_{100} [5-15]), and (III) no adequate template is found ($\Delta TM_{100} > 15$). The values represent percentages of targets in the CASP dataset for TBM (A), FM (B), and all (C) targets, respectively. Differences between TopThreader and the best primary threader for each subset is are highly significant ($p < 0.01$) according to the Ghent implementation of the Freeman-Halton exact test for 3x3 contingency tables [52]. For numerical values see Supporting Information Table S3 and S4.

The results in Figure 2 show that for the CASP subsets (A: TBM / B: FM / C: TBM+FM), TopThreader identifies the best template (category I, blue) as one of the top five templates in 92 %, 56 %, and 83% of the cases, respectively. Furthermore, an adequate template (category II, yellow) is found in 4 %, 24 %, and 9% of the cases, respectively. The best template (according to TM-Score) is not identified in only 4 %, 20 %, and 8 % of the cases is (category III, red). It also becomes clear that for FM targets it is more difficult to select the template with

the best TM-Score (Figure 2B) since all primary threaders and TopThreader fail to identify the best template for ~20% of targets. It is important to note, however, that for FM targets most TM-Scores are close to or below 0.4 even for the best template, and as such poorly reflect structural similarity in the first place, as two random structures will have a TM-Score of 0.17 when aligned [40].

In addition to evaluating absolute performance for all top five templates, we evaluated the difference in template TM-Score of each of the top five ranked templates by normalizing the TM-Score of a template with a given rank to the template with that rank if the templates had been ranked according to true (rather than predicted) TM-Score. These normalized scores were then averaged, resulting in values closer to 1 corresponding to a ranking similar on average to a perfect ranking by true rather than predicted TM-Score. The full results can be found in Table S5 and show that, in terms of ranking, TopThreader has a significantly better performance compared to the best primary threaders for TBM targets, with an average increase of 2 % across all top five template ranks. For FM targets, a large improvement is seen for the top ranked model (7 %), and lower performance than primary threaders for subsequent ranks. This is surprising considering that templates for FM targets are close to or below the 0.4 TM-Score limit used by TopThreader to distinguish true from false templates, and because according to CASP organizers these targets should have no templates available. This suggests that even for extremely remote structural similarities, TopThreader is able to distinguish between low quality templates and a random match to some degree, as is also reflected in Figure 2B. The lower ranking performance for FM targets for ranks other than the top ranked template is an effect of TopThreader requiring structural consensus between selected templates. Primary threaders do not require consensus, and can therefore rank multiple incompatible folds highly. This gives a higher chance that one of the lower ranked templates is the best, while TopModel only finds the best template if it is either ranked at the top or is structurally similar to the top-ranked template.

Evaluation of TopAligner

To evaluate the effect of using TopAligner to sample alignments with different template combinations and alignment programs, we compared models built from primary threading alignments (TopThreader step 7) with models from TopRefiner stage 1, which are selected from the TopAligner and TopThreader ensembles, but without modifying the models themselves. Model quality is evaluated in terms of GDT_TS score [53], which is used in CASP to evaluate model quality by comparing a model to the native structure and evaluates inter-model C_α atom distance conservation given different distance thresholds. We calculated the change in GDT_TS

score between the two alignment ensembles. However, as we are interested in the relative change in model quality, we calculate the percentage-wise difference denoted as the ΔGDT_TS. All models are built with TopBuilder and selected either with TopScoreSingle or according to the true GDT_TS score, and thus only the alignment ensemble used to generate the models differ. There is no bias from the composition or size of the model ensemble, since neither TopScoreSingle nor the true GDT_TS score depend on composition or size of the model ensemble. This allows us comparing the use of an ensemble of multi-template and single-template alignments, to the use of an ensemble of only single-template threading alignments. The results are shown in Figure 3A.

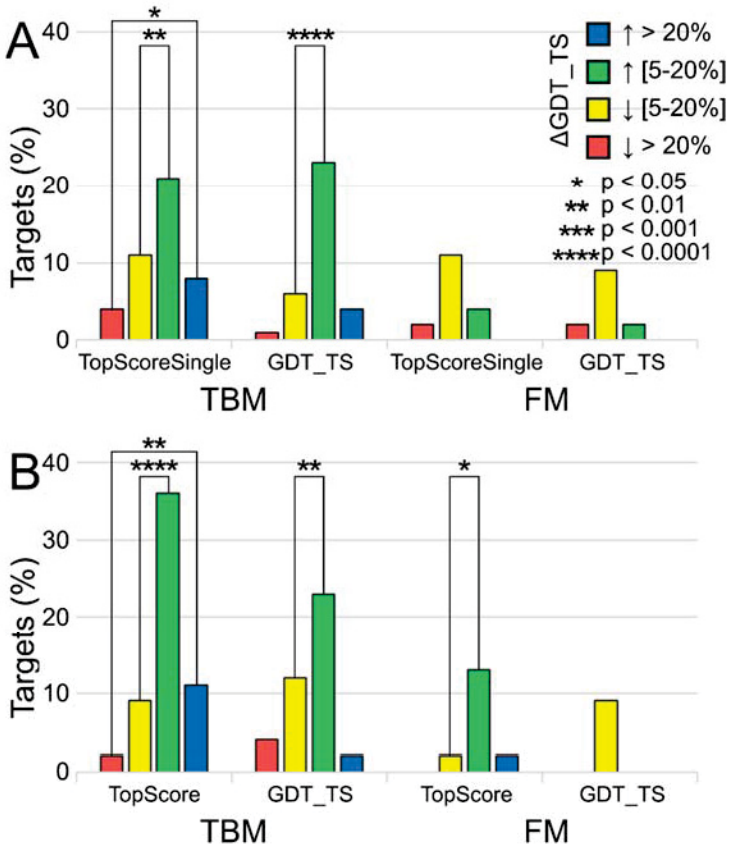


Figure 3. Impact of using TopAligner and TopRefiner on model quality. The relative change in GDT_TS score (ΔGDT_TS) is calculated by comparing a model selected before and after running TopAligner (A) or TopRefiner (B), respectively. **A.** Difference in model quality when selected from a multi/single-template model ensemble from TopAligner/TopThreader compared to selection from a single-template pairwise primary threader model ensemble. **B.** Difference in model quality when selected from the first stage of TopRefiner (before refinement) compared to selection from the last stage of TopRefiner (after refinement). The models are selected either by true GDT_TS or by TopScoreSingle (A) or TopScore (B). Five categories are defined based on the ΔGDT_TS: No change (ΔGDT_TS < 5%), small increase/decrease (ΔGDT_TS ↑/↓ [5%-20%]; green/yellow), large increase/decrease (ΔGDT_TS ↑/↓ > 20%; blue/red). The "No change" category is the most abundant and is not shown as it reflects

no significant change in model quality. Significance is calculated using a one-tailed *t*-test between corresponding increase/decrease categories (blue-red and green-yellow, respectively). The null hypothesis is that the probability of model quality increase of a given amount (5-20% or >20% ΔGDT_TS) is the same as the probability of quality decrease by the same amount. Pairwise comparisons where this hypothesis can be rejected are indicated with brackets and corresponding p-values (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$). The number of samples used is the number of CASP targets in the TBM (140) and FM (46) categories, respectively.

These findings indicate that sampling different alignments and combinations of templates using TopAligner in the majority of cases (56% and 82% of TBM and FM targets, respectively, if selected with TopScoreSingle) leads to little change in GDT_TS score. This result is expected, as for most targets, the different templates cover similar residues or are so similar that model quality is comparable. Furthermore, FM targets rarely have many similar templates identified by TopThreader, since TopThreader requires all identified templates to have the same fold as the top ranked template, which is rarely the case for FM targets. For TBM targets, using multiple templates leads to a decrease in GDT_TS score in 9% and 5 % of cases if selected by TopScoreSingle or by best GDT_TS, respectively. This indicates that in a small number of cases model quality decreases by using multiple templates, usually due to introduction of alignment errors when aligning poor templates with good ones. More importantly, however, for 22% of TBM targets the GDT_TS score improves by 5-20%, and for 9% of targets it improves by >20%. This shows an over three times higher chance that using TopAligner to sample different multi-template alignments will increase model quality. These findings are in line with previous work, showing that using multiple templates and sampling alternate alignments can improve model accuracy [41].

Evaluation of TopRefiner

TopRefiner has three aims: (1) Selection of a small ensemble of good models built by TopThreader and TopAligner, to be used for model combination and refinement, (2) combination of selected models to generate an ensemble of models converging on the correct fold, (3) selecting the best possible model as the final TopModel prediction.

To evaluate the achievement of the first goal, we calculate the ΔGDT_TS between the best model from the stage 1 ensemble and the best model achieved at any previous step of the TopModel workflow from any alignment of any template. We find that in just 6% of TBM targets this distance is more than 5 GDT_TS units (26% for FM targets). The cases in which this distance is large are primarily ones in which template selection with TopThreader fails to select the best template. This confirms that the models selected for refinement and model

combination represent good models compared to ones generated at earlier steps in the pipeline.

To see how well the second goal is achieved, the models from TopRefiner stage 1, which are not refined but simply selected from the TopThreader/TopAligner model ensembles, are compared with models from TopRefiner stage 4, which is after refinement. If TopRefiner is successful, significantly more targets should see an increase in GDT_TS compared to ones with a decrease. The result of this comparison is shown in Figure 3B and demonstrates that in 42% of TBM targets (92% for FM targets) ΔGDT_TS is < 5%, indicating that in these cases no significant change in GDT_TS score is observed, either because the starting models are too far from the true fold to be refined (most FM targets), or because the starting models are so close to the true structure that no improvement is seen (most TBM targets). However, for TBM targets, we find that there is a significant advantage of refinement, with over two times as many systems showing an increase in GDT_TS rather than a decrease. It is interesting to see that model selection with TopScore shows a larger improvement than according to true GDT_TS. This shows that part of the benefit of refinement is an improved ability to select the best model, and not only improving the models themselves, indicating that for many targets convergence to the native fold is a key part of refinement.

Comparison to previous CASP performances

To evaluate the performance of the entire TopModel pipeline, the final TopModel models from the CASP datasets are compared with the highest ranked CASP stage 2 models (CASP stage 2 consists of the top 200 automated server models for each target) from four established CASP servers: The Zhang [42] server (best automated server in CASP8-13) and Baker server [38], both of which use domain parsing and *ab initio* folding as part of their pipeline, and the HHPred [54] and Zhou [27] servers, which do not. Since TopModel has no *ab initio* folding module, and does not parse the target sequence into domains, servers that include such methods are expected to be at an advantage. To evaluate the performance based on the part of the target structure that was solved experimentally, rather than the sequence submitted for prediction, only experimentally resolved residues were evaluated. For each target, the GDT_TS score was calculated for the final model produced by TopModel and the top ranked model from each of the servers mentioned above, as well as the distribution of all server submissions in the stage 2 dataset. As we are interested in the absolute difference in model quality, we classify each CASP target based on the absolute difference in GDT_TS score (ΔGDT_TS_{abs}) between the final model from TopModel and the top ranked model from each server. The results can be found in Figures 4A and B for TBM and FM targets, respectively.

As of yet, TopModel has no domain parsing module to cut the input sequence into domains before modeling. Therefore, in the cases where multiple domains have good templates but no template covers the whole sequence, TopModel will match the best (often largest) domain template, leaving the other domains without template. Therefore TopModel is at a disadvantage for large multi-domain targets for which no template is found that covers all domains. We expect this to be particularly detrimental for FM targets, most of which have multiple domains. To estimate the hypothetical performance that TopModel could achieve if multi-domain targets were modeled domain-wise, and combined in the correct way, the CASP domain annotations (released after the end of each competition) were used to parse the sequences of multi-domain targets into their respective domains. Each domain was then submitted to TopModel separately, given the same restrictions as for regular targets to emulate previous CASP rounds. For each target, a weighted average (by number of residues) of the GDT_TS scores of the respective domains is calculated as the hypothetical accuracy if domain parsing and combination was used. We then compare the GDT_TS score of models built from the CASP sequence released for prediction with this weighted average, and evaluate the change in ΔGDT_TS$_{abs}$ ($\Delta\Delta$GDT_TS$_{abs}$). If $\Delta\Delta$GDT_TS$_{abs}$ is positive, domain parsing improves model quality relative to other servers, and if negative, it deteriorates model quality. The results are depicted in Figures 4C and D for TBM and FM targets, respectively.

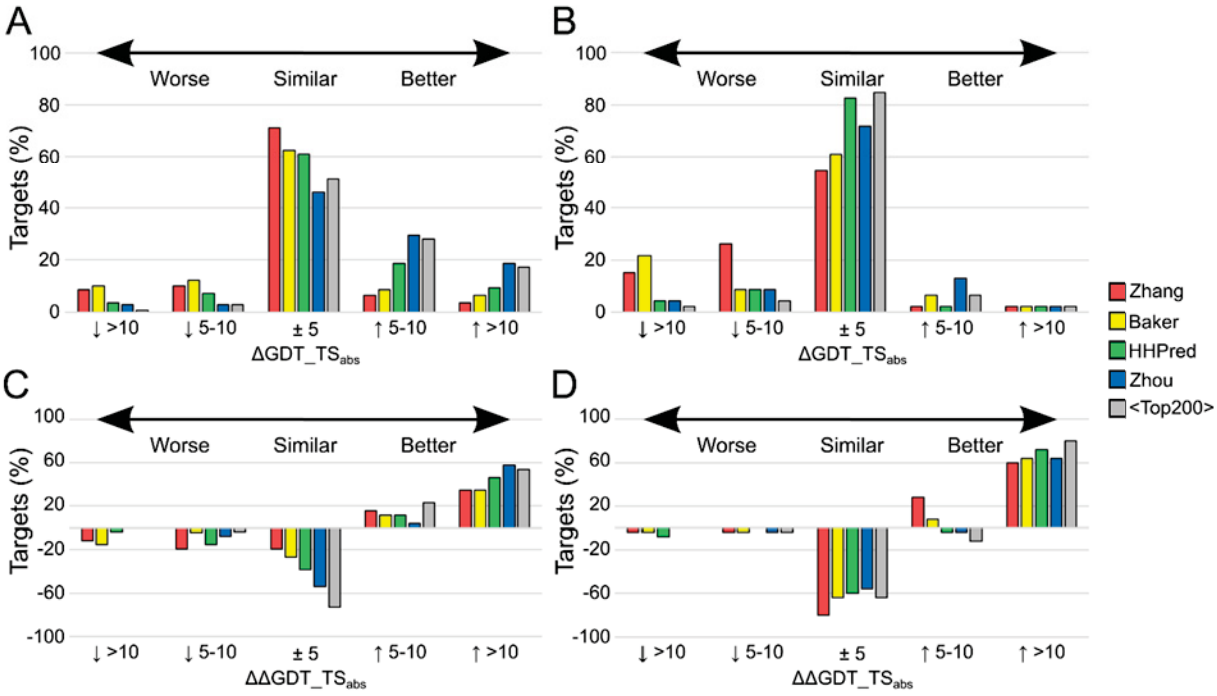


Figure 4. GDT_TS comparisons between TopModel and CASP servers. The bars represent comparison between TopModel and one of four established CASP servers (the Zhang Server (red), the Baker Server (yellow), the HHPred server (green), the Zhou Server (blue)) as well as the average of the top 200 server submissions for

each target (grey). The Zhang server and Baker server both make use of *ab initio* folding and domain parsing, putting them at an advantage over TopModel. **A.** ΔGDT_TS_{abs} for CASP TBM targets indicates for how many of CASP TBM targets TopModel shows similar, worse, or better model quality than other established servers. **B.** ΔGDT_TS_{abs} for CASP FM targets indicates for how many of CASP FM targets TopModel shows similar, worse, or better model quality than other established servers. **C.** $\Delta\Delta GDT\_TS_{abs}$ for multi-domain TBM CASP targets shows the change in the number of targets for which TopModel performs worse, similar, or better than established servers, if domain parsing, domain-wise modeling, and domain recombination was used. A large shift from worse/similar model qualities to better model qualities (compared to established servers) is seen. **D.** $\Delta\Delta GDT\_TS_{abs}$ for multi-domain FM CASP targets shows the change in the number of targets for which TopModel performs worse, similar, or better than established servers, if domain parsing, domain-wise modeling, and domain recombination was used. A large shift from worse/similar model qualities to better model qualities (compared to established servers) is seen.

Our findings shows that despite being at a disadvantage compared to the Zhang and Baker servers due to lack of domain parsing and *ab initio* folding, 60-70% of TBM target models from TopModel are of comparable quality, with 10-15% of targets having higher quality and 19-22% of targets having lower quality (Figure 4A). Compared to pure template based servers such as HHPred and Zhou Servers, on the other hand, TopModel has a clear advantage, with 28-48% of TBM targets having higher quality and 6-11% having lower quality. For FM targets, despite having no *ab initio* module, TopModel shows comparable accuracy to the Zhang and Baker server for 54-61% of targets (Figure 4B), but a lower accuracy for 30-41% of targets, which is not surprising given the lack of *ab initio* folding and domain prediction in TopModel (most FM targets are multi-domain targets).

The results in Figures 4C and 4D show that a large improvement is possible for multi-domain targets if the sequence is parsed into domains, predicted separately, and combined into a full-chain model. When compared to the Zhang Server, for example, for TBM targets the percentage of multi-domain targets for which TopModel is worse than the Zhang server drops by 31 points, while the percentage of targets for which TopModel is better than the Zhang server increases by 51 points. For FM targets the same trend is seen, with the percentage of worse models dropping by 8 points and the percentage of better models increasing by 88 points. Similar trends are observed for the other three investigated servers. This indicates that correctly parsing the input sequence into domains has a large impact on the quality of multi-domain models, in particular for FM targets.

We speculate that the reason behind this is that accurately identifying a partially matching template for a large multi-domain protein is difficult, especially for methods that have been trained to identify templates for single domains. As such, many FM targets may have been

classified as such due to a failure to detect templates using the full sequence as a query, and not because of an actual lack of templates. These results show that, when properly parsed into domains and searching for each domain, template detection is easier and distant structural homologues become detectable for many targets that would traditionally be considered without templates. Thus, a large model quality improvement is achievable by predicting domain boundaries and combining the domains into a final model. However, in order to achieve such accuracy on prospective targets, accurate domain prediction and domain combination is required, which is therefore the focus of our future work.

The results in Figure 4 show that TopModel has comparable or better performance than the average server submission (grey) for the majority of targets (97% for TBM, 93% for FM) and performs significantly better than template-based servers without *ab initio* folding such as the Zhou and HHPred servers. TopModel even shows comparable or better performance than the Baker and Zhang servers for 82% and 78% of TBM targets and 59% and 70% of FM targets, respectively. These data show the benefit of using a top-down consensus rather than majority voting, and the benefit of combining threading scores, model quality, and structural alignment using deep neural networks for ranking and selecting templates.

It is interesting to examine a case such as T0742 from CASP10. For this target the vast majority of predictions from CASP servers, including the consensus based MULTICOM server, fail to identify the best template (PDB ID 3TZG, identity = 14%, coverage = 70%, GDT_TS = 0.31) and instead predict a fold based on the wrong template identified by the majority of threaders. TopModel, however, identifies PDB ID 3TZG as the best template, a direct effect of its ability to discard wrong templates even when the consensus is indicating that they should be correct. A similar effect is seen for the prospective modeling of *Sa*NSR (see below).

Hard cases

Although TopModel correctly folds most CASP TBM targets and has better template selection and alignment than any of its primary threaders (Figures 1-3), there are cases where it fails to predict the best template when comparing GDT_TS scores to those of other competing servers. Aside from the issues of simulating previous CASP rounds mentioned earlier, manual inspection indicated three main types of such cases where TopModel is at a disadvantage compared to servers such as those of Zhang and Baker.

First, for several targets, no template is found that covers all domains of the target. Based on CASP annotations released after the competitions, 39% of targets in the CASP dataset are

multi-domain targets (18% of TBM, 98% of FM). Additionally, there are several targets (including T0721, T0737 and T0755) that are non-consecutive multi-domain targets annotated as single-domain by the CASP organizers, for which TopModel is either only able to match one domain (such as for T0755) or finds a slightly different and more favorable (better score from TopScore) conformation, resulting in a lower GDT_TS score (such as for T0922 and T0833).

Second, there are many targets (in particular FM targets), for which the sequence submitted for prediction differs significantly from that of the resolved native structure. In most such cases, the native structure covers only a small fraction of the residues submitted for prediction. This makes structure prediction much more difficult, since threading algorithms focus on templates that cover as much of the target sequence as possible, when in fact only a small fraction of it can be resolved. For these targets, servers that use domain prediction have an advantage as they mitigate the inherent threader bias towards high target coverage by cutting the sequence into predicted domains.

Third, there are a few cases in which TopThreader discards the best template for TBM targets as a false positive. Three such cases (T0678, T0700, and T0818) were identified. To examine these cases, the best template, the highest ranked template by *Initial Score*, and the highest ranked template identified by *Filtering Score* are compared to the native structure in terms of GDT_TS and TM scores. The results are shown in Table 1.

Table 1: Inspection of hard TopThreader TBM targets

ID	Template	Threaders	Identity↑	Coverage↑	Initial Score↑	TopScore Single↓	GDT_TS↑	TM↑
T0700	2OVR*	HHSearch	17%	79%	0.64	0.59	0.49	0.51
	3V7D (I)	HHBlits HHSearch	21%	52%	0.66	0.59	0.35	0.47
	1EZJ (F)	pDomThreader	21%	60%	0.46	0.49	0.21	0.22
T0812	4BQ2*	LOMETS	10%	87%	0.45	0.60	0.28	0.45
	1H6Y (I)	RAPTORX	12%	60%	0.51	0.60	0.12	0.20
	3LY6 (F)	RAPTOR-X	18%	59%	0.48	0.60	0.14	0.33
T0818	4HYZ*	HHBlits HHSearch	15%	55%	0.64	0.72	0.25	0.37
	3H51(I)	SPARKSX FFAS03	13%	80%	0.68	0.69	0.14	0.34
	4CE4(F)	LOMETS	13%	91%	0.56	0.62	0.15	0.25

Table 1: Inspection of hard TopThreader TBM targets. Summary of scores for the CASP TBM targets for which TopThreader fails to select the best templates. * indicates the best template according to lowest GDT_TS for a model built from that template. (I) indicates the highest ranked template according to *Initial Score*, which is a prediction of template TM-Score using only sequence-derived features from primary threaders. (F) Indicate the highest ranked template according to the *Filtering Score,* which is a prediction of template TM-Score using both sequence-derived features from primary threaders and predicted error in the resulting model according to TopScoreSingle. The GDT_TS and TM columns indicate structural similarity between the best model from a given template and the native structure (not the TM-Score of the

template). The arrows ↑↓ indicate if a score gets better with increasing or decreasing values, respectively.

T0700: For this target the best template (PDB ID 2OVR) is discarded because it scores much worse by TopScoreSingle, which lowers the *Filtering Score*. This shows that despite higher coverage, models built from such a template will not always exhibit a better model quality, and as such, selection by model quality alone does not guarantee that the best template is found.

T0812: For this target the best template (PDB ID 4BQ2) has a lower *Initial Score* than both the false positive templates PDB ID 1H6Y and PDB ID 3LY6. All three templates result in models with identical scores from TopScoreSingle. This shows that using scores from primary threaders alone does not guarantee that the best template is found.

T0818: For this target the best template (PDB ID 4HYZ) has lower coverage and consequently also worse TopScoreSingle score than both the best ranked template according to *Initial Score* (PDB ID 3H51) and *Filtering Score* (PDB ID 4CE4), leading to a false positive template being selected due to higher coverage. This is similar to T0700 in the sense that a higher weight on the *Initial Score* would have led to a better model, but in this case the template with lower coverage is better, unlike for T0700 and T0812 where the higher coverage templates are better.

Analyzing the few TBM cases for which TopThreader does not select the best template shows that template selection is a complex task and that no single feature is likely to result in a flawless prediction for every target. However, the performance of TopThreader (Figure 2) shows that taking features from both primary threaders and model quality into account using deep neural networks significantly improved template selection. We expect that using predicted residue-residue contacts can further improve the template selection to resolve such issues.

Prospective prediction and experimental validation of *Sa*NSR

Because TopModel uses a different consensus methodology than other methods, it can potentially go against the majority of threading results and give a prediction better than any of its constituent predictors. To illustrate the effect of this kind of consensus, we prospectively predicted the structure of *Sa*NSR (PDB ID 4Y68) prior to experimental structure determination and submission to the PDB. *Sa*NSR is a member of the S41 protease family, degrades the lantibiotic nisin, and that way contributes to the congenital resistance against nisin of *S. agalactiae*[51].

We then compared the model from TopModel to the distribution of primary threader models in terms of how close each model is to the experimental structure (measured by

GDT_TS) (Figure 5). The results reveal that models based on most of the primary threader alignments are of poor quality, with median GDT_TS scores of 38, while the model from TopModel is much more accurate, with a GDT_TS score of 55 and a C_α atom RMSD of 3.1 Å. The main reason for the failing of primary threaders in this case is that in most available templates, there is one or more large domain insertions. This causes the majority (82%) of threaders to thread the N-terminal helix bundle sequence onto the wrong domain (see the SPARKSX example in Figure 5) due to low (≤ 16%) sequence identity, incorrectly folding it into a β-sheet domain. However, because this β-sheet domain is scored poorly by TopScore and TopScoreSingle, the helix bundle is recovered in the model from TopModel. There is a minority of primary predictor models (18%) that show a correctly aligned helix bundle N-terminal domain. However, these contain significant differences in other parts of the model and are with traditional majority voting consensus far outweighed by the incorrect alignments, showing the benefit of using a top-down consensus approach rather than majority voting.

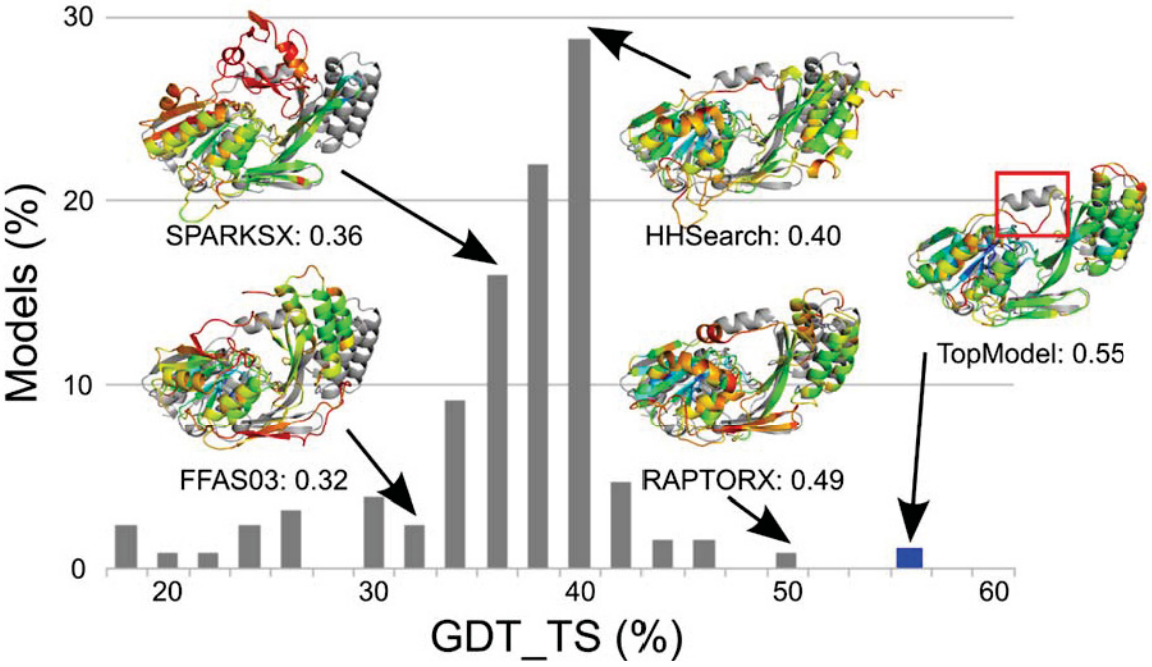


Figure 5. Prospective modeling of the *NSR* protein from *S. agalactiae*. The model quality distribution (in terms of GDT_TS score) of primary threader models for the *NSR* protein from *S. agalactiae* for prospective modeling before the release of the native structure (grey) to the PDB. The vast majority (82%) of models show an incorrectly threaded N-terminal domain (see SPARKSX model). A minority of models (18%) show a correctly threaded helical domain (HHSearch, RAPTORX, and FFAS03) on a few templates, often with large errors elsewhere in the model (such as β-sheets shown in red). Because TopModel does not use majority voting, the model produced (blue box) is of far better quality (GDT_TS = 55) than those produced by primary threaders (median GDT_TS = 38), while majority voting consensus would produce a model in the middle of the distribution at a GDT_TS of ~38. Model examples from the different bins are colored according to residue-wise lDDT score [55] to the native structure,

with red showing incorrectly modeled regions and blue showing perfect agreement with the crystal. The largest error in the TopModel model is the fact that the residues linking the helical bundle with the catalytic core of the protein do not fold into an α-helix (red box). This is because no model from any of the primary predictors correctly fold these residues into a helix, and as such TopRefiner has no fragment it can select during model fragmenting and refinement which would produce a helix for these residues; secondary structure prediction by PSIPRED[56] also fails to identify these residues as helical.

Prospective prediction and experimental validation of LipoP from *C. difficile*

Building on the previous successes, TopModel was used to prospectively predict the structure of LipoP from *C. difficile*. The templates identified by TopThreader (sequence identity in parenthesis; chain after the "_") are PDB IDs 5J7R_A (11%), 6GZ8_A (18%), 2JNV_A (18%), 5O5J_C (9%), and 3GKU_A (8%). Interestingly, the top ranked structure PDB ID 5J7R is a putative lipoprotein from *C. perfringens*, and as such is suggested to share the biological function with the homolog from *C. difficile*, despite the sequence identity being far below the 30% sequence identity limit generally considered the twilight-zone for template-based structure prediction [31]. The final model quality predicted by TopScore was 0.35, indicating about 35% error in the model. This shows that the model may not be highly confident, which is to be expected given the low sequence identity and the fact that the first 43 residues (28% of the protein) of the N-terminus (termed the tail region) are unstructured and therefore highly mobile (a description of the tail region is available in the Supporting Information Text T7). To validate the model and identify errors, NMR experiments were therefore carried out to determine the secondary structure and β-strand pairing, and small-angle X-ray scattering (SAXS) experiments were performed to estimate the shape and radius of gyration (R_G). A detailed description of the NMR and SAXS experiments can be found in the Supporting Information Text T7.

The initial model from TopModel has a good agreement with the secondary structure assignment and matches two out of three NOE β-strand pairings (strand 1/2, and strand 4/5) from NMR. The Matthews correlation coefficient (MCC) between the DSSP [57] secondary structure of the model from TopModel and the experimental assignment from NMR is 0.81 for β-strands, 0.68 for α-helices, and 0.66 for coil. However, there are still discrepancies between the predicted model and the experimental data. Four differences can be identified (Figure 6A, B): **(1)** α-helix 1 is eight residues shorter in the model than indicated by NMR, which is also indicated by TopScore showing the loop between α-helix 1 and β-strand 3 to have a high error (> 50% residue-wise error). This is also the reason that random coil and helical MCCs are much lower than those for β-strands. **(2)** β-strand 3 is indicated by NMR NOE measurements to be shifted by two residues, which produces a longer loop between α-helix 1 and β-strand 3, the

loop indicated by TopScore to contain high error. Since the shift is 2 residues however, the hydrophobic valines in this sheet are still buried. **(3)** According to NMR, the C-terminus of the protein is folded into a β-strand, which most likely pairs up with the previous stand (β-strand 5), which in the model is three residues too short. **(4)** α-helix 2 is scoring poorly according to TopScore, due to a difference in helix length of one residue and its proximity to errors 1 and 2.

All of the differences in the LipoP model are due to the fact that TopModel is a template-based structure prediction method which does not use *ab initio* folding. When none of the initial template-based models from pairwise or multi-template alignments produce correct structural fragments, TopRefiner is unable to select a fragment with the correct fold for such residues. A comparison of the final model from TopModel and the two highest ranked templates is shown in Figure S4 and illustrates this point. To correct differences such as these, *ab initio* folding is required, in order to supplement the template-based model ensemble with models from *ab initio* folding, and enable TopRefiner to select folded fragments not present in the templates. The Zhang server [42] had the same issue, which was remedied by the inclusion of *ab initio* models from QUARK [58]. It is important to note, however, that without the use of TopModel, the highest scoring model from primary threading alignments, generated by dPPAS2 from the LOMETS server, is of much lower quality (Figure 6C, D).

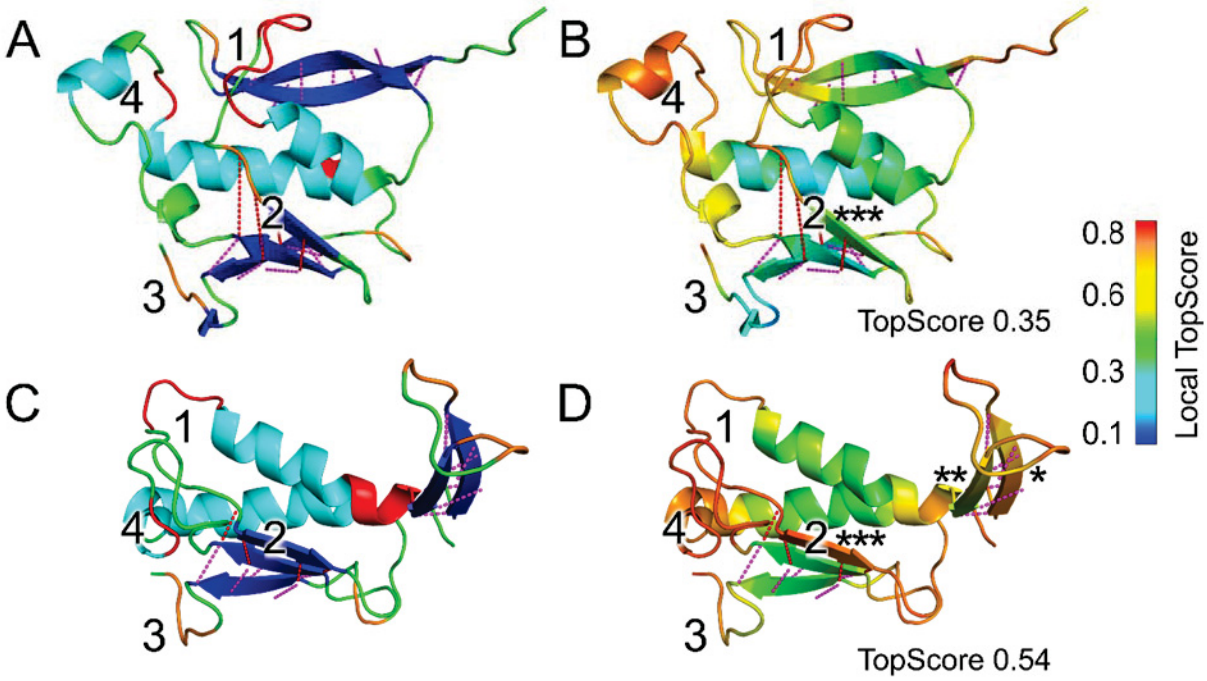


Figure 6. Prospective modeling of LipoP from *C. difficile* (disordered tail not shown for clarity). A. Agreement of the TopModel model with secondary structure assignments and NOE restraints from NMR. **Blue:** β-sheet residues showing agreement between model and NMR data. **Orange:** Residues identified to be in a β-strand in NMR but not found so in the model. **Cyan:** α-helical residues showing agreement between model and NMR data. **Red:** Residues identified to be α-helical in NMR but not found so in the model. **Magenta lines:** Experimental β-sheet NOE restraints showing agreement with the model. **Red Lines:** Experimental β-sheet NOE restraints showing a shift of two residue positions of β-strand 3 (**). **B.** The model colored according to residue-wise TopScore. Yellow/Red regions indicate regions with high residue-wise error (> 50%). **C.** The best model (according to TopScore) from primary predictors (dPPAS2 from the LOMETS server). The coloring scheme is the same as in A. **D.** The best model (according to TopScore) from primary predictors (dPPAS2 from the LOMETS server). The coloring scheme is the same as in B. The shift of β-strand 3 (***) is only 1 residue in this model, placing two hydrophobic valines on the wrong side of the sheet and exposing them to the solvent. Furthermore, β-strand 1(*) and 2 (**) are exposed to the solvent, exposing five hydrophobic isoleucines and one leucine to the solvent, all of which are buried in the TopModel prediction. Numbers 1-4 relate to the location of the errors described in the main text for panel A and B and corresponding locations in the best primary threader model in panels C and D. In panels A and B, these errors are caused by the fact that no template-based model from any primary predictor folds these regions correctly, which leaves TopRefiner unable to select a correctly folded fragment for these residues.

Since TopModel does not include any *ab initio* folding as of yet, we carried out molecular dynamics (MD) simulations of in total 600 ns starting from the TopModel model, either using only the folded domain or the full-length sequence including the disordered tail, in an attempt to improve agreement with the available experimental data in terms of NMR secondary structure assignment and radius of gyration (R_g) from SAXS. The best snapshot from the globular domain simulations and the best snapshot from the full length model simulations were

selected according to agreement with NMR and SAXS data, respectively. These two snapshots were combined with TopBuilder and energy minimized to create a final full-length refined model (Figure 6). A detailed description of the MD simulations, the selection protocol, and the structural refinement can be found in the Supporting Information Text T7. The final refined model shows a secondary structure MCC of 0.81 for β-sheets, 0.88 for α-helices, and 0.78 for random coils. The propensity for each residue to be helical or β-sheet across all simulations can be found in the Supporting Information Figure S5. The initial shape agreement with SAXS (see Table S6 and Figure S7 for experimental data) has a χ^2 of 49.8 (Figure S7A, B), which is high but not surprising given the highly mobile disordered tail. The model shows a radius of gyration of 26.7 Å, which compares favorably to the experimentally determined value of 24.3 Å (Table S6). Most interestingly, in the MD simulations, the loop between α-helix 1 and β-strand 3 shows some α-helix formation (see Figure S6 for a normalized distribution of secondary structure agreement with NMR across the MD simulations). After combining the two models agreeing best with NMR and SAXS, respectively, using TopBuilder, we find that error **(1)** (Figure 7A, B) has been mostly corrected, in that α-helix 1 has been extended to nearly the same length indicated by NMR. None of the other errors were significantly impacted by the MD refinement; however, one cannot expect MD simulations to be able to fix alignment errors on the time scales applied (20×30 ns).

To explore if the high χ^2 with respect to SAXS data is caused by the disordered tail, a truncated version of the protein was expressed in which the first 30 of the 43 disordered tail residues were removed. When SAXS measurements of the truncated protein are compared to the full-length model after MD refinement and combination with TopBuilder (Figure 7A) from which the same tail residues were removed, the shape agreement increases markedly as indicated by a drop in χ^2 from 49.8 to 3.9 (Figure 7C, D and Figure S7), confirming that the initial disagreement with the full-length SAXS data is indeed caused by the high mobility of the disordered tail, and that the shape of the folded domain shows a high agreement with experiment.

In short, the modeling of LipoP from *C. difficile* clearly demonstrates the value of close interplay between computational structure prediction with TopModel and the use of sparse experimental structural data to validate and improve the predicted model but also to identify structural parts that still lack accuracy.

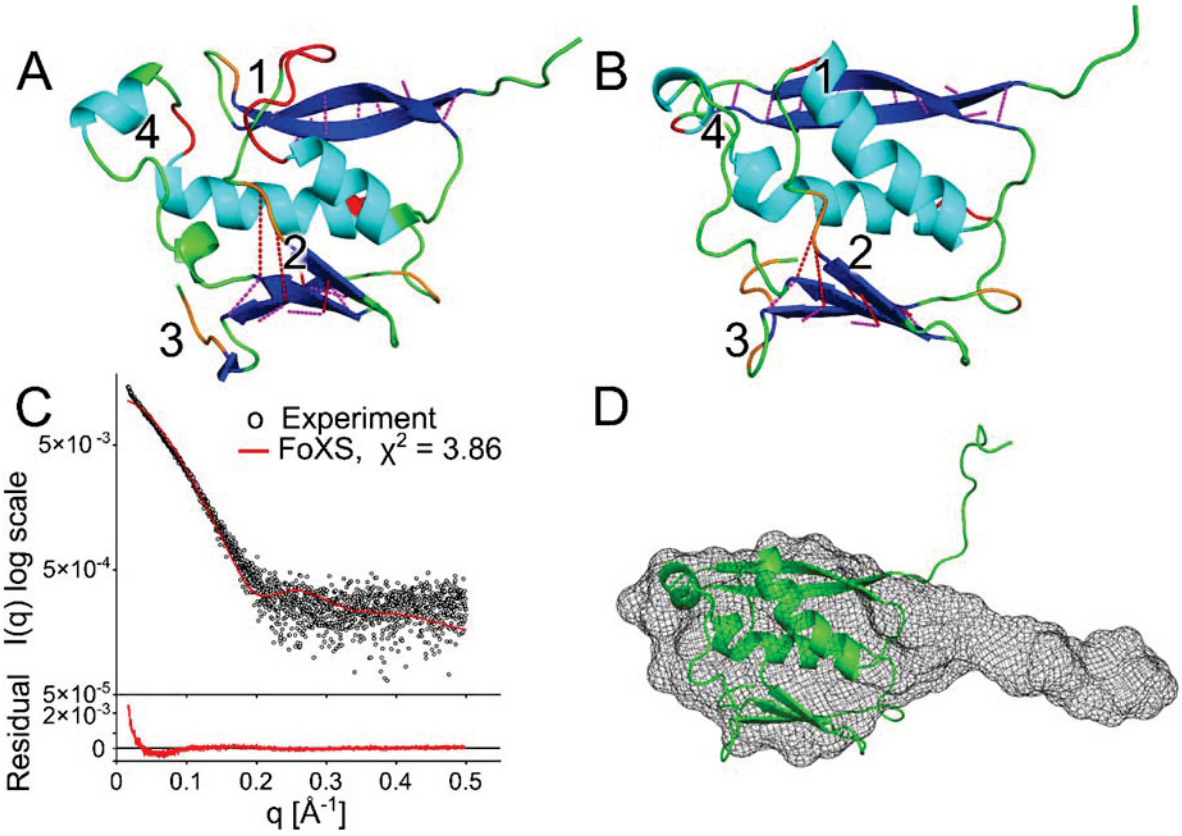


Figure 7. Model of LipoP from *C. difficile* after MD refinement and selection according to agreement with sparse experimental structural data. A. Agreement of the TopModel model with secondary structure assignments and NOE restraints from NMR. The numbers indicate the location of errors as described in the text previously. This panel is as in Figure 6A and again shown for ease of comparison here. **Blue:** β-sheet residues showing agreement between model and NMR data. **Orange:** Residues identified to be in a β-strand in NMR but not found so in the model. **Cyan:** α-helical residues showing agreement between model and NMR data. **Red:** Residues identified to be α-helical in NMR but not found so in the model. **Magenta lines:** Experimental β-sheet NOE restraints showing agreement with the model. **Red Lines:** Experimental β-sheet NOE restraints showing a shift of two residue positions of β-strand 3. **B.** Agreement between the model after MD refinement, selection according to agreement with experimental NMR and SAXS data, and combination with TopBuilder (see main text and Supporting Information Text T7 for details) and experimental NMR data, colors are following the same scheme as in panel A. The extension of α-helix 1 is seen. **C.** Agreement between the experimental scattering data from SAXS (black) and simulated scattering curve of the MD model (red); FoXS [59, 60] was used for simulating the scattering curve. The fit plots depict log-intensity *versus* q ($Å^{-1}$), the residuals plot shows the difference between experimental and computed intensity *versus* q ($Å^{-1}$). **D.** The volumetric envelope of LipoP, as calculated from the scattering data using GASBOR [61], is shown in gray mesh. The MD model of LipoP (green) was docked into the volumetric envelope using SUPCOMB [62]. Disagreement with SAXS is found mainly for the disordered tail of LipoP.

Preliminary competition in CASP13

Despite TopModel development not being finished at the time of the CASP13 competition, in particular lacking most of the TopRefiner module, we decided to compete as a human server.

The CASP12 and CASP13 competitions saw a huge impact of recent developments in *ab initio* folding in terms of highly accurate residue-residue contact and distance predictions, which also have had a large impact in template selection to remove false positive templates. CASP13 also had the highest number of multi-domain targets of any CASP competition to date, with some targets having more than 1000 residues. As such we did not expect TopModel to rank very well compared to servers that utilize these tools such as the Zhang and A7D servers. TopModel showed a very good performance for several targets, however, most notably targets T1016-D1 (Figure 8A), T1014-D2 (Figure 8B) and T0964-D1 (Figure 8C), for which the models produced by TopModel were ranked second. Overall, our findings in CASP13 confirm our conclusions from our own benchmarking on the CASP dataset, in that while deep learning does improve template-based structure prediction, *ab initio* folding and domain prediction is required for folding large multi-domain structures and structures without known templates.

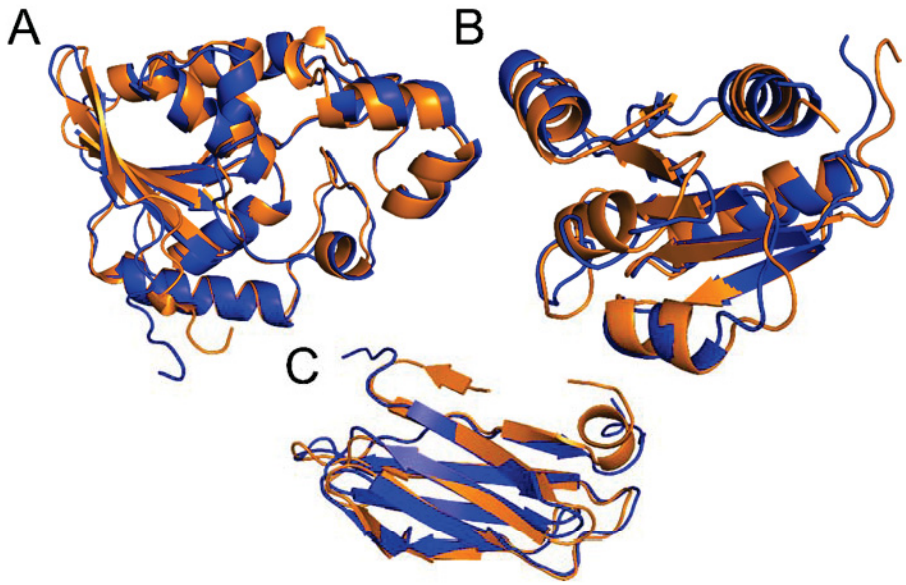


Figure 8. Examples of highly accurate structure predictions from TopModel in CASP13. A. T1016-D1: A7D predicted the best model (blue) and TopModel predicted the second best one (orange) ($GDT\_TS_{TopModel}$ = 81.9, GDT_TS_{Best} = 85.4, $C\alpha\text{-}RMSD_{TopModel\_to\_Best}$=1.1Å). **B.** T1014-D2: McGuffin predicted the best model (blue) and TopModel predicted the second best one (orange) ($GDT\_TS_{TopModel}$ = 76.4, GDT_TS_{Best} = 76.7, $C\alpha\text{-}RMSD_{TopModel\_to\_Best}$=1.5Å). **C.** T0964-D1: MESHI predicted the best model (blue) and TopModel predicted the second best one (orange) ($GDT\_TS_{TopModel}$ = 78.7, GDT_TS_{Best} = 80.0, $C\alpha\text{-}RMSD_{TopModel\_to\_Best}$=1.6Å). RMSD was calculated using the align function in PyMol[63]. The native structures were not released as of writing this manuscript.

Concluding remarks

In this study, we introduced TopModel, a fully automated meta-method for protein structure prediction, which improves template-based threading beyond any of the twelve evaluated primary predictors. Instead of using majority voting during template selection and model averaging during refinement as other approaches [41, 42], TopModel uses top-down consensus

and deep neural networks to select templates and identify and correct wrongly modeled regions. TopModel builds on numerous well-founded approaches to template-based structure prediction in terms of primary programs used for threading, alignment, model building, refinement, and model quality estimation. Yet, aside from the aspect of automation, TopModel offers several advantages over using these programs individually: We demonstrate a significant improvement for template selection and alignment accuracy due to sophisticated template selection with TopThreader, use of multiple alternate alignments between different combinations of templates with TopAligner, and model refinement with TopRefiner using TopScore and TopScoreSingle to detect wrongly modeled regions of the protein.

By applying TopModel to our CASP dataset, which includes targets from CASP10, 11, and 12 with released structures, we showed that TopModel consistently performs better than the average competing server, and outperforms established template-based servers such as the Zhou and HHPred servers. Yet, we identified two areas in which TopModel currently falls short of state-of-the art predictors, mainly in terms of *ab initio* structure prediction and domain prediction for multi-domain targets. As seen for top ranking servers in CASP12 and CASP13, such methods are required to be competitive for multi-domain targets for which no template is available that covers all domains, or for targets for which a correct template structure cannot be detected by threading.

Early versions of TopModel have been applied to several systems, including enzymes [2, 4], ethylene receptors [64], and restriction factors [65], and yielded good predictions that agreed with experimental results and/or allowed for guiding of biochemical experiments. Here, we applied TopModel to predict the structure of the *Sa*NSR protein *de novo*; subsequent experimental structure determination by X-ray crystallography showed that TopModel predicted the correct fold even when the vast majority of primary threaders produced incorrect alignments and models. Finally, we used TopModel to predict the structure of LipoP, which showed a good agreement with data from NMR spectroscopy and SAXS. The modeling of LipoP highlights the utility of the method and shows how the close interplay between computational structure prediction and sparse or low-resolution experimental data can be used synergistically to improve the final model.

Overall, we have shown that TopModel outperforms other stand-alone methods in the field with regards to template selection, template-target alignment, and model quality. However TopModel is at a disadvantage when compared to black-box automated online servers, which utilize recent developments in residue-residue contact predictions, *ab initio* folding, and domain predictions. Therefore, we are focusing future work on contact prediction, *ab initio* folding, and

domain prediction to improve the performance of TopModel for such targets. The TopModel suite is available as stand-alone program from https://cpclab.uni-duesseldorf.de/software.

Acknowledgements

We are grateful to the developers of all primary programs used in this work for making their methods available as standalone to the scientific community. In particular, we are thankful to the developers of Phyrestorm for providing their clustering tree and the developers of MergeAlign2 for providing matrices used for multiple alignment with MergeAlign2. The authors acknowledge access to the Jülich-Düsseldorf Biomolecular NMR Center. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 267205415 – SFB 1208 (project A03 to HG and B03 to PN) and by the Bundesministerium für Bildung und Forschung (BMBF) – Förderkennzeichen 031L0182 – InCelluloProtStruct to HG. We are grateful for computational support and infrastructure provided by the "Zentrum für Informations- und Medientechnologie" (ZIM) at the Heinrich Heine University Düsseldorf and the computing time provided by the John von Neumann Institute for Computing (NIC) to HG on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC) (user ID: HKF7). The Center for Structural Studies is funded by the Deutsche Forschungsgemeinschaft (DFG Grant number 417919780 and INST 208/761-1 FUGG).

Supporting information

The Supporting Information is available free of charge on the ACS Publications website at DOI: XXX.

The supporting information include: Detailed descriptions of TopThreader, TopAligner, TopBuilder, and TopRefiner, brief descriptions of the primary methods used as part of TopThreader and TopAligner, detailed description of the training and test datasets and the training of the deep neural networks for TopThreader, targets in the CASP evaluation dataset and numerical data for the evaluation of TopThreader performance on this dataset in terms of template selection and template ranking compared to primary predictors, detailed descriptions of the experimental validation, data collection and molecular dynamics simulations-based refinement for LipoP from *C. difficile*.

Author contributions

HG and DM jointly conceived the study. DM developed the method, performed computations, analyzed the results, and wrote the manuscript. NP performed the molecular dynamics

simulations and analyzed the disordered tail. HG supervised and managed the project, secured funding and resources for the project, and revised the manuscript. RC prepared samples for SAXS. JR performed SAXS measurement and data analysis. RC, LG, and SS prepared the NMR samples. PN recorded the NMR experiments, IA and PN analyzed the NMR spectra. All authors reviewed and approved the manuscript. The authors declare no competing interests.

References

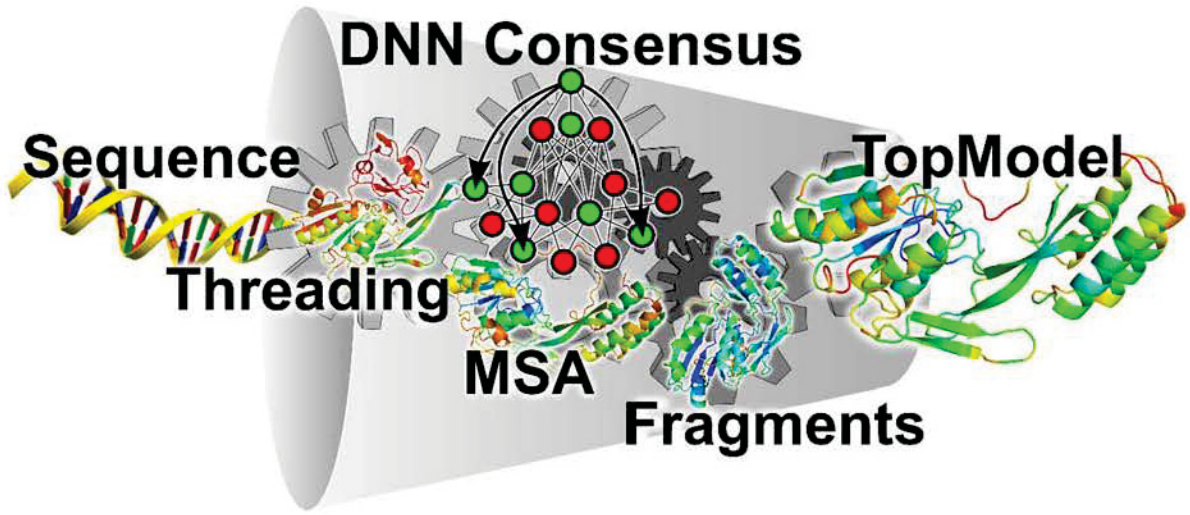
1. Rathi, P.C., Höffken, H.W., & Gohlke, H., Quality matters: Extension of clusters of residues with good hydrophobic contacts stabilize (hyper) thermophilic proteins. *Journal of chemical information and modeling*, **54**(2), 355-361 (2014).
2. Widderich, N., Pittelkow, M., Höppner, A., Mulnaes, D., Buckel, W., Gohlke, H., Smits, S.H.J., Bremer, E., Molecular dynamics simulations and structure-guided mutagenesis provide insight into the architecture of the catalytic core of the ectoine hydroxylase. *Journal of molecular biology*, **426**(3), 586-600 (2014).
3. Ingles-Prieto, A., Ibarra-Molero, B., Delgado-Delgado, A., Perez-Jimenez, R., Fernandez, J.M., Gaucher, E.A., Sanchez-Ruiz, J.M., Gavira, J.A., Conservation of protein structure over four billion years. *Structure*, **21**(9), 1690-1697 (2013).
4. Gohlke, H., Hergert, U., Meyer, T., Mulnaes, D., Grieshaber, M.K., Smits, S.H.J., Schmitt, L., Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion. *Journal of Chemical Information and Modeling*, **53**(10), 2493-2498 (2013).
5. Yang, J., A. Roy, & Y. Zhang, Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, **29**(20), 2588-95 (2013).
6. Janin, J., Assessing predictions of protein–protein interaction: the CAPRI experiment. *Protein Science*, **14**(2), 278-283 (2005).
7. Aehle, W., Sobek, H., Amory, A., Vetter, R., Wilke, D., Schomburg, D., Rational protein engineering and industrial application: Structure prediction by homology and rational design of protein-variants with improved 'washing performance' - the alkaline protease from Bacillus alcalophilus. *Journal of biotechnology*, **28**(1), 31-40 (1993).
8. Cavasotto, C.N. & Phatak, S.S., Homology modeling in drug discovery: current trends and applications. *Drug discovery today*, **14**(13), 676-683 (2009).
9. Roy, A., Yang, J., and Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research,* W471-7 (2012).
10. Roche, D.B., Buenavista, M.T., & McGuffin, L.J., The FunFOLD2 server for the prediction of protein–ligand interactions. *Nucleic acids research,* **41**(W1), W303-W307 (2013).
11. Zhang, Y., Protein structure prediction: when is it useful? *Current opinion in structural biology*, **19**(2), 145-155 (2009).
12. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research,* **25**(17), 3389-3402 (1997).
13. Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, **292**(2), 195-202 (1999).
14. Boratyn, G.M. Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., & Madden, T.L., Domain enhanced lookup time accelerated BLAST. *Biol Direct*,7(1), 12 (2012).
15. Panchenko, A.R., Finding weak similarities between proteins by sequence profile

comparison. *Nucleic acids research*, **31**(2), 683-689 (2003).
16. Rychlewski L., Jaroszewski, L., Li, W., & Godzik, A., Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science*, **9**(2), 232-241 (2000).
17. Lobley, A., Sadowski, M.I., & Jones, D.T., pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, **25**(14), 1761-1767 (2009).
18. Söding, J., Protein homology detection by HMM–HMM comparison. *Bioinformatics*, **21**(7), 951-960 (2005).
19. Eddy, S.R., Accelerated profile HMM searches. *PLoS Computational Biology*, **7**(10). e1002195 (2011).
20. Remmert, M., Biegert, A., Hauser, A., & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, **9**(2), 173-175 (2012).
21. Madera, M., Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics*, **24**(22), 2630-2631 (2008).
22. Karplus, K., SAM-T08, HMM-based protein structure prediction. *Nucleic acids research*, **37**, W492–W497 (2009).
23. Xu. D., Jaroszewski, L., Li, Z., Godzik, A. FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics*, 30(5), 660-667 (2013).
24. Chakravarty, S. & Varadarajan R., Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure*, 7(7), 723-732 (1999).
25. Wu, S. & Zhang Y., MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, and Bioinformatics*, **72**(2), 547-556 (2008).
26. Peng, J. & Xu, J., RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics*, **79**(S10), 161-171 (2011).
27. Yang, Y., Faraggi, E., Zhao, H., & Zhou, Y., Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, **27**(15), 2076-2082 (2011).
28. Zhou, H. & Zhou, Y., Fold recognition by combining sequence profiles derived from evolution and from depth□dependent structural alignment of fragments. *Proteins: Structure, Function, and Bioinformatics*, **58**(2), 321-328 (2005).
29. Fernandez-Fuentes N., Rai B.K., Madrid-Aliste, C.J., Fajardo, J.E. & Fiser, A., Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinformatics*, **23**(19), 2558-2565 (2007).
30. Moult, J., A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, **15**(3), 285-289 (2005).
31. Floudas, C.A., Fung, H.K., McAllister, S.R., Mönnigmann, M. & Rajgaria, R., Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, **61**(3), 966-988 (2006).
32. Rychlewski, L. & Fischer D., LiveBench□8: The large□scale, continuous assessment of automated protein structure prediction. *Protein Science*, **14**(1), 240-245 (2005).
33. Wu, S. & Zhang Z., LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research*, **35**(10), 3375-3382 (2007).
34. Wang, Z., Eickholt J. & Cheng, J., MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics*, **26**(7),

882-888 (2010).
35. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A., Critical assessment of methods of protein structure prediction (CASP) - round x. *Proteins: Structure, Function, and Bioinformatics*, **82**(S2), 1-6 (2014).
36. Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A. & Bonvin, A.M.J.J. Assessment of contact predictions in CASP12: Co□evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*, **86**(S1), 51-66 (2018).
37. Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U. & Sali, A. Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics*, Chapter 5:Unit-5.6 (2006).
38. Rohl, C.A., Strauss, C.E., Misura, K.M. & Baker, D., Protein structure prediction using Rosetta. *Methods in enzymology*, **383**, 66-93 (2004).
39. Mulnaes, D. & H. Gohlke, TopScore: Using Deep Neural Networks and Large Diverse Data Sets for Accurate Protein Model Quality Assessment. *Journal of chemical theory and computation*, **14**(11), 6117-6126 (2018).
40. Zhang, Y. & J. Skolnick, J., TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research,* **33**(7), 2302-2309 (2005).
41. Li, J., Deng, X., Eickholt, J., Cheng, J. Designing and benchmarking the MULTICOM protein structure prediction system. *BMC structural biology*, 2013. **13**(1), 2 (2013).
42. Zhang, Y., I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*, **9**(1), 40 (2008).
43. Kryshtafovych, A., Barbato, A., Fidelis, K., Monastyrskyy, B., Schwede, T. & Tramontano, A., Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins: Structure, Function, and Bioinformatics*, **82**(S2) 112-126 (2014).
44. Wang, Q., Canutescu, A.A., & Dunbrack R.L.J., SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nature protocols*, **3**(12), 1832 (2008).
45. Miao, Z., Cao, Y. & Jiang, T., RASP: rapid modeling of protein side chain conformations. *Bioinformatics*, **27**(22), 3117-3122 (2011).
46. Xu, D. & Zhang, Y., Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical journal*, **101**(10), 2525-2534 (2011).
47. Bhattacharya, D., Nowotny, J., Cao, R. & Cheng J., 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic acids research*, **44**(W1), W406-W409 (2016).
48. Cheng, J., A multi-template combination algorithm for protein comparative modeling. *BMC structural biology*, **8**(1), 18 (2008).
49. Wallner, B., Larsson, P., & Elofsson, A., Pcons. net: protein structure prediction meta server. *Nucleic acids research*, 2007. **35**(2), W369-W374 (2007).
50. Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R. & Schwede, T., Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins: Structure, Function, and Bioinformatics*, **86**, 387-398 (2018).
51. Khosa, S., Frieg B., Mulnaes, D., Kleinschrodt, D., Hoeppner, A., Gohlke, H., & Smits, S.H.J, Structural basis of lantibiotic recognition by the nisin resistance protein from Streptococcus agalactiae. *Scientific reports*, **6**, 18679 (2016).
52. Ghent, A.W., A method for exact testing of 2X2, 2X3, 3X3, and other contingency tables, employing binomial coefficients. *American Midland Naturalist*, 15-27 (1972).
53. Zemla, A., Venclovas, C., Moult, J. & Fidelis, K., Processing and analysis of CASP3

protein structure predictions. *Proteins: Structure, Function, and Bioinformatics*, **7**(S3), 22-29 (1999).
54. Söding, J., Biegert, A., & Lupas, A.N., The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, **33**(2), W244-W248 (2005).
56. Mariani, V., Biasini, M., Barbato, A. & Schwede, T., lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**(21), 2722-2728 (2013).
57. McGuffin, L.J., Bryson, K., & Jones, D.T., The PSIPRED protein structure prediction server. *Bioinformatics*, **16**(4), 404-405 (2000).
58. Kabsch, W. & Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen□bonded and geometrical features. *Biopolymers*, **22**(12), 2577-2637(1983).
59. Zhang, W., Yang, J., He, B., Walker, S.E., Zhang, H., Govindarajoo, B., Virtanen, J., Xue, Z., Shen, H.B., Zhang, Y., Integration of QUARK and I□TASSER for Ab Initio Protein Structure Prediction in CASP11. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1), 76-86 (2016).
60. Schneidman-Duhovny, D., Hammel, M., Tainer, J.A., Sali, A., Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophysical Journal*, **105**(4), 962-974 (2013).
61. Schneidman-Duhovny, D., Hammel, M., Tainer, J.A., Sali, A., FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res*, **44**(W1), W424-429 (2016).
62. Svergun, D.I., Petoukhov, M.V. & Koch, M.H., Determination of domain structure of proteins from X-ray solution scattering. *Biophysical Journal*, **80**(6), 2946-2953 (2001).
63. Kozin, M.B. & Svergun D.I., Automated matching of high- and low-resolution structural models. *Journal of Applied Crystallography*, **34**, 33-41 (2001).
64. DeLano, W.L., PyMOL. *CCP4 Newsletter On Protein Crystallography* (2002).
65. Milić, D., Dick, M., Mulnaes, D., Pfleger, C., Kinnen, A., Gohlke, H., Groth, G., Recognition motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1. *Scientific reports*, **8**(1), 3890 (2018).
66. Zhang, Z., Gu, Q., Vasudevan, A.A.J., Hain, A., Kloke, B.P., Hasheminasab, S., Mulnaes D., Sato, K., Cichutek, K., Häussinger, D., Bravo, I.G., Smits, S.H.J., Gohlke, H., Münk, C., Determinants of FIV and HIV Vif sensitivity of feline APOBEC3 restriction factors. *Retrovirology*, **13**(1), 46 (2016).

TOC Graphic



Supporting Information

TopModel: Template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks

Daniel Mulnaes[1], Nicola Porta[1], Rebecca Clemens[2], Irina Apanasenko[3,4], Jens Reiners[2,5], Lothar Gremer[3,4], Philipp Neudecker[3,4], Sander Smits[2,5], Holger Gohlke[1,4,6]*

[1]Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[2]Institute für Biochemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany.

[3]Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[4]Institute for Complex Systems - Structural Biochemistry (ICS-6) & Jülich Center for Structural Biology (JuStruct), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[5]Center for Structural Studies Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[6]John von Neumann Institute for Computing (NIC) & Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

*Address: Universitätsstr. 1, 40225 Düsseldorf, Germany.

Phone: (+49) 211 81 13662; Fax: (+49) 211 81 13847

E-mail: gohlke@uni-duesseldorf.de

Table of Content

Table S1. Primary methods used in TopModel

Threading	Alignment	Quality Assessment[3]
DeltaBLAST [1]	TCOFFEE [2] [2]	PROCHECK [3]
HMMER3 [4]	MAFFT7 [5]	MolProbity [6]
HHBlits [7]	MergeAlign2 [8]	ANOLEA [9]
HHSearch[10]	SAlign [11]	ProSA2003 [12]
FFAS03 [13]	PROMALS3D [14]	DOPE [15]
SPARKSX [16]	FORMATT [17]	GOAP [18]
RAPTORX [19]	MUSTANG [20]	ModFOLDClust2 [21]
LOMETS[1] [22]	3DCOMB [23]	Pcons [24]
pGenThreader [25]		SPICKER [26]
pDomThreader [25]		QMEAN6 [27]
FASTA [28]		ProQ2 [29]
SAMT2K [30]		ProQ2D[31]
		ProQ3D[31]
		SVMQA[32]
		SELECTPRO [33]

[1] The LOMETS software includes the threading algorithms PPAS, wPPAS, dPPAS, wdPPAS, PPAS2, dPPAS2, Env-PPAS, MUSTER, and wMUSTER.

[2] The following programs are used within the TCOFFEE suite as the default methods for calculating alignments: MUSTANG, ClustalW [34], POA [35], MUSCLE [36], ProbA [37], PCMA [38], ProbCons [39], DiAlign [40], SAP [41], and TM-Align [42].

[3] The model quality assessment programs are used as part of the TopScore module, which is also used extensively during the TopThreader and TopRefiner work flows[43].

Supporting Text T1: Detailed Description of TopThreader

The aim of TopThreader is to select templates most suited for structure prediction based on their structural similarity to the native structure. This requires a measure of structural similarity, which can be measured when the native structure is known, but has to be predicted for *de novo* cases. For this purpose, TopThreader uses the Template Modeling Score (TM-Score) [42] between a template and the native structure. The TM-score measures structural similarity using Levitt-Gerstein weights to emphasize small distances between residues and is normalized to be independent of protein size. Two structures with a TM-Score > 0.45 are considered to have the same fold, though our experience is that this cut-off is too strict, thus, a cut-off of 0.4 is used in TopThreader. Two random structures have a TM-Score of ~0.17. TopThreader predicts the TM-Score iteratively, using different input features from the threaders that identified the template, from pairwise alignment between the template and the target sequence, from initial structural models based on these alignments, and from the ensemble of alignments and templates found for the given target.

TopThreader uses a combination of deep neural networks (DNNs), model quality estimation using TopScore and TopScoreSingle[43], and pairwise structure alignments to predict TM-Scores, remove false positive templates, correct pairwise target-template alignments, and rank

templates according to their predicted TM-Score. TopThreader has eight over-all steps outlined next. A detailed workflow for TopThreader is shown in Fig. S1.

1. **Primary threaders.** Individual threaders may fail to identify templates due to differences in database composition (usually due to database clustering), search algorithm, and significance cutoffs. TopThreader seeks to avoid this by using twenty different primary threading algorithms from twelve different threaders (Table S1) and selects all alignments for the top five template structures from each threader. The top five templates were selected as the default to limit the amount of potential false positives, based on the assumption that the best template is found in the top five for at least one of the primary threaders. All threaders are run with default settings as described by the authors and uses multiple cores when possible.
2. **Pre-filtering.** Pre-filtering allows the user to discard templates according to imposed cutoffs to sequence identity, e-value, coverage, experimental method, and submission date. This feature is particularly useful in benchmarking to eliminate closely related templates. By default, templates with less than 30% coverage and computationally designed proteins (i.e. non-natural proteins) are removed during pre-filtering.
3. **Alignment fitting.** Since primary threaders use differently clustered databases and output formats, their alignments may not fit the template or target sequences exactly. The differences are generally small and arise from partial matches and non-standard or missing template residues. To alleviate this problem, the target and template sequences are fitted to the alignments produced by the threading programs using MAFFT7[5]. This fitting is also used to transfer alignments from highly similar templates identified by different threading programs (see Redundancy clustering subsection).
4. **Score templates using DNNs.** Different threading programs have different scores for ranking the alignments they produce, ranging from e-values to Z-scores or energy measures (see the references of the primary threaders in Table S1). However, these scores are not comparable between programs. Thus a template is usually found by only a subset of threaders. To overcome this, TopThreader first calculates a number of alignment features from the pairwise threading alignments, including sequence identity, sequence similarity, and target coverage. These are used to predict the TM-score, which is combined with those same input features as input for DNN's trained to impute the scores of the threaders that did not identify the template. The real and imputed scores are then used as input for DNN's to predict the TM-Score given the threader scores. This eliminates missing values for individual threaders if those threaders did not identify

a specific template, and is necessary to keep the input feature vector size constant, as the neural network cannot process missing values when a threader does not identify a template. All scores are then used as input for a final DNN to predict the TM-Score based on all pairwise alignment and threading features of the given template (real or imputed in the case of missing values) denoted as the *Initial Score* ($R^2 = 0.69$, $p < 0.05$).

5. **Redundancy clustering.** Differences in primary threader databases can lead to highly similar templates being identified. Thus, TopThreader clusters templates at 90% pairwise sequence identity to remove redundancy. From each cluster, the template with the highest predicted *Initial Score* is kept, and pairwise alignments from other templates are transferred to it using the alignment fitting mentioned in step 3. If a template with a TM-Score lower than 0.9 to the centroid but over 90% sequence identity to the centroid is found, both the alternate conformation and the centroid are kept and alignments are transferred to both, allowing TopModel to sample alternate conformations.
6. **False positive removal.** Removal of false positives is critical for both ensuring correct fold recognition and preventing structural alignment between templates from failing. Therefore, templates are clustered with the Phyrestorm clustering tree [44] to remove bias towards folds with many templates. For each cluster, models are built for each template and scored with TopScoreSingle to select the best centroid. For each cluster, the *Centroid Score* is predicted with a DNN using as input features the TopScoreSingle score[43], cluster size, PSIPRED [45] secondary structure agreement, and all features from step 4. The *Centroid Score* shows a better agreement with the true TM-Score ($R^2 = 0.85$, $p < 0.05$) than the *Initial Score* due to the inclusion of secondary structure agreement, cluster size, model quality, and scores from all cluster members (see Fig. S1). However, it does not reflect the TM-Score of the individual cluster members but only the centroid. To remove false positives, templates are compared to the best scoring centroid, and discarded if they are structurally dissimilar (TM-Score < 0.4).
7. **Consensus.** TopThreader attempts to correct pairwise threading alignments by using residue-wise weights indicative of alignment quality and combining pairwise threading alignments with a structural alignment between templates. First, models are built from all primary threader alignments and scored with TopScore. Three local and global scores are then calculated: (A) The global and local TopScore, (B) the local and global lDDT score [46] of a model compared to the highest ranked model, and (C) the global and local sequence similarity relative to all residues aligned to a specific position in the target sequence. These three scores are scaled from 0 (worst) to 1 (best) and applied as

weights to target-template residue pair restraints calculated from the pairwise alignments. The weighted target-template restraints are then added to a list of template-template residue restraints calculated from a MUSTANG [20] alignment to include template-template structural similarity. Template-template restraints are given a weight of 3 to enforce the structural alignment between templates, letting the much lower target-template weights (ranging from 0 to 1) determine the alignment between the target sequence and the template structures. For each of the three scores (A, B, and C described above) the combined set of template-template and template-target restraints is then converted into a multiple alignment using TCOFFEE[2], yielding three multiple alignments each maximizing one of the three scores A (structural quality), B (structural consensus), and C (sequence similarity to target). These three multiple alignments are combined with MergeAlign2 [8] to calculate a consensus alignment. For each template, the consensus alignment is extracted from the consensus multiple alignment, modeled with TopBuilder, and scored with TopScore and TopScoreSingle.

8. **Ranking.** The final template ranking is based on the predicted TM-Score from a DNN with input features from steps 4, 6, and 7. This prediction is termed the *TopThreader Score* ($R^2 = 0.77$, $p < 0.05$) and is used for the final ranking of templates.

Figure S1. Detailed TopThreader workflow

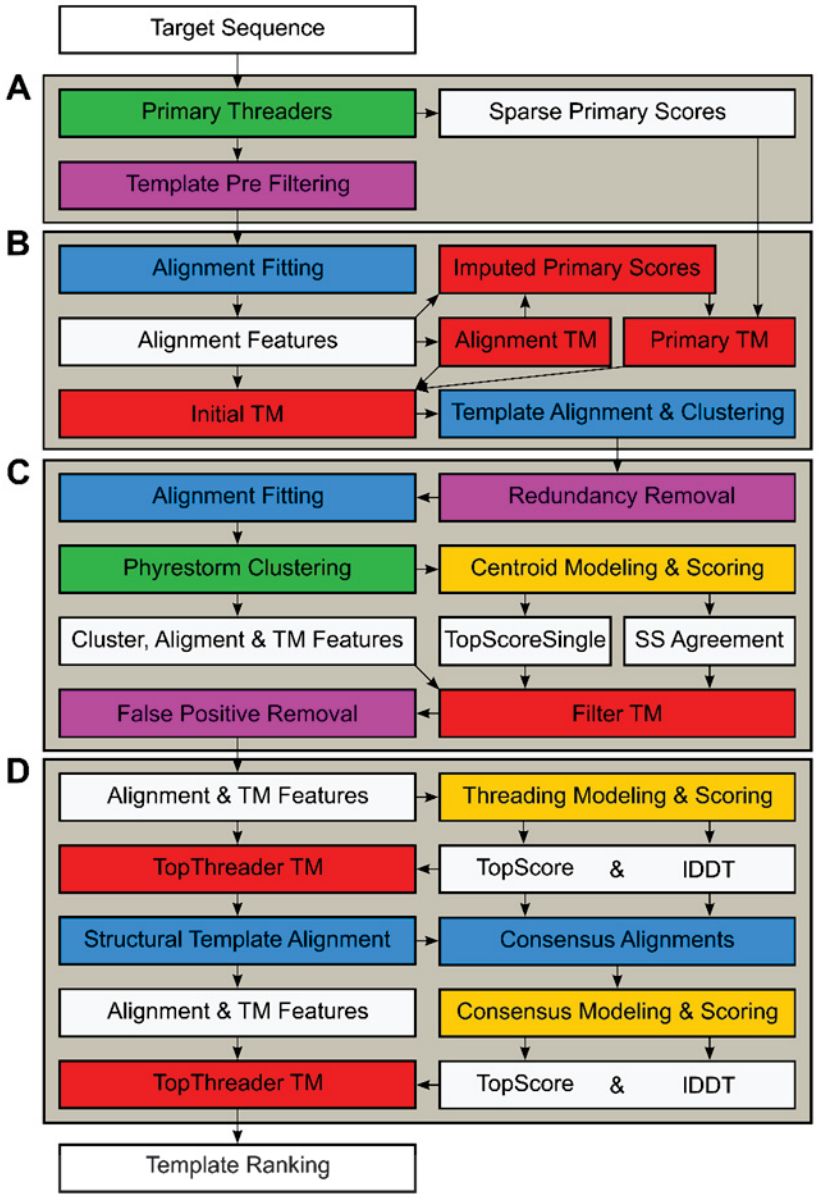


Figure S1. Detailed workflow of TopThreader. White boxes symbolize template features, calculated from threading programs, pairwise alignments or models of the pairwise alignments. **Red** boxes symbolize deep neural networks used to predict the TM-Score from the input features (white boxes). **Blue** boxes symbolize fitting of alignments or template-template alignment. **Yellow** boxes symbolize modeling of alignments and scoring of the models with TopScore and TopScoreSingle. **Magenta** boxes symbolize removal of templates because they are not meeting cut-offs or pre-filtering criteria, are structurally redundant, or are false positives. **Green** boxes symbolize threading with primary threaders and structural clustering of templates. **Step A:** The target sequence is threaded using primary threaders, the templates filtered, and the top five templates from each primary threader selected. **Step B:** The template alignments are fit to match the PDB structure and target sequence. Threading and alignment features are fed to DNNs to calculate the *Initial Score*. The templates are aligned and clustered. **Step C:** The redundant templates are removed and their alignments transferred to the cluster centroids. The non-redundant templates are clustered structurally, and the highest scoring centroid is modelled and scored. Then, DNNs predict the *Centroid Score*, and false positives are removed. **Step D:** Threading alignments from all templates are modeled and scored. DNNs then predict the TopThreader Score and rank models based on it for consensus calculation. Templates are then structurally aligned, and consensus alignments are calculated, modeled, and scored. Finally, all features are used to predict the *TopThreader Score* of each of the templates.

Training Dataset. The dataset used to train the TopThreader DNNs is a combination of the Top100 dataset [47] and the heterodimer docking cases in the ZDOCK5 dataset[48], using the receptor and the ligand chains as two separate entries. These datasets were chosen due to their high structural diversity and relatively limited size, because running TopThreader is computationally expensive. For each query, TopThreader was run with three different upper limits to the template-target pairwise sequence identity. These were chosen as 90%, 60% and 30% sequence identity to emulate trivial, easy, and difficult modeling cases, respectively. To increase the number of false positive templates, the top ten hits for each primary threading program was used instead of the default five. This was done to increase the DNNs performance at ranking false positives and the ability of TopThreader to discard them.

DNN training. All DNNs in TopThreader are built and trained in the same way, only differing in input features and architectures. The DNNs were trained using the python package SciKit-learn version 1.8.1 [49]. The data was first divided randomly into training and evaluation sets, leaving 80% of the data for training and 20% for final evaluation. The evaluation data was left out of the entire training procedure and was only used to evaluate the final DNNs. The IsolationForest method in SciKit-learn was used to remove the 1% most severe outliers from the training data to ensure a fit on the most representative data.

The DNNs were trained using the MLPRegressor method with the ADAM stochastic gradient descent algorithm [50] and default weight decay settings for L2 regularization to prevent overfitting. To estimate the meta-parameters of the DNNs, the training data was randomly sub-divided into training and test sets using the k-fold method in SciKit-learn to perform five-fold cross-validation. By using a grid-search, the DNN architecture and neuron type were varied. Architectures ranged from a single-hidden-layer perceptron to a three-hidden-layer perceptron with 10, 20, 40, 80, and 160 neurons in the first layer and subsequent layers having half the neurons of the previous layer. The tested transition functions were logistic, hyperbolic tangent, and rectified linear unit function. For each five-fold cross-validation split, the DNN was trained on 50% of the data and evaluated on the rest. To prevent over-training, the training was stopped early if the correlation between predicted and true TM-Score on the test half decreased. After selecting an optimal architecture and transition function, a DNN was trained on all the training data except for the outliers, again setting aside 50% of the data for testing and applying early stopping to prevent over-training. The final DNN performance was evaluated on the 20% of the data left out for evaluation at the beginning of the training. The performance of TopThreader training is shown in Figure S2.

Figure S2. TopThreader training performance

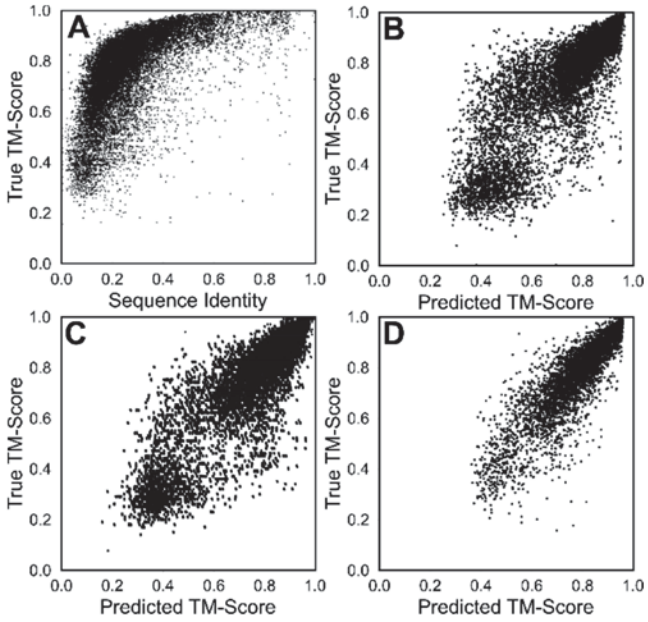


Figure S2. Distributions of templates found during the screening and used for training the TopThreader DNNs. A: Template sequence identity versus true TM-Score to native. **B:** Predicted versus true TM-Scores, predicted using only threading scores and alignment features (*Initial Score*). **C:** Predicted versus true TM-Scores of cluster centroids after Phyrestorm clustering, using input features from all cluster members, predicted model quality, and cluster size (*Centroid Score*). **D:** Predicted versus true TM-Score of remaining templates after false positive removal using all features (*TopThreader Score*).

Table S2. Targets in the CASP evaluation dataset

T0644	T0645	T0649	T0650	T0651	T0652	T0653	T0654	T0655	T0657
T0658	T0659	T0661	T0662	T0663	T0664	T0666	T0667	T0669	T0671
T0672	T0673	T0674	T0675	T0676	T0678	T0679	T0680	T0681	T0682
T0683	T0684	T0685	T0686	T0687	T0688	T0689	T0690	T0691	T0692
T0699	T0700	T0703	T0704	T0705	T0707	T0708	T0712	T0713	T0715
T0716	T0717	T0719	T0720	T0721	T0724	T0726	T0731	T0733	T0735
T0736	T0737	T0738	T0742	T0743	T0744	T0746	T0747	T0749	T0752
T0753	T0755	T0756	T0757	T0759	T0760	T0761	T0762	T0763	T0764
T0765	T0766	T0767	T0768	T0769	T0770	T0771	T0772	T0773	T0774
T0775	T0776	T0777	T0780	T0781	T0782	T0783	T0784	T0785	T0786
T0792	T0794	T0796	T0797	T0798	T0799	T0800	T0801	T0802	T0803
T0806	T0807	T0808	T0812	T0813	T0814	T0815	T0816	T0817	T0818
T0819	T0821	T0828	T0829	T0830	T0831	T0832	T0833	T0834	T0840
T0841	T0843	T0845	T0847	T0848	T0849	T0851	T0852	T0853	T0854
T0855	T0857	T0859	T0860	T0861	T0862	T0863	T0865	T0866	T0868
T0869	T0870	T0872	T0873	T0877	T0878	T0879	T0882	T0884	T0885
T0886	T0889	T0891	T0892	T0893	T0894	T0895	T0900	T0902	T0903
T0904	T0909	T0912	T0918	T0920	T0921	T0922	T0928	T0942	T0943
T0944	T0945	T0948							

Table S2. CASP Targets in the CASP evaluation dataset indicated with CASP ID

Table S3. Comparison of template selection by TopThreader and primary threaders

CASP subset	140 TBM targets			46 FM targets			186 All targets		
ΔTM_{100}	< 5	5-15	> 15	< 5	5-15	> 15	< 5	5-15	> 15
HMMER3	39%	11%	50%	7%	4%	89%	30%	10%	60%
FASTA	51%	11%	38%	15%	13%	72%	42%	11%	47%
SAMT2K	56%	11%	33%	26%	0%	74%	48%	8%	44%
DELTABLAST	52%	16%	32%	20%	6%	74%	44%	14%	42%
pDomThreader	35%	26%	39%	**48%**	**28%**	**24%**	38%	26%	36%
HHBlits	49%	9%	42%	11%	15%	74%	39%	11%	50%
HHSearch	76%	13%	11%	30%	15%	55%	65%	13%	22%
pGenThreader	69%	14%	19%	46%	35%	19%	63%	19%	18%
RAPTOR-X	72%	15%	13%	46%	17%	37%	66%	15%	19%
LOMETS	66%	24%	10%	37%	26%	37%	59%	24%	17%
SPARKS-X	**81%**	**8%**	**11%**	39%	31%	30%	**71%**	**13%**	**16%**
FFAS03	79%	14%	7%	41%	26%	33%	70%	17%	13%
TopThreader	**92%**	**4%**	**4%**	**56%**	**24%**	**20%**	**83%**	**9%**	**8%**

Comparison of template selection performance on the CASP dataset. Performance is evaluated based on the ΔTM_{100} score, which evaluates the difference between the best of the top five ranked templates of a given threader, and the best template found by any threader. The ΔTM_{100} score is calculated using the formula $\Delta TM_{100} = 100\% \cdot (\max[TM_{\text{all templates}}] - \max[TM_{\text{top5 templates}}])$. For each target, three categories are selected: (I) the best template is found ($\Delta TM_{100} < 5$), (II) an adequate template is found ($\Delta TM_{100} \in (5, 15)$), and (III) no adequate template is found ($\Delta TM_{100} > 15$). The values represent percentages of CASP dataset targets for template-based modeling (TBM) targets, template-free modeling (FM) targets, and all targets. For each column, the best primary threader and TopThreader are highlighted in bold. See Figure 2 in the main text for a bar plot representation.

Table S4. Pairwise comparison of TopThreader with primary threaders

Type	Threader	I / II	vs.	II / I	I / III	vs.	III / I	II / IIII	vs.	III / II
140 TBM targets	HMMER3 ****	10.0 %		0.0%	43.6 %		0.0 %	3.6 %		0.7 %
	FASTA ****	10.0 %		0.7 %	32.1 %		0.0 %	2.9 %		0.0 %
	SAMT2K ****	10.0 %		0.7 %	27.9 %		0.7 %	2.9 %		0.0 %
	DELTABLAST *****	15.7 %		0.0 %	24.3 %		0.0 %	3.6 %		0.6 %
	pDomThreader *****	20.0 %		0.0 %	37.9 %		0.7 %	0.7 %		2.1 %
	HHBlits *****	6.4 %		0.7 %	40.0 %		2.1 %	0.7 %		0.0 %
	HHSearch ****	11.4 %		2.1 %	7.9 %		1.4 %	0.7 %		0.0 %
	pGenThreader *****	10.7 %		0.7 %	14.3 %		0.7 %	1.4 %		0.7 %
	RAPTOR-X *****	12.9 %		0.7 %	8.6 %		0.7 %	1.4 %		0.0 %
	LOMETS *****	19.3 %		0.7 %	8.6 %		1.4 %	0.0 %		0.7 %
	SPARKS-X *****	6.4 %		1.4 %	6.4 %		0.7 %	1.4 %		0.0 %
	FFAS03 *****	10.0 %		0.7 %	5.7 %		1.4 %	0.7 %		1.4 %
46 FM targets	HMMER3 ***	4.3 %		0.0 %	45.7 %		0.0 %	23.9 %		0.0 %
	FASTA ***	10.9 %		0.0 %	32.6 %		2.2 %	21.7 %		0.0 %
	SAMT2K ****	0.0 %		0.0 %	30.4 %		0.0 %	23.9 %		0.0 %
	DELTABLAST ***	2.2 %		2.2 %	37.0 %		0.0 %	19.6 %		2.2 %
	pDomThreader ****	15.2 %		10.9 %	10.9 %		6.5 %	2.2 %		2.2 %
	HHBlits **	8.7 %		0.0 %	37.0 %		0.0 %	17.4 %		0.0 %
	HHSearch ***	10.9 %		4.3 %	23.9 %		4.3 %	15.2 %		0.0 %
	pGenThreader ***	17.4 %		10.9 %	8.7 %		4.3 %	4.3 %		8.7 %
	RAPTOR-X *****	6.5 %		10.9 %	15.2 %		0.0 %	4.3 %		2.2 %
	LOMETS *****	17.4 %		13.0 %	15.2 %		0.0 %	4.3 %		2.2 %
	SPARKS-X *****	17.4 %		8.7 %	10.9 %		2.2 %	4.3 %		2.2 %
	FFAS03 ****	13.0 %		8.7 %	15.2 %		4.3 %	4.3 %		2.2 %

Comparison of relative template selection performance on the CASP dataset. Performance is evaluated based on the ΔTM_{100} score, which evaluates the difference between the best of the top five ranked templates by a given threader, and the best template found by any threader. The ΔTM_{100} score is calculated using the formula $\Delta TM_{100} = 100 \cdot (\max[TM_{\text{all templates}}] - \max[TM_{\text{top5 templates}}])$. For each target, three categories are selected: (I) the best template is found ($\Delta TM_{100} < 5$), (II) an adequate template is found ($\Delta TM_{100} \in (5, 15)$), and (III) no adequate template is found ($\Delta TM_{100} > 15$). The values represent percentages of CASP dataset targets for template-based modeling (TBM) targets and template free modeling (FM) targets. Targets where both TopThreader and the primary threader are in the same category are not shown as they reflect no change. The first column shows the percent of targets where TopThreader is in category I but the primary threader is in category II, compared to when the opposite is the case. The other columns follow the same format but for different pairs of categories. Pairwise comparisons in which a primary predictor outperforms TopThreader are highlighted in gray. Significance is calculated by comparing contingency tables between each primary Threader and TopThreader with the Freeman-Halton exact test and indicated for each primary threader. In all cases, the difference between TopThreader and any primary threader is highly significant (**: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$).

CASP results. TopThreader was evaluated on the targets of our CASP dataset (Table S2) as described in the main paper. The results shown in Table S3 and S4 show that for CASP TBM targets, TopThreader outperforms all primary predictors for almost all categories significantly ($p < 0.01$). However, in a few cases a primary threader performs slightly better than TopThreader at selecting category II over category III templates. For the FM targets, two out of 36 pairwise comparisons shows a primary predictor being slightly better at selecting templates than TopThreader. Overall, TopThreader outperforms all primary threaders even for FM targets

despite these templates having true TM-Scores close to or below the 0.4 cutoff that TopThreader uses to distinguish between true and false positive templates.

To compare the ranking ability of TopThreader, for each of the CASP targets, the templates are ranked according to their true TM-Score to the native structure, and this ranking is compared to the ranking of templates by the individual threaders as well as TopThreader. For each ranking position, as well as for the best template in the top five ranked templates, the TM-Score of a template ranked by a threader is normalized to the template with that rank in the perfect ranking. For the TBM and FM targets, the mean of that normalized TM-Score are calculated, and the results are shown in Table S5.

Table S5. Template ranking of TopThreader compared to its primary threaders

Type		Best of top 5	1st ranked	2nd ranked	3rd ranked	4th ranked	5th ranked
140 TBM targets	HMMER3	0.61±0.10	0.58±0.07	0.53±0.08	0.33±0.08	0.18±0.07	0.04±0.03
	FASTA	0.73±0.09	0.70±0.07	0.55±0.08	0.33±0.08	0.15±0.06	0.01±0.01
	SAMT2K	0.70±0.10	0.66±0.06	0.61±0.08	0.56±0.08	0.54±0.09	0.53±0.09
	DELTABLAST	0.80±0.09	0.76±0.06	0.65±0.07	0.50±0.08	0.27±0.08	0.03±0.03
	pDomThreader	0.78±0.06	0.72±0.06	0.73±0.07	0.69±0.07	0.63±0.07	0.62±0.07
	HHBlits	0.56±0.09	0.53±0.07	0.50±0.07	0.46±0.08	0.38±0.08	0.25±0.07
	HHSearch	0.94±0.07	0.89±0.06	0.88±0.07	0.85±0.07	0.77±0.07	0.53±0.09
	pGenThreader	0.91±0.06	0.86±0.06	0.85±0.08	0.83±0.08	0.83±0.08	0.82±0.09
	RAPTOR-X	0.92±0.06	0.88±0.06	0.85±0.07	0.82±0.07	0.71±0.07	0.70±0.08
	LOMETS	0.91±0.06	0.86±0.06	0.76±0.07	0.64±0.07	0.45±0.08	0.21±0.07
	SPARKS-X	**0.95±0.06**	**0.91±0.06**	0.88±0.07	0.83±0.07	0.82±0.08	0.76±0.08
	FFAS03	0.95±0.06	0.90±0.06	**0.91±0.07**	**0.88±0.08**	**0.85±0.08**	**0.84±0.08**
	TopThreader	**0.97±0.05**	**0.94±0.06**	**0.93±0.06**	**0.88±0.07**	**0.87±0.07**	**0.87±0.08**
46 FM targets	HMMER3	0.20±0.14	0.20±0.13	0.10±0.10	0.11±0.11	0.05±0.08	0.00±0.00
	FASTA	0.44±0.18	0.43±0.16	0.33±0.18	0.13±0.13	0.07±0.10	0.00±0.00
	SAMT2K	0.37±0.22	0.30±0.16	0.33±0.19	0.32±0.18	0.36±0.21	0.35±0.21
	DELTABLAST	0.37±0.20	0.35±0.17	0.22±0.14	0.15±0.12	0.07±0.10	0.00±0.00
	pDomThreader	0.81±0.18	0.69±0.16	0.71±0.16	0.68±0.15	0.73±0.18	0.63±0.15
	HHBlits	0.37±0.18	0.35±0.15	0.31±0.15	0.26±0.16	0.25±0.17	0.19±0.16
	HHSearch	0.56±0.23	0.52±0.17	0.47±0.18	0.45±0.19	0.43±0.20	0.30±0.18
	pGenThreader	**0.85±0.18**	**0.76±0.17**	0.73±0.17	**0.81±0.18**	**0.82±0.20**	**0.67±0.19**
	RAPTOR-X	0.78±0.19	0.70±0.16	0.67±0.17	0.64±0.16	0.59±0.20	0.53±0.20
	LOMETS	0.80±0.19	0.70±0.16	0.65±0.16	0.60±0.17	0.39±0.17	0.19±0.16
	SPARKS-X	0.81±0.18	0.72±0.17	**0.76±0.18**	0.70±0.17	0.70±0.18	0.65±0.20
	FFAS03	0.78±0.19	0.69±0.16	0.73±0.17	0.64±0.15	0.61±0.18	0.56±0.20
	TopThreader	**0.87±0.20**	**0.83±0.19**	**0.69±0.18**	**0.60±0.19**	**0.55±0.21**	**0.53±0.22**

Mean normalized TM-Score, normalized to the TM-Score of a template at a given rank if the ranking was perfect according to TM-Score to the native structure. A value of 1 thus indicates that the average template with that rank is identical to the best possible template for that rank, if all identified templates were ranked according to true rather than predicted TM-Score. The 95% confidence interval is indicated after the ± sign. For each column, the best primary predictor values and TopThreader values are highlighted in bold.

Supporting Text T2: Primary Threader Description

PSIPRED [45] was developed by the Jones lab and predicts secondary structure from sequence profiles and PSSM's generated by PSI-BLAST using a three-state neural network.

DELTABLAST [1] was developed by Boratyn *et al.* and scans the conserved domain database to construct a PSSM that is used to increase sensitivity when running PSI-BLAST.

HMMER3 [4] was developed in the Eddy group and uses the Multiple Segment Viterbi algorithm to accelerate HMM-HMM comparison for HMM database alignment searches.

HHBLITS [7] was developed by the Söding group as an accelerated HHSEARCH. It constructs a HMM by adding context-specific pseudo-counts at each residue position and iteratively searches HMM databases using heuristic filters to remove false positives.

HHSEARCH [10] was developed by the Söding group. It uses HHBLITS to generate a target HMM from which secondary structure is predicted using PSIPRED and combined with the HMM, which is then used to search a template database by matching both sequence and secondary structure terms.

RAPTOR-X [19] was developed by the Xu group and uses a regression tree-based non-linear alignment scoring-function. It measures the profile information, based on which gap-penalties and sequence similarity are derived, from profile- and context-specific features including predicted secondary structure, solvent accessibility, amino acid identity, and residue hydropathy.

SPARKSX [16] was developed by the Zhou group and uses predictions of secondary structure, solvent accessibility, and dihedral angles from SPINE-X [51] combined with HMM-predicted probability of the predicted values on a residue-wise level. These scores are combined with PSSM's and profile-profile comparisons to increase alignment accuracy.

FFAS03 [13] was developed by the Godzik group and combines normalized PSSM's derived from PSI-BLAST profiles with predicted secondary structure, solvent accessibility, and residue depth. It uses a weighted dynamic programming algorithm from which alignment scores are calibrated by length and ranked according to a neural network-predicted MaxSub score.

LOMETS [22] was developed by the Zhang group and includes eight versions of the MUSTER algorithm with differently optimized scoring terms. It uses weighted dynamic programming with differently weighted features from closely and distantly homologous profiles as well as predicted secondary structure, solvent accessibility, backbone dihedral angles, hydropathy scoring matrices, and depth-dependent structure profiles to generate alignments.

pGenThreader and **pDomThreader** [25] were developed by the Jones Lab. pGenThreader uses profile-profile alignment based on PSI-BLAST PSSMs combined with secondary

structure-specific gap penalties, pairwise residue potentials, and hydrophobic burial scores. Weights for the individual terms were optimized using support vector machine regression to optimize template-target TM-Score. pDomThreader was trained on the same input but using support vector classification rather than regression to provide a clearer distinction between protein super-families.

SAMT2K [30] was developed by the Karplus group. It uses HMMs generated from profiles from BLAST searches against the non-redundant sequence database to detect templates and performs both forward and reverse alignment of the query sequence to the HMM to improve confidence in the ranking of hits.

FASTA [28] was developed by Pearson as one of the first rapid sequence-sequence comparison methods. It uses word search heuristics to identify identical fragments and joins the ten largest regions based on their word matches. These regions are then joined using a penalty analogous to a gap penalty. Sequences that score well according to this rapid heuristic are aligned using standard dynamic programming.

Supporting Text T3: Detailed Description of TopAligner

The templates identified by TopThreader should belong to the same fold, but may occasionally differ too much to be accurately aligned by primary alignment programs. To prevent templates with a different fold from being included in the multiple alignment, TopAligner clusters the templates and uses only templates that share the same structural fold (TM-Score > 0.5) as the majority of the top five-ranked templates for multiple alignment. TopThreader then calculates an ensemble of different pairwise and multiple alignments using eight different primary methods for constructing multiple alignments (see Table S1). Aside from the alignments produced by the primary MSA programs, TopAligner uses the MSA generated by TopThreader, as well as all pairwise threading and consensus alignments.

After building a multiple alignment of the templates from a specific primary alignment program, TopAligner uses the local and global scores calculated by TopThreader to weight the pairwise primary threading and consensus alignments calculated by TopThreader. These weighted pairwise alignments are used to add the target sequence to the multiple template alignment using TCOFFEE (see also TopThreader step 7). From the multiple template-target alignment calculated by the primary alignment program, every possible combination of the top five-ranked templates are then extracted and added to the alignment ensemble. Then the procedure is repeated for the next primary alignment program. Finally, the pairwise threading and consensus alignments of the top five-ranked templates are added to the alignment ensemble.

The primary alignment programs used by TopAligner can be found in Table S1 and a detailed description of them in the following section. They vary in methodologies but can be broadly divided into one or more of the following categories:

1. **Horizontal-first methods.** These methods [2, 5, 8, 11, 14, 20] progressively merge pairwise alignments using a guide-tree in the order of pairwise sequence similarity, which is fast but may introduce alignment errors and requires iterative refinement.
2. **Vertical-first methods.** These methods [17, 23] identify similar fragment blocks across all templates and expand the alignment between blocks to generate a full MSA, which is more costly but seeks to eliminate pairwise alignment errors.
3. **Structure-improved sequence alignment.** These methods [5, 11, 14] use 3D- and/or secondary structure information to improve sequence alignment methods, for example, by changing gap penalties in secondary structure elements.
4. **Sequence-improved structure alignment.** These methods [17, 23] use sequence information to improve structural alignment by using regional alignment of flexible parts of the structure with sequence- rather than structure-based methods.

Supporting Text T4: Primary Aligner Description

TM-Align [42] was designed by the Zhang group. It performs pairwise structure alignments using rigid-body superposition by iteratively optimizing the TM-score. The TM-score measures structural similarity using Levitt-Gerstein weights to emphasize small distances between residues and is normalized to be independent of protein size.

MAFFT7 [5] was designed by the Katoh group. It represents residues as vectors of polarity and volume and uses fast Fourier transformations to calculate pairwise alignments, which are combined progressively using a guide-tree and iteratively refined using tree-dependent restricted partitioning.

MergeAlign2 [8] was designed by the Kelly group to make a consensus MSA from an ensemble of MSA's. It represents the ensemble as a directed acyclic graph, weights nodes by their prevalence in the ensemble, and uses dynamic programming to find the highest-weighted path corresponding to the consensus. In the context of calculating MSA's, MergeAlign2 uses 91 different substitution matrices with MAFFT7 to generate alignments from which a consensus alignment is constructed.

PROMALS3D [14] is a horizontal-first multiple sequence/structure alignment tool developed by the Grishin group. It combines PSI-BLAST homologue detection, with PSSM's, SS-prediction, sequence clustering, profile-profile alignment, and constraints from TM-Align [42]

and FAST [52].

SALIGN [11] is a horizontal-first multiple sequence/structure alignment tool developed by the Sali group. It uses a tree constructed from all combinations of pairwise alignments and serially aligns closest branches to each other using sequence and structure features with heuristic weights and context specific gap penalties.

MUSTANG [20] is a horizontal-first multiple structure alignment software developed by the Lesk group. It calculates the largest structural fragments that can be rigidly superimposed to generate pairwise alignments that consider flexibility without the need for gap-penalties. These are progressively combined using a neighbor-joining guide tree into a full MSA.

TCOFFEE [2] is a horizontal-first multiple sequence/structure alignment meta-tool developed by the Notredame group. It combines libraries of residue matches from pairwise alignments from multiple structure and sequence alignment methods. From this library, a PSSM and a guide-tree are calculated from which a MSA is calculated using dynamic programming. In TopModel, the default alignment methods for TCOFFEE is the 3DCOFFEE mode, consisting of MUSTANG [20], ClustalW [34], POA [35], MUSCLE [36], ProbA [37], PCMA [38], ProbCons [39], DiAlign [40], SAP [41], and TM-Align [42] as input aligners.

3DCOMB [23] is a vertical-first multiple structure alignment tool developed by the Xu group. It combines local and global structure features with Conditional Random Field probabilistic modeling to identify highly similar fragment blocks in all templates and uses these as anchors from which the alignment is extended.

FORMATT [17] is a vertical-first multiple structural alignment method developed by the Cowen group. It uses MATT [53] to identify and align highly similar fragment blocks and aligns residues that do not belong to a block using default settings of MAFFT7. In TopModel, these regions are aligned using MergeAlign2 as described above, rather than MAFFT7.

Supporting Text T5: Detailed Description of TopBuilder

TopBuilder functions as a front and back end to Modeller 9 [54]. Template-covered residues are built with the default Modeller procedure. Folding of loops without template have by default constraints, based on secondary structure predictions from PSIPRED[45], imposed to guide folding. Loop refinement is based on loop size, where tiny loops (1-2 residues) are not refined, medium loops (3-15 residues) are refined using the DOPE potential [15], large loops (16-25) are refined with Modeller's geometric potential, and massive loops (> 25 residues) are skipped because convergence of loops of this size is unlikely.

TopBuilder uses a knot detection algorithm inspired by Khatib *et al.* [55] to identify

knots in the models. If knots are identified, template-based restraints are removed for the knotted region, and the whole model is then re-built up to five times with different random seeds to generate a model without knots. Should a modeling fail to converge (usually due to large regions not covered by a template), it is restarted up to three times with different random seeds. After model construction, additional refinement is done using either SCWRL5[56], RASP[57], i3Drefine[58], or Modrefiner[59], with RASP being the default option.

Supporting Text T6: Detailed Description of TopRefiner

The goal of TopRefiner is to select the best models from previous modeling steps, combine and refine these models, and present a single high-quality model to the user. This is accomplished in a four-step workflow which is outlined below and in Figure S3:

1. Model selection. For TopAligner there are two possible scenarios: Either the alignment and model quality improves when using multiple templates, or it deteriorates. If the former is the case, models converge on the same fold, and TopScore and TopScoreSingle scores are highly correlated. In the latter case, TopScore performance at selecting the best model declines because the clustering methods bias TopScore towards multi-template models, whereas TopScoreSingle performance remains unaffected. Therefore, the correlation of TopScore and TopScoreSingle determines which of the two is used for model selection, a selection scheme denoted as TopScoreMix. For each template combination from TopAligner's multi-template model ensemble, the best-scoring model is selected according to this scheme. Additionally, the top five single-template models according to TopScore and TopScoreSingle are selected from the single-template and consensus steps of TopThreader. The selected models are scored, and each model is compared to the highest ranked model using TM-Align[42]. Outlier models with a TM-Score < 0.5 are removed as they are too different from the best model to be used for the fragment combination and model hybridization.

2. Model fragment recombination. The models from step 1 are fragmented by removing bad regions predicted to contain local errors according to TopScore and TopScoreSingle. Three rules are used to select whether a given residue is good or bad: (I) Residues with an error larger than 0.7 are bad (0.6 if the residue is in an unstructured terminus). (II) Residues with a local error below 0.4 are good if they have non-loop secondary structure. (III) Residues are good if their local error is within 0.5 median absolute deviations of the best scoring residue at this position across all models. Limits are imposed to ensure that no deleted region is smaller than 4 residues, and no fragment smaller than 7 residues is kept. The fragmented models with bad regions removed are then used as input templates for the ROSETTA [60] template-based hybridization

protocol ROSETTACM. Fifty models are generated of which the top 20 are selected according to ROSETTA score. If all ROSETTA models are very different from the best model from stage 1 (TM score < 0.5), the model fragments are instead used as input templates for TopBuilder to produce a set of 20 models. Models that are not consistent with the best model from step 1 (TM score < 0.5) are removed. The remaining models are scored with TopScore and TopScoreSingle. This step combines good regions from different models while letting ROSETTA or TopBuilder reconstruct regions that are predicted to be bad in all models.

3. Template hybridization. The top five best models from step 1 and 2 according to TopScoreMix are selected for model/template hybridization. These models are structurally aligned to each template identified by TopThreader with a TM-Score > 0.5 to the model using MUSTANG[20]. This produces new pairwise alignments with the highest structural agreement between good scoring models and the templates. These pairwise alignments are used with the templates as input for the ROSETTACM protocol. Fifty models are generated of which the top 20 are selected according to ROSETTA score, and scored with TopScore and TopScoreSingle. This step allows for information from all true templates to be included, not just the top five, and allows for new alignments between good structural models and the templates to be used, instead of threading-based sequence-structure alignments.

4. Model selection and final refinement. The best five models from steps 1, 2, and 3 according to TopScoreMix are selected, and outliers are filtered as in step 1 (TM-score > 0.5). From these models bad regions are removed as in step 2 and the fragmented models used as input templates for TopBuilder to yield 10 average models which are added to the pool of input models. All models are then refined with Modrefiner [59], which uses fragment-guided MD to refine the models. Finally, the models are scored with TopScore. The best model according to TopScore from the largest cluster according to SPICKER [26] is selected as the final model and returned to the user. This step lets the best models from previous steps be averaged and refined before final scoring and model selection. TopScore is used rather than TopScoreSingle since the final model ensemble is converged due to the filtering of outliers and averaging of structures.

Figure S3. Detailed TopRefiner workflow

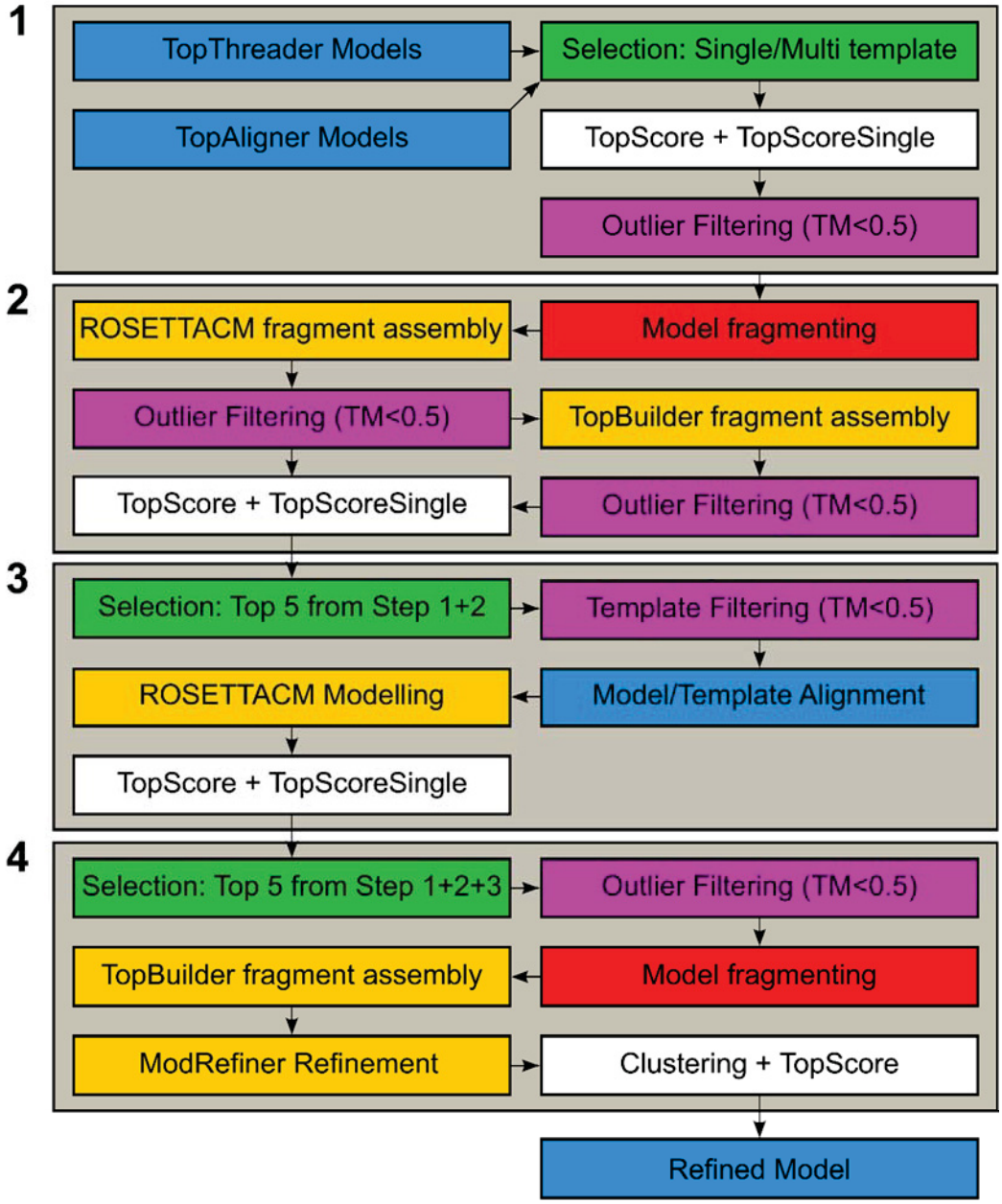


Figure S3. Detailed workflow of TopRefiner. White boxes symbolize model scoring, calculated using TopScore and TopScoreSingle. **Red** boxes symbolize model fragmenting, in which regions of the models containing errors according to TopScore and TopScoreSingle are deleted resulting in model fragments. **Blue** boxes symbolize input or output models or the structural alignment between good scoring models and input templates. **Yellow** boxes symbolize assembly of fragments into improved models, modeling of new alignments to the templates, or refinement of the models with ModRefiner. **Magenta** boxes symbolize removal of models or templates because they have a TM-Score less than 0.5 compared to the best scoring model. **Green** boxes symbolize selection of models from previous steps using TopScore and TopScoreSingle. **Step 1:** An initial model ensemble is selected from the pairwise single-template ensemble from TopThreader and the multi-template ensemble from TopAligner. **Step 2:** The selected models are fragmented by deleting regions predicted by TopScore or TopScoreSingle to contain errors and the resulting fragments are used as templates for template-based modeling with ROSETTACM, if all resulting models don't pass the outlier filter, the fragment assembly is carried out by TopBuilder instead. The resulting models are filtered for outliers and scored with TopScore and TopScoreSingle. **Step 3:** The top 5 best models from steps 1 and 2 are selected and structurally aligned to all identified templates that has a TM-Score larger than 0.5 to the model. These alignments are modeled with ROSETTACM and scored with TopScore and TopScoreSingle. **Step 4:** The top 5 best models from steps 1, 2 and 3 are selected and outliers are removed. The remaining models are fragmented and TopBuilder is used for fragment assembly to generate improved models. All models are then refined with ModRefiner and scored with TopScore. The best model according to TopScore is selected from the largest cluster according to SPICKER.

Supporting Text T7: Validation of LipoP from *C. difficile*

Experimental SAXS data collection

At the EMBL-Lab outstation Grenoble, the lipoprotein LipoP (UniProt ID Q18BL3) was freshly purified with a Superdex 200 10/300 column (GE Healthcare) pre-equilibrated with SAXS-buffer (25 mM MES pH 6.5, 250 mM NaCl, 5 % glycerol) at a flow rate of 0.5 mL/min.

SAXS data for the full length LipoP was collected on beamline BM29 at the ESRF Grenoble [61, 62], equipped with a PILATUS 1M detector (Dectris) with a fixed distance of 2.867 m. The achievable q-range under these conditions was 0.025 - 5 nm^{-1}, and the maximum measurable radius of gyration (R_g) of the investigated particles was 20 nm. All measurements were performed at 4°C with protein concentrations between 0.51 and 7.68 mg/mL. For each sample, ten frames with an exposer time of one second were collected. By comparing these frames, we excluded the possibility of radiation damage during the measurement.

SAXS data for the truncated LipoP was collected on beamline P12, PETRA III at the DESY Hamburg[63], equipped with a PILATUS 6M detector (Dectris) with a fixed distance of 3.0 m. The achievable q-range under these conditions was 0.02 - 6 nm^{-1}. All measurements were performed at 4°C with protein concentrations between 0.9 and 8.8 mg/mL. We collected frames with an exposer time of 0.045 seconds.

All used programs for data processing were part of the ATSAS Software package (Version 2.8.1) [64]. The primary data reduction was performed with the program PRIMUS [65]. With the Guinier approximation [66] (implemented in PRIMUS [65]), we determined the forward scattering *I(0)* and R_g. We estimated the maximum particle dimension (D_{max}) with the pair-distribution function $p(r)$, computed with the program GNOM [67]. Low resolution *ab initio* density models were calculated with GASBOR [68]. Superimposing of the predicted model into the SAXS density was done with the program SUPCOMB [69]. We used a reference solution of bovine serum albumin (66 kDa) to determine the molecular weight of the protein from the forward scattering. A summary of the SAXS data collection can be found in Table S6.

Experimental NMR data collection for secondary structure determination

The NMR samples contained 0.63 mM [U-^{15}N] or 0.63 mM [U-^{13}C,^{15}N] $(His)_{10}$-LipoP, 100 mM NaCl, 5 mM NaN_3, 25 mM MES (pH 6.5) in 10% (v/v) D_2O. NMR experiments were recorded at 30.0°C on Bruker AVANCE III HD 600 MHz, Bruker AVANCE III HD 700 MHz, Varian VNMRS 800 MHz, or Varian VNMRS 900 MHz NMR spectrometers equipped with room temperature (900 MHz) or cryogenically cooled (600 MHz, 700 MHz, 800 MHz) triple or quadruple resonance probes with *z*-axis pulsed field gradient capabilities. The sample

temperature was calibrated using methanol-d_4[70]. Sequence-specific assignments for the backbone resonances were obtained from TROSY [71-73] versions of the following 2D and 3D triple-resonance experiments [74, 75]: [^{1}H,^{15}N] TROSY [73] [^{1}H,^{13}C] CT-HSQC[76], [^{1}H,^{15}N] TOCSY-HSQC [77] with a 10.0 kHz DIPSI-2rc mixing scheme ([78]; 60 ms mixing time), [^{1}H,^{15}N] NOESY-TROSY ([77]; 120 ms mixing time), TROSY-HNCO [72], TROSY-HN(CO)CA [79], TROSY-HN(CO)CACB [79], TROSY-HNCA [79], TROSY-HN(CA)CO [79], TROSY-HNCACB [79], and H(CCO)NH-TROSY and C(CO)NH-TROSY [80] with a 16.7 kHz FLOPSY-16 mixing scheme ([81]; 14 ms mixing time). The ^1H_2O resonance was suppressed by gradient coherence selection, with quadrature detection in the indirect dimensions achieved by States-TPPI [82] and the echo-antiecho method [83, 84]. All NMR spectra were processed with NMRPipe [76] software and analyzed with NMRViewJ [85]. ^{1}H chemical shifts were referenced with respect to external DSS in D_2O; ^{13}C and ^{15}N chemical shifts were referenced indirectly [86]. ^{1}HN and ^{15}N amide group chemical shifts were obtained from the peak positions of the TROSY multiplet components by subtracting out the scalar coupling contribution of $-|^1J_{NH}|/2$ and $+|^1J_{NH}|/2$, respectively, assuming a uniform scalar coupling constant of $^1J_{NH}$ = -93 Hz. Random Coil Index [87] backbone order parameters, S_{RCI}^2, and confidence levels for helical (H) or extended/strand (E) secondary structure, P_H or P_E, respectively, were calculated from the backbone chemical shifts using TALOS-N[88] with the default parameters.

Table S6: Summary of experimental SAXS data collection for LipoP

Data collection parameters	Full length LipoP	Truncated LipoP
Beamline	BM29, ESRF Grenoble [61, 62]	P12, PETRA III, DESY Hamburg[63]
Detector	PILATUS 1 M	PILATUS 6 M
Detector distance (m)	2.867	3.0
Beam size (µm x µm)	700 x 700	12 x 200
Wavelength (Å)	0.99	1.24
Sample environment	Quartz glass capillary, 1 mm ø	Quartz glass capillary, 1 mm ø
s range (nm^{-1})‡	0.025–5.0	0.02-6.0
Temperature (K)	277	277
Exposure time per frame (s)	1s	0.45s
Mode of measurement	Static	Static
Protein concentration range (mg/ml)	0.51 – 7.68	0.9-8.8
Structural parameters		
I(0) from P(r)	14.31	0.013
R_g (real-space from P(r)) (nm)	2.54	2.49
$I(0)$ from Guinier fit	14.25	0.013
s-range for Guinier fit (nm^{-1})	0.154 – 0.453	0.27-0.53
R_g (from Guinier fit) (nm)	2.43	2.45
D_{max} (nm)	10.14	8.13
POROD volume estimate (nm^3)	35.12	35.70
Molecular mass from POROD volume (kDa)	20.66	21.00
Molecular mass from I(0) (kDa)	14.28	18.0
Molecular mass from sequence (kDa)	18.63	17.4
Software		
Primary data reduction	PRIMUS [65]	
Data processing	GNOM [67]	
Ab initio modelling	GASBOR [68]	
Superimposing	SUPCOMB [69]	
Model visualization	PyMOL [89]	

‡$s = 4\pi \sin(\theta)/\lambda$, 2θ – scattering angle, λ – X-ray wavelength

Figure S4. Top ranked templates and TopModel model of LipoP from *C. difficile*

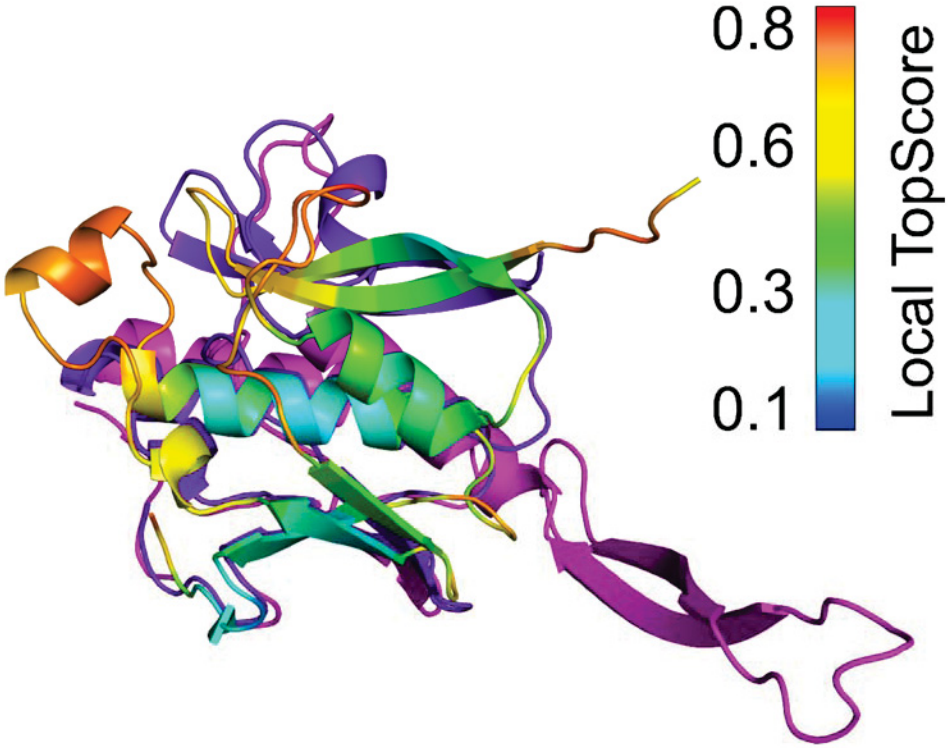


Figure S4. Top ranked templates and TopModel model of LipoP from *C. difficile*. The top ranked template with PDB ID 5J7R is shown in magenta. The protruding β-sheet makes contacts to other monomers in the tetrameric structure. The second highest-ranked template with PDB ID 6GZ8 is shown in purple, with the β-sheet folded up against the core of the protein. TopModel favors the positioning of the β-sheet up against the core of the protein. The model from TopModel is colored according to residue-wise TopScore. Yellow/Red regions indicate regions with high residue-wise error (> 50%). The disordered tail is not shown before residue 43.

The intrinsically disordered tail region of LipoP from *C. difficile*

In the starting model from TopModel, residue 43 separates the unstructured N-terminal (denoted as the tail region) from the rest of the protein (denoted as the folded domain). In MD simulations, the tail region has a much higher mobility than the folded domain (Figure S5). In the tail region, only transient secondary structure elements with low consistency between individual replicas are observed (Figure S5). Furthermore, the enrichment of disorder-promoting residues (e.g., K and S) [90, 91] support the hypothesis that the tail region is intrinsically disordered. The ability of disordered regions to bind, and exert a function, is mainly attributed to segments called molecular recognition features (MoRFs), which undergo a disorder-to-order transition upon binding [92, 93]. Residues 8-SISAVELV-15 are highly likely (53%-78%) a MoRF region as predicted by MoRFPred [94].

Molecular dynamics simulations

To increase agreement between the LipoP model from TopModel and the experimental data, we subjected the model to all-atom molecular dynamics (MD) simulations in explicit solvent.

Because we have experimental data from both NMR and SAXS, we simulated both the full model and a truncated version of it, without the unstructured tail. In the first case, with the aim of finding a tail conformation that results in a radius of gyration (R_g) in agreement with SAXS, and in the latter, in order to select snapshots of the folded domain that agree best with secondary structure determined from NMR. Both models were prepared for MD simulations with LEaP [95]. Sodium counter ions were added to establish charge neutrality, and each system was placed in a truncated octahedral box of TIP3P water [96] with a minimum distance between the solute and the border of the box of 11 Å.

Structural relaxation, thermalization, and production runs of MD simulations were conducted with pmemd.cuda [97] of Amber 16 [98] using the force field ff14SB [99] for the protein and Joung-Chetham parameters for ions, as reported previously [100]. Ten independent replicas of 30 ns were performed for both systems, with an aggregate simulation time of 2×300 ns. To setup independent replicas, the temperature was set to ten slightly different values in the thermalization (between 299.5 K and 300.4 K, offset 0.1 K), resulting in diverse starting structures for MD production runs, which were then performed in the NVT ensemble at 300 K.

The MD trajectories were analyzed with cpptraj [101]. To remove global translational and rotational motions, snapshots were extracted every nanosecond and fit on the folded domain using the first frame as reference. To measure protein compactness, we calculated R_g, defined as the root mean square distance of the collection of atoms from their common center of gravity. Secondary structure was analyzed using DSSP[102]. The eight DSSP secondary structure classes were reduced to three classes using the scheme: I/H/G→H, B/E→E, S/T/C→C to allow for comparison with NMR data. To summarize the secondary structure information of each snapshot from the MD simulation, collected with 1 ns in between snapshots, for each residue was assigned to a class H, E, or C if the propensity of the secondary structure was > 0.5. The secondary structure propensity was calculated as the fraction of snapshots exhibiting a given secondary structure class (H, E or C) across all replicate simulations for all systems. The results of this analysis is shown in Figure S5.

Figure S5. Secondary structure propensity across all MD simulations of LipoP

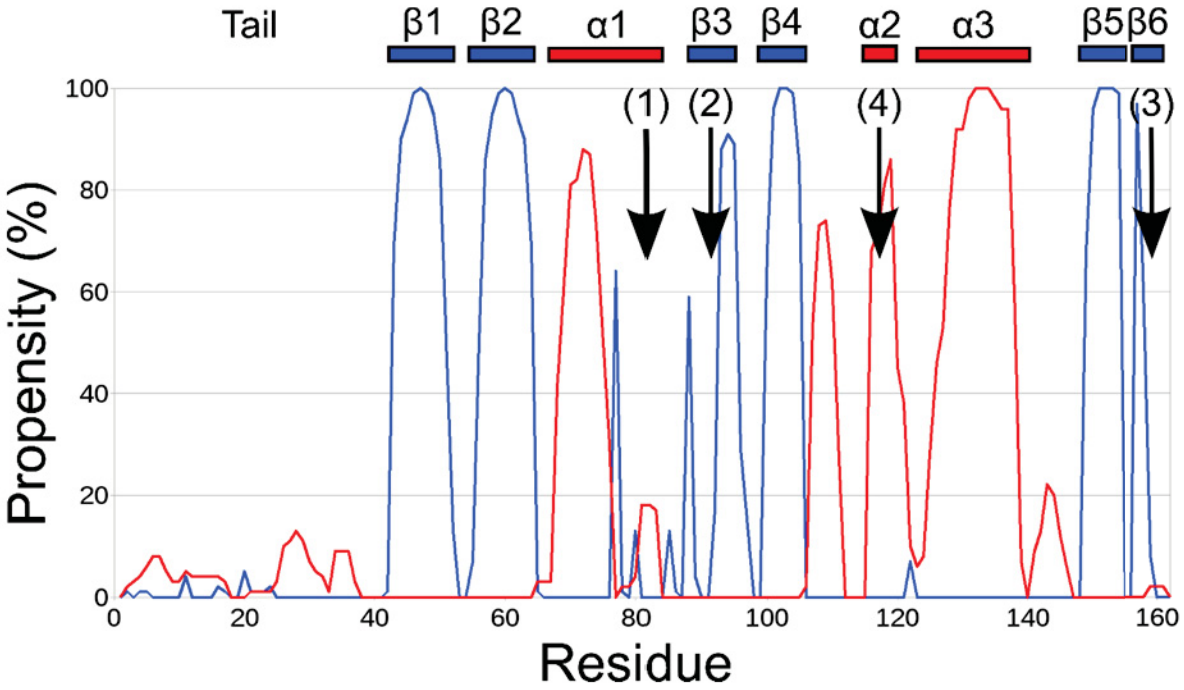


Figure S5. Secondary structure propensity across all MD simulations of LipoP. The average propensity for each residue to be in a helix (red) or β-strand (blue) is shown. The disordered tail barely shows secondary structure content apart from sparse helix formation. Secondary structure assignment according to NMR data is shown on top. Four differences between NMR secondary structure assignment and secondary structure propensity from MD simulations are visible, and numbered according to importance as discussed in detail in the main text: (1) α-helix 1 is longer according to NMR than found in the initial model, but shows an extension in some of the snapshots of the simulations; (2) β-strand 3 is shifted by two residues towards the C terminus in the initial model, which causes some strand formation earlier in the sequence and makes β-strand 3 less stable; (3) β-strand 6 is 3 residues too short in the initial model compared to NMR; (4) α-helix 2 is one residues shorter at the C-terminal end.

Model selection and combination

To build a model from the MD simulations-generated ensembles that agrees best with the experimental NMR-based secondary structure assignment, the Matthews correlation coefficient (MCC) was calculated for each secondary structure type and normalized to a standard Z-score using the mean and standard deviation across all snapshots of all replica simulations, with or without tail. These distributions are shown in Figure S6. As expected, during the MD simulations, the agreement with the NMR-determined secondary structure mostly deteriorates due to atomic fluctuations, however, some models show a better agreement with experimental secondary structure, indicating that these models are closer to a more native-like structure. The top ten snapshots from the MD trajectories in terms of average MCC for helix, strand, and coil were then selected from the non-tail residues of all MD simulations. These snapshots were subsequently energy minimized, using the same procedure applied prior to MD simulations, consisting of three steps. First, harmonic restraints with a force constant of 5 kcal·mol^{-1}·Å^{-2} were applied to all protein atoms (500 cycles steepest descent (SD) and 2000 cycles conjugate

gradient (CG) minimization). Second, we reduced the harmonic restraints and applied a force constant of 1 kcal·mol^{-1}·Å^{-2} (2000 cycles SD and 8000 cycles CG minimization). Finally, the positional restraints were removed completely, and all atoms were free to move (1000 cycles SD and 4000 cycles CG minimization). After minimization, the model with the highest average MCC for helix, strand, and coil was selected as the model for the folded domain.

From the MD simulations with the unstructured tail, all models with a radius of gyration within ±1 Å of the experimentally determined value from SAXS were selected. Of these models, the one with the highest agreement with NMR-determined secondary structure was selected according to average MCC for helix, strand, and coil. Finally, the best model of the folded domain was combined with the best model with the disordered tail by using TopBuilder with both models as templates for building the full target sequence. Subsequently the combined model was energy minimized as the final MD-refined model.

Figure S6. Distributions of standard z-scores of secondary structure agreement for LipoP.

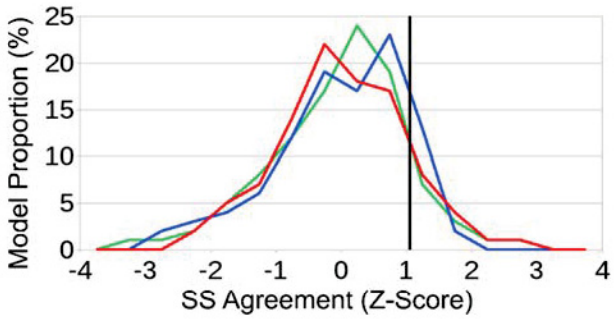


Figure S6. Distributions of standard z-scores of secondary structure agreement for LipoP. The black line indicated where the initial model from TopModel is located relative to the distribution from MD. Distributions of standard z-scores for α-helix (red), β-sheet (blue) and random-coil (green). Despite the majority of MD snapshots showing a lower agreement (lower z-score) with experimentally determined secondary structure, some snapshots (right of the black line) show an increased agreement.

For the final MD-refined model, we calculated the R_g value and computed theoretical scattering profiles with FoXS [103, 104]. The profile is computed using the Debye formula for spherical scatterers and plotted as intensity (log scale) *versus* the momentum of transfer (q). Finally, the goodness of fit between scattering profiles was measured in terms of χ^2 [105]. The final model after MD refinement shows a secondary structure MCC of 0.81 for β-sheets, 0.88 for α-helices, and 0.78 for random coils. The final shape agreement with SAXS shows a χ^2 of 49.8, which is high, but not surprising given the mobile, disordered tail, and is still an improvement compared to the initial model from TopModel with a χ^2 of 81.5. The model shows a radius of gyration of 26.7 Å compared to the experimentally determined value of 24.3 Å, due to slight drift of the disordered tail during model combination and energy minimization.

To explore if the high χ^2 from SAXS is indeed caused by the mobile, disordered tail, a truncated

version of the protein was expressed in which the first 30 of the 43 disordered tail residues were removed. When SAXS measurements of the truncated protein are compared to a the final model in which the same residues were removed, the shape agreement increases markedly as indicated by a drop in χ^2 from 49.8 to 3.86, confirming that the initial disagreement with the SAXS data for the full-length protein is indeed caused by the high mobility of the disordered tail, and that the shape of the folded domain shows a high agreement with experiment.

To conclude, generating conformational ensembles by MD simulations starting from a model structure obtained by homology modelling allowed us to improve its agreement with secondary structure assignment and SAXS data. The tail region is highly mobile and disordered, which likely results in the elongated shape density seen in SAXS experiments on the full-length protein. At present, it is unclear what causes the elongated shape and if the underlying tail conformation(s) is (are) functionally relevant.

Figure S7. Comparison of SAXS data for full-length and truncated LipoP from *C. difficile*

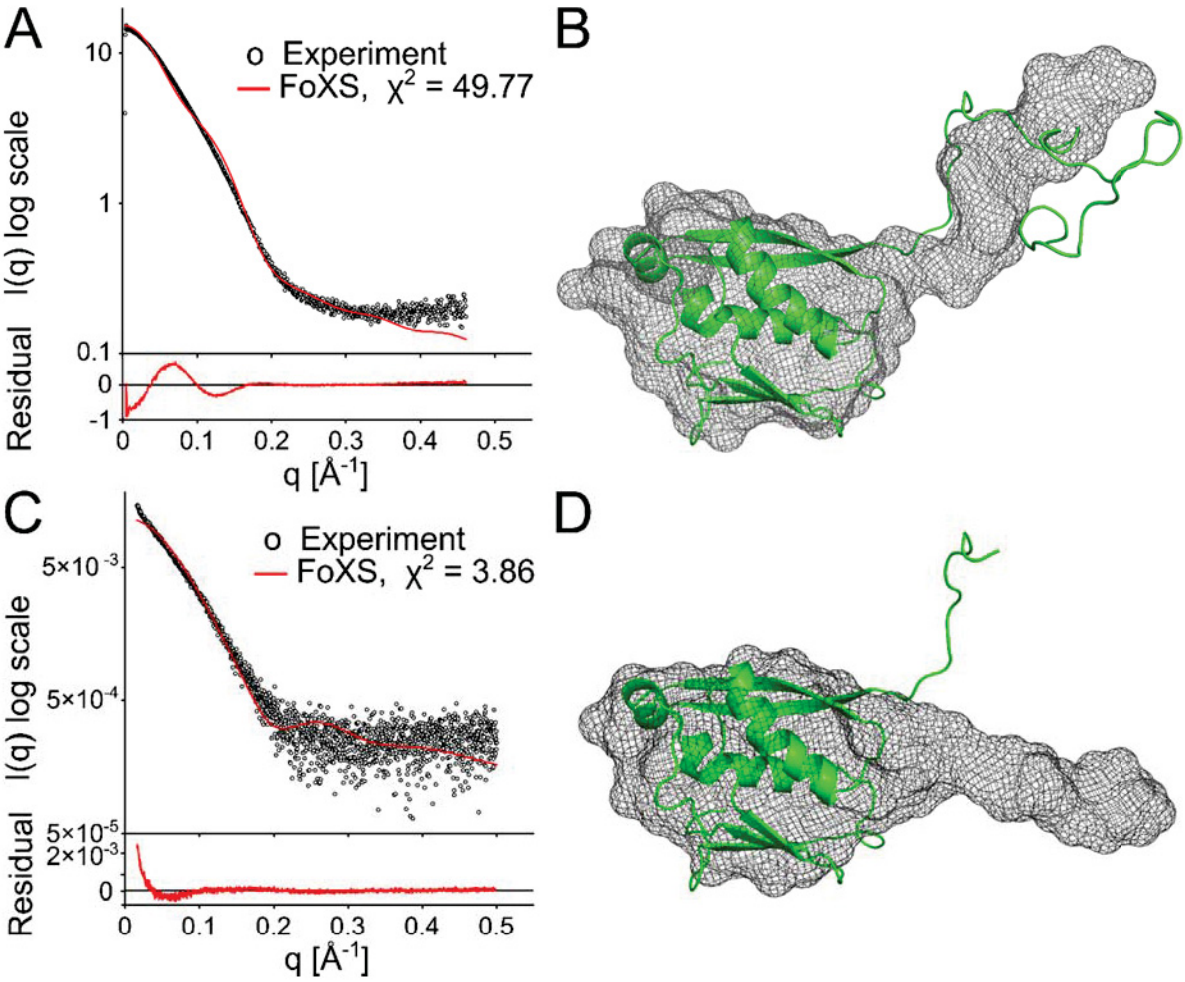


Figure S7. Comparison of SAXS data for full-length and truncated LipoP from *C. difficile*

Comparison of the experimental and simulated scattering profiles of the full-length (A) and truncated (C) LipoP and their calculated density maps from the full-length (B) and truncated (D) proteins as calculated with GASBOR. The SAXS densities show how the disordered tail occupies different average conformations in the full-length and truncated versions of the proteins. The χ^2 values from FoXS show that the shape agreement is greatly increased once the majority of the disordered tail is removed. In panel B, the final model after MD refinement is shown in green. For the truncated version in D, the first 30 residues of the 43 residues of the disordered tail were removed.

Supplemental References

1. Boratyn, G.M. Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., & Madden, T.L., Domain enhanced lookup time accelerated BLAST. *Biol Direct*,7(1), 12 (2012)
2. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D.G., Notredame, C., 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *Journal of molecular biology*, **340**(2),385-395 (2004).
3. Laskowski, R.A., MacArthur, M.W., Moss, D.S., &Thornton, J.M., PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography*, **26**(2), 283-291 (1993).
4. Eddy, S.R., Accelerated profile HMM searches. *PLoS Computational Biology*, **7**(10). e1002195 (2011).
5. Katoh, K. & Standley D.M., MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, **30**(4), 772-780 (2013).
6. Chen, V.B., Arendall, W.B.3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., Richardson, D.C., MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, **66**(1), 12-21 (2009).
7. Remmert, M., Biegert, A., Hauser, A., & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, **9**(2), 173-175 (2012).
8. Collingridge, P.W. & Kelly S., MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC bioinformatics*, **13**(1), 117 (2012).
9. Melo, F. & Feytmans, E., Novel knowledge-based mean force potential at atomic level. *Journal of molecular biology*, **267**(1), 207-222 (1997).
10. Söding, J., Biegert, A., & Lupas, A.N., The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, **33**(2), W244-W248 (2005).
11. Madhusudhan, M.S., Webb, B.M., Marti-Renom, M.A., Eswar, N., Sali, A., Alignment of multiple protein structures based on sequence and structure features. *Protein Engineering Design and Selection*, **22**(9), 569-574 (2009).
12. Sippl, M.J., Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Genetics*, **17**(4), 355-362 (1993).
13. Rychlewski L., Jaroszewski, L., Li, W., & Godzik, A., Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science*, **9**(2), 232-241 (2000).
14. Pei, J., Kim, B.H., & Grishin, N.V., PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research*, **36**(7), 2295-2300 (2008).
15. Shen, M.Y. & Sali, A., Statistical potential for assessment and prediction of protein structures. *Protein science*,. **15**(11), 2507-2524 (2006).
16. Yang, Y., Faraggi, E., Zhao, H., & Zhou, Y., Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, **27**(15), 2076-2082 (2011).
17. Daniels, N.M., Nadimpalli, S., & Cowen, L.J., Formatt: Correcting protein multiple structural alignments by incorporating sequence alignment. *BMC bioinformatics*, **13**(1), 259 (2012).
18. Zhou, H. & Zhou Y., Distance□scaled, finite ideal□gas reference state improves structure□derived potentials of mean force for structure selection and stability

prediction. *Protein science*, **11**(11), 2714-2726 (2002).
19. Peng, J. & Xu, J., RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics*, **79**(S10), 161-171 (2011).
20. Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., Lesk, A.M., MUSTANG: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, **64**(3), 559-574 (2006).
21. McGuffin, L.J. & Roche, D.B., Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*, **26**(2), 182-188 (2010).
22. Wu, S. & Zhang Y., LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research*, **35**(10), 3375-3382 (2007).
23. Wang, S., Peng, J. & Xu, J., Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. *Bioinformatics*, **27**(18), 2537-2545 (2011).
24. Wallner, B. & Elofsson, A., Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Science*, **15**(4), 900-913 (2006).
25. Lobley, A., Sadowski, M.I., & Jones, D.T., pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, **25**(14), 1761-1767 (2009).
26. Zhang, Y. & Skolnick, J., SPICKER: A clustering approach to identify near□native protein folds. *Journal of computational chemistry*, **25**(6), 865-871 (2004).
27. Benkert, P., Schwede, T., & Tosatto, S.C., QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. *BMC structural biology*, **9**(1), 35 (2009).
28. Pearson, W.R., Finding protein and nucleotide similarities with FASTA. *Current protocols in bioinformatics*, 3.9: 1-25 (2004).
29. Ray, A., Lindahl E.,& Wallner, B., Improved model quality assessment using ProQ2. *BMC bioinformatics*, **13**(1), 224 (2012).
30. Karplus, K., SAM-T08, HMM-based protein structure prediction. *Nucleic acids research,* **37**, W492–W497 (2009).
31. Uziela, K., Menéndez-Hurtado, D., Shu, N., Wallner, B. & Elofsson, A., ProQ3D: improved model quality assessments using deep learning. *Bioinformatics*. **33**(10), 1578-1580 (2017).
32. Manavalan, B. & Lee, J., SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics*, **33**(16), 2496-2503 (2017).
33. Randall, A. & Baldi, P., SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *BMC structural biology*, **8**(1), 52 (2008).
34. Thompson, J.D., Higgins, D.G. & Gibson, T.J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, **22**(22), 4673-4680 (1994).
35. Lee, C., Grasso, C. & Sharlow, M.F, Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**(3), 452-464 (2002).
36. Edgar, R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**(5), 1792-1797 (2004).
37. Sierk, M.L., Smoot, M.E., Bass, E.J., Pearson, W.R., Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. *BMC bioinformatics*, **11**(1), 146 (2010).

38. Pei, J., Sadreyev, R., & Grishin, N.V., PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**(3), 427-428 (2003).
39. Do C.B., Mahabhashyam, M.S., Brudno, M. & Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research*, **15**(2), 330-340 (2005).
40. Al-Ait, L., Yamak, Z.& Morgenstern, B., DIALIGN at GOBICS - multiple sequence alignment using various sources of external information. *Nucleic acids research*, **41**(W1), W3-W7 (2013).
41. Taylor, W.R., Protein structure comparison using iterated double dynamic programming. *Protein Science*, **8**(03), 654-665 (1999).
42. Zhang, Y. & Skolnick, J., TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, **33**(7), 2302-2309 (2005).
43. Mulnaes, D. & H. Gohlke, TopScore: Using Deep Neural Networks and Large Diverse Data Sets for Accurate Protein Model Quality Assessment. *Journal of chemical theory and computation*, **14**(11), 6117-6126 (2018).
44. Mezulis, S., Sternberg, M.J.& Kelley, L.A, PhyreStorm: A web server for fast structural searches against the PDB. *Journal of molecular biology*, **428**(4), 702-708 (2016).
45. McGuffin, L.J., Bryson, K. & Jones, D.T., The PSIPRED protein structure prediction server. *Bioinformatics*, **16**(4), 404-405 (2000).
46. Mariani, V., Biasini, M., Barbato, A. & Schwede, T., lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**(21), 2722-2728 (2013).
47. Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S., Richardson, D.C. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *Journal of molecular biology*, **285**(4), 1711-1733 (1999).
48. Vreven, T., Moal, I.H., Vangone, A., Pierce, B.G., Kastritis, P.L., Torchala, M., Chaleil, R., Jiménez-García, B., Bates, P.A., Fernandez-Recio, J., Bonvin, A.M. & Weng, Z., Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology*, **427**(19), 3031-3041 (2015).
49. Pedregosa, F., Varoquaux, G., Gramfort A., Michel, V. & Thirion, B., Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830 (2011).
50. Kingma, D. & Ba, J., Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980 (2014).
51. Faraggi, E., Zhang, T., Yang, Y., Kurgan, L.& Zhou,Y., SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *Journal of computational chemistry*, **33**(3), 259-267 (2012).
52. Zhu, J. & Weng, Z., FAST: a novel protein structure alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, **58**(3), 618-627 (2005).
53. Menke, M., Berger, B. & Cowen, L., Matt: local flexibility aids protein multiple structure alignment. *PLoS computational biology*, **4**(1), e10 (2008).
54. Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U. & Sali, A. Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics*, Chapter 5:Unit-5.6 (2006).
55. Khatib, F., Weirauch, M.T.& Rohl, C.A., Rapid knot detection and application to protein structure prediction. *Bioinformatics*, **22**(14), e252-e259 (2006).
56. Wang, Q., Canutescu, A.A., & Dunbrack R.L.J., SCWRL and MolIDE: computer

programs for side-chain conformation prediction and homology modeling. *Nature protocols*, **3**(12), 1832 (2008).
57. Miao, Z., Cao, Y. & Jiang, T., RASP: rapid modeling of protein side chain conformations. *Bioinformatics*, **27**(22), 3117-3122 (2011).
58. Bhattacharya, D., Nowotny, J., Cao, R. & Cheng J., 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic acids research*, **44**(W1), W406-W409 (2016).
59. Xu, D. & Zhang, Y., Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical journal*, **101**(10), 2525-2534 (2011).
60. Rohl, C.A., Strauss, C.E., Misura, K.M. & Baker, D., Protein structure prediction using Rosetta. *Methods in enzymology*, **383**, 66-93 (2004).
61. Pernot, P., Theveneau, P., Giraud, T., Nogueira Fernandes, R., Nurizzo, D., Spruce, D., Surr, J., McSweeney, S., Round, A., Felisaz, F., Foedinger, L., Gobbo, A., Huet, J., Villard, C. & Cipriani, F. New beamline dedicated to solution scattering from biological macromolecules at the ESRF. *Journal of Physics: Conference Series*, **247**(1), 012009 (2010).
62. Pernot, P., Round, A., Barrett, R., De Maria Antolinos, A., Gobbo, A, Gordon, E., Huet, J., Kieffer, J., Lentini, M., Mattenet, M., Morawe, C., Mueller-Dieckmann, C., Ohlsson, S., Schmid, W., Surr, J., Theveneau, P., Zerrad, L. & McSweeney, S. Upgraded ESRF BM29 beamline for SAXS on macromolecules in solution. *Journal of Synchrotron Radiation*, **20**(4), 660-664 (2013).
63. Blanchet, C.E., Spilotros, A., Schwemmer, F., Graewert, M.A., Kikhney, A., Jeffries, C.M., Franke, D., Mark, D., Zengerle, R., Cipriani, F. & Fiedler, S. Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *Journal of Applied Crystallography*, **48**(2), 431-443 (2015).
64. Franke, D., Petoukhov, M.V., Konarev, P.V., Panjkovich, A., Tuukkanen, A., Mertens, H.D.T., Kikhney, A.G., Hajizadeh, N.R., Franklin, J.M., Jeffries, C.M. and Svergun, D.I. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *Journal of Applied Crystallography*, **50**(4), 1212-1225 (2017).
65. Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch M.H.J. & Svergun, D.I., PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography*, **36**, 1277-1282 (2003).
66. Guinier, A., Diffraction of x-rays of very small angles-application to the study of ultramicroscopic phenomenon. *Annales de Physique*, **12**, 161-237 (1939).
67. Svergun, D.I., Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. *Journal of Applied Crystallography*, **25**, 495-503 (1992).
68. Svergun, D.I., Petoukhov, M.V. & Koch M.H.J., Determination of domain structure of proteins from X-ray solution scattering. *Biophysical Journal*, **80**(6), 2946-2953 (2001).
69. Kozin, M.B. & Svergun, D.I., Automated matching of high- and low-resolution structural models. *Journal of Applied Crystallography*, **34**, 33-41 (2001).
70. Findeisen, M., Brand T. & Berger S., A 1H□NMR thermometer suitable for cryoprobes. *Magnetic Resonance in Chemistry*, **45**(2), 175-178 (2007).
71. Salzmann, M., Wider, G., Pervushin, K.& Wüthrich, K., Improved sensitivity and coherence selection for [15N, 1H]-TROSY elements in triple resonance experiments. *Journal of biomolecular NMR*, **15**(2), 181-184 (1999).
72. Yang, D. & Kay L.E., Improved 1HN-detected triple resonance TROSY-based

experiments. *Journal of biomolecular NMR*, **13**(1), 3-10 (1999).
73. Nietlispach, D., Suppression of anti-TROSY lines in a sensitivity enhanced gradient selection TROSY scheme. *Journal of biomolecular NMR*, **31**(2), 161-166 (2005).
74. Sattler, M., Schleucher, J. & Griesinger, C., Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution. *Progress in nuclear magnetic resonance spectroscopy*, **34**, 93-158 (1999).
75. Cavanagh, J., Skelton, N., Fairbrother, W., Rance, M., Palmer, A., Protein NMR spectroscopy: principles and practice. *Elsevier* (1995).
76. Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. & Bax, A., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of biomolecular NMR*, **6**(3), 277-293 (1995).
77. Zhang, O., Kay, L.E., Olivier, J.P., Forman-Kay, J.D., Backbone 1 H and 15 N resonance assignments of the N-terminal SH3 domain of drk in folded and unfolded states using enhanced-sensitivity pulsed field gradient NMR techniques. *Journal of biomolecular NMR*, **4**(6), 845-858 (1994).
78. Cavanagh, J. & Rance M., Suppression of cross-relaxation effects in TOCSY spectra via a modified DIPSI-2 mixing sequence. *Journal of Magnetic Resonance*, **96**(3), 670-678 (1992).
79. Yang, D. & Kay L.E., TROSY triple-resonance four-dimensional NMR spectroscopy of a 46 ns tumbling protein. *Journal of the American Chemical Society*, **121**(11), 2571-2575 (1999).
80. Kovacs, H. & Gossert A., Improved NMR experiments with 13 C-isotropic mixing for assignment of aromatic and aliphatic side chains in labeled proteins. *Journal of biomolecular NMR*, **58**(2), 101-112 (2014).
81. Kadkhodaie, M., Rivas, O., Tan, M., Mohebbi, A., Shaka, A.J., .Broadband homonuclear cross polarization using flip-flop spectroscopy. *Journal of magnetic resonance*, **91**, 437-443 (1991).
82. Marion, D., Ikura, M., Tschudin, R., Bax, A., Rapid recording of 2D NMR spectra without phase cycling. Application to the study of hydrogen exchange in proteins. *Journal of Magnetic Resonance*, **85**, 393-399 (1989).
83. Kay, L., Keifer, P., & Saarinen, T., Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *Journal of the American Chemical Society*, **114**(26), 10663-10665 (1992).
84. Schleucher, J., Sattler, M. & Griesinger, C., Coherence Selection by Gradients without Signal Attenuation: Application to the Three□Dimensional HNCO Experiment. *Angewandte Chemie International Edition*, **32**(10), 1489-1491 (1993).
85. Johnson, B.A. & Blevins R.A., NMR View: A computer program for the visualization and analysis of NMR data. *Journal of biomolecular NMR*, **4**(5), 603-614 (1994).
86. Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E., Wüthrich, K. Recommendations for the presentation of NMR structures of proteins and nucleic acids. *Journal of molecular biology*, **280**(5), 933-952 (1998).
87. Berjanskii, M.V. & Wishart D.S., A simple method to predict protein flexibility using secondary chemical shifts. *Journal of the American Chemical Society*, **127**(43), 14970-14971 (2005).
88. Shen, Y. & Bax, A., Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *Journal of biomolecular NMR*, **56**(3), 227-241 (2013).
89. DeLano, W.L., PyMOL. *CCP4 Newsletter On Protein Crystallography* (2002).
90. Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W.,

Garner, E.C., Obradovic, Z. Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, **19**(1), 26-59 (2001).

91. Theillet, F.X., Kalmar, L., Tompa, P., Han, K.H., Selenko, P., Dunker, A.K., Daughdrill, G.W. & Uversky, V.N. The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. *Intrinsically Disord Proteins*, **1**(1), e24360 (2013).
92. Mohan, A., Oldfield, C.J., Radivojac, P., Vacic, V., Cortese, M.S., Dunker, A.K. & Uversky, V.N., Analysis of molecular recognition features (MoRFs). *Journal of Molecular Biology*, **362**(5), 1043-1059 (2006).
93. Vacic, V., Oldfield, C.J., Mohan, A., Radivojac, P., Cortese, M.S., Uversky, V.N. & Dunker, A.K., Characterization of molecular recognition features, MoRFs, and their binding partners. *Journal of Proteome Research*, **6**(6), 2351-2366 (2007).
94. Oldfield, C.J., Uversky, V.N., & Kurgan L., Predicting Functions of Disordered Proteins with MoRFpred. *Methods in Molecular Biology*, **1851**, 337-352 (2019).
95. Schafmeister, C.E.A.F., Ross, W.S. & Romanovski, V. LEaP. *University of California, San Francisco* (1995).
96. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics*, **79**(2), 926-935 (1983).
97. Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S. & Walker, R.C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation*, **9**(9), 3878-3888 (2013).
98. Case, D.A., Betz, R.M., Cerutti, D.S., Cheatham, T.E., Darden, T.A., Duke, R.E., Giese, T.J., Gohlke, H., Goetz, A.W., Homeyer, N., Izadi, S., Janowski, P., Kaus, J., Kovalenko, A., Lee, T.S., LeGrand, S., Li, P., Lin, C., Luchko, T., Luo, R., Madej, B., Mermelstein, D, Merz, K.M., Monard, G., Nguyen, H., Nguyen, H.T., Omelyan, I., Onufriev, A., Roe, D.R., Roitberg, A., Sagui, C., Simmerling, C.L., Botello-Smith, W.M., Swails, J., Walker, R.C., Wang, J., Wolf, R.M., Wu, X., Xiao, L. & Kollman, P.A., AMBER 2016, *University of California, San Francisco*(2016).
99. Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E. & Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, **11**(8), 3696-3713 (2015).
100. Frieg, B., Görg, B., Homeyer, N., Keitel, V., Häussinger, D., Gohlke, H., Molecular Mechanisms of Glutamine Synthetase Mutations that Lead to Clinically Relevant Pathologies. *PLoS Computational Biology*, **12**(2), e1004693 (2016).
101. Roe, D.R. & Cheatham T.E., 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, **9**(7), 3084-3895 (2013).
102. Kabsch, W. & Sander C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**(12), 2577-2637 (1983).
103. Schneidman-Duhovny, D., Hammel, M., Tainer, J.A., Sali, A., Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophysical Journal*, **105**(4), 962-974 (2013).
104. Schneidman-Duhovny, D., Hammel, M., Tainer, J.A., Sali, A., FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res*, **44**(W1), W424-429 (2016).
105. Bevington, P.R. & Robinson D.K., Data reduction and error analysis for the physical sciences., *New York; Montreal: McGraw-Hill*,(1992).

19. PUBLICATION III

Binding region of Alanopine Dehydrogenase predicted by unbiased Molecular Dynamics simulations of ligand diffusion

Holger Gohlke[1], Ulrike Hergert[2], Tatu Meyer[2], Daniel Mulnaes[1], Manfred K. Grieshaber[2], Sander H. J. Smits[2], and Lutz Schmitt[2]

[1]Institute of Pharmaceutical and Medicinal Chemistry,

Heinrich-Heine University Düsseldorf, Düsseldorf, Germany

[2]Institute of Biochemistry, Department of Mathematics and Natural

Sciences, Heinrich-Heine University, 40204 Düsseldorf, Germany

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Letter

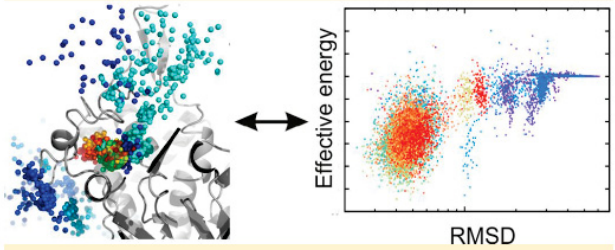
pubs.acs.org/jcim

Binding Region of Alanopine Dehydrogenase Predicted by Unbiased Molecular Dynamics Simulations of Ligand Diffusion

Holger Gohlke,*[,†] Ulrike Hergert,[‡] Tatu Meyer,[‡] Daniel Mulnaes,[†] Manfred K. Grieshaber,[‡] Sander H. J. Smits,[‡] and Lutz Schmitt*[,‡]

[†]Institute for Pharmaceutical and Medicinal Chemistry and [‡]Institute of Biochemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine-University, 40204 Düsseldorf, Germany

Supporting Information



ABSTRACT: Opine dehydrogenases catalyze the reductive condensation of pyruvate with L-amino acids. Biochemical characterization of alanopine dehydrogenase from *Arenicola marina* revealed that this enzyme is highly specific for L-alanine. Unbiased molecular dynamics simulations with a homology model of alanopine dehydrogenase captured the binding of L-alanine diffusing from solvent to a putative binding region near a distinct helix-kink-helix motif. These results and sequence comparisons reveal how mutations and insertions within this motif dictate the L-amino acid specificity.

Unbiased MD simulations of ligand binding have become possible only recently due to advances in the simulation algorithms and hardware. In addition to identifying the binding region of a ligand, they can reveal (un)binding pathways, identify metastable intermediate states, and provide quantitative estimates of binding affinities and on- and off-rates.[1−5] To the best of our knowledge, unbiased MD simulations of ligand binding have not yet been applied for investigating determinants of substrate specificity starting from comparative protein models. Such an application should be widely interesting for other areas of structure-based life sciences as well. In this context, we investigate L-alanine binding to alanopine dehydrogenase of *Arenicola marina* (AlaDH*Am*)[6] by means of comparative modeling in combination with unbiased molecular dynamics (MD) simulations and a biochemical characterization of AlaDH*Am*.

AlaDH*Am* is a member of the family of opine dehydrogenases (OpDHs), which catalyze the reductive condensation of pyruvate with an L-amino acid in the presence of NADH to so-called opines during anaerobic glycolysis.[7] The best characterized enzyme of this family is octopine dehydrogenase (OcDH), which catalyzes the reductive condensation of pyruvate with L-arginine to D-octopine.[8−10] Structures of OcDH determined by X-ray crystallography in complex with NADH, with NADH and L-arginine as well as with NADH and pyruvate[11] demonstrated that domain I of OcDH binds the cofactor NADH, whereas the main binding site of the amino acid substrate is located in domain II. The binding of L-arginine induces a rotational movement of domain II toward domain I.[11−13] AlaDH*Am* catalyzes the reductive condensation of pyruvate with L-alanine to alanopine (*N*-(1-D-carboxylethyl)-L-alanine).[14,15] With lower efficiency also glycine can be used as amino acid precursor, which results in the formation of strombine (*N*-(carboxymethyl)-D-alanine)[7] In contrast to OcDH, but in agreement with *N*-(1-D-carboxylethyl)-L-norvaline dehydrogenase (CENDH),[16] AlaDH*Am* contains a characteristic insertion at position 209 in the helix-kink-helix motif located at the N-terminal part of domain II (AlaDH*Am* numbering is used throughout this study; Figure S1 in the Supporting Information (SI)). In OpDHs having an N209 insertion, almost exclusively valine is found at position 208 then, whereas in OpDHs lacking N209, aspartate, arginine, lysine, or tyrosine is found at position 208 depending on the respective L-amino acid substrate.[11] CENDH has been crystallized only in the *apo* form, and no structural information is available for AlaDH*Am*. Thus, the role of sequence positions 208 and 209 in the helix-kink-helix motif in determining the specificity for the L-amino acid substrate in OpDHs has remained elusive.

In order to elucidate this role, first, we biochemically characterized AlaDH*Am*. Cloning and expression of the gene of AlaDH*Am* (Uniprot entry: B5D5P2_AREMA) was performed as described for OcDH from *P. maximus*[10] (see SI for details). The final preparation contained a single homogeneous protein of approximately 45 kDa (SI Figure S2), in agreement with the sequence-based calculated mass and the molecular mass estimated by size-exclusion chromatography using standard proteins (results not shown). The AlaDH*Am* followed standard Michaelis−Menten kinetics for the substrates used (Figure 1 and Table 1). Substrate inhibition was observed for L-alanine as well glycine, a feature observed for many other OpDHs.[17] For L-alanine, a K_m of 14.8 ± 2.1 mM and a V_{max} of 1513.0 ± 144.5 U/mg was found (Table 1). Thus, AlaDH*Am* is highly active, in contrast to other AlaDH characterized:[18] The AlaDH from *M. sanguinea* displays an almost 20-fold reduced catalytic efficiency (k_{cat} for AlaDH is 1084.6 and 51.70 s^{-1} for *A. marina* and *M. sanguinea*, respectively). Furthermore, AlaDH*Am* displays a

Published: September 5, 2013

dx.doi.org/10.1021/ci400370y | J. Chem. Inf. Model. 2013, 53, 2493 2498

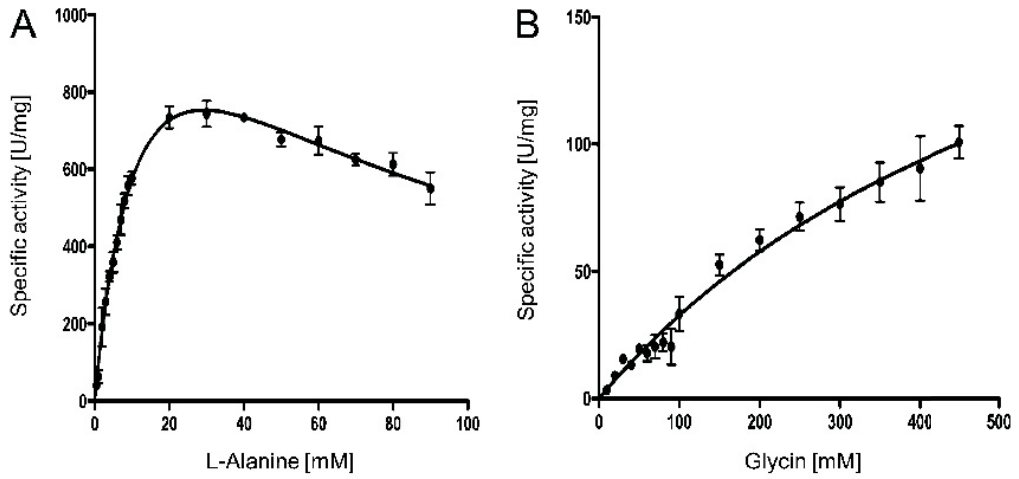


Figure 1. Michaelis–Menten kinetics of the alanopine reaction. Plotted is the specific activity of the AlaDH*Am* against increasing amounts of (A) L-alanine and (B) glycine. The enzymatic activity of AlaDH*Am* was measured spectrophotometrically at 25 °C following the decrease in absorbance at 340 nm. Standard assays were carried out using 3 mM pyruvate, 0.16 mM NADH in 50 mM triethanolamine buffer pH 7.0. The reaction was started by the addition of L-alanine or glycine. Activities were calculated using a specific absorbance coefficient ε = 6.31 mM^{-1} cm^{-1} for NADH. One unit is defined as the amount of enzyme catalyzing the oxidation of 1 μmol NADH per minute.

Table 1. Kinetic Parameters for the Forward Reaction (NADH Oxidation) Catalyzed by AlaDH*Am*[a]

substrate	L-alanine	glycine
K_m [mmol L^{-1}]	14.8 ± 2.1	655.1 ± 45.2
V_m [U mg^{-1}]	1513.0 ± 144.5	246.1 ± 32.1
K_i [mmol L^{-1}]	58.0 ± 10.5	*b*
k_{cat} [s^{-1}]	1084.6 ± 78.1	176.4 ± 24.1
catalytic efficiency [mol^{-1} s^{-1}]	7.3×10^4	*b*

[a]For the determination of the kinetic constants, the initial velocities at different substrate concentrations of L-alanine or glycine were recorded spectrophotometrically at 340 nm. Kinetic parameters were obtained using nonlinear least-squares analysis of the data fitted to the Michaelis–Menten rate equation ($v = V_{max}[S]/K_m + [S]$) or the Michaelis–Menten equation corrected for uncompetitive substrate inhibition ($v = V_{max}[S]/K_m + [S](1 + [S]/K_i)$ where v is the velocity, V_{max} is the maximum velocity, [S] is the substrate concentration, K_m is the Michaelis constant, and K_i is the inhibition constant, using the enzyme kinetic module 2.0 of Sigma-plot 9.0 (Systat Software, Erkrath, Germany). [b]Not determined.

high specificity toward L-alanine: When glycine was used as a substrate, activity dropped at least 3- to 4-fold with a significantly higher K_m value suggesting a significantly lower affinity (Table 1), whereas for other small amino acids tested, e.g. L-serine, L-threonine, L-cysteine, or L-valine, no or only negligible activities were found (data not shown). In contrast, the AlaDH from *M. sanguinea* displayed a broader substrate specificity allowing also other small amino acids to form the corresponding opine.[18] This suggests that the binding site for the amino acid has been optimized in AlaDH*Am* to preferentially bind L-alanine with high efficiency.

In order to structurally elucidate the binding region of L-alanine in AlaDH*Am*, a model of the protein was generated by comparative modeling using the in-house workflow TopModel (D. Mulnaes and H. Gohlke, unpublished results), which is based on the Modeler program[19] (see SI for details). Pursuing a multitemplate modeling strategy, a pBLAST[20] search on the Protein Data Base[21] revealed three suitable template structures, two of which are OcDHs bound to NADH and either L-arginine (PDB code 3C7C)[11] or agmatine (3IQD).[12] The third template was CENDH (1BG6).[16] The sequence identities of AlaDH*Am* with OcDH and CENDH are 46% and 20%, respectively.

A multiple sequence alignment using structural information from the templates revealed a high degree of residue conservation for sequence positions in the vicinity of the substrate-binding region identified in OcDH and respective sequence positions in AlaDH*Am* (SI Figure S1). In particular, E141 of domain I and W279 of domain II are conserved in OpDHs and are present in AlaDH*Am*, too. Thus, the substrate specificity toward L-alanine of AlaDH*Am* cannot be mediated by these amino acids. Y208 of OcDH from *P. maximus* located in the kink of the helix-kink-helix motif is the only amino acid involved in substrate binding that differs between the OpDHs. In addition, in AlaDH*Am*, N209 is inserted. As position 209 is also located in the kink, N209 can be accommodated in the model structure of AlaDH*Am* without disturbing the overall structure (SI Figure S3A; see the Homology modeling section in the SI for an evaluation of the structural quality of the model). An overlay of the L-arginine bound OcDH structure with the AlaDH*Am* model indicated that the inserted amino acid would sterically interfere with the binding position of L-arginine (Figure S3B). Accordingly, while for the generation of an AlaDH*Am*/NADH/L-alanine model the coordinates of NADH could be copied from the OcDH/NADH/L-arginine complex structure without steric clashes, a geometry optimization was required to reduce steric clashes of the L-alanine substrate initially placed at the position of the backbone of the arginine substrate (see Supporting Information for details). The optimized binding pose of L-alanine is shifted by ~3 Å with respect to the starting location (Figure S3B).

To further refine the AlaDH*Am*/NADH/L-alanine complex structure, the structure was subjected to three independent molecular dynamics (MD) simulations in explicit solvent of 200 ns length each (see Supporting Information for details on the protocols of the MD simulation). Overall, only moderate deviations of the AlaDH*Am* structures from the starting structure were observed (root mean-square deviations (rmsd) of the C_α atoms in all trajectories between 2.5 and 4 Å, in rare cases also up to 4.5 Å; SI Figure S4A and B), which are

dx.doi.org/10.1021/ci400370y | J. Chem. Inf. Model. 2013, 53, 2493–2498

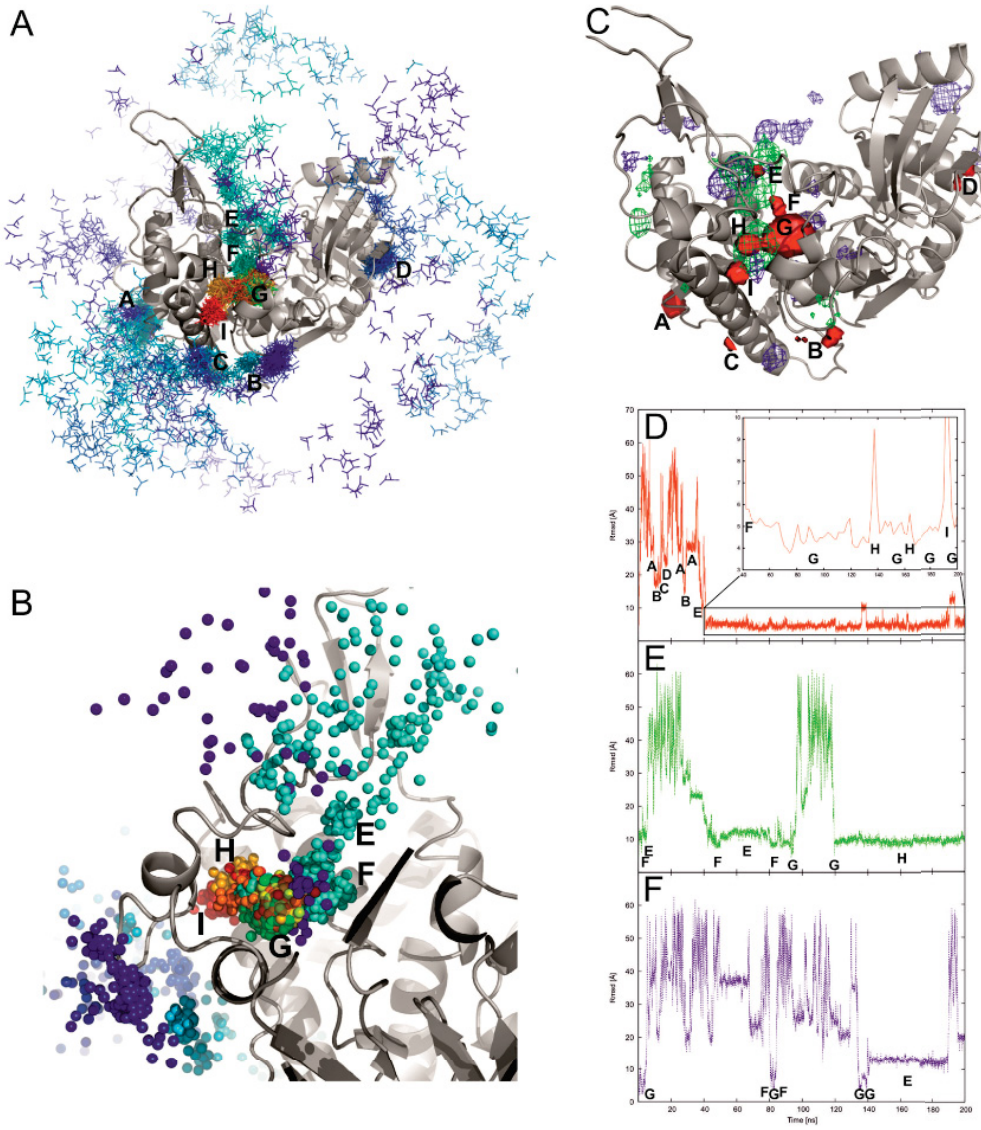


Figure 2. Unbiased MD simulations of L-alanine diffusion. (**A**–**F**) Black letters indicate regions of high density of L-alanine during the MD simulation 1 as identified in panel C. Region G is the predicted binding region. (**A**) Traces of L-alanine extracted from the trajectory 1 generated by MD simulations of 200 ns length of the AlaDH*Am*/NADH/L-alanine system in explicit water (see Supporting Information for details); L-alanine reaches the predicted binding region after ~40 ns (see panel D). The time evolution of the MD simulation is color coded from blue (0 ns) to red (200 ns). For clarity, only a conformation closest to the average conformation of AlaDH*Am* is shown (gray cartoon). (**B**) Close-up view of the predicted binding region shown in panel A with the trace of C_α atoms of L-alanine extracted from trajectory 1 shown as spheres. See panel A regarding the color coding. (**C**) Overlay of density maps extracted from trajectories 1 (red isocontour surface), 2 (green isocontour mesh), and 3 (blue isocontour mesh) showing the frequency of interactions of L-alanine on the surface of AlaDH*Am*; the contour level is 3 sigma. Regions of high density identified from trajectory 1 are labeled with black letters. The protein conformation is as in panel A. (**D**–**F**) Root mean square deviations (rmsds) of the L-alanine atoms during the course of the MD simulations 1 (panel D), 2 (panel E), and 3 (panel F) with respect to the modeled starting structure (see SI Figure S3) after superimposing AlaDH*Am* based on its C_α atoms.

comparable to those observed when MD simulations of 100 ns length are started from one of the crystal structures used as a template (PDB ID 3C7C; rmsd = 1.5–3.5 Å; data not shown). NADH remained at its binding position in all three simulations (Figure S4). However, despite a careful thermalization of the complex structures, the ligand, L-alanine, left the initial binding region after at most 5 ns in all three simulations and escaped into the solvent (rmsd up to 60 Å; Figure 2A–F). Yet, L-alanine spontaneously returned to this region after 40 ns in MD simulation 1 (Figure 2A, B, and D; the binding region is marked with a "G") and remained bound for almost all of the remaining simulation time. Similar returns are observed in MD simulation 2 (after 95 and 120 ns) and 3 (after 82 and around 137 ns) with residence times of L-alanine of at most 6 ns (Figure 2E and F). Such short residence times do not contradict expectations arising from the knowledge of the very weak binding affinity observed for L-alanine to AlaDH*Am* (K_i = 58.0 ± 10 mM; Table 1). These MD results are remarkable for three reasons: (I) initially, the ligand completely escapes to the solvent (see the trace of L-alanine in MD simulation 1 in Figure 2A) and diffuses there for at least 40 ns before rebinding such that the rebinding should not be

influenced by the starting position; (II) the observed binding events occurred from unbiased MD simulations, i.e. no prior knowledge of the binding region was applied during the MD simulations; (III) it is reassuring that in all three independent MD simulations L-alanine does bind again. In all, this makes our MD simulations one of the few examples[1−5] known to date that capture binding of a ligand diffusing from solvent to the bound state.

In order to provide quantitative estimates of the binding thermodynamics and kinetics[4] substantially more observed unbinding and binding events would have been required. In particular, sampling of the unbound state is not converged after 200 ns of simulation time as demonstrated by nonoverlapping regions of highest frequency of L-alanine interactions on the outer surface of AlaDH*Am* in MD simulations 1−3 (Figure 2C). Still, the simulations provide suggestions for energetically favorable interaction "hot spots" on the protein's outside, as exemplarily shown for L-alanine "hopping" between regions A−D in MD simulation 1 (Figure 2A, C, and D). In contrast, all three MD simulations yield overlapping densities of the frequency of L-alanine interactions with AlaDH*Am* when L-alanine approaches the bound state, i.e., for regions E, F, G, and H (Figure 2B and C). When mapped onto the MD trajectories, these findings suggest that L-alanine consistently unbinds from (and binds to) region G via regions F and E (Figure 2D−F). The effective energy of binding $\Delta G_{effective}$ (i.e., the sum of gas-phase and solvation free energy) computed along MD trajectory 1 by the MM-GBSA method (see Supporting Information for details)[22] corroborates this view, which shows a global effective energy minimum corresponding to L-alanine binding to region G accompanied by several local minima corresponding to L-alanine in non-native poses (the most pronounced of which refers to region E) (Figure 3A; SI Figure S5A). In total, a funnel-shaped landscape[23−25] of the binding effective energy emerges, which is similar to landscapes of the binding effective energy observed for the binding of kinase inhibitors to the Src kinase.[1] Contributions due to the changes in the configurational entropy of the solute molecules upon binding are not considered in the effective energy calculations. Considering that bound L-alanine shows a considerable amount of residual motions as judged from the observation of configurational fluctuations of L-alanine of ∼3 Å in region G (Figures 2D and 3A), adverse contributions to binding due to changes in the configurational entropy are expected to be small (see SI Supplemental Results for an estimate). Thus, the overall shape of a landscape of the free energy of binding should not differ qualitatively from our landscape. Finally, from these calculations, three energetically most favorable and non-distinguishable L-alanine positions ($\Delta G_{effective}$ = −27.34, −27.38, −27.25 kcal mol^{-1}) are identified which all reside in region G (Figure 3A; Figure S5A). When computing the effective energy of binding for MD trajectories 2 and 3, $\Delta G_{effective}$ = −16.73 and −13.4 kcal mol^{-1} are found for L-alanine positions in region G, respectively (SI Figure S6); these values are among the most favorable effective energies computed in both cases. Global minima are found at $\Delta G_{effective}$ = −22.23 and −24.68 kcal mol^{-1} for these trajectories, respectively; the corresponding L-alanine positions belong to regions H and E (Figures S6 and 2E and F). The time series of $\Delta G_{effective}$ values identify these cases as singletons, however, suggesting that the energy wells associated with these minima are narrow and that, accordingly, adverse contributions to binding due to changes in the configurational entropy should

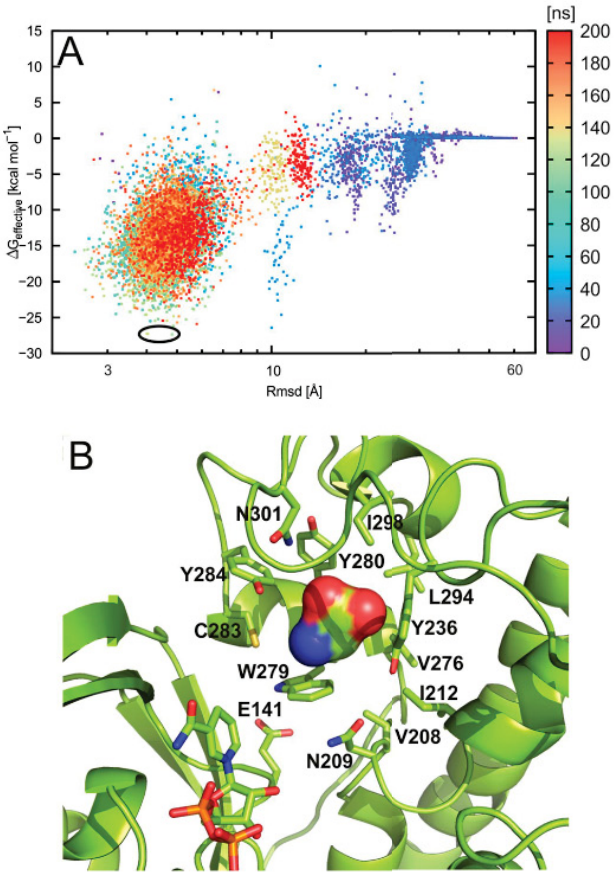


Figure 3. Effective binding energy calculations and predicted binding region. (A) Effective energies (i.e., the sum of gas-phase and solvation free energies) of L-alanine binding to AlaDH*Am* calculated by the MM-GBSA approach (see Supporting Information for details) along the trajectory 1 as a function of the L-alanine rmsd with respect to the starting structure. The time evolution along the trajectory is color coded from blue (0 ns) to red (200 ns) (see also color scale). The black circle indicates the three most favorable and energetically indistinguishable AlaDH*Am*/L-alanine configurations belonging to region G depicted in Figure 2, panels A−D (see also SI Figure S5). A second minimum at ∼10 Å rmsd refers to binding in region E. (B) Close-up view of the binding region with one of the three most favorable and energetically indistinguishable AlaDH*Am*/L-alanine configurations (see Figure S5B for the full structure) obtained from MD simulation 1. The bound L-alanine is depicted by a surface representation to indicate its residual configurational fluctuations of ∼3 Å. NADH is depicted as sticks as are residues surrounding the binding region and/or involved in enzymatic function. The side chain of Y304 has been omitted for clarity. Label numbers refer to the AlaDH*Am* sequence.

be pronounced (Figure S6). Furthermore, these global minima are at least 2.5 kcal mol^{-1} higher than those found in MD simulation 1. Thus, in addition to showing the most favorable effective energy of binding found in all MM-GBSA calculations, region G is also most frequently populated across all three MD simulations. These two independent results strongly suggest that region G is the substrate-binding region of AlaDH*Am* (Figures 3B and SI Figure S5B and C).

Figure 3B reveals that L-alanine is accommodated in a pocket mainly formed by residues Y236, V276, W279, Y280, Y284, L294, N301, and Y304 of domain II, five of which are strictly conserved across OcDH, CENDH, and AlaDH*Am* (SI Figure

dx.doi.org/10.1021/ci400370y | J. Chem. Inf. Model. 2013, 53, 2493−2498

S1). L-Alanine sits with its amino moiety "above" W279, allowing for favorable cation-π interactions, thereby pointing with its amino moiety toward a putative binding region of pyruvate; W279 itself is stabilized by weak hydrogen bond interactions with E141 of domain I (occupancy along the trajectory: 16%). E141 and W279 both have been found to be involved in L-arginine binding in OcDH, too.[7] However, in OcDH, the guanidinium moiety of arginine is placed "below" W279 (SI Figure S3); this position is sterically precluded for L-alanine in AlaDH*Am* by the inserted N209, however. Thus, sequence position 209 is decisive in determining the substrate specificity of AlaDH*Am*. In OcDH, which lacks position 209, this role is taken over by Y208, which generates a large binding site for the L-arginine side chain.[11] Accordingly, in other OpDHs lacking N209, aspartate, arginine, lysine, or tyrosine are found at position 208 depending on the respective amino acid substrate, while almost exclusively valine is found at this position in OpDHs that do have an N209 insertion.[11]

The carboxylate moiety of L-alanine is surrounded by the polar groups of Y280, Y284, N301, and Y304. However, in neither case are strong hydrogen bonds formed as judged from the distances (3.5−4.5 Å); this is in line with the observation of residual mobility of L-alanine when bound to region G (see above). Finally, C_β of L-alanine points to a wall of aromatic and aliphatic carbons in close proximity. This may explain why larger amino acids such as L-cysteine, L-serine, L-threonine, or L-valine cannot bind to AlaDH*Am* (see above). In turn, glycine binding may lack the hydrophobic interactions formed by the methyl group of L-alanine, which may explain why the K_m is ~50-fold higher for glycine than for L-alanine (Table 1).

The binding of L-alanine is accompanied by a rotation of domain II toward the NADH binding domain (domain I; SI Figure S7A). After superimpositioning domain I, this results in an rmsd of the C_α atoms of domain II of ~3.7 Å with respect to the starting structure. A similar closing movement has been observed when L-arginine binds to an OcDH/NADH complex[11−13] and an even more pronounced closing after pyruvate binding to OcDH/NADH.[7] A bound pyruvate together with the additional closing movement can be expected to restrict the residual mobility observed for L-alanine when bound to region G (see above), which may otherwise hamper an efficient catalysis.

The closing movement of domain II also leads to an enclosed binding region of L-alanine (SI Figure S7B). This together with the (un)binding pathway of L-alanine toward (from) there via the regions E and F (Figure 2A−F) can also provide an explanation at an atomic level as to the observed substrate inhibition of AlaDH*Am*. Assuming that after the enzymatic reaction alanopine needs to escape the binding region on the same pathway as L-alanine accesses it, this escape will be the more hampered the more L-alanine occupies the energetically favorable (Figures 3A and S5A) regions E and F. The assumption of the same access and escape pathways seems to be justified particularly for the transition between regions F and G, given the narrow, gorge-like character of that region (SI Figure S7B). This narrowness has also been implicated[7] as the reason for the observed binding of substrates to OcDH in a sequential, ordered manner (first L-arginine, then pyruvate).[12,13] Capturing AlaDH*Am* in an alanopine bound state that way would result in a decrease of the enzymatic reaction velocity that becomes more pronounced with increasing L-alanine concentration.

In summary, we presented an initial biochemical characterization of AlaDH*Am*, which catalyzes the reductive condensation of L-alanine with pyruvate to alanopine. AlaDH*Am* displays a high catalytic efficiency and substrate specificity, although it is prone to substrate inhibition. Unbiased MD simulations captured the binding of L-alanine diffusing from solvent to the putative binding region. This binding region is located at the helix-kink-helix motif, as observed for binding of L-arginine to OcDH from *P. maximus*. At the same time, the observed binding of L-alanine provides an explanation for the role of amino acids 208 and 209 in substrate specificity, the only amino acids within the binding region that differ between OpDHs with different substrates. Finally, the presence of energetically favorable non-native ligand binding states in the vicinity of the binding region can provide an explanation for the substrate inhibition of AlaDH*Am*.

■ ASSOCIATED CONTENT

Supporting Information

Material and methods, results on estimating changes in the configurational entropy upon binding, and figures of the alignment of sequences of AlaDH*Am* and the three templates, Coomassie stained SDS-PAGE of AlaDH*Am* from different purification steps, modeled AlaDH*Am*/NADH/L-alanine starting structure vs OcDH/NADH/L-arginine crystal structure, structural deviations during the MD simulations, effective binding energies and energetically most favorable AlaDH*Am*/L-alanine configurations, time course of effective energies of binding of L-alanine to AlaDH*Am* for MD simulations 2 and 3, and movement of domain II of AlaDH*Am* relative to domain I in the course of L-alanine binding. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Phone: +49(0)221-81-13662. E-mail: gohlke@hhu.de (H.G.).
*Phone: +49(0)221-81-10773. E-mail: lutz.schmitt@hhu.de (L.S.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge support (and training) from the International NRW Research School BioStruct, granted by the Ministry of Innovation, Science and Research of the State North Rhine-Westphalia, the Heinrich-Heine-University of Düsseldorf (HHU), and the Entrepreneur Foundation at the HHU. H.G. is grateful to the initiative "Fit for Excellence" of the HHU for financial support and to the "Zentrum für Informations- und Medientechnologie" (ZIM) at the HHU for providing computational support. Some of this work has been supported by the DFG (grant GR 456/23-1 to M.K.G).

■ ABBREVIATIONS

AlaDH, alanopine dehydrogenase; AlaDH*Am*, alanopine dehydrogenase from *Arenicola marina*; CENDH, *N*-(1-D-carboxylethyl)-L-norvaline dehydrogenase; MD, molecular dynamics; MM-GBSA, molecular mechanics generalized Born surface area; OpDH, opine dehydrogenase; OcDH, octopine dehydrogenase; rmsd, root mean-square deviation

REFERENCES

(1) Shan, Y. B.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* **2011**, *133* (24), 9181−9183.
(2) Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhani, D. W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D. E. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (32), 13118−23.
(3) Giorgino, T.; Buch, I.; De Fabritiis, G. Visualizing the Induced Binding of SH2-Phosphopeptide. *J. Chem. Theory. Comput.* **2012**, *8* (4), 1171−1175.
(4) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (25), 10184−10189.
(5) Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Wess, J.; Kobilka, B. K. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **2012**, *482* (7386), 552−6.
(6) Zebe, E.; Schiedeck, D. The lugworm Arenicola marina: a model of physiological adaptation to life in intertidal sediments. *Helgoländer Meeresunters* **1996**, *30*, 37−68.
(7) Grieshaber, M. K.; Hardewig, I.; Kreutzer, U.; Portner, H. O. Physiological and metabolic responses to hypoxia in invertebrates. *Rev. Physiol. Biochem. Pharmacol.* **1994**, *125*, 43−147.
(8) van Thoai, N.; Huc, C.; Pho, D. B.; Olomucki, A. Octopine dehydrogenase. Purification and catalytic properties. *Biochim. Biophys. Acta* **1969**, *191* (1), 46−57.
(9) Pho, D. B.; Olomucki, A.; Huc, C.; Thoai, N. V. Spectrophotometric studies of binary and ternary complexes of octopine dehydrogenase. *Biochim. Biophys. Acta* **1970**, *206* (1), 46−53.
(10) Muller, A.; Janssen, F.; Grieshaber, M. K. Putative reaction mechanism of heterologously expressed octopine dehydrogenase from the great scallop, Pecten maximus (L). *Febs. J.* **2007**, *274* (24), 6329−6339.
(11) Smits, S. H.; Mueller, A.; Schmitt, L.; Grieshaber, M. K. A structural basis for substrate selectivity and stereoselectivity in octopine dehydrogenase from Pecten maximus. *J. Mol. Biol.* **2008**, *381* (1), 200−11.
(12) Smits, S. H.; Meyer, T.; Mueller, A.; van Os, N.; Stoldt, M.; Willbold, D.; Schmitt, L.; Grieshaber, M. K. Insights into the mechanism of ligand binding to octopine dehydrogenase from Pecten maximus by NMR and crystallography. *PLoS One* **2010**, *5* (8), e12312.
(13) van Os, N.; Smits, S. H. J.; Schmitt, L.; Grieshaber, M. K. Control of D-octopine formation in scallop adductor muscle as revealed through thermodynamic studies of octopine dehydrogenase. *J. Exp. Biol.* **2012**, *215* (9), 1515−1522.
(14) Siegmund, B.; Grieshaber, M.; Reitze, M.; Zebe, E. Alanopine and Strombine Are End Products of Anaerobic Glycolysis in the Lugworm, Arenicola-Marina-L (Annelida, Polychaeta). *Comp. Biochem. Physiol. B: Biochem. Mol. Biol.* **1985**, *82* (2), 337−345.
(15) Siegmund, B.; Grieshaber, M. K. Determination of meso-alanopine and D-strombine by high pressure liquid chromatography in extracts from marine invertebrates. *Hoppe Seylers Z. Physiol. Chem.* **1983**, *364* (7), 807−12.
(16) Britton, K. L.; Asano, Y.; Rice, D. W. Crystal structure and active site location of N-(1-D-carboxylethyl)-L-norvaline dehydrogenase. *Nat. Struct. Biol.* **1998**, *5* (7), 593−601.
(17) Gade, G.; Grieshaber, M. K. Pyruvate Reductases Catalyze the Formation of Lactate and Opines in Anaerobic Invertebrates. *Comp. Biochem. Physiol. B: Biochem. Mol. Biol.* **1986**, *83* (2), 255−272.
(18) Endo, N.; Kan-No, N.; Nagahisa, E. Purification, characterization, and cDNA cloning of opine dehydrogenases from the polychaete rockworm Marphysa sanguinea. *Comp. Biochem. Physiol. B: Biochem. Mol. Biol.* **2007**, *147* (2), 293−307.
(19) Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234* (3), 779−815.
(20) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403−410.
(21) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(22) Gohlke, H.; Kiel, C.; Case, D. A. Insights into protein-protein binding by free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J. Mol. Biol.* **2003**, *330*, 891−913.
(23) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **1995**, *21* (3), 167−195.
(24) Wang, J.; Verkhivker, G. M. Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. *Phys. Rev. Lett.* **2003**, *90* (18), 188101.
(25) Tsai, C.-J.; Kumar, S.; Ma, B.; Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci.* **1999**, *8*, 1181−1190.

Supporting Information

Binding region of alanine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion

Holger Gohlke,[1] Ulrike Hergert,[2] Tatu Meyer,[2] Daniel Mulnaes,[1] Manfred K. Grieshaber,[2] Sander H.J. Smits,[2] and Lutz Schmitt[2]

[1]Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine-University, Düsseldorf, Germany

[2] Institute for Biochemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

Supplemental Material & Methods

Cloning and expression of alanopine dehydrogenase from *Arenicola marina* (AlaDH*Am*)

For large scale expression, *E. coli* cells ER2566 harboring the plasmid pTYB1-AlaDH*Am*, introducing a hexa-histidine tag at the C-terminus, were grown at 37°C in 8 liter LB medium containing 100 µg mL-1 ampicillin until an OD_{600} of 0.6 was reached. Expression of AlaDH*Am*-His6 was induced by adding 0.2 mM IPTG, and cultivation continued at 18°C for 24 hours. Cells were harvested, resuspended in lysis buffer (50 mM Na-phosphate buffer pH 8.0 containing 300 mM NaCl and 10 mM imidazole) and disrupted by sonification (Bandelin, Berlin, Germany). After centrifugation for 60 min at 22.000g and 4°C, the supernatant was subjected to a Ni^{2+} nitrilo-triacetic acid column (Ni-NTA, Qiagen, Hilden, Germany). Unbound proteins were washed off the column with a Na-phosphate buffer (pH 8.0, 300 mM NaCl, 10 mM imidazole) followed by a second washing step with the same buffer, but including 20 mM imidazole. Bound AlaDH*Am* was eluted using a linear gradient ranging from 40 to 250 mM imidazole. Fractions with the highest AlaDH*Am* activity were pooled and concentrated using an Amicon cell with YM-10 filter membranes (10.000 MWCO; Millipore, Eschborn, Germany). The resulting solution was applied on a Sephadex G-100 column, equilibrated with 50 mM K-phosphate buffer pH 7.5 containing 2 mM EDTA, 10% (v/v) glycerol, and 0.1% (v/v) mercapto-ethanol. From 1 l of cell culture 3.2 mg homogenous AlaDH*Am* was obtained.

Homology modeling of the AlaDH*Am* structure

Protein structures to be used as templates in the homology modeling were searched by pBLAST [1] on the Protein Data Base.[2] Requiring a sequence coverage and identity > 20%, a resolution < 3Å, and an expectation value < 10^{-3} yielded three templates with PDB codes 3C7A (chain A), 3IQD (chain B), and 1BG6 (chain A). The template sequences were aligned repetitively three times using structural information by the Modeller [3] salign class (Figure S1). Each re-alignment takes into account more features of the template structures and had 1D gap-penalties for initialization and extension of -450 and -50 and an rmsd cut-off of 3.5 Å.

Then the AlaDH*Am* sequence is aligned to the template alignment by using a progressive pairwise alignment with 1D initialization and extension penalties of -450 and 0. The 2D penalties were set to 0.35 for sequence/structure, 1.2 for α-helix, 0.9 for β-sheet, 1.2 for

sidechain accessibility, 0.6 for straightness, 8.6 for pairwise Cα-Cα distances, 1.2 for local conformation and 0 otherwise.

For generating homology models, Modeller's automodel class is used with a three-fold slow refinement for 300 iterations with a molpdf ("energy" computed by Modeller's objective function) limit of 10^6. Then the loops in each model are refined using the dope_loopmodel [4] class with the slow md_level, giving 25 models in total.

Each model is processed using Anolea [5], Dope [6], and Procheck [7] to determine structure quality. For Procheck a residue-wise score is calculated as the maximum deviation of dihedral-angles, bond-angles and bond-lengths from optimal values for that residue. To compare each model to the rest, an "Ensemble Z-score" is calculated for each score. These residue-wise Ensemble Z-scores are averaged over a window of nine residues and divided by the standard deviation of the model to avoid bias between score types. In the same way a "Template Z-score" is calculated, using the template ensemble to compare model residues to the aligned template residues. "Composite Scores" are then computed as the average of the respective Z-scores, showing the quality of the modelling of each residue compared to the reference (ensemble structures or templates). "Global Scores" are computed as the residue-average of the Composite Scores for the respective reference, and the "best" model is defined as the one with the lowest average of the Global Ensemble and Global Templates Score. This model is characterized by an average Anolea energy per residue of -3.487, a global Dope score of -0.9386, and a Ramachandran plot with 92.9% of the residues in the core region, 5.6% in the allowed region, 1.2% in the generously allowed region, and 0.3% in the disallowed region. This model was used for further generation of the AlaDH*Am*/L-alanine complex structure.

Generation of the AlaDH*Am*/NADH/L-alanine starting structure

For generation of the starting structure for the molecular dynamics (MD) simulations, the modeled AlaDH*Am* structure was root mean-square fitted onto the octopine dehydrogenase (OcDH) structure from PDB code 3C7C. Coordinates of the NADH of the latter structure where copied without steric clashes to the AlaDH*Am* model. The L-alanine ligand for AlaDH*Am* was initially modeled from the bound L-arginine in the OcDH structure. Keeping all residues of the AlaDH*Am* model fixed but the sidechains immediately surrounding the L-alanine ligand, the complex was then minimized using the MAB force field [8] as implemented in *Moloc*; this resulted in a shift of L-alanine with respect to the backbone region of L-

arginine of ~3 Å rmsd (Figure S3; see also main text for details). This complex structure was used as a starting structure for the MD simulations.

MD simulations of the AlaDH*Am*/NADH/L-alanine complex

The MD simulations were performed with the AMBER 11 suite of programs [9] using the GPU accelerated code of *pmemd* [10] together with the ff99SB force field.[11, 12] Atomic charges for L-alanine in the zwitterionic form were derived by the RESP procedure;[13] bonded and nonbonded parameters for NADH were taken from refs. [14, 15]. The complex consisting of AlaDH*Am*, NADH, and L-alanine was neutralized by adding Na^+ counter ions, and the systems were then solvated in a truncated octahedral periodic box of TIP3P water [16] with a distance between the edges of the box and the closest solute atom of at least 11 Å. This resulted in system sizes of ~$5*10^4$ atoms.

The particle mesh Ewald (PME) method was used to treat long-range electrostatic interactions [17], and a direct-space non-bonded cutoff of 8 Å was applied. Bond lengths involving bonds to hydrogen atoms were constrained using SHAKE.[18] The time-step for all MD simulations was 2 fs. The systems were initially minimized by 500 steps applying harmonic restraints with force constants of at least 5 kcal mol^{-1} $Å^{-2}$ to all solute atoms. Applying harmonic restraints with force constants of 5 kcal mol^{-1} $Å^{-2}$ to all solute atoms, NVT-MD was carried out for 50 ps, during which the system was heated from 100 K to 300 K. Subsequent NPT-MD was used for 150 ps to adjust the solvent density. Finally, the force constants of the harmonic restraints on solute atom positions were gradually reduced to zero during 100 ps of NVT-MD. Three NVT-MD simulations for production were spawned at that point, at 300.0, 300.1, and 300.2 K, respectively. Trajectories of 200 ns length were generated with conformations extracted every 20 ps for analysis. Structural analyses were performed with the program *ptraj* of the AMBER 11 suite. All figures and images were generated by *gnuplot* [19] and *pymol*.[20]

Calculation of effective binding energies

For calculation of effective energies (i.e., the sum of gas-phase and solvation free energies) of binding, the single-trajectory MM-GBSA approach was employed via the *mm_pbsa.pl* script on the 10^4 conformations extracted from MD simulation 1.[21-23] All water molecules were deleted as were all counter ions. Gas-phase energies (MM) were calculated by summing up contributions from internal energies, electrostatic energies, and van der Waals energies using

the ff99SB [11, 12] force-field with no cutoff. Solvation free energies were computed as the sum of polar and non-polar contributions. The polar contribution was calculated using the GB^{OBC} generalized Born model [24] as implemented in AMBER 11 (*igb = 5*) together with mbondi2 dielectric radii and a concentration of 1-1 mobile counterions of 100 mM. The nonpolar contribution was calculated by a solvent-accessible surface area (SASA)-dependent term using a value of 0.0072 kcal mol^{-1} $Å^{-2}$ for the surface tension and a zero offset. The SASA was determined with the LCPO method [25] implemented in AMBER 11.

Supplemental Results

Estimating changes in the configurational entropy upon binding

Applying the rigid-rotor harmonic oscillator approximation commonly used in the context of end-point free energy calculations [22, 26], we estimated contributions to the binding at $T = 300$ K due to changes in the configurational entropy in the following way: I) Assuming a bound volume $V_b \approx 3^3$ $Å^3$ from the configurational fluctuations observed for the ligand when bound to region G (see main text) and considering the standard-state volume $V^0 = 1661$ $Å^3$ at 1 M concentration, an adverse effect to binding due to the loss in the translational entropy of the ligand of $T\Delta S = RT \ln(V_b / V^0) = -2.5$ kcal mol^{-1} is estimated;[27] II) the adverse contribution due to the loss in the rotational entropy of the ligand can be at most $T\Delta S = -1.2$ kcal mol^{-1}, as the absolute value equals the rotational entropy contribution of the ligand in the gas-phase as determined from classical statistical thermodynamics;[28] III) although probably the most difficult to determine,[22, 26] we think that it is safe to assume that contributions to the binding due to changes in the vibrational entropy of the solutes will be small here, given that L-alanine is rather rigid and does not make strong interactions with AlaDH*Am* (see below).

Supplemental Figures

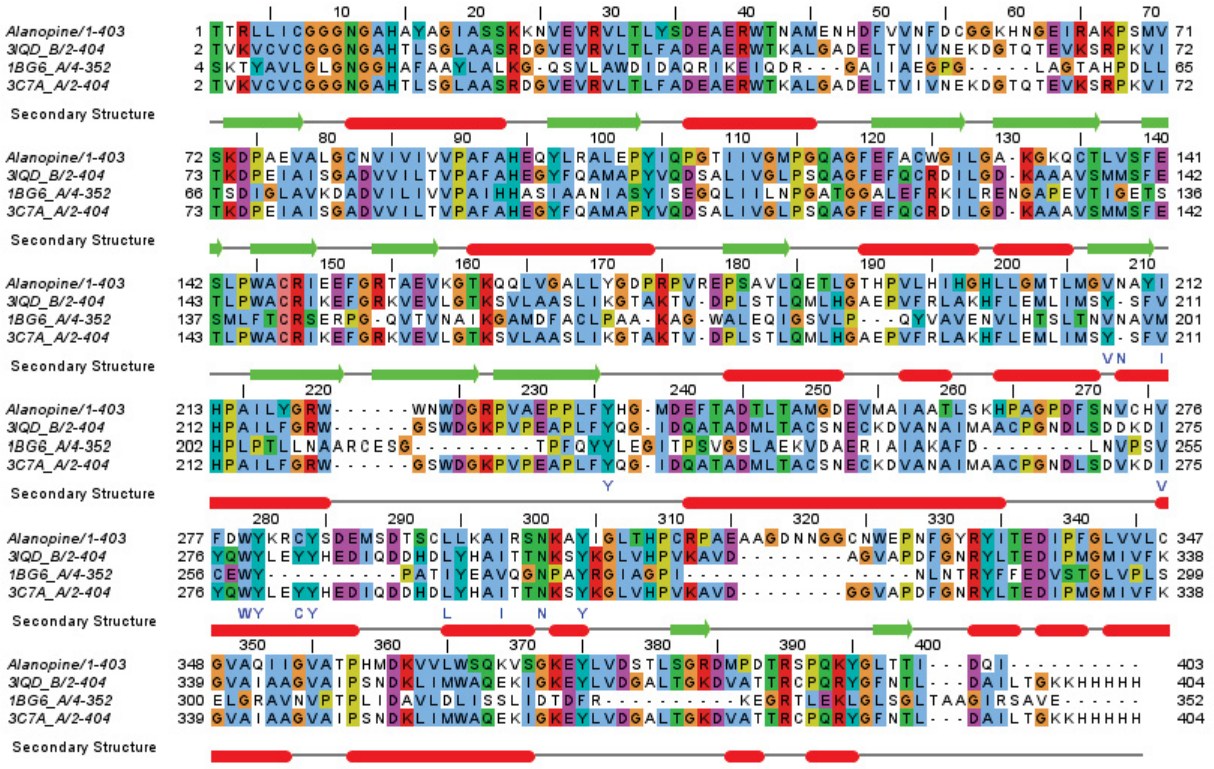


Figure S1: Alignment of sequences of AlaDH*Am* and the three templates.
See chapter "Homology modeling of the alanine dehydrogenase structure" for how the alignment was generated. Red bars and green arrows indicate α-helices and β-strands, respectively, of the AlaDH*Am* model generated as determined by DSSP.[29] Numbers provided on the left and right refer to positions in the respective sequences; numbers provided on top refer to the positions in the AlaDH*Am* sequence. The amino acids are colored according to the ClustalW criteria in Jalview (orange: G; yellow: P; cyan: H and Y; blue: hydrophobic amino acids (A, I, L, M, F, W, V, C); green: polar amino acids (N, Q, S, T); red: positively charged amino acids (K, R); magenta: negatively charged amino acids (D, E)) if the amino acid profile of the alignment at that position meets a minimum criterion specific for the residue type.

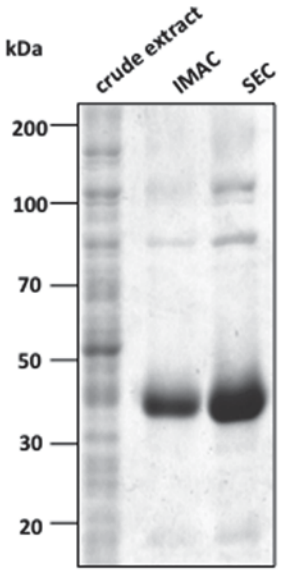


Figure S2: Coomassie stained SDS-PAGE of AlaDH*Am* from different purification steps.
Loaded on the SDS gel are the crude extract, pooled elution from the IMAC column, and elution fraction from the size exclusion chromatography column.

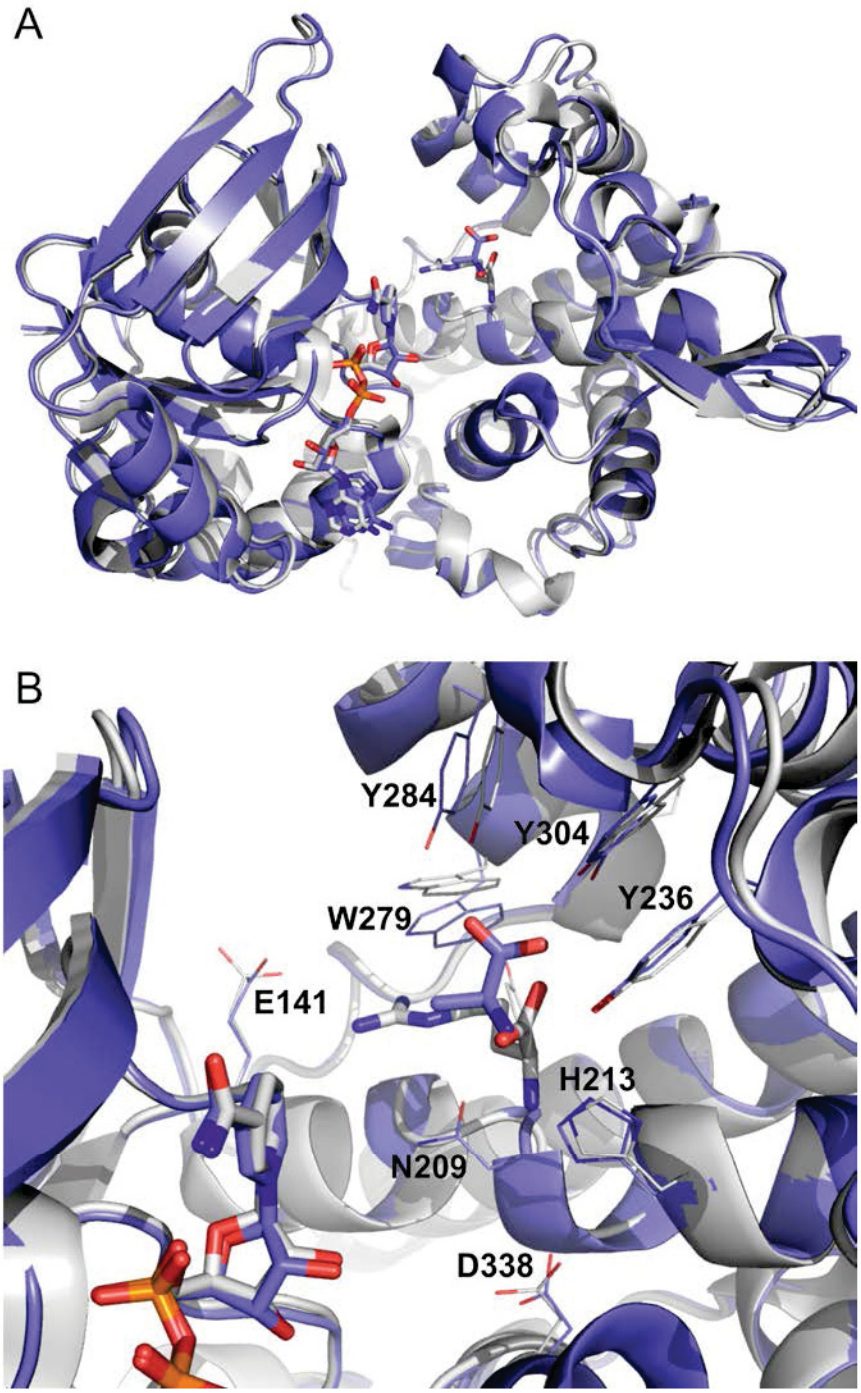


Figure S3: Modeled AlaDH*Am*/NADH/L-alanine starting structure vs. OcDH/NADH/L-arginine crystal structure.

A, B: Overlay of the modeled starting structure of L-alanine bound to AlaDH*Am* (blue) and L-arginine bound to OcDH (white; PDB code: 3C7C);[30] the bound amino acids and NADH are depicted as sticks. Panel B is a close-up view of the binding region with residues discussed in the text depicted as lines. Label numbers refer to the AlaDH*Am* sequence.

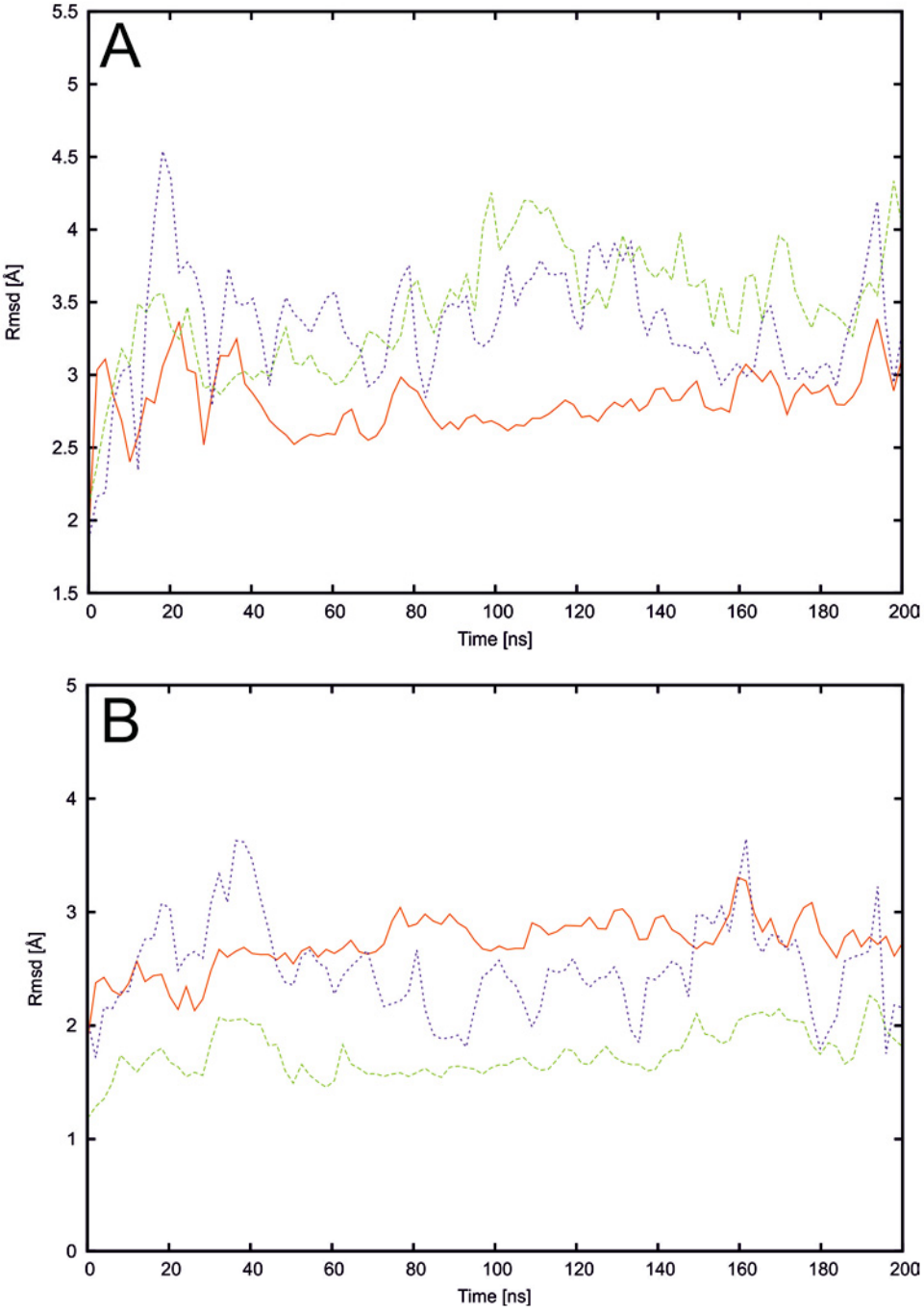


Figure S4: Structural deviations during the MD simulations.
A: Root mean-square deviations (rmsd) of the C_α atoms of AlaDH*Am* during the course of the MD simulations 1 (red), 2 (green), and 3 (blue) with respect to the modeled starting structure (see Figure S3) after superimposing AlaDH*Am* based on its C_α atoms. **B:** Rmsd of the C_α atoms of the C-terminal domain (red) and the N-terminal domain (green) of AlaDH*Am* as well as of all atoms of NADH (blue) during the course of the MD simulation 1. For computing the rmsd of the domains, the domains were superimposed onto themselves, respectively; the rmsd of NADH was computed after superimposing the protein conformations.

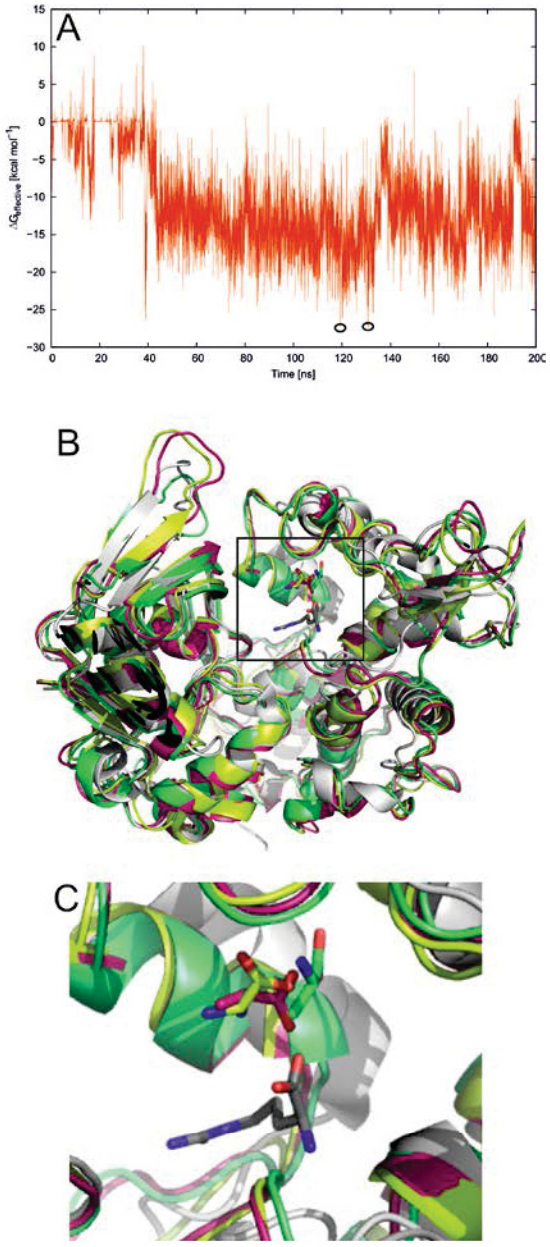


Figure S5: Effective binding energies and energetically most favorable AlaDH*Am*/L-alanine configurations.

A: Time course of effective energies of binding of L-alanine to AlaDH*Am* during the MD simulation 1. The two black circles indicate the three most favorable and energetically indistinguishable AlaDH*Am*/L-alanine configurations belonging to region G in panels A-D of Figure 2 (see also Figure 3A). **B:** The three most favorable and energetically indistinguishable AlaDH*Am*/L-alanine configurations (lime, green, red) as identified from panel A (see also Figure 3A); a close-up view of the lime structure is shown in Figure 3B. In white, the OcDH/L-arginine complex crystal structure (PDB code 3C7C) [30] is shown for comparison. The L-alanine and L-arginine ligands are depicted as sticks. **C:** Blowup of the region marked by the black square in panel B.

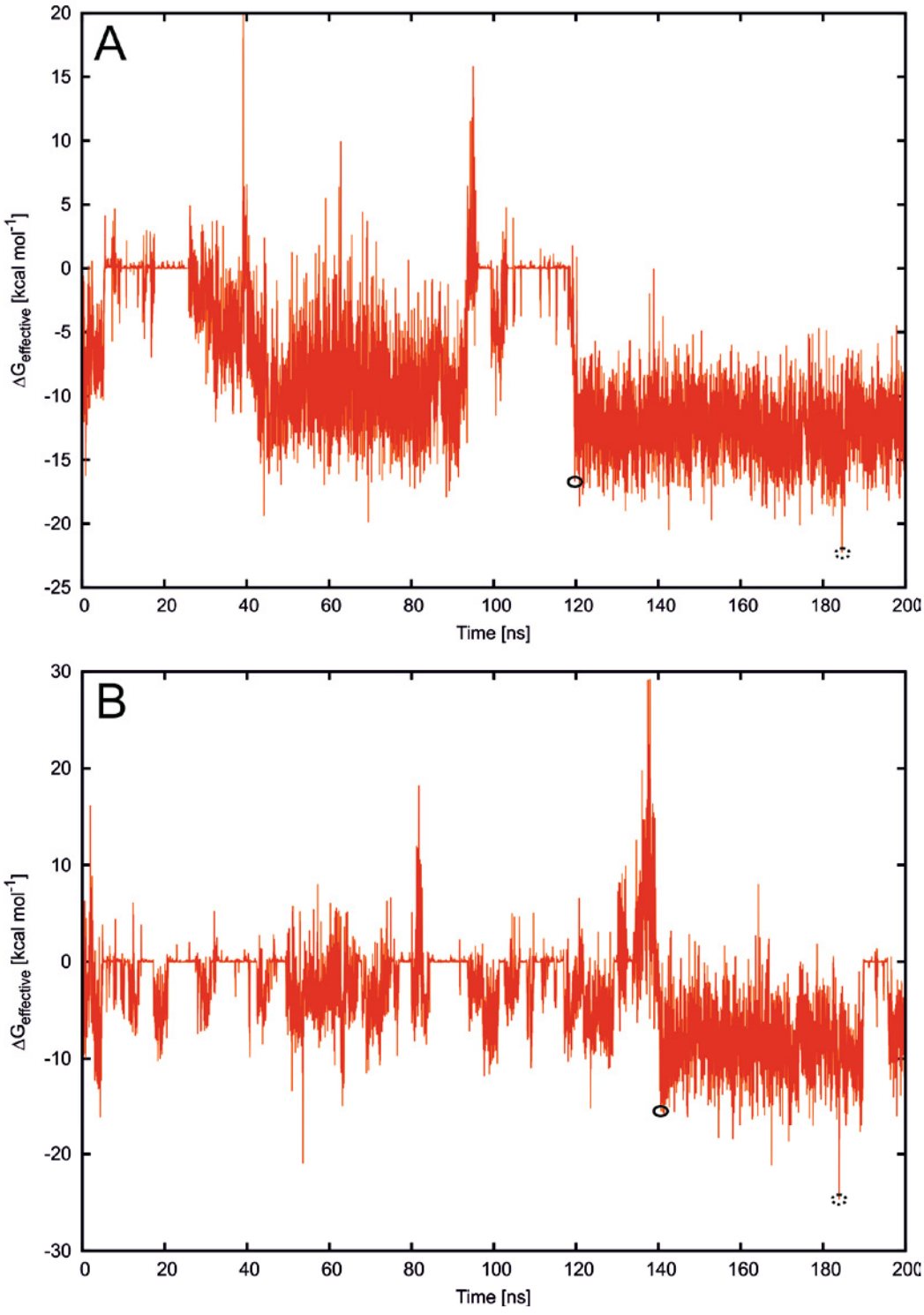


Figure S6: Time course of effective energies of binding of L-alanine to AlaDH*Am*.
A: The effective energies of binding were calculated from MD simulation 2. **B:** The effective energies of binding were calculated from MD simulation 3. In both panels A and B, the black circles with straight lines indicate the most favorable effective energy found for L-alanine binding to region G (see also Figure 2E, F); the black circles with dashed lines mark the respective global minima.

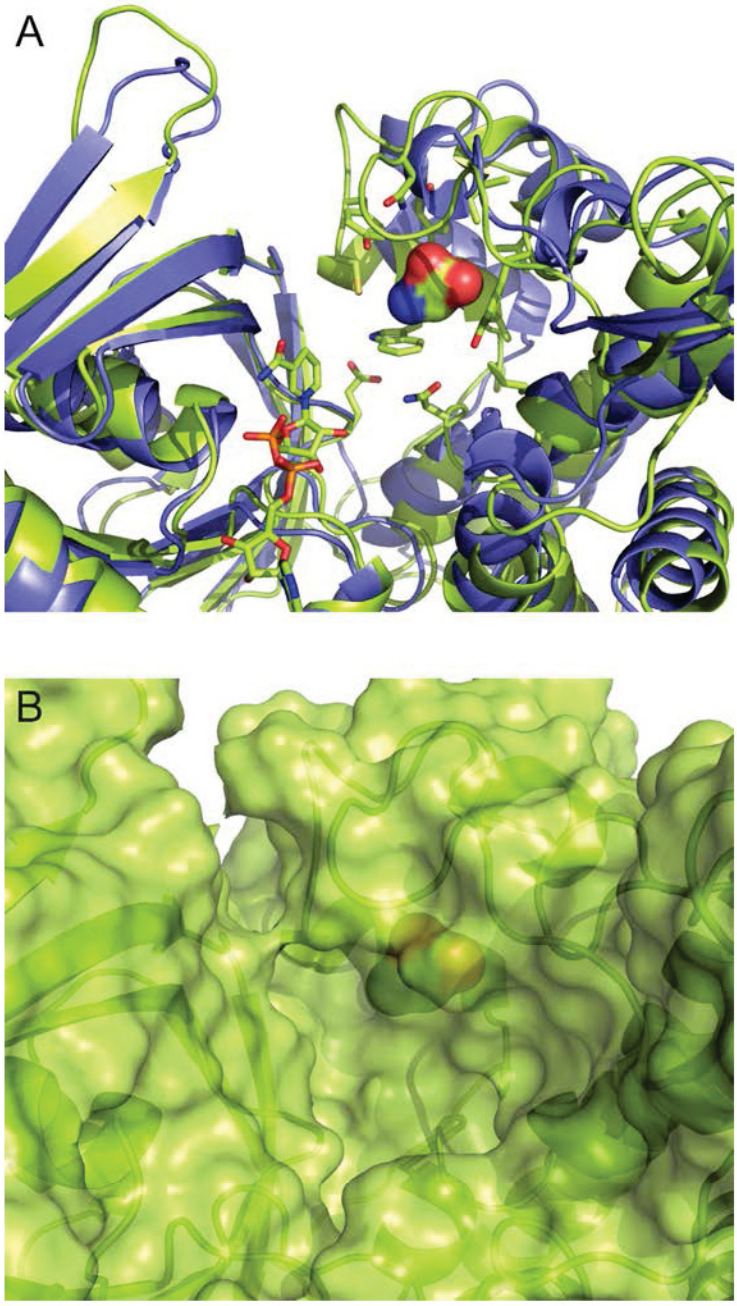


Figure S7: Movement of domain II of AlaDH*Am* relative to domain I in the course of L-alanine binding.

A: Overlay of the modeled starting structure of AlaDH*Am* (blue) and one of the three most favorable and energetically indistinguishable AlaDH*Am*/L-alanine configurations (lime) as identified from Figure S5A for MD simulation 1 (see also Figure 3A). The bound L-alanine is depicted by a surface representation. NADH is depicted as sticks as are residues surrounding the binding region and/or involved in enzymatic function (for labels see Figure 3A). The sidechain of Y304 has been omitted for clarity. **B:** One of the three most favorable and energetically indistinguishable AlaDH*Am*/L-alanine configurations as in panel A except that now the protein including NADH are represented with a transparent molecular surface. The bound L-alanine (dark blob at the center) is occluded by Y304.

Supplemental References

1. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990,** *215* (3), 403-410.

2. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28* (1), 235-242.

3. Sali, A.; Blundell, T. L., Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993,** *234* (3), 779-815.

4. Fiser, A.; Do, R. K. G.; Sali, A., Modeling of loops in protein structures. *Protein Sci.* **2000,** *9* (9), 1753-1773.

5. Melo, F.; Feytmans, E., Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.* **1998,** *277* (5), 1141-1152.

6. Shen, M. Y.; Sali, A., Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006,** *15* (11), 2507-24.

7. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M., PROCHECK: A program to check the stereochemical quality of protein structures. *J Appl Cryst* **1993,** *26*, 283-291.

8. Gerber, P. R.; Müller, K., MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J. Comput. Aided Mol. Des.* **1995,** *9*, 251-268.

9. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005,** *26* (16), 1668-1688.

10. Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput* **2012,** *8* (5), 1542-1555.

11. Cornell, W. D.; Cieplak, C. I.; Bayly, I. R.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995,** *117*, 5179-5197.

12. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006,** *65* (3), 712-25.

13. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem. B* **1993,** *97*, 10269-10280.

14. Walker, R. C.; de Souza, M. M.; Mercer, I. P.; Gould, I. R.; Klug, D. R., Large and fast relaxations inside a protein: Calculation and measurement of reorganization energies in alcohol dehydrogenase. *J. Phys. Chem. B* **2002,** *106* (44), 11658-11665.

15. Pavelites, J. J.; Gao, J. L.; Bash, P. A.; Mackerell, A. D., A molecular mechanics force field for NAD(+), NADH, and the pyrophosphate groups of nucleotides. *J. Comput. Chem.* **1997,** *18* (2), 221-239.

16. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983,** *79* (2), 926-935.

17. Cheatham, T. E., III; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A., Molecular dynamics simulations on solvated biomolecular systems: The Particle Mesh Ewald Method leads to stable trajectories of DNA, RNA, and proteins. *J Am Chem Soc* **1995,** *117*, 4193-4194.

18. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977,** *23* (3), 327-341.

19. Williams, T.; Kelley, C., gnuplot 4.2. **2009**.

20. PyMOL *The PyMOL Molecular Graphics System, Version 1.3r1, Schrödinger, LLC*, 2010.

21. Gohlke, H.; Kiel, C.; Case, D. A., Insights into protein-protein binding by free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J. Mol. Biol.* **2003,** *330*, 891-913.

22. Gohlke, H.; Case, D. A., Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* **2004,** *25*, 238-250.

23. Homeyer, N.; Gohlke, H., Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method. *Molecular Informatics* **2012,** *31* (2), 114-122.

24. Onufriev, A.; Bashford, D.; Case, D. A., Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B.* **2000,** *104*, 3712-3720.

25. Weiser, J.; Shenkin, P. S.; Still, W. C., Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comput. Chem.* **1999,** *20*, 217-230.

26. Homeyer, N.; Gohlke, H., Free energy calculations by the Molecular Mechanics Poisson−Boltzmann Surface Area method. *Molecular Informatics* **2012,** *31*, 114-122.

27. Buch, I.; Giorgino, T.; De Fabritiis, G., Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2011,** *108* (25), 10184-10189.

28. McQuarrie, D. A., *Statistical mechanics*. Harper & Row: New York, 1976.

29. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-637.

30. Smits, S. H.; Mueller, A.; Schmitt, L.; Grieshaber, M. K., A structural basis for substrate selectivity and stereoselectivity in octopine dehydrogenase from Pecten maximus. *J. Mol. Biol.* **2008,** *381* (1), 200-11.

20. PUBLICATION IV

Determinants of FIV and HIV Vif sensitivity of feline APOBEC3 restriction factors

Zeli Zhang[1], Qinyong Gu[1], Ananda Ayyappan Jaguva Vasudevan[1], Anika Hain[1], Björn-Philipp Kloke[2,8], Sascha Hasheminasab[1], Daniel Mulnaes[4], Kei Sato[5,6], Klaus Cichutek[2], Dieter Häussinger[1], Ignacio G. Bravo[7], Sander H. J. Smits[3], Holger Gohlke[4] and Carsten Münk[1]

[1]Clinic for Gastroenterology, Hepatology, and Infectiology, Medical Faculty,

Heinrich-Heine University Düsseldorf, Germany.

[2]Department of Medical Biotechnology, Paul-Ehrlich-Institute, Langen, Germany.

[3]Institute of Biochemistry, Heinrich Heine University Düsseldorf, Germany.

[4]Institute of Pharmaceutical and Medicinal Chemistry,

Heinrich-Heine University Düsseldorf, Germany.

[5]Laboratory of Viral Pathogenesis, Institute for Virus Research,

Kyoto University, Kyoto, Japan.

[6]CREST, Japan Science and Technology Agency, Saitama, Japan.

[7]National Center of Scientific Research (CNRS), Montpellier, France.

[8]BioNTech RNA Pharmaceuticals GmbH, Mainz, Germany.

RESEARCH Open Access

Determinants of FIV and HIV Vif sensitivity of feline APOBEC3 restriction factors

Zeli Zhang[1], Qinyong Gu[1], Ananda Ayyappan Jaguva Vasudevan[1], Anika Hain[1], Björn-Philipp Kloke[2,8], Sascha Hashemínasab[1], Daniel Mulnaes[4], Kei Sato[5,6], Klaus Cichutek[2], Dieter Häussinger[1], Ignacio G. Bravo[7], Sander H. J. Smits[3], Holger Gohlke[4] and Carsten Münk[1*]

Abstract

Background: Feline immunodeficiency virus (FIV) is a global pathogen of Felidae species and a model system for Human immunodeficiency virus (HIV)-induced AIDS. In felids such as the domestic cat (*Felis catus*), APOBEC3 (A3) genes encode for single-domain A3Z2s, A3Z3 and double-domain A3Z2Z3 anti-viral cytidine deaminases. The feline A3Z2Z3 is expressed following read-through transcription and alternative splicing, introducing a previously untranslated exon in frame, encoding a domain insertion called linker. Only A3Z3 and A3Z2Z3 inhibit Vif-deficient FIV. Feline A3s also are restriction factors for HIV and Simian immunodeficiency viruses (SIV). Surprisingly, HIV-2/SIV Vifs can counteract feline A3Z2Z3.

Results: To identify residues in feline A3s that Vifs need for interaction and degradation, chimeric human–feline A3s were tested. Here we describe the molecular direct interaction of feline A3s with Vif proteins from cat FIV and present the first structural A3 model locating these interaction regions. In the Z3 domain we have identified residues involved in binding of FIV Vif, and their mutation blocked Vif-induced A3Z3 degradation. We further identified additional essential residues for FIV Vif interaction in the A3Z2 domain, allowing the generation of FIV Vif resistant A3Z2Z3. Mutated feline A3s also showed resistance to the Vif of a lion-specific FIV, indicating an evolutionary conserved Vif–A3 binding. Comparative modelling of feline A3Z2Z3 suggests that the residues interacting with FIV Vif have, unlike Vif-interacting residues in human A3s, a unique location at the domain interface of Z2 and Z3 and that the linker forms a homeobox-like domain protruding of the Z2Z3 core. HIV-2/SIV Vifs efficiently degrade feline A3Z2Z3 by possible targeting the linker stretch connecting both Z-domains.

Conclusions: Here we identified in feline A3s residues important for binding of FIV Vif and a unique protein domain insertion (linker). To understand Vif evolution, a structural model of the feline A3 was developed. Our results show that HIV Vif binds human A3s differently than FIV Vif feline A3s. The linker insertion is suggested to form a homeo-box domain, which is unique to A3s of cats and related species, and not found in human and mouse A3s. Together, these findings indicate a specific and different A3 evolution in cats and human.

Keywords: APOBEC3, FIV, Gene evolution, HIV, Homeobox, Homology modelling, Restriction factor, SIV, Vif

Background

APOBEC3 (A3) cytidine deaminases are anti-viral restriction factors containing either one or two zinc (Z)-binding domains found in different clade-specific gene numbers and gene arrangements in placental mammals [1–4]. For example, primates have seven genes (A3A–A3D, A3F–A3H), while cats encode four genes (A3Z2a–A3Z2c, A3Z3) [3, 5]. These A3 proteins target broadly viruses and mobile genetic elements that depend on reverse transcription, but also show antiviral activity against unrelated viruses (for recent reviews see [6, 7]). Some retroviruses express viral A3-counteracting proteins, such as Vif of lentiviruses, Bet of foamy viruses,

*Correspondence: carsten.muenk@med.uni-duesseldorf.de
[1] Clinic for Gastroenterology, Hepatology, and Infectiology, Medical Faculty, Heinrich-Heine-University Düsseldorf, Building 23.12.U1.82, Moorenstr. 5, 40225 Düsseldorf, Germany
Full list of author information is available at the end of the article

the nucleocapsid of *Human T cell leukemia virus type 1* (HTLV-1), and the glycosylated (glyco)-Gag of *Murine leukemia virus* (MLV) [8–13]. The Vif protein prevents encapsidation of host-cell derived A3 proteins into nascent viral particles. In the absence of Vif, encapsidated A3s inhibit lentiviruses during infection by deamination of cytidines in the single-stranded DNA formed during reverse transcription, by introducing G-to-A mutations in the coding strand. Additionally, some A3s inhibit virus replication by reducing reverse transcription and integration via non-editing mechanisms [14–19].

The domestic cat *Felis catus* (Fca) is the host to many diverse retroviruses, such as the lentivirus *Feline immunodeficiency virus* (FIV), gammaretroviruses of the *Feline leukemia virus* (FeLV) group, and the spumaretrovirus *Feline foamy virus* (FFV) (for reviews see [20–23]). In a small proportion of naturally infected domestic cats, FIV causes an immunodeficiency disease similar to *Human immunodeficiency virus type 1* (HIV-1)-induced AIDS [24]. However, highly pathogenic FIV isolates can cause mortality up to 60 % under experimental conditions [25–27]. Thus, FIV infection of cats is a valuable animal model to study HIV-1 and AIDS [28–30]. In addition to the domestic cat, species-specific FIVs that might cause disease in some natural hosts have been isolated in many *Felidae* [31]. FFVs replicate in domestic cats and in other *Felidae* and are not causing disease [32–34]. In contrast, FeLVs are pathogenic and induce in domestic cats serious diseases such as lymphomas and anemia [24], but are rarely found in other *Felidae* [31].

The domestic cat, and likely all other *Felidae*, encode four A3 genes, three closely related A3Z2 genes (A3Z2a, A3Z2b, A3Z2c) and one A3Z3 gene [4, 35]. Besides the four canonical A3 proteins, the cat genome can express, by read-through transcription and alternative splicing, a fifth A3 protein, namely the double-domain A3Z2Z3, with two detected variants A3Z2bZ3 and A3Z2cZ3 (Fig. 1a). A3Z2Z3s are also found in big cats (Pantherinae), indicating evolutionary conserved gene regulation [4, 36]. FIV Vif induces proteasome-dependent degradation of feline A3Z2s, A3Z3, and A3Z2Z3 [4, 37]. The double-domain feline A3Z2Z3 contains two FIV Vif interaction regions, one in each Z-domain [36]. Interestingly, and currently unexplained, FIVΔ*vif* can be inhibited by feline AZ3 and A3Z2Z3, but not by A3Z2s [4, 36]. A reverse observation was made with FFVΔ*bet*, where feline A3Z2s act as major inhibitors while A3Z3 and A3Z2Z3 only moderately reduce the infectivity of FFVΔ*bet* [4, 10, 38, 39]. Recent data indicate that certain polymorphisms in feline A3Z3 genes correlate with the susceptibility to FIV and/or FeLV infections [40].

FIV Vif induces the poly-ubiquitination of feline A3s and bridges A3s to an E3 ubiquitin ligase complex containing Cullin5 (Cul5), Elongin B/C (EloB/C), and RING-box protein RBX2 [37]; HIV-1 Vif forms a similar E3-ligase complex [41–43]. However, while HIV-1 Vif needs to additionally interact with the CBF-β protein to be stabilized and form this multiprotein complex [44, 45], FIV Vif does not bind CBF-β, and the FIV Vif-induced degradation of feline A3s does not require CBF-β to be expressed [46–49]. HIV-1 Vif cannot counteract feline A3s, and HIV-1 is therefore inhibited to various degrees by all feline A3s, with A3Z2Z3 displaying the strongest anti-HIV activity [36, 50–52]. The mechanistic reason preventing HIV-1 Vif from degrading feline A3s is unclear, especially because HIV-1 Vif and feline A3Z2Z3 are recovered together using co-immunoprecipitation assays [51]. In contrast to the Vif protein of HIV-1, Vif of Simian immunodeficiency virus from macaques (SIVmac) induces degradation of feline A3s [46, 51]. To assess the feasibility of generating an animal model for the human system based on FIV, we and others cloned FIV *vif* into HIV-1 and proved that in feline cell lines the A3 proteins are the dominant restriction factors against HIV-1 [36, 51].

In order to understand the FIV Vif interaction with feline A3 proteins, we identified in this study important A3 residues and used a homology model of feline A3Z2Z3 to describe the structure–function relationship of these potential FIV Vif binding amino acids.

Results

FIV and HIV-2/SIVmac/smm Vif induced degradation of felines A3s

In order to identify the molecular interaction of the FIV Vif protein and feline A3 proteins, we used FIV of domestic cats (*Felis catus*, Fca), hereafter referred as FIV. Co-transfection experiments of cat-derived A3s and FIV Vif expression plasmids were performed in 293T cells. All A3 constructs expressed the corresponding A3 protein as a C-terminal HA-tag, whereas Vif was expressed as a C-terminal V5-tag fusion protein. In addition, we also studied Vifs derived from HIV-1, HIV-2, SIVmac, and SIVsmm. Immunoblots of protein extracts from cells co-expressing both A3 and Vif were used as a read-out for degradation of the respective A3 protein. Results in Fig. 1b show that FIV Vif induces degradation of single-domain feline A3Z2a, A3Z2b, A3Z2c, A3Z3, and double-domain A3Z2bZ3 and A3Z2cZ3 in agreement with previous reports [4, 36, 37, 51]. The double-domain feline A3Z2bZ3 and A3Z2cZ3 were degraded by SIVmac Vif as seen before [46, 51], as well by the Vifs of SIVsmm and HIV-2. For subsequent experiments we used the expression plasmid FcaA3Z2bZ3, hereafter referred to as feline A3Z2Z3 for simplicity.

To understand, whether FIV Vif binds directly to feline A3s, we expressed A3Z2 and A3Z3 as GST fusion

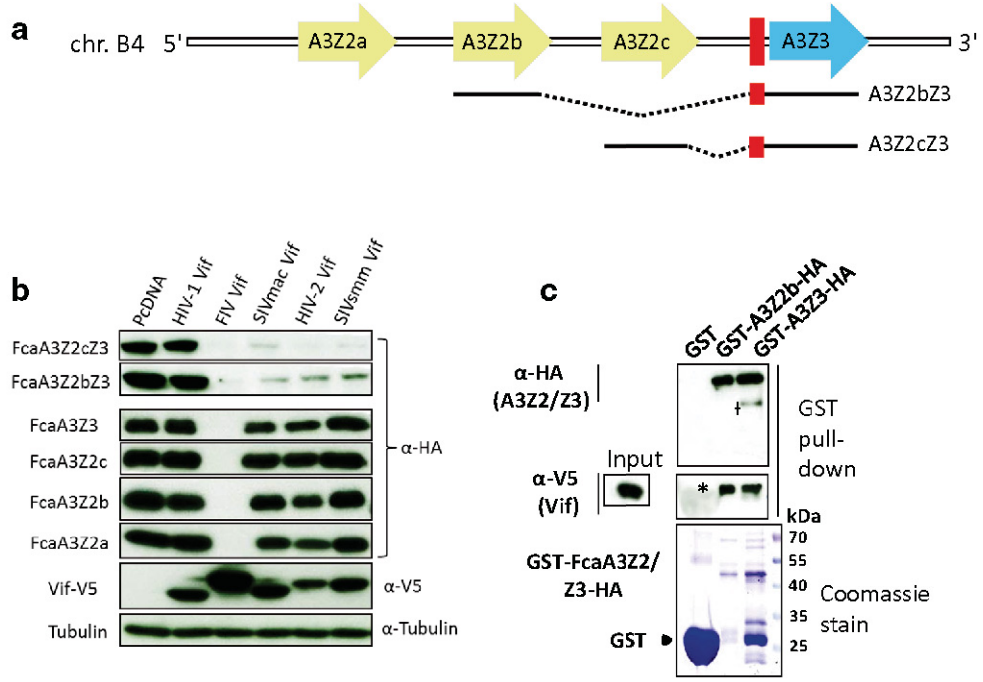


Fig. 1 The interaction of feline APOBEC3s with FIV Vif. **a** Representation of APOBEC3 (A3) genes in the genome of *Felis catus*, chromosome (chr.) B4. Coding regions of the A3 genes (A3Z2a, A3Z2b, A3Z2c, A3Z3) shown as *arrows*. *Red rectangle* exon 2 of A3Z3 that is untranslated in the A3Z3 mRNA, however translated ("linker domain") in readthrough transcripts A3Z2bZ3 and A3Z2cZ3. Spliced-out introns (*dashed lines*) are indicated, mRNAs for A3Z2s and A3Z3 not shown. For details, see references [4, 36]. **b** 293T cells were transfected with expression plasmids for FcaA3Z2a, FcaA3Z2b, FcaA3Z2c, FcaA3Z3, FcaA3Z2bZ3 and FcaA3Z2cZ3 together with HIV-1, SIVmac, HIV-2 and SIVsmm or FIV Vif or no Vif (replaced by pcDNA3.1). FcaA3s, Vifs and tubulin were visualized by immunoblot using anti-HA, anti-V5 and anti-tubulin antibodies. **c** Immunoprecipitation of *E. coli* expressed FcaA3Z2 and -Z3 (GST fusion proteins) after mixing with 293T-derived FIV Vif (Vif-V5). *Asterisk* indicates the impression of GST blob on the blot and *Dagger* denotes for a possible degradation product of FcaA3Z3

proteins in *E. coli*. Recombinant A3s were purified by affinity chromatography and mixed with lysates of 293T cells expressing FIV Vif. Following GST pulldown, immunoblots showed Vif binding to GST-A3Z2 and to GST-A3Z3 but not to GST (Fig. 1c). We further explored the interaction of FIV Vif with feline A3s by analyzing the cellular distribution in co-expressing cells. HOS cells were transfected either with plasmids encoding for feline A3Z2, A3Z3, or A3Z2Z3 alone or together with a plasmid encoding for FIV Vif-TLQAAA. The TLQ to AAA mutation in the Vif putative BC-box prevents its interaction with the E3 complex [37]. Feline A3 proteins showed a mostly cytoplasmic localization with no or very little nuclear A3, and feline A3Z3 localized in addition to the nucleoli (Additional file 1: Fig. S1, compare to Fig. S4). Nucleolar localization of A3Z3 proteins derived from humans and horses had been described before [53]. Very similar to the A3s, FIV Vif-TLQAAA showed a cytoplasmic distribution with little presence in the nucleus. Under these experimental conditions, strong co-localization of Vif and A3s was detected in cytoplasmic areas near the nucleus (Additional file 1: Fig. S1).

Identification of feline A3Z3 residues important for FIV Vif induced degradation

Feline A3Z3 and A3Z2Z3 are the restriction factors for FIVΔ*vif*, whereas A3Z2s are not active against FIVΔ*vif* [4, 36, 37, 51]. To characterize the Vif interaction with residues in feline A3Z3, A3Z3s derived from humans (A3H haplotype II, HsaA3H) and big cats (tiger, *Panthera tigris*, Pti; lion, *Panthera leo*, Ple; lynx, *Lynx lynx*, Lly; puma, *Puma concolor*, Pco) (protein alignments are highlighted in Additional file 1: Fig. S2) were used in co-transfection experiments with FIV Vif. A3s derived from tiger, lion, lynx, and puma were efficiently degraded by FIV Vif (Fig. 2a). Because A3H was resistant to FIV Vif-induced degradation, the construction of Hsa–Fca chimeric A3Z3s promised a rational approach to identify the A3Z3/FIV-Vif binding region. The chimeras Z3C1 and Z3C2 spanned respectively amino acids 1–22 and 1–50 of feline A3Z3, with the remaining part being derived from A3H, whereas Z3C6 and Z3C7 were mostly feline A3Z3 with residues 1–22 or 1–50 derived from A3H (Fig. 2b). Among the four A3Z3 chimeras, Z3C2 and Z3C6 were efficiently degraded by FIV Vif, while Z3C1

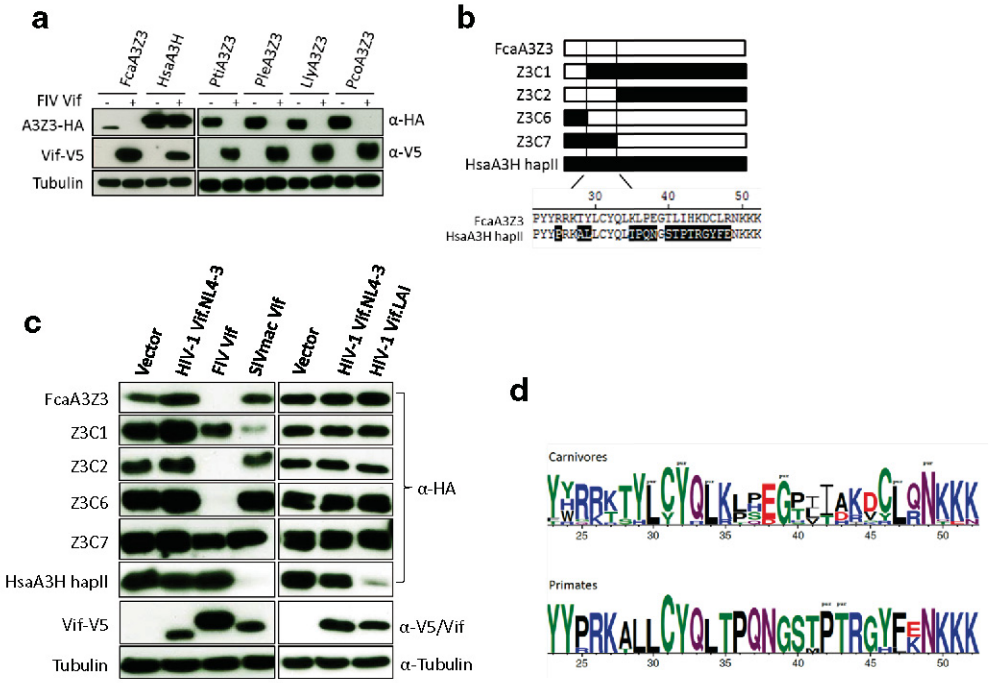


Fig. 2 Identification of residues in FcaA3Z3 for FIV Vif induced degradation. **a** Expression plasmids for FcaA3Z3, HsaA3H and big cat A3Z3s were co-transfected with FIV Vif into 293T cells. Cell lysates were analyzed by immunoblot. A3s contain a C-terminal HA-tag, Vifs contain a C-terminal V5-tag. FcaA3s represent domestic cat (*Felis catus*) APOBEC3s. Pti, Ple, Lly and Pco represent *Panthera tigris corbetti; Panthera leo bleyenberghi; Lynx lynx; Puma concolor.* **b** Scheme of FcaA3Z3/HsaA3H.hapII chimeras (Z3C1, -C2, –C6, -C7). *Highlighted* sequence diversity between HsaA3H and FcaA3Z3 in an N-terminal region. **c** 293T cells were co-transfected with expression plasmids for FcaA3Z3, Z3C1, Z3C2, Z3C6, Z3C7 or HsaA3H hapII and FIV Vif, HIV-1 (NL4-3 or LAI) or SIVmac Vif. The expression of chimeras and Vif proteins were detected by using anti-HA and anti-V5 antibodies, respectively. Tubulin served as loading control. **d** Amino acid logo for the N-terminus in 15 A3Z3 sequences from ten Carnivores species (*upper panel*) and for eight A3Z3 sequences from eight Primates species (*lower panel*). Residues identified to evolve under purifying selection are labelled with "pur". No residue was identified to evolve under diversifying selection in this A3Z3 stretch

and Z3C7 showed resistance to degradation (Fig. 2c). HIV-1 Vif (derived from clones NL4-3 or LAI) could not degrade any of the A3Z3 chimeras, but LAI Vif degraded A3H as reported before [54], and SIVmac Vif degraded Z3C1 and A3H but not Z3C2, Z3C6 and Z3C7 (Fig. 2c).

Our findings indicate that feline-derived residues shared by Z3C2 and Z3C6 (positions 23–50) are essential for FIV Vif interaction. This A3 stretch contained a number of positions evolving exclusively under purifying selection, for both carnivores and for primates (Fig. 2d). Globally, diversity among A3Z3 from carnivores was higher than among the primates' orthologs (respectively 0.24 ± 0.02 vs 0.054 ± 0.007, overall average pairwise nucleotide distance ± bootstrap standard error estimate) (Additional file 1: Fig. S3A). During this analysis, we identified for the first time duplicated A3Z3s in the same genome (i.e. in-paralogs [55]) retrieved from different lineages within Caniformia (Ursidae, the giant panda and the polar bear; Phocidae, the Weddell seal; and Odobenidae, the walrus) but we could neither identify the two A3Z3 in-paralogs in Canidae (dog) nor in Mustelidae (ferret) genomes. By contrast, in all Felidae genomes that we have screened we could only identify one of these in-paralogs (Additional file 1: Fig. S3A).

The A3Z3 region position 23–50 differs in 16 amino acids between human and feline A3Z3s, and contains certain highly conserved amino acid positions (Fig. 2b, d). We mutated thus most feline-specific residues in feline A3Z3, in positions 35–38 and 40–48. Residues in position 35 + 36 (KL), 37 + 38 (PE), 41 + 42 (LI) and 43 (H) in A3Z3 were substituted by the corresponding ones found in A3H. Additionally, we exchanged the A3Z3 residues at position 45 + 46 (DC), 47 + 48 (LR) and 41 + 42 (LI) against AA (Fig. 3a). These mutated A3s were characterized for resistance to degradation by co-expression with FIV Vif. We found that only A3Z3s mutated at position 41 + 42 (LI ≫ TP and LI ≫ AA) showed partial resistance to degradation by Vif (Fig. 3b). A65I in feline A3Z3 has been described in Brazilian cats and discussed to be a relevant resistance mutation against FIV [40, 56]. Under our experimental conditions, A3Z3 mutated in position 65 (A65I) displayed only little

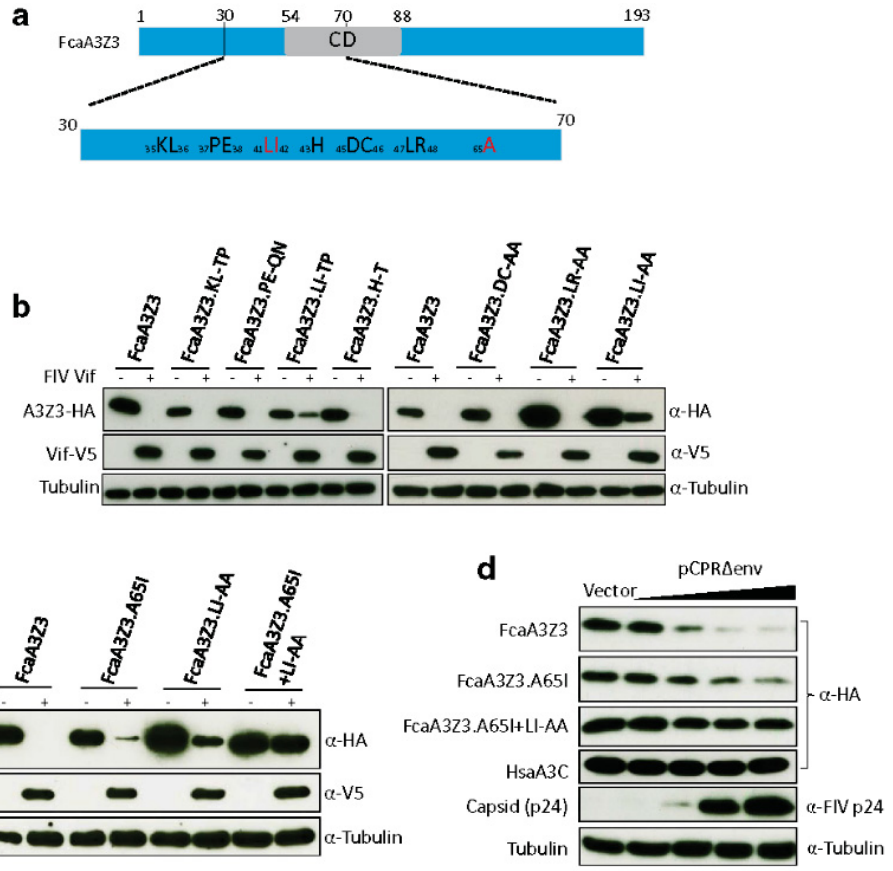


Fig. 3 Generation of FIV Vif resistant FcaA3Z3. **a** Representation of FcaA3Z3 protein. Residues investigated are shown. *CD* cytidine deaminase domain. **b**, **c** Several mutants at N-terminal region of FcaA3Z3 were generated. To analyze the sensitivity of FcaA3Z3 mutants to FIV Vif, 293T cells were co-transfected with expression plasmids for FcaA3Z3 wild-type (FcaA3Z3) or indicated mutants and FIV Vif. 48 h later, FcaA3Z3, Vif and tubulin proteins were detected by immunoblot. **d** Increasing amounts of FIV *gag-pol-vif* helper plasmid pCPRΔ*env* that expresses virus-typical levels of FIV Vif without a tag were analyzed for degradation of HsaA3C, FcaA3Z3, FcaA3Z3.A65I and FcaA3Z3.A65I + LI-AA. 293T cells were co-transfected with 150 ng HsaA3C, FcaA3Z3, or A3Z3 mutants and pCPRΔ*env* (0, 10, 50, 150 or 250 ng). 48 h later, A3 expression was analyzed by immunoblots by using anti-HA antibody. Anti-FIV p24 detected expression of FIV helper plasmid pCPRΔ*env*, tubulin detection confirmed equal protein loading

resistance to Vif-mediated degradation (Fig. 3c). However, very important, the combination of mutations, A65I and L41A-I42A, resulted in an A3Z3 variant that showed complete resistance to FIV Vif degradation (Fig. 3c). We wondered whether experimental overexpression of the V5-tagged FIV Vif could mask the potency of the natural A65I variant to resist degradation. To address this question, we used as a source for Vif expression the replication-deficient FIV packaging construct pCPRΔ*env* [57]. Expression of increasing levels of pCPRΔ*env* in the presence of constant amounts of A3 revealed that the A65I mutation was degraded less efficiently than the wild-type A3Z3 (Fig. 3d). As a control we used A3C and A3Z3. A65I + LI-AA, which both showed no degradation by Vif derived by pCPRΔ*env*. Together, these findings indicate that the A65I mutation in feline A3Z3 mediates a partial protection, and that a combination with L41A-I42A resulted in enhanced resistance to Vif.

The stretch involved in the interaction with Vif encompassed a number of highly conserved residues between A3Z3s from carnivores and primates, as well as residues under purifying selection (Fig. 2d). The L41–I42 residues in cat A3Z3 identified to interact with Vif are strictly conserved (L|I) in A3Z3 from felids, to the extent that even the codons used are also strictly conserved (CTT|ATT) for the five Felidae species analyzed. Interestingly, the two A3Z3 paralogs in Canifornia display different amino acid profiles in this Vif-binding region (Additional file 1: Fig. S3A), and albeit chemically related, amino acid residues in these positions are variable (I/L/V|I/T). Finally, this A3Z3 stretch is very different in the corresponding positions in A3Z3 from primates (T/M|P). Altogether,

evolutionary relationships for these two residues could thus at least partly explain species-specificity of the interaction between felidae A3Z3 and FIV Vif, reflecting adaptation and specific targeting.

Generation of a FIV Vif resistant feline A3Z2Z3

Our results demonstrate that feline A3Z2 can also be efficiently degraded by FIV Vif, thus implying a specific interaction between both proteins (Fig. 1b). In order to generate an A3Z2Z3 protein resistant to FIV Vif, we decided to mutate as well the A3Z2 moiety. To identify residues important for FIV Vif interaction with feline A3Z2, chimeric A3s of A3Z2 and human A3C, called Z2C1, -C4, -C5 and -C30 (Fig. 4a), were co-expressed with FIV Vif. The chimeras Z2C1, Z2C4 and Z2C5 spanned the 1–22, 1–131 and 1–154 amino acids of feline A3Z2, respectively, the remaining parts being derived from A3C. Chimera Z2C30 was feline A3Z2, with amino acids 132–154 derived from A3C. Chimeras Z2C1 and Z2C4 showed moderately reduced protein levels when FIV Vif was co-expressed, chimera Z2C5 resistance to degradation, and chimera Z2C30 was efficiently degraded by FIV Vif (Fig. 4b). As controls, we investigated all chimeras for degradation by HIV-1 and SIVmac Vifs. HIV-1 Vif induced degradation of Z2C1 only, and SIVmac Vif completely degraded Z2C1, Z2C4 and Z2C30, and mostly Z2C5 (Fig. 4b). Because the Z2C5 chimera, in which the C-terminal 37 residues were of A3C origin, was resistant to FIV

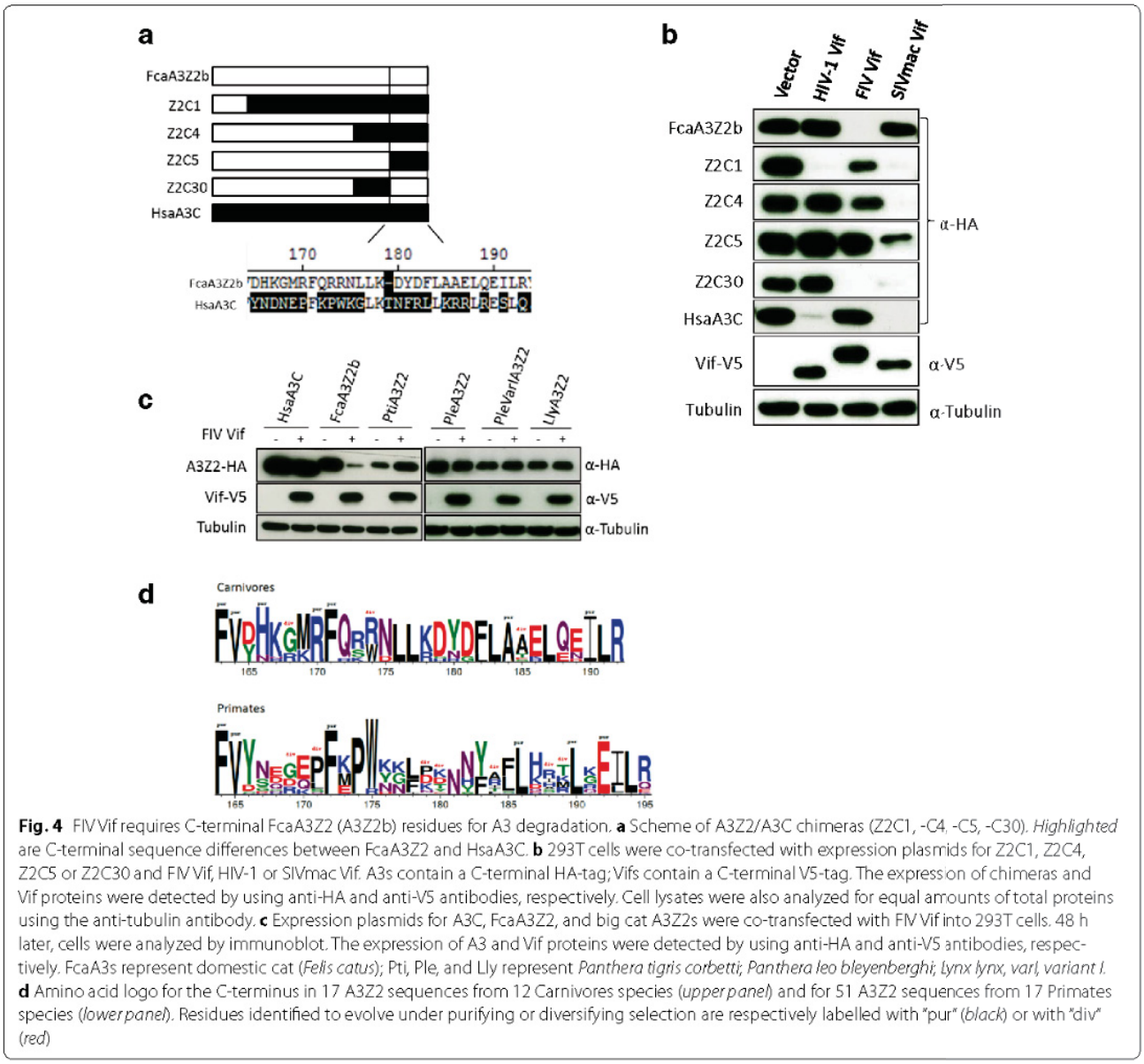


Fig. 4 FIV Vif requires C-terminal FcaA3Z2 (A3Z2b) residues for A3 degradation. **a** Scheme of A3Z2/A3C chimeras (Z2C1, -C4, -C5, -C30). *Highlighted* are C-terminal sequence differences between FcaA3Z2 and HsaA3C. **b** 293T cells were co-transfected with expression plasmids for Z2C1, Z2C4, Z2C5 or Z2C30 and FIV Vif, HIV-1 or SIVmac Vif. A3s contain a C-terminal HA-tag; Vifs contain a C-terminal V5-tag. The expression of chimeras and Vif proteins were detected by using anti-HA and anti-V5 antibodies, respectively. Cell lysates were also analyzed for equal amounts of total proteins using the anti-tubulin antibody. **c** Expression plasmids for A3C, FcaA3Z2, and big cat A3Z2s were co-transfected with FIV Vif into 293T cells. 48 h later, cells were analyzed by immunoblot. The expression of A3 and Vif proteins were detected by using anti-HA and anti-V5 antibodies, respectively. FcaA3s represent domestic cat (*Felis catus*); Pti, Ple, and Lly represent *Panthera tigris corbetti, Panthera leo bleyenberghi, Lynx lynx, varl, variant I.* **d** Amino acid logo for the C-terminus in 17 A3Z2 sequences from 12 Carnivores species (*upper panel*) and for 51 A3Z2 sequences from 17 Primates species (*lower panel*). Residues identified to evolve under purifying or diversifying selection are respectively labelled with "pur" (*black*) or with "div" (*red*)

Vif, we speculated that the C-terminal region of cat A3Z2 could be important for FIV Vif-induced degradation. SIVmac Vif, which cannot degrade feline A3Z2 (Fig. 4b), interacts presumably with C-terminal human-derived sequences spanning A3C sequences present in Z2C5 and Z2C30 (Fig. 4b). In addition we analyzed the degradation sensitivity of A3Z2 proteins from big cats and found that FIV Vif did not induce degradation of A3Z2 from tiger, lion or lynx (Fig. 4c). These felid A3Z2s are very similar to FcaA3Z2 as they share 89–93 % identically conserved residues (Additional file 1: Figs. S2, S3B, Fig. 4d), whereas cat A3Z2 and human A3C are much more diverse and share only 47 % identical amino acids. Thus, we identified four positions in which all big cat A3Z2s differed from FcaA3Z2, in positions N18, T44, D165 and H166 (Additional file 1: Fig. S2, Fig. 5a). We mutated accordingly

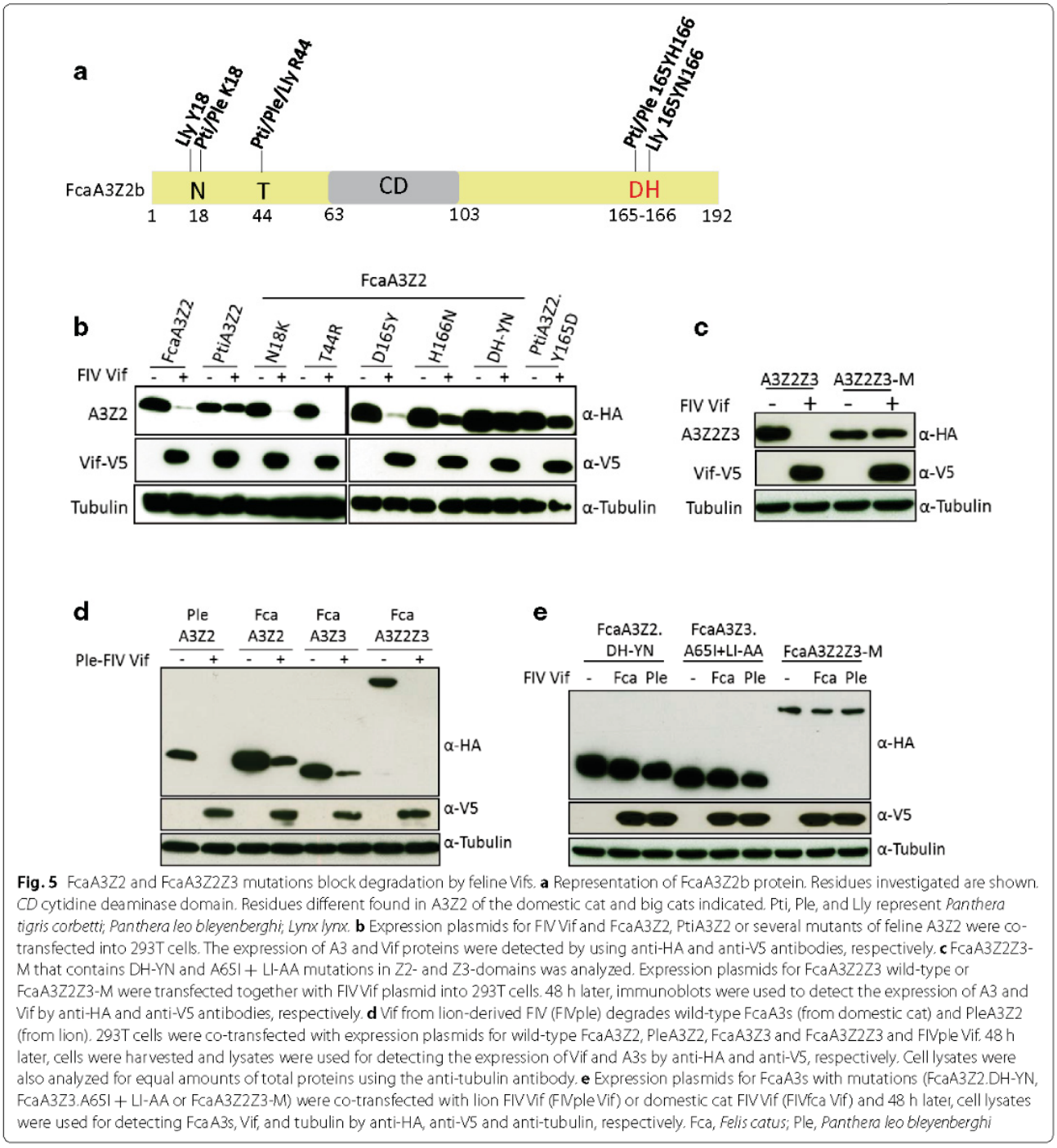


Fig. 5 FcaA3Z2 and FcaA3Z2Z3 mutations block degradation by feline Vifs. **a** Representation of FcaA3Z2b protein. Residues investigated are shown. *CD* cytidine deaminase domain. Residues different found in A3Z2 of the domestic cat and big cats indicated. Pti, Ple, and Lly represent *Panthera tigris corbetti; Panthera leo bleyenberghi; Lynx lynx.* **b** Expression plasmids for FIV Vif and FcaA3Z2, PtiA3Z2 or several mutants of feline A3Z2 were co-transfected into 293T cells. The expression of A3 and Vif proteins were detected by using anti-HA and anti-V5 antibodies, respectively. **c** FcaA3Z2Z3-M that contains DH-YN and A65I + LI-AA mutations in Z2- and Z3-domains was analyzed. Expression plasmids for FcaA3Z2Z3 wild-type or FcaA3Z2Z3-M were transfected together with FIV Vif plasmid into 293T cells. 48 h later, immunoblots were used to detect the expression of A3 and Vif by anti-HA and anti-V5 antibodies, respectively. **d** Vif from lion-derived FIV (FIVple) degrades wild-type FcaA3s (from domestic cat) and PleA3Z2 (from lion). 293T cells were co-transfected with expression plasmids for wild-type FcaA3Z2, PleA3Z2, FcaA3Z3 and FcaA3Z2Z3 and FIVple Vif. 48 h later, cells were harvested and lysates were used for detecting the expression of Vif and A3s by anti-HA and anti-V5, respectively. Cell lysates were also analyzed for equal amounts of total proteins using the anti-tubulin antibody. **e** Expression plasmids for FcaA3s with mutations (FcaA3Z2.DH-YN, FcaA3Z3.A65I + LI-AA or FcaA3Z2Z3-M) were co-transfected with lion FIV Vif (FIVple Vif) or domestic cat FIV Vif (FIVfca Vif) and 48 h later, cell lysates were used for detecting FcaA3s, Vif, and tubulin by anti-HA, anti-V5 and anti-tubulin, respectively. Fca, *Felis catus*; Ple, *Panthera leo bleyenberghi*

position 18 (N18K) and 44 (T44R) in FcaA3Z2, but found both mutants to be efficiently degraded by FIV Vif (Fig. 5b). Very similar, A3Z2.D165Y was depleted when co-expressed with FIV Vif. Interestingly, mutation of residue 166 (H166N) generated a partially Vif-resistant A3Z2 protein. We speculated that the adjacent D165 might enhance the Vif-resistance seen in the H166N variant. Indeed, the A3Z2.DH-YN mutant showed complete resistance to FIV Vif (Fig. 5b). We also analyzed tiger A3Z2.Y165D but could not reverse the resistance to degradation by FIV Vif (Fig. 5b). We conclude that D165-H166 in the C-terminal region of cat A3Z2 are important for Vif-mediated degradation together with other residues that remain to be characterized.

Finally, we constructed A3Z2Z3-M containing D165Y, H166N in Z2 and A65I + L41A, I42A in Z3. Co-expression experiments of A3Z2Z3-M with FIV Vif showed that this A3 variant was Vif-resistant (Fig. 5c). Importantly, the mutations that generated Vif-resistance did not impact the subcellular localization of the feline A3, as demonstrated by confocal microscopy of transiently transfected HOS cells (Additional file 1: Fig. S4). We also studied lion specific FIV (FIVple) Vif, which shares only 52 % identical residues with domestic cat FIV Vif (Additional file 1: Fig. S2C). FIVple Vif was able to induce degradation of PleA3Z2 and of FcaA3Z2, A3Z3 and A3Z2Z3 (Fig. 5d). Interestingly, FIVple Vif could not induce degradation of the mutated cat A3s A3Z2.DH-YN, A3Z3.A65I + LIAA and A3Z2Z3-M (Fig. 5e). These findings suggest that Vifs from lion and from domestic cat FIVs interact with identical residues in the domestic cat A3s.

To check whether the FIV Vif-resistant mutant A3s displayed modified binding to Vif, wild-type and mutated A3s together with FIV Vif-TLQAAA were co-expressed and analyzed by anti-HA immuno-precipitation (Fig. 6). Wild-type cat A3Z3 precipitated FIV Vif (Fig. 6a), consistent with a direct interaction of both proteins (Fig. 1c). Only very little Vif bound to A3Z3.A65I and no Vif was detected in precipitations of A3Z3.LI-AA and A3Z3.A65I + LI-AA (Fig. 6a). However, when we examined wild-type A3Z2 and the DH-YN mutant, we detected similar amounts of Vif in both precipitations. Wild-type A3Z2Z3 bound high levels of Vif, and this binding was much reduced by the mutated variant A3Z2Z3-M (Fig. 6b). Globally, our observations suggest that A65I and LI-AA mutations in A3Z3 abolished FIV Vif binding, while hitherto not identified residues mediate Vif binding in A3Z2.

Structural analysis of feline A3Z2Z3

To identify the position of residues in feline A3s that, when mutated, prevent binding of FIV Vif and A3Z2Z3 degradation, a structural model of feline A3Z2Z3 was generated, initially aligning its sequence to the human full-length A3G model [58]. Surprisingly, the alignment indicated a large insertion in the Z2–Z3 linker region in the feline sequence that is not present in the human counterpart (Additional file 1: Fig. S5). This domain insertion spans 46 residues, extending the feline A3Z2Z3 linker to 83 residues compared to 27 residues in humans. The structure of the 83-residue linker was predicted using TopModel [59, 60]. Although the five identified template structures show only a low sequence identity with respect to the linker (up to 19.4 %; see "Methods" section), they all share a homeo-box domain fold [61]. The best three templates were aligned to the linker sequence

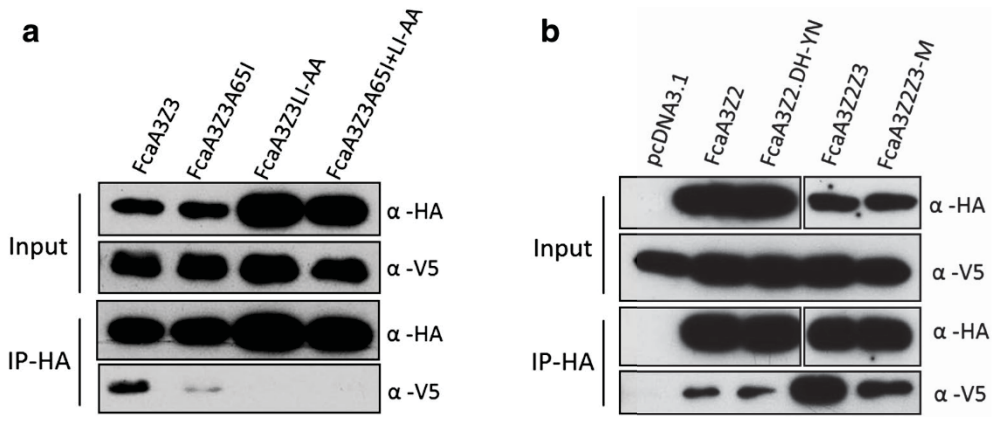


Fig. 6 *Differential binding of FIV Vif to wild-type and mutant feline A3s.* **a** Expression plasmids for FcaA3Z3s wild-type and mutants (all with HA-tag) and FIV Vif-TLQAAA (V5 tag) were co-transfected into 293T cells. The proteins were immunoprecipitated by α-HA beads and analyzed by immunoblots using anti-HA and anti-V5 antibodies. **b** Expression plasmids for FcaA3s (FcaA3Z2, FcaA3Z2.DH-YN, FcaA3Z2Z3 and FcaA3Z2Z3-M, all with HA-tag) and FIV Vif-TLQAAA (V5-tag) and were co-transfected into 293T cells, pcDNA3.1 (+) served as an A3-free control. 48 h later, cells were harvested, proteins were immunoprecipitated by α-HA beads. The FcaA3s and FIV Vif proteins were detected by anti-HA and anti-V5 antibodies, respectively

(Additional file 1: Fig. S5) and used for structure prediction. The rest of the feline A3Z2Z3 protein was predicted using the homology model of human A3G [58] as a template. Finally, the linker domain and the rest of the feline A3Z2Z3 protein were manually docked, sequentially connected, and unstructured parts of the linker domain were energy minimized (Fig. 7a). While this cannot be expected to result in an exact structural model, it provides a representation where the linker domain insertion could be located with respect to the rest of the feline A3Z2Z3 protein.

The five residues in feline A3s that, when mutated, prevent binding of FIV Vif and A3Z2Z3 degradation (D165, H166, L285, I286, A309; the last three corresponding to L41, I42 and A65 of A3Z3), are located opposite to the putative location of the linker domain and are at the boundary between the Z2 and Z3 domains (Fig. 7b). The predicted HIV-1 Vif binding regions in human A3G, A3C and A3H are additionally depicted in Fig. 7b–d, respectively. For A3C and A3H, the predicted HIV-1 Vif binding regions are spatially clearly separated from the respective five residues identified here in feline A3s (Fig. 7c, d). One may thus speculate that our findings indicate a FIV Vif binding region in feline A3 different from the ones described for HIV-1 Vif in human A3s.

FIV Vif-resistant feline A3s are antiviral

In the next set of experiments, we investigated whether feline A3s carrying the putative Vif-binding mutations displayed antiviral activity and resistance against Vif in FIV infections. We generated FIV luciferase reporter viruses by co-expression with either no A3 or with A3Z3, A3Z3.A65I, A3Z3.LI-AA or A3Z3.A65I + LI-AA and increasing levels of the FIV Vif plasmid (0–160 ng). Vector particles were normalized for reverse transcription (RT) activity, and luciferase activity was quantified 2 days post infection (Fig. 8a). All feline A3Z3s, either wild-type or mutants, were able to inhibit to the same extent Vif-deficient FIV, demonstrating that the described mutations do not hinder the potential for antiviral activity. Wild-type feline A3Z3 was fully counteracted by the lowest amount of Vif plasmid (40 ng) (Fig. 8a), matching well complete degradation observed in the lysates of FIV-producing cells (Fig. 8b). Opposite to the homogenous behavior in the absence of Vif, mutated A3Z3s showed variable resistance to Vif-counteraction, as was obvious in the levels of remaining A3 signal in the cell lysates of the FIV-producing cells (Fig. 8b). Intermediate amounts of *vif*-encoding plasmid (40–80 ng) partially counteracted the inhibition of A3Z3.A65I or A3Z3.LI-AA mutants, and higher levels of Vif (160 ng plasmid) recovered infectivity of FIVs produced in the presence of A3Z3.A65I and A3Z3.LI-AA. However, even the highest levels of Vif were not able to counteract the antiviral activity of A3Z3.A65I + LI-AA (Fig. 8a, b). The importance of Z2- and Z3-mutations in feline A3Z2Z3-M was characterized with 100 ng of FIV Vif plasmid. FIV luciferase viruses were produced and examined as described above using A3Z2Z3 and A3Z2Z3-M. FIV Vif restored the infectivity to levels similar to those in the absence of A3Z2Z3, while A3Z2Z3-M strongly inhibited FIV, either with or without Vif expression (Fig. 8c). The immunoblots of the corresponding FIV producing cells confirmed protein expression and Vif-dependent degradation of the wild-type A3Z2Z3 protein (Fig. 8d). To explore whether stable expression of A3Z2Z3-M can impact spreading infection, human HOS.CD4.CCR5 cells expressing either wild-type or mutated A3Z2Z3 were established (Fig. 8e, Additional file 1: Fig. S6) and infected by HIV-1 expressing FIV Vif (HIV-1vif_{FIV}) [36]. FIV could not be investigated directly, because there are no feline cell lines known to be negative for A3 expression, and FIV cannot replicate in human cell lines. Whereas HIV-1vif_{FIV} was detected at day six in the supernatant cells with wild-type A3Z2Z3, HOS cells expressing the A3Z2Z3-M showed a much delayed kinetic of viral replication (Fig. 8f). This observation suggests that the engineered A3Z2Z3-M protein also gained the capacity to restrict FIV Vif during multi-rounds of replication.

Encapsidation of A3 proteins in nascent virions is required for their antiviral activity. We investigated first whether mutated feline A3s could be differentially encapsidated into nascent virions. For this, we produced Vif-deficient FIV particles during expression of the different A3 proteins, measured the differential infectivity of the particles (Additional file 1: Fig. S7A) and subjected virus lysates to immunoblot analysis (Additional file 1: Fig. S7B). Results showed that wild-type and mutated feline A3s were detected in the concentrated FIVs (VLPs). However, while wild-type A3Z3 was less efficiently packaged compared with the A3Z3.A65I + LI-AA mutant, the wild-type A3Z2Z3 was detected in virions in higher abundance than the A3Z2Z3-M variant (Additional file 1: Fig. S7B). We addressed then the question whether encapsidated mutated feline A3s effectively exerted their cytidine deaminase activity onto the FIV genome in the virion. To tackle this question, cells were infected with FIV produced during cellular expression of feline A3Z3, A3Z3.A65I + LI-AA, A3Z2Z3 or A3Z2Z3-M, in the absence of A3 expression as a negative control, or during expression of human A3G as positive control. Total DNA was isolated from infected cells and subjected to differential DNA denaturing PCR (3D-PCR) [62] on the viral vector encoded luciferase gene 12 h post infection. Based on

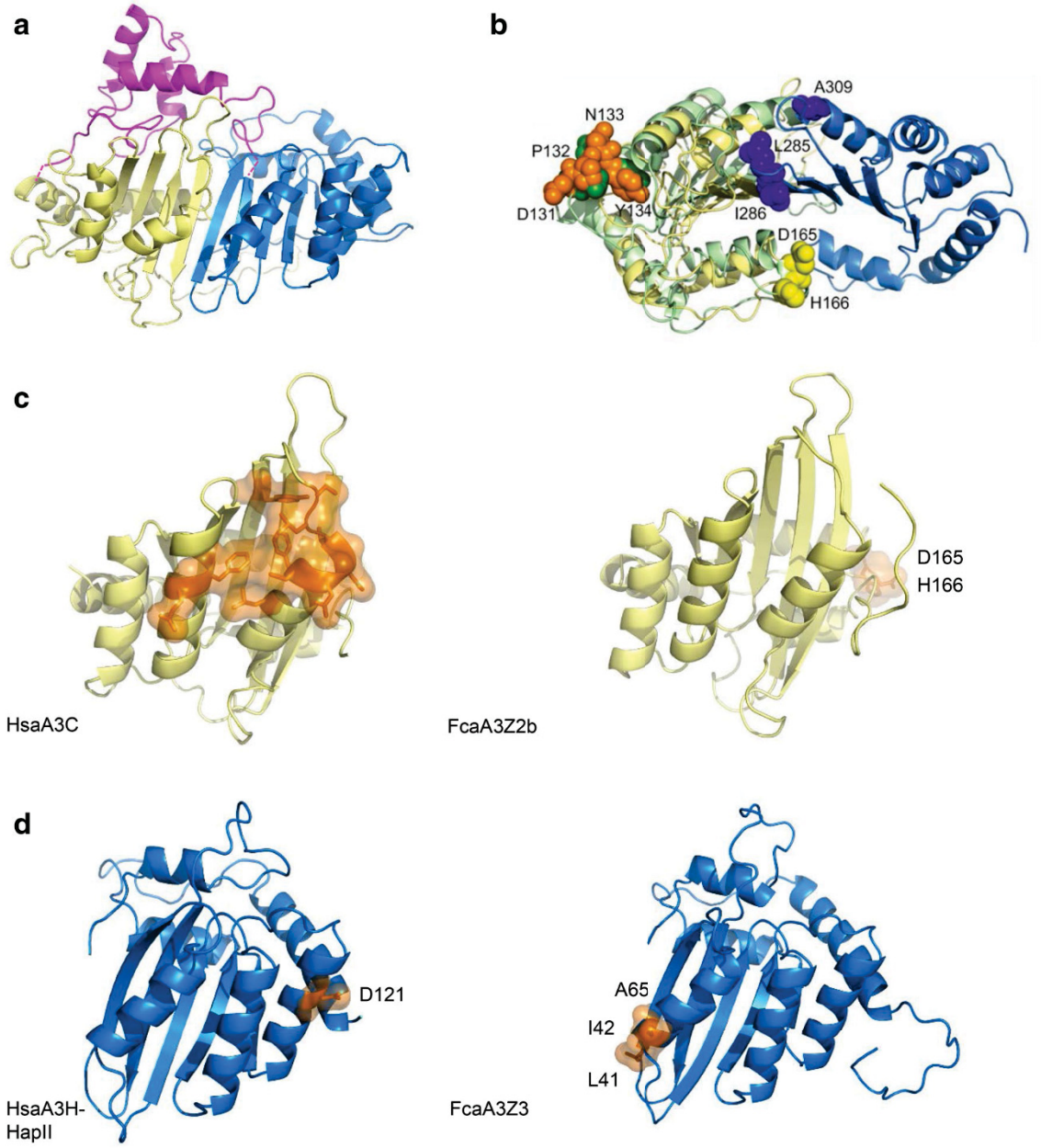


Fig. 7 Composite model of feline APOBEC3 and locations of residues mediating Vif binding. **a** Structural model of FcaA3Z2Z3 including the linker (*pink*) at a putative location above the Z2 (*yellow*) and Z3 (*blue*) domains; the linker connections to the Z2 and Z3 domains are highlighted by *dashed lines*. [69]. **b** Structural model of FcaA3Z2Z3 rotated by 90° with respect to **a** (Z2 domain: *yellow*; Z3 domain: *blue*; the linker region and parts of the N-terminus of Z2 and the C-terminus of Z3 for which no structure could be modeled were omitted for clarity). Residues in sphere representation in *yellow* (D165/H166), those in *blue* (L285/I286/A309), and those in *orange* (D131-Y134). Residues sequentially equivalent to the latter in the soluble N-terminal Vif-binding domain (sNTD) of A3G (PDB ID 2MZZ; *pale green*) are colored in green; these residues are part of the Vif-binding regions of the sNTD [68]. **c** Crystal structure of human A3C (PDB ID: 3VOW) and structural model of feline A3Z2b depicting the positions of respective HIV-1 Vif and FIV Vif binding sites. The domains are orientated as the Z2 domain in **a**. **d** Structural model of human A3H-HapII and feline A3Z3 depicting the positions of respective HIV-1 Vif and FIV Vif binding sites. The domains are orientated as the Z3 domain in **a**. Key residues involved in Vif binding are labelled (except human A3C), represented in sticks and highlighted with its surface in *orange color*

the overall nucleotide content, 3D-PCR amplifies PCR products at different denaturing temperatures (Td), with amplicons with higher A + T content displaying lower denaturing temperatures than amplicons with higher G + C content. The net effect of the cytidine deaminase A3 activity is thus expected to lower the denaturing

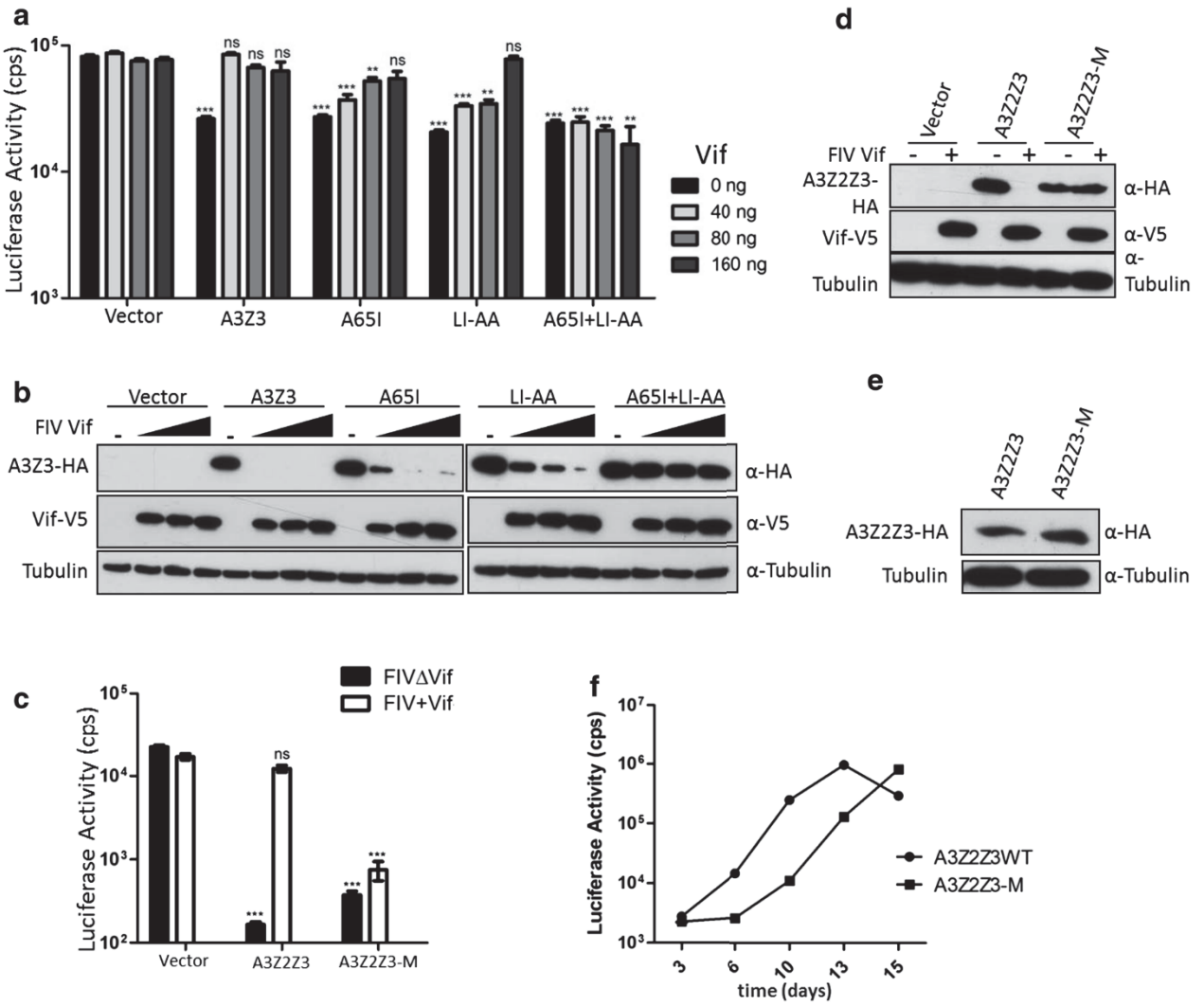


Fig. 8 The impact of mutations in FcaA3s on FIV infectivity. **a** FIVΔ*vif* luciferase reporter virions were produced in the presence of feline A3 expression plasmids (FcaA3Z3 wild-type, FcaA3Z3.A65I, FcaA3Z3.LI-AA or FcaA3Z3.A65I + LI-AA, 250 ng plasmid) with increasing amounts of FIV Vif (0, 40, 80 and 160 ng plasmid), pcDNA3.1 (+) was added as a control (vector). Infectivity of reporter vectors was determined by quantification of luciferase activity in 293T cells transduced with vector particles. **b** Cell lysates of FIV producer cells examined in **a** were used to detect the expression of FcaA3Z3 and FIV Vif by anti-HA and anti-V5 antibodies, respectively. Cell lysates were also analyzed for equal amounts of total proteins using anti-tubulin antibody. **c** The impact of FcaA3Z2Z3 and FcaA3Z2Z3-M on infectivity of FIVΔ*vif* luciferase reporter vectors in the presence of FIV Vif. 293T cells were co-transfected FIVΔ*vif* luciferase reporter viruses with 1000 ng FcaA3s and 100 ng FIV Vif expression plasmids. 48 h later, FIV particles were used to infect 293T cells, and infectivity was determined by quantification of luciferase activity. **d** Cell lysates for FIV producer cells used in **c** were analyzed by immunoblots to detect the expression of FcaA3Z3s and FIV Vif by anti-HA and anti-V5 antibodies, respectively. Cell lysates were also analyzed using anti-tubulin antibody. **e**, **f** Spreading replication of HIV-1 expressing FIV Vif (NL-Bal.*vif*$_{FIV}$) is inhibited by feline A3Z2Z3-M. **e** Immunoblot analysis of HOS.CD4.CCR5 cells stable expressing feline A3Z2Z3 proteins. A3 proteins were detected by anti-HA antibody. Anti-tubulin served to demonstrate equal protein loading. **f** HOS.CD4.CCR5 cells expressing either wild-type A3Z2Z3 or A3Z2Z3-M were infected by NL-Bal.*vif*$_{FIV}$ with an MOI of 0.01 and virus replication was monitored by using the cell culture supernatant for infection of TZM-bl luciferase reporter cells. *Asterisks* represent statistically significant differences: ***$p < 0.001$; **$0.001 < p < 0.01$; *$0.01 < p < 0.05$; ns, $p > 0.05$ (Dunnett *t* test)

temperature of the target DNA, leading to lower Tds values. Indeed, FIV virions produced in the absence of A3s yielded 3D-PCR products with the lowest Td of 86.3 °C, whereas all FIV virions produced during A3 expression resulted in 3D-PCR products with decreased Tds (as low as 84.2 °C) (Additional file 1: Fig. S7C). This indicates that the wild-type and mutant feline A3s display enzymatic deamination activities.

The linker in feline A3Z2Z3 is targeted by HIV-2 and SIVmac/smm Vifs

We and others have observed that SIVmac Vif can induce degradation of feline A3Z2Z3 (Fig. 1b) [46, 51]. Figure 1b demonstrates that the Vifs of SIVsmm and HIV-2 also display this phenotype and are able to degrade feline A3Z2Z3. To elucidate this unexpected capacity of primate lentiviruses to counteract feline A3s in the context of viral infections, we generated luciferase reporter viruses for SIVmac and HIV-2 (Fig. 9). SIVmac or SIVmacΔ*vif* luciferase reporter viruses [63] were produced in the absence or presence of human A3G, feline A3Z2a, A3Z2b, A3Z2c, A3Z2bZ3, A3Z2cZ3 or A3Z2bZ3s that included polymorphic residues found in exon 4 of different *F. catus* breeding lines (Birman, Japanese Bobtail, British Shorthair, Turkish Van [36]). The Vif proficient virus SIVmac-Luc expresses Vif in its natural expression context; however Vif lacks a tag for detection. Viral particles were normalized for RT activity and luciferase activity of infected cells was quantified 2 days post infection (Fig. 9a). We found that double-domain feline A3s strongly inhibited Vif-deficient SIVmac, and that Vif expression fully counteracted this antiviral activity, showing therefore a similar pattern to human A3G. However, Vif expression did not affect inhibition of SIVmac by single domain A3s (Fig. 9a). The corresponding immunoblots of the virus producing cells showed Vif-dependent degradation of human A3G and of all feline A3Z2Z3s inspected. Feline A3Z2s and A3Z3 displayed a resistance to degradation by Vif proficient SIVmac (Fig. 9c). We performed a similar experiment using a HIV-2 luciferase reporter virus [64], which is a three-plasmid lentiviral vector system that requires Vif to be co-expressed from a separate plasmid (Fig. 9b). Using this system, we found that HIV-2 Vif counteracted the antiviral activity of human A3G, feline A3Z2bZ2 and of A3Z2cZ3. Again, the antiviral activity of feline single-domain A3s could not be inhibited by HIV-2 Vif (Fig. 9b). The immunoblots of the virus producing cells showed a Vif-dependent depletion of human A3G as well as of the feline double-domain A3s (Fig. 9d).

The Vif-mediated degradation profile exclusive to A3Z2Z3s may indicate that the HIV-2/SIVmac/smm Vifs require for interaction with the feline A3Z2Z3 a protein domain that is absent in the single-domain A3Z2 or A3Z3. We speculated that the homeo-box domain insertion (linker region) could play a central role in these Vif interactions. To test our hypothesis, three constructs were assayed: an A3Z2Z3 in which the linker was deleted (ΔLinker); and two versions of A3Z2Z3 in which either residues 223–240 (Δ222) or residues 211–240 (Δ210) in the linker were removed (Fig. 10a). All these constructs successfully expressed protein upon transfection, and FIV Vif was able to degrade all of them. Only the linker truncations Δ222 and Δ210 were efficiently degraded by Vif of HIV-2/SIVmac/smm, whereas the ΔLinker construct showed very little degradation (Fig. 10b). We extended this experiment and analyzed the degradation with increasing levels (0, 20, 50, 150 or 250 ng) of SIVmac or HIV-2 Vif expression plasmid (Additional file 1: Fig. S8). Interestingly, the A3Z2Z3 lacking the linker domain (ΔLinker) showed dose-dependent moderate degradation, while mutants Δ222 and Δ210 showed a HIV-2/SIV Vif-dependent degradation similar as the wildtype A3Z2Z3 protein (Fig. 10b, Additional file 1: Fig. S8). To characterize the linker mutant A3s for functional antiviral activity, FIVΔ*vif* and SIVmacΔ*vif* luciferase reporter viruses were generated in the presence of wildtype and mutated A3s (Fig. 10c, d, Additional file 1: Fig. S9). Immunoblots of the viral particles showed that all A3s were encapsidated (Additional file 1: Fig. S9). Consistently in both viral systems, A3Z2Z3 moderately lost antiviral activity when part or the complete linker was deleted (ΔLinker, Δ210, Δ222) (Fig. 10c, d). Together, our results suggest that the linker domain enhances the antiviral activity of feline A3Z2Z3 but is not essentially required for it and that the linker is important for HIV-2/SIVmac/smm Vif degradation of feline A3Z2Z3. Whether the linker domain forms part of the HIV-2/SIV Vif interaction surface will be an important future question.

Because the linker insertion is absent in human A3s, we tried to learn more about the evolution of this unique domain. The DNA sequences in A3Z3 exon 2 encoding for the linker region in the double-domain A3Z2Z3 proteins are extremely conserved among members of Felinae and Pantherinae. The linker sequence is indeed more conserved than the corresponding Z2 and Z3 domains, the evolutionary distances being 0.044 ± 0.006 for the Z2 stretch, 0.011 ± 0.006 for the linker and 0.018 ± 0.004 for the Z3 stretch (overall average pairwise nucleotide distance ± bootstrap standard error estimate). The evolutionary origin of the linker remains nevertheless obscure, as systematic BLASTn and BLAT searches using this linker sequence as seed did not retrieve hits beyond spurious matches. However, tBLASTn successfully retrieved hits associated with A3 genes: in the 5′ untranslated region of the A3 gene (XR_434780) in the Weddell seal genome, in the 5′ untranslated region of the A3 gene (FJ716808) in the camel genome, as well as in the 5′ regulatory region of the A3 gene (FJ716803) in the pig genome. The only very remote hit in primates with linker-similar sequence could be located in an A3 gene of tarsier (XM_008049574.1), but similarity levels do not allow in this case claiming common ancestry.

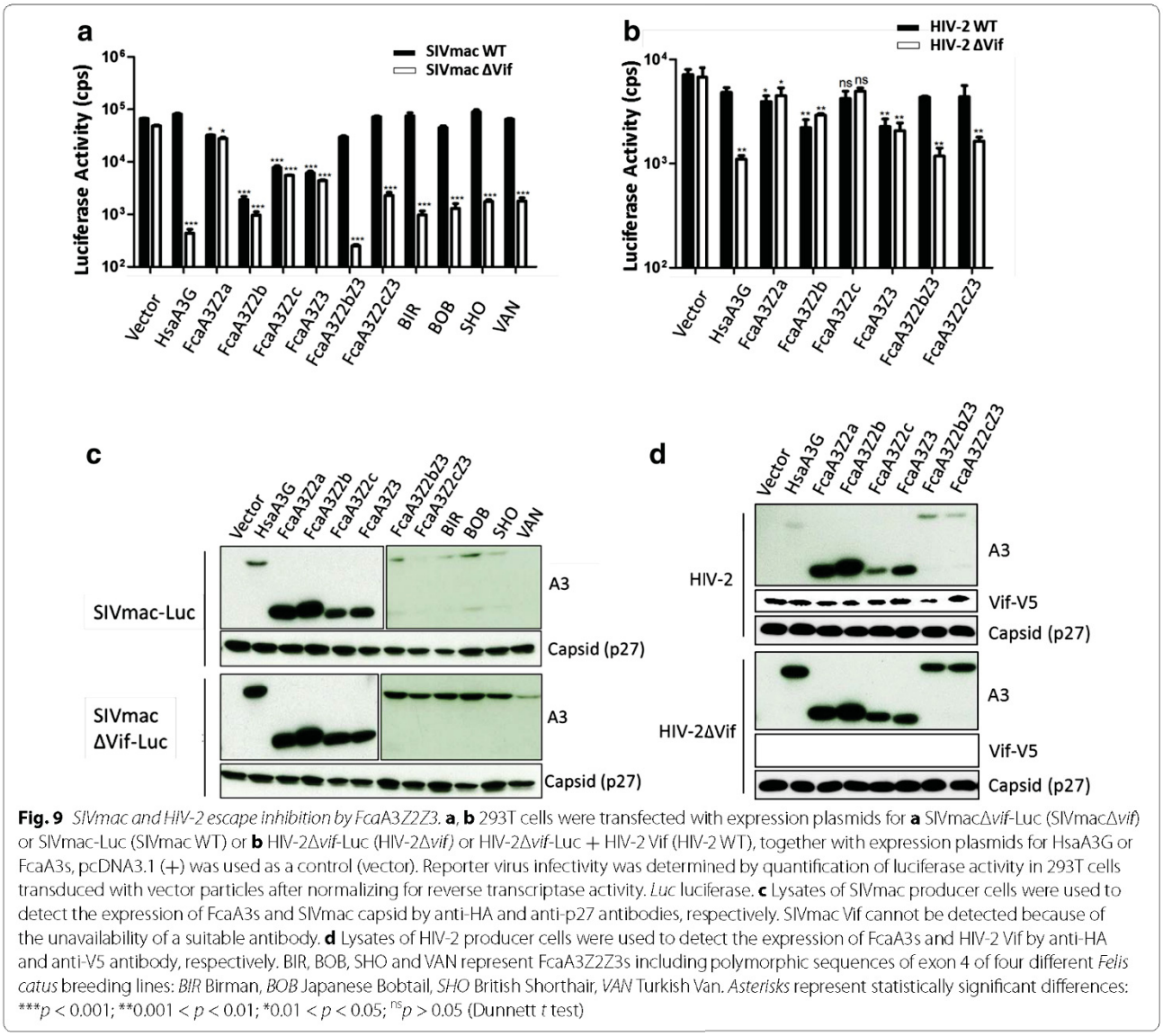


Fig. 9 *SIVmac and HIV-2 escape inhibition by FcaA3Z2Z3.* **a**, **b** 293T cells were transfected with expression plasmids for **a** SIVmacΔ*vif*-Luc (SIVmacΔ*vif*) or SIVmac-Luc (SIVmac WT) or **b** HIV-2Δ*vif*-Luc (HIV-2Δ*vif*) or HIV-2Δ*vif*-Luc + HIV-2 Vif (HIV-2 WT), together with expression plasmids for HsaA3G or FcaA3s, pcDNA3.1 (+) was used as a control (vector). Reporter virus infectivity was determined by quantification of luciferase activity in 293T cells transduced with vector particles after normalizing for reverse transcriptase activity. *Luc* luciferase. **c** Lysates of SIVmac producer cells were used to detect the expression of FcaA3s and SIVmac capsid by anti-HA and anti-p27 antibodies, respectively. SIVmac Vif cannot be detected because of the unavailability of a suitable antibody. **d** Lysates of HIV-2 producer cells were used to detect the expression of FcaA3s and HIV-2 Vif by anti-HA and anti-V5 antibody, respectively. BIR, BOB, SHO and VAN represent FcaA3Z2Z3s including polymorphic sequences of exon 4 of four different *Felis catus* breeding lines: *BIR* Birman, *BOB* Japanese Bobtail, *SHO* British Shorthair, *VAN* Turkish Van. *Asterisks* represent statistically significant differences: ***$p < 0.001$; **$0.001 < p < 0.01$; *$0.01 < p < 0.05$; $^{ns}p > 0.05$ (Dunnett *t* test)

HIV-1 Vif weakly interacts with feline A3Z2Z3

The finding that HIV-2 Vif counteracts one of the feline A3s reinforces the view that the initially described species-specificity of Vifs [63] is not absolute [36, 52]. For the generation of an HIV-1 animal model based on the cat, it would of advantage to understand whether feline A3 proteins are structurally accessible for HIV-1 Vif. We show here that HIV-1 fails to degrade feline A3 proteins (Figs. 1b, 10b) and appears only to bind weakly to the feline A3Z2Z3 protein compared to FIV Vif (Additional file 1: Fig. S10A).

The structural model of feline A3Z2Z3 was used to rationalize the binding of HIV-1 Vif to A3Z2Z3. When comparing the amino acid sequences of A3G and feline A3s, we noticed that the HIV-1 Vif binding domain 124-YYFWDPDY-131 is conserved in feline A3Z2 (Additional file 1: Fig. S10B). This domain spans amino acid residues with a well-characterized role in Vif-binding, such as 128-DPD-130 [65] and the recently characterized Y125 [66] in the β4-α4 loop of human A3G [67]. In the feline A3Z2 domain we find DPN instead of the DPD motif; however, in human A3G DPN binds to HIV-1 Vif as wild-type DPD [65]. As our structural model of A3Z2Z3 in comparison to the soluble N-terminal domain (sNTD) of A3G [68] revealed that the two regions around these residues are similarly accessible (Fig. 7b), we attempted to restore binding of HIV 1 Vif to feline A3Z2Z3 by a N133D mutation, resulting in a YYFWDPD133Y motif sequentially identical to the one in A3G (Additional file 1: Fig. S10b). We did not observe degradation of A3Z2Z3.

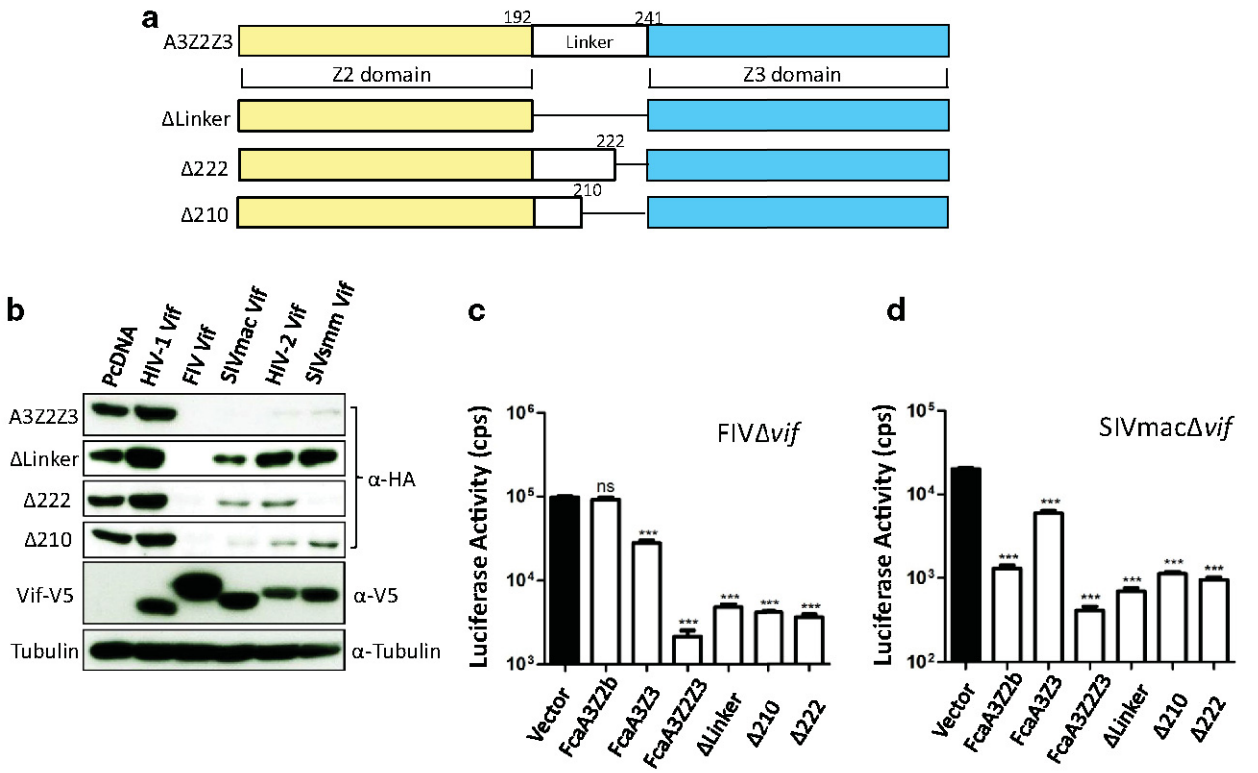


Fig. 10 The linker region in FcaA3Z2Z3 is important for HIV-2/SIVmac/smm Vif induced degradation. **a** Schematic representation of FcaA3Z2Z3 mutants (ΔLinker, Δ222, Δ210). Z2 and Z3 domains are shown as *yellow* or *blue rectangle*, respectively. The linker from amino acid 192–241 is shown as *white rectangle*. **b** The Vif sensitivity of FcaA3Z2Z3 mutants was analyzed by co-transfecting expression plasmids of the A3 mutants together with lentiviral Vifs (from HIV-1, FIV, SIVmac, HIV-2 or SIVsmm) into 293T cells. The expression of FcaA3s and Vifs were confirmed by immunoblots and anti-HA and anti-V5 antibodies, respectively. Cell lysates were also analyzed for equal amounts of total proteins using anti-tubulin antibody. **c** Test of antiviral activity of linker mutants against FIVΔ*vif*-luciferase and **d** against SIVmacΔ*vif*-luciferase. Infectivity of virions generated with wild-type and mutated A3 was measured on 293T cells transduced by RT activity normalized particles. *CPS* counts per second. *Asterisks* represent statistically significant differences: ***$p < 0.001$; **$0.001 < p < 0.01$; *$0.01 < p < 0.05$; $^{ns}p > 0.05$ (Dunnett *t* test)

N133D by HIV-1 Vif (Additional file 1: Fig. S10C), however; neither were mutations of P132 to introduce additional side chain interactions successful in that respect (Additional file 1: Figs. S10B, S10C). As to a possible explanation, for A3C, which is structurally highly similar to the sNTD of A3G [68], another motif of residues critical for Vif binding was found, centering on F75, Y86, F107, and H111 [69] (Fig. 7c). The sequentially equivalent residues of A3Z2Z3 are F78, Y89, F110, and Y114 such that the exchange of His versus Tyr may explain the failing of the binding of HIV-1 Vif. Another possible explanation for A3Z2Z3 is given by the occlusion of space required for HIV-1 Vif binding due to the presence of the predicted linker domain, where the long unstructured regions at the beginning and the end of the structured linker part may make it possible that the linker domain tips over the Z2 domain, that way shielding the putative HIV-1 Vif binding region (Fig. 7a, b).

Discussion

The A3 restriction factors are of extraordinary importance for the evolution and pathogenicity of lentiviruses and likely also of most other retroviruses. Here we identified A3 residues that are relevant for the FIV Vif interaction with both single-domain A3s, A3Z2 and A3Z3 (results are summarized in Table 1). In addition, we analyzed a unique A3 protein insertion domain called linker present in the feline A3Z2Z3 protein. The linker is suggested to form a homeo-box domain and mediates the sensitivity of A3Z2Z3 to degradation by Vifs of the HIV-2/SIVmac/smm group of primate lentiviruses.

Our knowledge about the interaction regions of A3s and of human and non-human lentivirus Vifs is limited. It was discussed that Vif is not simply a linker between the substrate A3 and the E3 ubiquitin ligase [70, 71]. In our study we investigated the interaction of three groups of Vif proteins (FIV, HIV-2/SIV, HIV-1) with feline A3s.

Table 1 Summary Vif-mediated A3 degradation

Feline A3[a]	Degradation by Vif				Rescue of infection by Vif[b]		
	FIV	HIV-1	HIV-2	SIVmac/smm	FIV	HIV-2	SIVmac
A3Z2	++	−	−	−	ND	−	−
A3Z2.DH-YN	−	ND	ND	ND	ND	ND	ND
A3Z3	++	−	−	−	++	−	−
A3Z3.A65I	±	ND	ND	ND	±	ND	ND
A3Z3.LI-AA	±	ND	ND	ND	±	ND	ND
A3Z3.A65I + LI-AA	−	ND	ND	ND	−	ND	ND
A3Z2Z3	++	−	++	++	++	++	++
A3Z2Z3-M	−	ND	++	++	−	ND	ND
A3Z2Z3ΔLinker	++	−	±	±	ND	ND	ND

Degradation of A3 by Vif: ++, mostly degraded; ±, partial degradation with high amount of Vif; −, no degradation

Rescue of infection by Vif: ++, complete rescue; ±, partial rescue; −, no rescue

ND not done

[a] Feline A3: A3s from domestic cat *Felis catus*

[b] Experiments to rescue the infection were not done with HIV-1 and SIVsmm

Previous experimental evidence described residue A65 in feline A3Z3 in modulating the sensitivity to FIV Vif [56]. We identified here two additional residues (L41, I42) in feline A3Z3 whose combined mutation resulted in an A3 protein that was resistant even to degradation by very high amounts of co-expressed FIV Vif. The mutated feline A3Z3 protein clearly showed reduced binding to FIV Vif, supporting the model that Vif binding to A3 is needed for A3 degradation. In feline A3Z2 residues D165 and H166 were also found to regulate the FIV Vif induced degradation, but mutations in these positions did not block the binding to FIV Vif in co-immunoprecipitation assays. This observation demonstrates that Vif binding to A3s is not sufficient for A3 degradation. Supporting evidence that Vif interaction is necessary but not sufficient is coming from reports describing that HIV-1 NL4-3 Vif binds A3C mutants, A3B and A3H without inducing APOBEC3 degradation [71–73]. The qualitative co-immunoprecipitation assays used in our study did not much differentiate the binding strength of individual Vif-A3 pairs, and it is very well possible that a weak interaction of e.g. HIV-1 Vif with feline A3Z2Z3 is below a threshold to form a stable E3 ligase complex. However, the binding of mutated feline A3Z2.DH-YN to FIV Vif appeared to be robust, indicating a more complex mechanism. Studies on HIV-1 Vif binding to human A3B and A3H similarly concluded that the interaction strength is not the only determinant for complete Vif-mediated degradation, and the individual interfaces of the A3-Vif pair additionally regulate degradation [72].

Recently, Richards et al. [74] presented a wobble model of the evolution of the Vif-A3 interaction. This model implicates that Vif forms several interactions, of which some are essential and some provide additional stabilizing contacts. Based on this idea, only if Vif forms a sufficient network of interactions with its A3 binding partner, a functional interaction is made. Suboptimal, destabilized interactions could be restored by the evolution of compensatory changes in Vif–A3 interface. It is thus possible that in feline A3Z3 residue A65 and L41, I42 are major independent interactions in the Vif-A3 interface, whereas in feline A3Z2 D165 and H166 represent one of the relevant interacting points for FIV Vif complex formation, while additional contact points still exist. Such a suboptimal Vif-A3 interaction might, for example, not be sufficient to facilitate E3 ligase conjugation of K48-linked polyubiquitin chains that are generally recognized by the proteasome.

The exact Vif-A3 interfaces are not known, because high-resolution structures have been only solved of single proteins such as of Z1- and Z2-domain human proteins (A3A, A3C), of the N-terminal Z2- and C-Terminal Z1-domain of human A3G, of the C-terminal Z2-domain of A3F and of HIV-1 Vif [67, 75, 76]. The structures of the full-length double domain A3s are unknown, however. Human A3Z1s and A3Z2s are globular proteins with six α-helices and five β-sheets arranged in a characteristic motif (α1-β1-β2/2′-α2-β3-α3-β4-α4-β5-α5-α6) [67, 76]. In human A3C, A3D and A3F, the HIV-1 Vif binding site is conserved and located in a hydrophobic cavity and on the surrounding surface of the α2, α3 and α4 helices [69, 77, 78]. In human A3G, HIV-1 Vif binds a surface different to the binding region in A3C/D/F, with residues Y125, 128-DPN-130 in the β4-α4 loop being important for HIV-1 Vif binding [65, 66]. In the human Z3 protein A3H, binding of HIV-1 Vif is mediated by residue 121

(either E or D) [79, 80]. Based on our structural model of feline A3Z2Z3 (Fig. 7b), we locate the residues important for FIV Vif binding in feline A3Z2Z3 at the domain boundary of the Z2 and the Z3 domains, distant to the binding motifs in human A3s (Fig. 7c, d).

In feline A3Z2, the presumed HIV-1 Vif-binding domain of human A3G, the β4-α4 loop, is conserved. Nevertheless, HIV-1 Vif fails to degrade feline A3Z2 or A3Z2Z3 despite the presence of the well-characterized residues DPN (in A3G residues 128–130) and Y125 [65, 66]. Based on our structural model, we suggest that the β4-α4 loop of feline A3Z2 is surface exposed. This suggests that the Z2-domain of human A3G contains in addition to the Y125, 128-DPN-130 motif residues for HIV-1 Vif binding that are absent or hidden in feline A3Z2 or A3Z2Z3. Indeed, the presences of such important residues outside this motif in A3G were recently postulated [68, 81]. In addition to FIV Vif, we and others found previously that HIV-2/SIVmac/smm Vifs induce degradation of feline A3Z2Z3 [46, 51] by possibly targeting the unique linker domain. The previously called linker, a domain insertion in feline A3Z2Z3, is not found in any double-domain A3 protein of human or mouse origin. Our modelling results suggest that the insertion forms a homeo-box domain-like structure that protrudes the Z2-Z3 structure.

In general, it appears that double-domain A3 proteins display stronger antiviral activities than single-domain A3s. The evolution of double-domain encoding A3 genes could thus have been most likely adaptive, as it significantly increased the host fitness against retroviral infections. Our results suggest that primates and felids could have evolved double-domain A3s through different routes. The sequence of the linker insertion is located in 5′UTR of the felid A3Z3 gene in exon 2, which is exclusively translated in read-through transcripts spanning the A3Z2 and A3Z3 genes in felines (Fig. 1a). The sequence encoding for exon 2 seems to be restricted to members of Felinae and Pantherinae. In this sense, the A3Z2Z3 linker region resembles an orphan domain specific to Feliformia, and the linker could thus be a synapomorphy of this clade. Nevertheless, homology searches identified an enrichment of significantly remote tBLASTn hits associated with regulatory or non-coding regions of A3 genes in the genomes of different species, in the carnivore Weddell seal and in the artyodactyls pig and camel. This concentration of sequences with a possible common origin with the feline A3Z2Z3 linker found in the close vicinity of the A3 genes in other species within Laurasiatheria suggests that the linker could have been recruited as a coding sequence into the feline A3Z2Z3 mature mRNA from a pre-existent non-coding possibly regulatory sequence, in an example of gain of function. This sequence could have been recruited after point mutation/s resulting in stop codon removal, introduction of frameshifts or unmasking previously cryptic functional sites [82] during the evolution of carnivores, after the split Caniformia/Feliformia but before the split Pantherinae/Felinae. In the case of primates and of rodents there are no descriptions of read-through transcripts of single domain A3s resulting in mRNAs encoding double-domain A3s. Instead, the human heterologous double domain A3s (i.e. A3B and A3G, both being A3Z2Z1) or homologous double domain A3s (i.e. A3D and A3F, both being A3Z2Z2) could have evolved after the fusion of head-to-tail duplicated genes, as the several rounds of gene duplication in the evolutionary history of the A3 locus in primates suggest [3].

During our evolutionary analysis of the A3Z3 genes, we found here for the first time duplicated A3Z3 genes. A3Z3 duplications were identified in the genomes of different carnivores (the giant panda, the polar bear, the Weddell seal and the walrus), but were not found in dog and ferret and also not in any felid. The most parsimonious hypothesis would be that a duplication event occurred within Caniformia, after the basal split of Canidae. However, given the inferred position of the most recent common ancestor of all A3Z3 in carnivores, and given the within-clades and between-clades evolutionary distances (Additional file 1: Fig. S3), we propose that an ancient A3Z3 duplication event may have occurred prior to the Caniformia/Feliformia split. One of the in-paralogs would have disappeared in the Felidae ancestor, and at least in the dog genome, while both copies would have been maintained in most lineages within Caniformia (the absence in the ferret genome should be confirmed when better quality data are available).

Conclusions

Host-virus arms races formed the Vif-A3 interactions. Our data support that the evolution of HIV-1, HIV-2 and FIV follow intrinsic currently unexplained evolutionary pathways adapting to the antiviral A3 repertoire. This study also revealed that the A3 gene evolution included newly identified duplications (in-paralogs) of A3Z3 genes in some caniformia and the inclusion of a homeobox-domain in the feline A3Z2Z3 protein. This homeobox domain insertion may reflect a transitional situation (read-through transcription) of the evolutionary development of double Z-domain containing A3 proteins. Further resolution of the interaction surface of feline A3s with Vif proteins will help us to understand the biochemistry of these interactions and may give us tools to explore the HIV-1 Vif interaction with human A3s.

Methods

Cells and transfections

HEK293T (293T, ATCC CRL-3216), HOS (ATCC CRL-1543) and TZM-bl cells (NIH AIDS Reagent program [83, 84]) were maintained in Dulbecco's high-glucose modified Eagle's medium (DMEM, Biochrom, Berlin, Germany) supplemented with 10 % fetal bovine serum (FBS), 2 mM L-glutamine, penicillin (100 U/ml), and streptomycin (100 µg/ml). Stable A3 expressing cells: FcaA3Z2Z3 wild type and mutant pcDNA-constructs were digested by BglII, and then were transfected into HOS.CD4.CCR5 cells using Lipofectamine LTX (Thermo Fisher Scientific, Schwerte, Germany) according to manufacturer's instruction, cells stably express feline A3s were selected by 750 µg/ml G418 (Biochrom, GmbH) in the following 3 weeks. The A3s degradation experiments were performed in 24-well plates, 1×10^5 293T cells were transfected with 250 ng A3s expression plasmids together with 250 ng HIV-1, HIV-2, SIVmac and SIVsmm Vif expression plasmids or 20 ng codon-optimized FIV Vif expression plasmid, pcDNA3.1 (+) (Life Technologies) was used to fill the total plasmid to 500 ng. To produce FIV-luciferase viruses, 293T cells were co-transfected with 0.6 µg FIV packaging construct, 0.6 µg FIV-luciferase vector, 1 µg A3 expression plasmid, 0.1 µg VSV-G expression plasmid; in some experiments pcDNA3.1 (+) (Life Technologies) was used instead of Vif or A3 expression plasmids. For HIV-2 and SIVmac-luciferase transfections, 1.2 µg HIV-2-Luc and SIVmac-Luc plasmids were used instead of FIV plasmids. At 48 h post transfection, cells and supernatants were collected.

Vif and A3 plasmids

FIV-34TF10 (codon-optimized), HIV-1, HIV-2, SIVmac and SIVsmm Vif genes were inserted into pcWPRE containing a C-terminal V5 tag [36]. HIV-1 Vif represents always HIV-1 Vif from clone NL4-3, except specifically stated LAI. HIV-1 Vif LAI is a gift from Viviana Simon and does not contain a protein tag [54]. pCPRΔ*env* FIV gag-pol plasmid that in addition expresses Vif was described previously [57]. FIV-Lion Vif gene (FIV_{Ple} subtype B, accession number EU117991) was synthesized and codon-optimized. FIV-Lion Vif expression plasmid was generated by cloning codon-optimized FIV-Lion Vif fragment containing a V5 tag into pcWPRE using EcoRI and NotI. All A3s are expressed a carboxy-terminal hemagglutinin (HA) tag. Domestic cat and big cat (*Pantherinae*) A3s were described previously [36]. Human A3C (HsaA3C) and feline A3Z2b (FcaA3Z2b) chimeras were made by overlapping extension PCR. HsaA3C/FcaA3Z2 chimera Z2C1, Z2C4 and Z2C5 contain residues 1–22, 1–131 and 1–154 of FcaA3Z2, respectively; the remaining C-terminal fragments are derived from human A3C. The 5′ and 3′ fragments were amplified separately by using primer pairs (Additional file 2: Table S1); two fragments were then mixed and amplified with the two external primers (Additional file 2: Table S1). To make HsaA3C/FcaA3Z2 chimera Z2C30, the first fragment was amplified by primers feApo3.fw and hufe3C 485.rv using chimera Z2C4 as a template, the second fragment was amplified by primers hufe3C 485.fw and HA-rv using FcaA3Z2 as a template, the two fragments were mixed and amplified with the two external primers. The FcaA3Z2b mutants were generated by fusion PCR using primer pairs described in Additional file 2: Table S1. The final products of HsaA3C/FcaA3Z2 chimeras and FcaA3Z2 mutants were cloned into pcDNA3.1 (+) using HindIII and XhoI restriction sites. The HsaA3H/FcaA3Z3 chimeras were constructed by the same method using primer pairs listed in Additional file 2: Table S2. To make FcaA3Z2bZ3-M, the PCR products of FcaA3Z2b DH-YN and FcaA3Z3 A65I + LI-AA were fused, and then inserted into pcDNA3.1 (+) by EcoRI and NotI restriction sites. The FcaA3Z2Z3 mutation constructs were generated by using the primers shown in Additional file 2: Table S3.

Viruses and infection

To produce FIV single-cycle luciferase viruses (FIV-Luc), 293T cells were co-transfected with the replication deficient packaging construct pFP93, a gift from Eric M. Poeschla [85], which only expresses *gag*, *pol*, and *rev*; the FIV luciferase vector pLinSin [4]; a VSV-G expression plasmid pMD.G; FcaA3s expression plasmids; FIV Vif expression plasmid; or empty vector pcDNA3.1 (+). To produce SIV-Luc viruses, 293T cells were co-transfected with SIVmac-Luc (R-E-); or SIVmac-Luc (R-E-) Δ*vif* [63]; and FcaA3s expression plasmids. HIV-2-Luc was produced by co-transfecting 293T cells with HIV-2 packaging plasmid pHIV2Δ4 [86]; transfer vector plasmid HIV-2-luc (SV40) [64]; pMD.G, together with FcaA3s expression plasmids or empty vector pcDNA3.1 (+) and HIV-2 Vif-V5 expression plasmid or pcDNA3.1 (+) empty vector. The reverse transcriptase (RT) activity of FIV, SIVmac and HIV-2 were quantified by using the Cavidi HS lenti RT kit (Cavidi Tech, Uppsala, Sweden). For reporter virus infection, 293T cells were seeded in 96-well plate 1 day before transduction. After normalizing for RT activity, the same amounts of viruses were used for infection. Three days post transduction, firefly luciferase activity was measured with the Steadylite HTS reporter gene assay system (Perkin-Elmer, Cologne, Germany) according to the manufacturer's instructions on a MicroLumat Plus luminometer (Berthold Detection Systems, Pforzheim, Germany). Each sample was performed transduction in triplicates; the error bar of each triplicate

was shown. Replication-competent HIV-1 plasmids NL-BaL.vif_{FIV} were described previously [36]. NL-BaL.vif_{FIV} virus stocks were prepared by collecting the supernatant of transfected 293T cells. The kinetics of viral spreading replication was determined with HOS.CD4.CCR5.FcaA3s cells by infection with MOI 0.01 of NL-BaL.vif_{FIV}. Spreading virus replication was quantified over 15 days by infecting 10 µl supernatant to TZM-bl cells. All experiments were repeated independently at least three times.

Immunoblot analysis

Transfected 293T cells were lysed in radioimmunoprecipitation assay (RIPA) buffer (25 mM Tris–HCl [pH7.6], 150 mM NaCl, 1 % NP-40, 1 % sodium deoxycholate, 0.1 % sodium dodecyl sulfate [SDS], protease inhibitor cocktail set III [Calbiochem, Darmstadt, Germany]). The expression of FcaA3s and lentivirus Vif were detected by mouse anti-hemagglutinin (anti-HA) antibody (1:7500 dilution, MMS-101P; Covance, Münster, Germany) and mouse anti-V5 antibody (1:4500 dilution, MCA1360, ABDserotec, Düsseldorf, Germany) separately, the tubulin and SIV capsid protein were detected using mouse anti-α-tubulin antibody (1:4000, dilution, clone B5-1-2; Sigma-Aldrich, Taufkirchen, Germany), HIV Vif LAI was detected by HIV-1 Vif monoclonal antibody (#319) (NIH AIDS Reagent Program [87]) and mouse anti-capsid p24/p27 MAb AG3.0 (1:50 dilution [88]) separately, followed by horseradish peroxidase-conjugated rabbit anti-mouse antibody (α-mouse-IgG-HRP; GE Healthcare, Munich, Germany), and developed with ECL chemiluminescence reagents (GE Healthcare). Encapsidation of FcaA3 proteins into FIV particles: HEK293T cells were transfected with 600 ng pFP93, 600 ng of pLinSin, 100 ng pMD.G and 1000 ng of FcaA3 constructs. Viral supernatants were collected 48 h later, overlaid on 20 % sucrose and centrifuged for 4 h at 14,800 rpm in a table top centrifuge. Viral pellet was resuspended in RIPA buffer, boiled at 95 °C for 5 min with Roti load reducing loading buffer (Carl Roth, Karlsruhe, Germany) and resolved on a SDS-PAGE gel. The FcaA3s and tubulin proteins were detected as the above method. VSV-G and FIV p24 proteins were detected using mouse anti-VSV-G antibody (1:10,000 dilution; clone P5D4; Sigma-Aldrich) and mouse anti-FIV p24 antibody (1:2000 dilution; clone PAK3-2C1; NIH AIDS REPOSITORY) separately, followed by horseradish peroxidase-conjugated rabbit anti-mouse antibody (α-mouse-IgG-HRP; GE Healthcare, Munich, Germany), and developed with ECL chemiluminescence reagents (GE Healthcare).

Immunofluorescence and flow cytometry

HOS cells grown on polystyrene coverslips (Thermo Fisher Scientific, Langenselbold, Germany) were transfected with expression plasmids for FcaA3 wild-type and mutants or together with FIV Vif-TLQAAA using Lipofectamine LTX (Life Technologies). At day one post transfection, cells were fixed in 4 % paraformaldehyde in PBS for 30 min, permeabilized in 0.1 % Triton X-100 in PBS for 15 min, incubated in blocking buffer (FBS in PBS) for 1 h, and then cells were stained by mouse anti-HA antibody in a 1:1000 dilution in blocking solution for 1 h. Donkey anti-mouse Alexa Fluor 488 (Life Technologies) was used as a secondary antibody in a 1:300 dilution in blocking solution for 1 h. FIV Vif-TLQAAA was stained by rabbit anti-V5 antibody in a 1:1000 dilution in blocking solution for 1 h. Donkey anti-rabbit Alexa Fluor 594 (Life Technologies) was used as a secondary antibody in a 1:300 dilution in blocking solution for 1 h. Finally, DAPI was used to stain nuclei for 2 min. The images were captured by using a 40× objective on a Zeiss LSM 510 Meta laser scanning confocal microscope (Carl Zeiss, Cologne, Germany). To analyze CD4 and CCR5 expression level of HOS.CD4.CCR5.FcaA3s, cells were stained by α-hCD4 PE mouse $IgG1_k$ (Dako, Hamburg, Germany) and α-hCCR5 FITC (BD Bioscience, Heidelberg, Germany) separately according to the manufacturer's instruction. The measurement was carried out by BD FACSanto (BD Bioscience). Data analysis was done with the Software FlowJo version 7.6 (FlowJo, Ashland, USA).

Immunoprecipitation

To determine Vif and A3 binding, 293T cells were co-transfected with 1 µg FIV Vif TLQAAA-V5 and 1 µg FcaA3 wild-type or mutants or pcDNA3.1 (+). 48 h later, the cells were lysed in IP-lysis buffer (50 mM Tris/HCl pH 8, 1 mM PMSF, 10 % Glycerol, 0.8 % NP-40, 150 mM NaCl, and protease inhibitor cocktail set III (Calbiochem, Darmstadt, Germany). The lysates were cleared by centrifugation. The supernatant were incubated with 20 µl α-HA Affinity Matrix Beads (Roche) at 4 °C for 2 h. The samples were washed 5 times with lysate buffer on ice. Bound proteins were eluted by boiling the beads for 5 min at 95 °C in SDS loading buffer. Immunoblot analysis and detection were done as described.

3D-PCR

293T cells (5×10^5 cells/well in a 6-well plate) were transfected with 600 ng pFP93, 600 ng pLinSin, 100 ng pMD.G and 1000 ng FcaA3s expression plasmids or pcDNA3.1 (+) as a control. 48 h later, the viral supernatant was harvested, filtered (0.45 µm) and treated with DNase I (Life Technologies) at 37 °C for 1 h. 200 µl of supernatant was used for infecting 293T cells. 12 h post transduction, 293T cells were washed with PBS and DNA was isolated using DNeasy blood and tissue kit (Qiagen, Hilden, Germany). A 714-bp fragment of within the spliced luciferase

gene was amplified using the primers 5′-GATATGTGGATTTCGAGTCGTC-3′ and 5′-GTCATCGTCTTTCCGTGCTC-3′. For selective amplification of the hypermutated products, the PCR denaturation temperature were lowered stepwise from 87.6 to 83.5 °C (83.5, 84.2, 85.2, 86.3, 87.6 °C) using a gradient thermocycler. The PCR parameters were as follows: (1) 95 °C for 5 min; (2) 40 cycles, with 1 cycle consisting of 83.5–87.6 °C for 30 s, 55 °C for 30 s, 72 °C for 1 min; (3) 10 min at 72 °C. PCRs were performed with recombination *Taq* DNA polymerase (Thermo Fisher Scientific).

Purification of GST tagged proteins and pull down assay

Feline A3Z2 and A3Z3 coding sequences were cloned in pGEX-6P2 vector (GE healthcare) with a C terminal HA tag to produce fusion proteins GST-FcaA3Z2-HA and GST-FcaA3Z3-HA (PCR primer in Additional file 2: Table S4). GST alone and fusion proteins were overexpressed in *E. coli* Rosetta (DE3) cells (EMD Millipore, Darmstadt, Germany) and purified by affinity chromatography using Glutathione Sepharose 4B beads (GE healthcare). After the culture of transformants until 0.6 OD_{600}, cells were induced with 1 mM isopropyl-beta-D-thiogalactopyranoside (IPTG) and 1 μM $ZnSO_4$ and cultured at 18 °C overnight. GST and Feline A3Z2/Z3 harboring cells were washed with PBS and lysed with 1× Bug buster protein extraction reagent (EMD Millipore) containing 50 mM Tris (pH 7.0), 10 % glycerol, and 1 M NaCl clarified by centrifugation and the soluble protein fraction was mixed with pre-equilibrated glutathione Sepharose beads. After 3 h incubation at 4 °C in end-over-end rotation, the beads were washed thrice with wash buffer containing 50 mM Tris (pH 8.0), 10 % glycerol and 500 mM NaCl and a single wash with the mild lysis buffer (50 mM Tris (pH 8), 1 mM PMSF, 10 % glycerol, 0.8 % NP-40, 150 mM NaCl and 1× complete protease inhibitor). These GST protein bound beads are used for the subsequent binding assay. GST pull down assay to detect direct binding with Vif of FIV: The protocol of protein–protein interactions was adapted from a previously described procedure [89]. HEK293T cells were transfected with 1.5 μg of FIV Vif-V5 coding plasmid and incubated for 48 h. Soluble protein fraction of HEK293T cells were obtained by lysing the cells with mild lysis buffer (50 mM Tris (pH 8), 1 mM PMSF, 10 % glycerol, 0.8 % NP-40, 150 mM NaCl, and 1× complete protease inhibitor (Calbiochem) and a 30 min centrifugation at 14,800 rpm. A fraction of the supernatant was kept for immunoblots; remaining lysates were equally added on the bead samples GST, GST-FcaA3Z2-HA and GST-FcaA3Z3-HA and incubated overnight at 4 °C in end-over-end rotation. Next day, the beads were washed thrice with the mild lysis buffer and the GST protein and protein complexes were eluted by adding wash buffer containing 25 mM reduced glutathione. A fraction of the eluted proteins (equal amount) were boiled at 95 °C for 5 min with Roti load reducing loading buffer (Carl Roth) and resolved on a SDS-PAGE gel. FIV Vif and GST-FcaA3s were detected by anti-V5 and -HA antibody, respectively. Coomassie brilliant blue stained gel was also added to show the purity of GST and FcaA3 fusion proteins.

Evolutionary analyses

The initial set of A3 sequences was taken from Münk and coworkers [3]. These sequences were used as seeds for BLASTn, tBLASTn and BLAT searches to recover additional A3 sequences from genomes in Carnivora. The final dataset (closed on November 2015) contained four A3Z1 sequences from four Caniformia species, six A3Z2 sequences from six Caniformia species, eleven A3Z2 sequences from six Feliformia species, ten A3Z3 sequences from five Caniformia species and five A3Z3 sequences from five Feliformia species. Sequences were aligned at the amino acid level with MUSCLE [90]. The final alignment encompassed 629 and 269 alignment patterns at the nucleotide level and at the amino acid level, respectively. Phylogenetic inference was performed with RAxML_v8.2 [91, 92] using the GTR + 4Γ model at the nucleotide level and LG + Γ model at the amino acid level, the model choice done after initial maximum likelihood searches with RAxML. Additional phylogenetic inference was performed separately for the A3Z2 and A3Z3 genes using the same settings. In all cases, no significant differences between amino acid and nucleotide tree topologies were observed using the Shimodaira–Hasegawa test [93]. Phylogenetic supernetworks were constructed with SplitsTree_v4 [94] using 1000 either nucleotide or amino acid bootstrapped maximum likelihood trees. Selection on individual codons was inferred under a Bayesian framework with SELECTON V2.4 (http://selecton.tau.ac.il/) [95] contrasting the M8 and M8a models, and with DATAMONKEY (http://www.datamonkey.org/) using the Random Effects Likelihood (REL) model [96].

Statistical analysis

Data are represented as the mean with SD in all bar diagrams. Statistically significant differences between two groups were analyzed using the unpaired Student's t test with GraphPad Prism version 5 (GraphPad software, San Diego, CA, USA). Validity of the null hypothesis was verified with significance level at α value = 0.05.

Homology modelling of feline A3Z2Z3 protein

The homology modeling of the linker region of the feline A3Z2Z3 was performed in several steps: First, the

in-house meta-tool TopModel [59, 60] was used to compute a consensus alignment for the feline A3 sequences to the structural model of the human A3G [58] using 13 different alignment programs (Additional file 2: Table S5). From the consensus alignment, the feline A3 linker was identified and then submitted to TopModel for automated structure prediction using eight state-of-the-art threading programs (Additional file 2: Table S5). The identified templates (2YS9, chain A (19.4 %); 2MMB, chain A (17.1 %); 2DA4, chain A (14.7 %); 2LFB, chain A (9.2 %) and 1FTZ, chain A (12.9 %); sequence identities with respect to the linker are given in parentheses) were aligned to the linker sequence with TopModel using threading, sequence, and structural alignment programs, to produce a large alignment ensemble from every combination of the top three ranked templates and the target sequence. These alignments were modeled using Modeller9.1 [97], refined with RASP [98], and ranked using the in-house meta-tool for model quality assessment TopScore (D. Mulnaes, H. Gohlke, unpublished results), which combines quality assessments from eight different model quality assessment programs (Additional file 2: Table S5). The top ranked models for each template combination were refined with ModRefiner [99] and used as templates for a second round of modeling where bad scoring regions were removed. The resulting models were re-ranked and refined, and the top ranking model was selected as the linker representative. The model of the rest of the feline A3Z2Z3 was made with TopModel in a similar fashion using the feline–human consensus alignment. The linker domain was manually positioned near the linker region gap, unstructured parts were connected to the rest of the feline A3Z2Z3 and minimized using the MAB force field [100] as implemented in Moloc, thereby keeping all other protein atoms fixed.

Homology modelling of human A3H and feline A3Z2b and A3Z3 proteins

The models of the three proteins were built using the default settings in TopModel and all possible combinations of the top three ranked templates in each case. For the human A3H model, the templates were: PDB ID 4 J4 J, chain A, 35 % identity/96 % coverage; 2KBO, chain A, 37/95 %; 2RPZ, chain A, 30/94 %, resulting in a model with 84 % accuracy according to TopScore. For the feline A3Z2b model, the templates were: 3VM8, chain A, 42/94 %; 2KBO, chain A, 39/94 %; 1M65, chain A, 10/87 %, resulting in a model with 88 % accuracy according to TopScore. For the feline A3Z3 model, the templates were: 4J4J, chain A, 31/91 %; 2KBO, chain A, 36/90 %; 2RPZ, chain A, 24/92 %, resulting in a model with 84 % accuracy according to TopScore.

Additional files

Additional file 1: Figure S1. *Cellular localization of feline A3s and FIV Vif.* HOS cells were transfected with FcaA3Z2b, FcaA3Z3, or FcaA3Z2Z3 (all with HA-tag), together with FIV Vif-TLQAAA. To detect A3 (green) immunofluorescence, staining was performed with an anti-HA antibody. To detect FIV Vif (red) immunofluorescence, staining was performed with an anti-V5 antibody. Nuc ei (blue) were visualized by DAPI staining. **Figure S2**. *Comparison of protein sequences of A3s and Vif.* (A, B) The sequence alignment of (A) FcaA3Z2 (FcaA3Z2b), (B) FcaA3Z3 and big cat A3 proteins. The D165-H166 and L40-I41 + A65 domains that are essential for FIV Vif induced degradation are marked by red boxes. (C) Sequence alignment of domestic cat FIV Vif (FIVfca subtype 34TF10) and lion FIV (FIVple subtype E) Vif. The C187 and C190 that are essential for induced FcaA3s degradation and marked the presumed BC box (TLQ/SLQ) marked by red boxes. (D) Sequence alignment of HIV-1 (strain NL4-3) and HIV-2 (strain RodA) Vif. The CUL5 box (HCCH) and BC box (SLQ) were marked by red boxes. Pti, Ple, Lly and Pco represent *Panthera tigris corbetti; Panthera leo bleyenberghi; Lynx lynx; Puma concolor*. **Figure S3**. *Evolutionary supernetwork of A3 sequences retrieved from carnivores.* The network was constructed with SplitsTree_v4 using 1,000 maximum likelihood bootstrapped trees created with RAxML_v8.2. Scale bar is given in substitutions per site. The approximate position of the root obta ned using maximum likelihood inference with all A3Z1, A3Z2 and A3Z3 sequences from carnivores is indicated in grey. (A) The evolutionary distances among A3Z3 sequences within the two in-paralogs within Canifcrmia (upper branches and left branch) and within Feliformia (right b·anches) are indicated as overall average pairwise nucleotide distance ± bootstrap standard error estimate. For each tip, the actual sequence orthologous to positions 38-44 in the *F. catus* A3Z3 gene are given in parentheses. The inset displays the evolutionary relationships among the Carnivcra species for which we have identified A3Z3 paralogs. (B) The evoluticnary distances among A3Z2 sequences within Canifornia (upper branches) and within Feliformia (lower branches) are indicated as overall aver:ge pairwise nucleotide distance ± bootstrap standard error estimate. For each tip, the actual sequence orthologous to positions 165-170 in the *F. catus* A3Z2 genes are given in parentheses. **Figure S4**. *The mutations in FcaA3 that cause resistance to FIV Vif do not alter the cellular distribution of FcaA3s.* HOS cells were transfected with FcaA3Z2 (FcaA3Z2b), FcaA3Z2.DH-YN, FcaA3Z3, FcaA3Z3.A65I + LI-AA, FcaA3Z2Z3 or FcaA3Z2Z3-M (all with HA-tag). To detect A3 (green) immunofluorescence, staining was performed with an anti-HA antibody. Nuclei (blue) were visualized by DAPI staining. **Figure S5**. *Sequence alignment of feline APOBEC3.* Sequence aligrment of feline A3Z2Z3 and human A3G as generated by the TopModel approach. The Z2 and Z3 domains are underlined in yellow and blue, respectively. The sequence of the linker domain is underlined in magenta. Helical regions and β-strands are depicted as red helices and green arrows, respectively. In addition, the alignments of three template structures used to model the structure of the linker domain are given (PDB IDs 2YS9, 2DA4, 2MMB). **Figure S6**. *Expression of CD4 and CCR5 receptors on the surface of HOS (red) and HOS.CD4.CCR5 cells (blue) expressing feline A3Z2Z3 or A3Z2Z3-M.* CD4 and CCR5 were detected by flow cytometry and anti-CD4 and anti-CCR5 antibodies. Numbers indicate the percentage of positive cells. HOS cells served as background control. **Figure S7**. *The mutated FcaA3s are encapsidated and inhibit FIV by cytidine deamination.* (A) The mutated FcaA3s can inhibit the infectivity of FIVΔ*vif* reporter viruses. 293T cells were co-transfected with plasmids for FIVΔ*vif* luciferase together with FcaA3s. 48 h later, supernatant normalized for reverse transcriptase activity was used to transduce 293T cells. Luciferase activity was determined two days post transduction. Asterisks represent statistically significant differences: ***, $p < 0.001$; **, $0.001 < p < 0.01$; *, $0.01 < p < 0.05$; ns, $p > 0.05$ [Dunnett *t* test]. (B) Immunoblot of FIV producer cells and VLPs used for (A). Encapsidation of wild-type and mutated feline A3s into FIVΔ*vif* virus like particles (VLPs), A3 proteins were detected by anti-HA antibody. Tubulin detection for equal loading of cell lysate was done using anti-tubulin, for demonstration of equal loading of FIV VLPs VSV-G and FIV p24 proteins were detected by anti-VSV-G and anti-FIV p24 antibodies separately. (C) Encapsidated wild-type and mutated FcaA3s deaminate cytidines FIV genomes. FIVΔ*vif* was produced in the absence and presence of wild-type and mutant FcaA3s (FcaA3Z3,

FcaA3Z3.A65I + LI-AA, FcaA3Z2Z3, FcaA3Z2Z3-M) or HsaA3G. The vector particles were used to infect 293T cells. 12 h later, the total cellular DNA was extracted and differential DNA denaturation PCR (3D-PCR) was performed. Td: denaturing temperature. **Figure S8**. *Vif titration on FcaA3 linker mutants.* Co-transfection of increasing amounts of expression plasmids for (A) HIV-2 Vif and (B) SIVmac Vif with constant amounts of the indicated A3 expression plasmids. The expression of FcaA3s and Vifs were analyzed by anti-HA and anti-V5 antibodies, respectively. Cell lysates were also analyzed for equal amounts of total proteins using anti-tubulin antibody. **Figure S9**. *Expression and encapsidation of feline A3 linker mutants using (A) FIVΔvif and (B) SIVmacΔvif.* Immunoblots of corresponding experiments shown in Fig. 10C and 10D. Immunoblots of lysates of virus producer cells (cell) and virus particles (VLP). A3s were detected by anti-HA antibodies, cell lysates were also analyzed for equal amounts of total proteins using anti-tubulin antibody and VLP lysates using anti-VSV-G antibody. **Figure S10.** *HIV-1 Vif cannot target the "YYFWDPN/DY" domain in FcaA3.* (A) CO-IP of feline A3Z2Z3 (HA tag) with either HIV-1 Vif (V5 tag) or FIV Vif (TLQAAA mutant, V5 tag). A3Z2Z3 immune precipitated and detected by anti-HA antibody, co-precipitated Vif was detected by anti-V5 antibody. (B) Comparison of the "YYFWDPN/DY" domain in HsaA3G and FcaA3Z2Z3 and derived mutations generated in FcaA3Z2Z3, the mutated residues shown in bold. (C) FcaA3Z2Z3 mutants were investigated for being sensitive for degradation by HIV-1 Vif. Expression plasmids of FcaA3Z2Z3 mutants of HsaA3G were co-transfected together with HIV-1 Vif into 293T cells. 48 h later, Cell lysates were used to detect the expression of FcaA3Z2Z3 and HIV-1 Vif by anti-HA and anti-V5 antibodies, respectively. Cell lysates were also analyzed for equal amounts of total proteins using anti-tubulin antibody.

Additional file 2: Table S1. Primer list used for HsaA3C/FcaA3Z2 chimeras and FcaA3Z2 mutants. **Table S2**. Primer list used for HsaA3H/FcaA3Z3 chimeras and FcaA3Z3 mutants. **Table S3**. Primer list used for FcaA3Z2Z3 mutants. **Table S4**. Primer used to clone GST fusion constructs. **Table S5**. The software used in TopModel for threading, alignment and model quality estimation[a].

Abbreviations

FIV: feline immunodeficiency virus; HIV: human immunodeficiency virus; SIV: simian immunodeficiency virus; APOBEC3: apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like; Vif: viral infectivity factor; UTR: untranslated region.

Authors' contributions

ZZ, QG, AAJV, AH, BPK, and SH conducted experiments. DM, SHJS and HG generated the feline A3 homology model. IGB analyzed the evolutionary origin of feline A3s. KS, KC, DH, SHJS, HG and CM analyzed data and conceived experiments. IGB, SHJS, HG and CM wrote the manuscript. CM conceived the study. All authors read and approved the final manuscript.

Author details

[1] Clinic for Gastroenterology, Hepatology, and Infectiology, Medical Faculty, Heinrich-Heine-University Düsseldorf, Building 23.12.U1.82, Moorenstr. 5, 40225 Düsseldorf, Germany. [2] Department of Medical Biotechnology, Paul-Ehrlich-Institute, Paul-Ehrlich-Str. 51-59, 63225 Langen, Germany. [3] Institute of Biochemistry, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany. [4] Institute of Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany. [5] Laboratory of Viral Pathogenesis, Institute for Virus Research, Kyoto University, Kyoto 6068507, Japan. [6] CREST, Japan Science and Technology Agency, Saitama 3220012, Japan. [7] MIVEGEC (UMR CNRS 5290, IRD 224, UM), National Center of Scientific Research (CNRS), 34394 Montpellier, France. [8] Present Address: BioNTech RNA Pharmaceuticals GmbH, Anderen Goldgrube 12, 55131 Mainz, Germany.

Acknowledgements

We thank Wioletta Hörschken for excellent technical assistance, Rebecca Clemens for discussion and sharing of preliminary data, Yannick Bulliard and Didier Trono for the model of the full-length A3G. We thank Nathanial R. Landau, Garry Nolan, Eric Poeschla, and Viviana Simon for reagents. The following reagents were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: Monoclonal antibody to HIV-1 p24 (AG3.0) from Jonathan Allan, TZM-bl cells from John C. Kappes, Xiaoyun Wu and Tranzyme Inc., HIV-1 Vif monoclonal antibody (#319)) from Michael H. Malim and anti-FIV p24 monoclonal (PAK3-2C1).

Competing interests

The authors declare that they have no competing interests.

Funding

ZZ and QG are supported by a scholarship from China Scholarship Council; CM is supported by the Heinz Ansmann foundation.

Received: 8 March 2016 Accepted: 9 June 2016

Published online: 01 July 2016

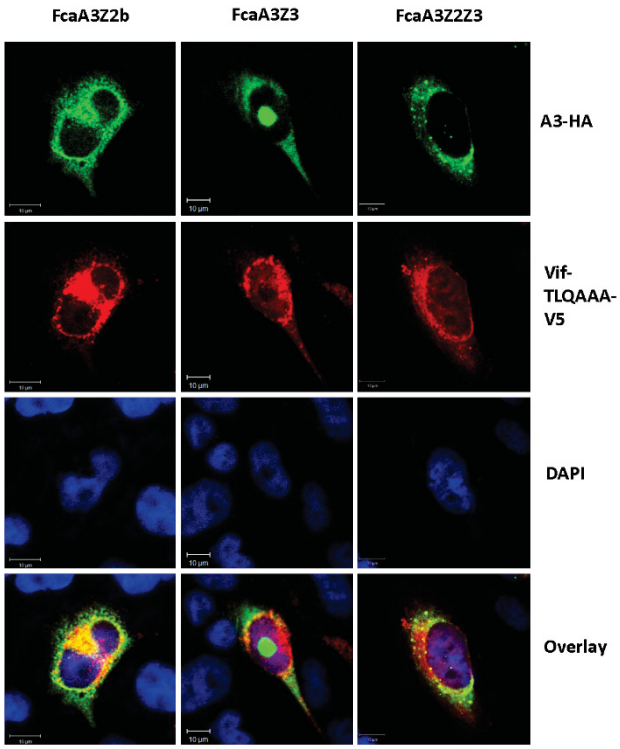
References

1. LaRue RS, Jonsson SR, Silverstein KA, Lajoie M, Bertrand D, El-Mabrouk N, Hotzel I, Andresdottir V, Smith TP, Harris RS. The artiodactyl APOBEC3 innate immune repertoire shows evidence for a multi-functional domain organization that existed in the ancestor of placental mammals. BMC Mol Biol. 2008;9:104.
2. LaRue RS, Andresdottir V, Blanchard Y, Conticello SG, Derse D, Emerman M, Greene WC, Jonsson SR, Landau NR, Löchelt M, et al. Guidelines for naming nonprimate APOBEC3 genes and proteins. J Virol. 2009;83(2):494–7.
3. Münk C, Willemsen A, Bravo IG. An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. BMC Evol Biol. 2012;12:71.
4. Münk C, Beck T, Zielonka J, Hotz-Wagenblatt A, Chareza S, Battenberg M, Thielebein J, Cichutek K, Bravo IG, O'Brien SJ, et al. Functions, structure, and read-through alternative splicing of feline APOBEC3 genes. Genome Biol. 2008;9(3):R48.
5. Jarmuz A, Chester A, Bayliss J, Gisbourne J, Dunham I, Scott J, Navaratnam N. An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. Genomics. 2002;79(3):285–96.
6. Willems L, Gillet NA. APOBEC3 interference during replication of viral genomes. Viruses. 2015;7(6):2999–3018.
7. Harris RS, Dudley JP. APOBECs and virus restriction. Virology. 2015;479–480:131–45.
8. Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature. 2002;418(6898):646–50.
9. Derse D, Hill SA, Princler G, Lloyd P, Heidecker G. Resistance of human T cell leukemia virus type 1 to APOBEC3G restriction is mediated by elements in nucleocapsid. Proc Natl Acad Sci USA. 2007;104(8):2915–20.
10. Löchelt M, Romen F, Bastone P, Muckenfuss H, Kirchner N, Kim YB, Truyen U, Rosler U, Battenberg M, Saib A, et al. The antiretroviral activity of APOBEC3 is inhibited by the foamy virus accessory Bet protein. Proc Natl Acad Sci USA. 2005;102(22):7982–7.
11. Stavrou S, Nitta T, Kotla S, Ha D, Nagashima K, Rein AR, Fan H, Ross SR. Murine leukemia virus glycosylated Gag blocks apolipoprotein B editing complex 3 and cytosolic sensor access to the reverse transcription complex. Proc Natl Acad Sci USA. 2013;110(22):9078–83.
12. Rosales Gerpe MC, Renner TM, Belanger K, Lam C, Aydin H, Langlois MA. N-linked glycosylation protects gammaretroviruses against deamination by APOBEC3 proteins. J Virol. 2015;89(4):2342–57.
13. Kolokithas A, Rosenke K, Malik F, Hendrick D, Swanson L, Santiago ML, Portis JL, Hasenkrug KJ, Evans LH. The glycosylated Gag protein of a murine leukemia virus inhibits the antiretroviral function of APOBEC3. J Virol. 2010;84(20):10933–6.
14. Holmes RK, Koning FA, Bishop KN, Malim MH. APOBEC3F can inhibit the accumulation of HIV-1 reverse transcription products in the absence of hypermutation. Comparisons with APOBEC3G. J Biol Chem. 2007;282(4):2587–95.
15. Iwatani Y, Chan DS, Wang F, Maynard KS, Sugiura W, Gronenborn AM, Rouzina I, Williams MC, Musier-Forsyth K, Levin JG.

Deaminase-independent inhibition of HIV-1 reverse transcription by APOBEC3G. Nucleic Acids Res. 2007;35(21):7096–108.
16. Gillick K, Pollpeter D, Phalora P, Kim EY, Wolinsky SM, Malim MH. Suppression of HIV-1 infection by APOBEC3 proteins in primary human CD4(+) T cells is associated with inhibition of processive reverse transcription as well as excessive cytidine deamination. J Virol. 2013;87(3):1508–17.
17. Wang X, Ao Z, Chen L, Kobinger G, Peng J, Yao X. The cellular antiviral protein APOBEC3G interacts with HIV-1 reverse transcriptase and inhibits its function during viral replication. J Virol. 2012;86(7):3777–86.
18. Mbisa JL, Barr R, Thomas JA, Vandegraaff N, Dorweiler IJ, Svarovskaia ES, Brown WL, Mansky LM, Gorelick RJ, Harris RS, et al. Human immunodeficiency virus type 1 cDNAs produced in the presence of APOBEC3G exhibit defects in plus-strand DNA transfer and integration. J Virol. 2007;81(13):7099–110.
19. Mbisa JL, Bu W, Pathak VK. APOBEC3F and APOBEC3G inhibit HIV-1 DNA integration by different mechanisms. J Virol. 2010;84(10):5250–9.
20. Kenyon JC, Lever AM. The molecular biology of feline immunodeficiency virus (FIV). Viruses. 2011;3(11):2192–213.
21. Pecon-Slattery J, Troyer JL, Johnson WE, O'Brien SJ. Evolution of feline immunodeficiency virus in Felidae: implications for human health and wildlife ecology. Vet Immunol Immunopathol. 2008;123(1–2):32–44.
22. Willett BJ, Hosie MJ. Feline leukaemia virus: half a century since its discovery. Vet J. 2013;195(1):16–23.
23. Rethwilm A, Bodem J. Evolution of foamy viruses: the most ancient of all retroviruses. Viruses. 2013;5(10):2349–74.
24. Hartmann K. Clinical aspects of feline immunodeficiency and feline leukemia virus infection. Vet immunol immunopathol. 2011;143(3–4):190–201.
25. de Rozieres S, Mathiason CK, Rolston MR, Chatterji U, Hoover EA, Elder JH. Characterization of a highly pathogenic molecular clone of feline immunodeficiency virus clade C. J Virol. 2004;78(17):8971–82.
26. Diehl LJ, Mathiason-Dubard CK, O'Neil LL, Obert LA, Hoover EA. Induction of accelerated feline immunodeficiency virus disease by acute-phase virus passage. J Virol. 1995;69(10):6149–57.
27. Obert LA, Hoover EA. Feline immunodeficiency virus clade C mucosal transmission and disease courses. AIDS Res Hum Retrovir. 2000;16(7):677–88.
28. Lehman TL, O'Halloran KP, Hoover EA, Avery PR. Utilizing the FIV model to understand dendritic cell dysfunction and the potential role of dendritic cell immunization in HIV infection. Vet Immunol Immunopathol. 2010;134(1–2):75–81.
29. Yamamoto JK, Sanou MP, Abbott JR, Coleman JK. Feline immunodeficiency virus model for designing HIV/AIDS vaccines. Curr HIV Res. 2010;8(1):14–25.
30. Elder JH, Lin YC, Fink E, Grant CK. Feline immunodeficiency virus (FIV) as a model for study of lentivirus infections: parallels with HIV. Curr HIV Res. 2010;8(1):73–80.
31. O'Brien SJ, Troyer JL, Brown MA, Johnson WE, Antunes A, Roelke ME, Pecon-Slattery J. Emerging viruses in the Felidae: shifting paradigms. Viruses. 2012;4(2):236–57.
32. German AC, Harbour DA, Helps CR, Gruffydd-Jones TJ. Is feline foamy virus really apathogenic? Vet Immunol Immunopathol. 2008;123(1–2):114–8.
33. Winkler IG, Lochelt M, Flower RL. Epidemiology of feline foamy virus and feline immunodeficiency virus infections in domestic and feral cats: a seroepidemiological study. J Clin Microbiol. 1999;37(9):2848–51.
34. Phung HT, Ikeda Y, Miyazawa T, Nakamura K, Mochizuki M, Izumiya Y, Sato E, Nishimura Y, Tohya Y, Takahashi E, et al. Genetic analyses of feline foamy virus isolates from domestic and wild feline species in geographically distinct areas. Virus Res. 2001;76(2):171–81.
35. Zielonka J, Münk C. Cellular restriction factors of feline immunodeficiency virus. Viruses. 2011;3(10):1986–2005.
36. Zielonka J, Marino D, Hofmann H, Yuhki N, Löchelt M, Münk C. Vif of feline immunodeficiency virus from domestic cats protects against APOBEC3 restriction factors from many felids. J Virol. 2010;84(14):7312–24.
37. Wang J, Zhang W, Lv M, Zuo T, Kong W, Yu X. Identification of a Cullin5-ElonginB-ElonginC E3 complex in degradation of feline immunodeficiency virus Vif-mediated feline APOBEC3 proteins. J Virol. 2011;85(23):12482–91.
38. Chareza S, Slavkovic LD, Liu Y, Rathe AM, Münk C, Zabogli E, Pistello M, Löchelt M. Molecular and functional interactions of cat APOBEC3 and feline foamy and immunodeficiency virus proteins: different ways to counteract host-encoded restriction. Virology. 2012;424(2):138–46.
39. Münk C, Hechler T, Chareza S, Löchelt M. Restriction of feline retroviruses: lessons from cat APOBEC3 cytidine deaminases and TRIM5alpha proteins. Vet Immunol Immunopathol. 2010;134(1–2):14–24.
40. de Castro FL, Junqueira DM, de Medeiros RM, da Silva TR, Costenaro JG, Knak MB, de Matos Almeida SE, Campos FS, Roehe PM, Franco AC. Analysis of single-nucleotide polymorphisms in the APOBEC3H gene of domestic cats (*Felis catus*) and their association with the susceptibility to feline immunodeficiency virus and feline leukemia virus infections. Infect Genet Evol. 2014;27:389–94.
41. Mehle A, Goncalves J, Santa-Marta M, McPike M, Gabuzda D. Phosphorylation of a novel SOCS-box regulates assembly of the HIV-1 Vif-Cul5 complex that promotes APOBEC3G degradation. Genes Dev. 2004;18(23):2861–6.
42. Yu Y, Xiao Z, Ehrlich ES, Yu X, Yu XF. Selective assembly of HIV-1 Vif-Cul5-ElonginB-ElonginC E3 ubiquitin ligase complex through a novel SOCS box and upstream cysteines. Genes Dev. 2004;18(23):2867–72.
43. Yu X, Yu Y, Liu B, Luo K, Kong W, Mao P, Yu XF. Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. Science. 2003;302(5647):1056–60.
44. Jager S, Kim DY, Hultquist JF, Shindo K, LaRue RS, Kwon E, Li M, Anderson BD, Yen L, Stanley D, et al. Vif hijacks CBF-beta to degrade APOBEC3G and promote HIV-1 infection. Nature. 2012;481(7381):371–5.
45. Zhang W, Du J, Evans SL, Yu Y, Yu XF. T-cell differentiation factor CBF-beta regulates HIV-1 Vif-mediated evasion of host restriction. Nature. 2012;481(7381):376–9.
46. Yoshikawa R, Takeuchi JS, Yamada E, Nakano Y, Ren F, Tanaka H, Munk C, Harris RS, Miyazawa T, Koyanagi Y, et al. Vif determines the requirement for CBF-beta in APOBEC3 degradation. J Gen Virol. 2015;96(Pt 4):887–92.
47. Kane JR, Stanley DJ, Hultquist JF, Johnson JR, Mietrach N, Binning JM, Jonsson SR, Barelier S, Newton BW, Johnson TL, et al. Lineage-specific viral hijacking of non-canonical E3 Ubiquitin ligase cofactors in the evolution of Vif anti-APOBEC3 activity. Cell Rep. 2015;11(8):1236–50.
48. Ai Y, Zhu D, Wang C, Su C, Ma J, Ma J, Wang X. Core-binding factor subunit beta is not required for non-primate lentiviral Vif-mediated APOBEC3 degradation. J Virol. 2014;88(20):12112–22.
49. Han X, Liang W, Hua D, Zhou X, Du J, Evans SL, Gao Q, Wang H, Viqueira R, Wei W, et al. Evolutionarily conserved requirement for core binding factor beta in the assembly of the human immunodeficiency virus/simian immunodeficiency virus Vif-cullin 5-RING E3 ubiquitin ligase. J Virol. 2014;88(6):3320–8.
50. Münk C, Zielonka J, Constabel H, Kloke BP, Rengstl B, Battenberg M, Bonci F, Pistello M, Lochelt M, Cichutek K. Multiple restrictions of human immunodeficiency virus type 1 in feline cells. J Virol. 2007;81(13):7048–60.
51. Stern MA, Hu C, Saenz DT, Fadel HJ, Sims O, Peretz M, Poeschla EM. Productive replication of Vif-chimeric HIV-1 in feline cells. J Virol. 2010;84(14):7378–95.
52. LaRue RS, Lengyel J, Jonsson SR, Andresdottir V, Harris RS. Lentiviral Vif degrades the APOBEC3Z3/APOBEC3H protein of its mammalian host and is capable of cross-species activity. J Virol. 2010;84(16):8193–201.
53. Zielonka J, Bravo IG, Marino D, Conrad E, Perkovic M, Battenberg M, Cichutek K, Münk C. Restriction of equine infectious anemia virus by equine APOBEC3 cytidine deaminases. J Virol. 2009;83(15):7547–59.
54. Binka M, Ooms M, Steward M, Simon V. The activity spectrum of Vif from multiple HIV-1 subtypes against APOBEC3G, APOBEC3F, and APOBEC3H. J Virol. 2012;86(1):49–59.
55. Sonnhammer EL, Koonin EV. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. 2002;18(12):619–20.
56. Yoshikawa R, Izumi T, Yamada E, Nakano Y, Misawa N, Ren F, Carpenter MA, Ikeda T, Münk C, Harris RS, et al. A naturally occurring domestic cat APOBEC3 variant confers resistance to FIV infection. J Virol. 2015;90(1):474–85.
57. Curran MA, Kaiser SM, Achacoso PL, Nolan GP. Efficient transduction of nondividing cells by optimized feline immunodeficiency virus vectors. Mol Ther. 2000;1(1):31–8.
58. Bulliard Y, Turelli P, Röhrig UF, Zoete V, Mangeat B, Michielin O, Trono D. Functional analysis and structural modeling of human APOBEC3G

reveal the role of evolutionarily conserved elements in the inhibition of human immunodeficiency virus type 1 infection and Alu transposition. J Virol. 2009;83(23):12611–21.
59. Gohlke H, Hergert U, Meyer T, Mulnaes D, Grieshaber MK, Smits SH, Schmitt L. Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion. J Chem Inf Model. 2013;53(10):2493–8.
60. Widderich N, Pittelkow M, Höppner A, Mulnaes D, Buckel W, Gohlke H, Smits SH, Bremer E. Molecular dynamics simulations and structure-guided mutagenesis provide insight into the architecture of the catalytic core of the ectoine hydroxylase. J Mol Biol. 2014;426(3):586–600.
61. Gehring WJ. The homeobox in perspective. Trends Biochem Sci. 1992;17(8):277–80.
62. Suspene R, Henry M, Guillot S, Wain-Hobson S, Vartanian JP. Recovery of APOBEC3-edited human immunodeficiency virus G → A hypermutants by differential DNA denaturation PCR. J Gen Virol. 2005;86(Pt 1):125–9.
63. Mariani R, Chen D, Schröfelbauer B, Navarro F, König R, Bollman B, Münk C, Nymark-McMahon H, Landau NR. Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. Cell. 2003;114(1):21–31.
64. Bähr A, Singer A, Hain A, Vasudevan AA, Schilling M, Reh J, Riess M, Panitz S, Serrano V, Schweizer M, et al. Interferon but not MxB inhibits foamy retroviruses. Virology. 2015;488:51–60.
65. Huthoff H, Malim MH. Identification of amino acid residues in APOBEC3G required for regulation by human immunodeficiency virus type 1 Vif and Virion encapsidation. J Virol. 2007;81(8):3807–15.
66. Letko M, Booiman T, Kootstra N, Simon V, Ooms M. Identification of the HIV-1 Vif and human APOBEC3G protein interface. Cell Rep 2015;13(9):1789–99.
67. Vasudevan AA, Smits SH, Hoppner A, Häussinger D, Koenig BW, Münk C. Structural features of antiviral DNA cytidine deaminases. Biol Chem. 2013;394(11):1357–70.
68. Kouno T, Luengas EM, Shigematsu M, Shandilya SM, Zhang J, Chen L, Hara M, Schiffer CA, Harris RS, Matsuo H. Structure of the Vif-binding domain of the antiviral enzyme APOBEC3G. Nat Struct Mol Biol. 2015;22(6):485–91.
69. Kitamura S, Ode H, Nakashima M, Imahashi M, Naganawa Y, Kurosawa T, Yokomaku Y, Yamane T, Watanabe N, Suzuki A, et al. The APOBEC3C crystal structure and the interface for HIV-1 Vif binding. Nat Struct Mol Biol. 2012;19(10):1005–10.
70. He Z, Zhang W, Chen G, Xu R, Yu XF. Characterization of conserved motifs in HIV-1 Vif required for APOBEC3G and APOBEC3F interaction. J Mol Biol. 2008;381(4):1000–11.
71. Zhang W, Huang M, Wang T, Tan L, Tian C, Yu X, Kong W, Yu XF. Conserved and non-conserved features of HIV-1 and SIVagm Vif mediated suppression of APOBEC3 cytidine deaminases. Cell Microbiol. 2008;10(8):1662–75.
72. Baig TT, Feng Y, Chelico L. Determinants of efficient degradation of APOBEC3 restriction factors by HIV-1 Vif. J Virol. 2014;88(24):14380–95.
73. Marin M, Golem S, Rose KM, Kozak SL, Kabat D. Human immunodeficiency virus type 1 Vif functionally interacts with diverse APOBEC3 cytidine deaminases and moves with them between cytoplasmic sites of mRNA metabolism. J Virol. 2008;82(2):987–98.
74. Richards C, Albin JS, Demir O, Shaban NM, Luengas EM, Land AM, Anderson BD, Holten JR, Anderson JS, Harki DA, et al. The binding interface between human APOBEC3F and HIV-1 Vif elucidated by genetic and computational approaches. Cell Rep. 2015;13(9):1781–8.
75. Salter JD, Morales GA, Smith HC. Structural insights for HIV-1 therapeutic strategies targeting Vif. Trends Biochem Sci. 2014;39(9):373–80.
76. Aydin H, Taylor MW, Lee JE. Structure-guided analysis of the human APOBEC3-HIV restrictome. Structure. 2014;22(5):668–84.
77. Land AM, Shaban NM, Evans L, Hultquist JF, Albin JS, Harris RS. APOBEC3F determinants of HIV-1 Vif sensitivity. J Virol. 2014;88(21):12923–7.
78. Siu KK, Sultana A, Azimi FC, Lee JE. Structural determinants of HIV-1 Vif susceptibility and DNA binding in APOBEC3F. Nat Commun. 2013;4:2593.
79. Ooms M, Letko M, Binka M, Simon V. The resistance of human APOBEC3H to HIV-1 NL4-3 molecular clone is determined by a single amino acid in Vif. PLoS One. 2013;8(2):e57744.
80. Zhen A, Wang T, Zhao K, Xiong Y, Yu XF. A single amino acid difference in human APOBEC3H variants determines HIV-1 Vif sensitivity. J Virol. 2010;84(4):1902–11.
81. Lavens D, Peelman F, Van der Heyden J, Uyttendaele I, Catteeuw D, Verhee A, Van Schoubroeck B, Kurth J, Hallenberger S, Clayton R, et al. Definition of the interacting interfaces of Apobec3G and HIV-1 Vif using MAPPIT mutagenesis analysis. Nucleic Acids Res. 2010;38(6):1902–12.
82. Tautz D, Domazet-Loso T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011;12(10):692–702.
83. Wei X, Decker JM, Liu H, Zhang Z, Arani RB, Kilby JM, Saag MS, Wu X, Shaw GM, Kappes JC. Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy. Antimicrob Agents Chemother. 2002;46(6):1896–905.
84. Derdeyn CA, Decker JM, Sfakianos JN, Wu X, O'Brien WA, Ratner L, Kappes JC, Shaw GM, Hunter E. Sensitivity of human immunodeficiency virus type 1 to the fusion inhibitor T-20 is modulated by coreceptor specificity defined by the V3 loop of gp120. J Virol. 2000;74(18):8358–67.
85. Loewen N, Barraza R, Whitwam T, Saenz DT, Kemler I, Poeschla EM. FIV Vectors. Methods Mol Biol. 2003;229:251–71.
86. Morris KV, Gilbert J, Wong-Staal F, Gasmi M, Looney DJ. Transduction of cell lines and primary cells by FIV-packaged HIV vectors. Mol Ther. 2004;10(1):181–90.
87. Simon JH, Southerling TE, Peterson JC, Meyer BE, Malim MH. Complementation of vif-defective human immunodeficiency virus type 1 by primate, but not nonprimate, lentivirus vif genes. J Virol. 1995;69(7):4166–72.
88. Simm M, Shahabuddin M, Chao W, Allan JS, Volsky DJ. Aberrant Gag protein composition of a human immunodeficiency virus type 1 vif mutant produced in primary lymphocytes. J Virol. 1995;69(7):4582–6.
89. Jaguva Vasudevan AA, Perkovic M, Bulliard Y, Cichutek K, Trono D, Häussinger D, Münk C. Prototype foamy virus Bet impairs the dimerization and cytosolic solubility of human APOBEC3G. J Virol. 2013;87(16):9030–40.
90. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
91. Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 2005;21(4):456–63.
92. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–13.
93. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 1999;16(8):1114.
94. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006;23(2):254–67.
95. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res. 2007;35(Web Server issue):W506–11.
96. Pond SK, Muse SV. Site-to-site variation of synonymous substitution rates. Mol Biol Evol. 2005;22(12):2375–85.
97. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234(3):779–815.
98. Miao Z, Cao Y, Jiang T. RASP: rapid modeling of protein side chain conformations. Bioinformatics. 2011;27(22):3117–22.
99. Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J. 2011;101(10):2525–34.
100. Gerber PR, Muller K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. J Comput Aided Mol Des. 1995;9(3):251–68.

Suppl. Fig. S1

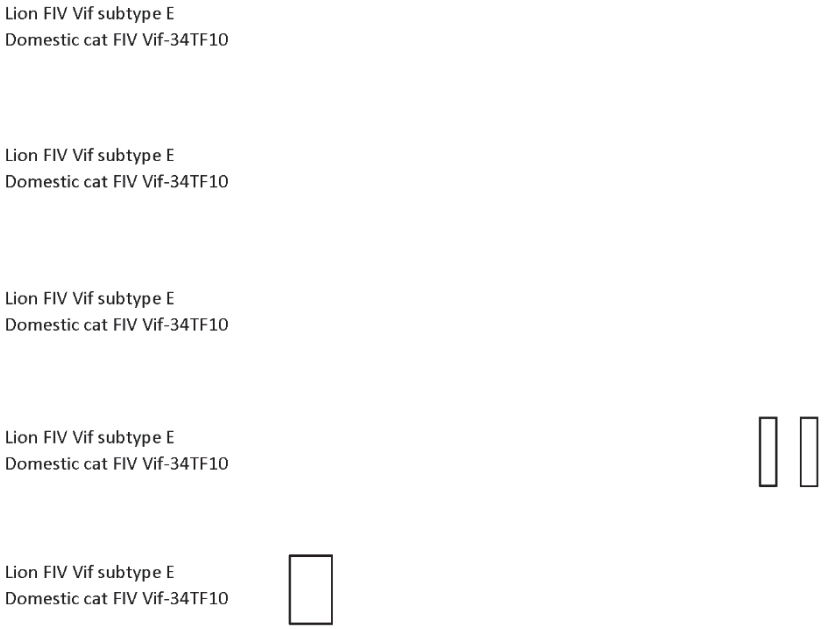

FcaA3Z2b
FcaA3Z3
FcaA3Z2Z3
A3-HA
Vif-TLQAAA-V5
DAPI
Overlay

Suppl. Fig. S2

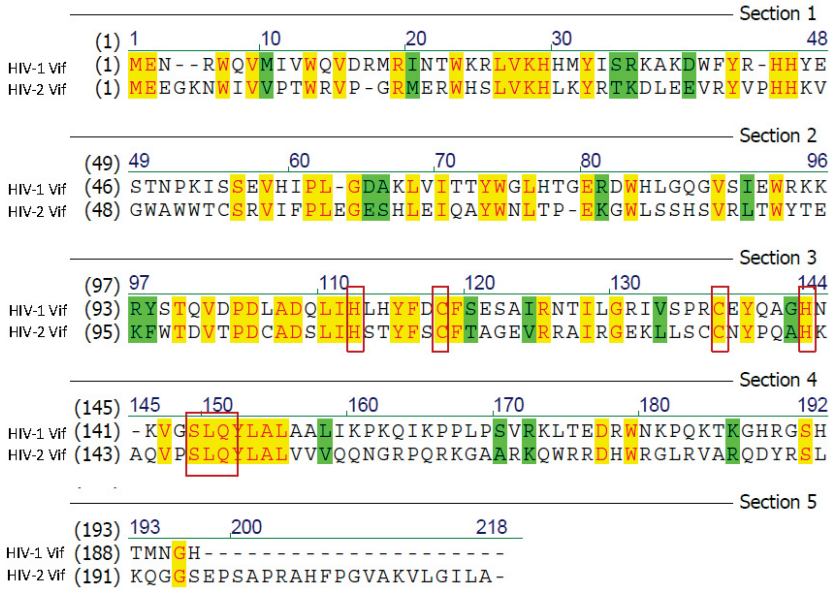
(A)

(B)

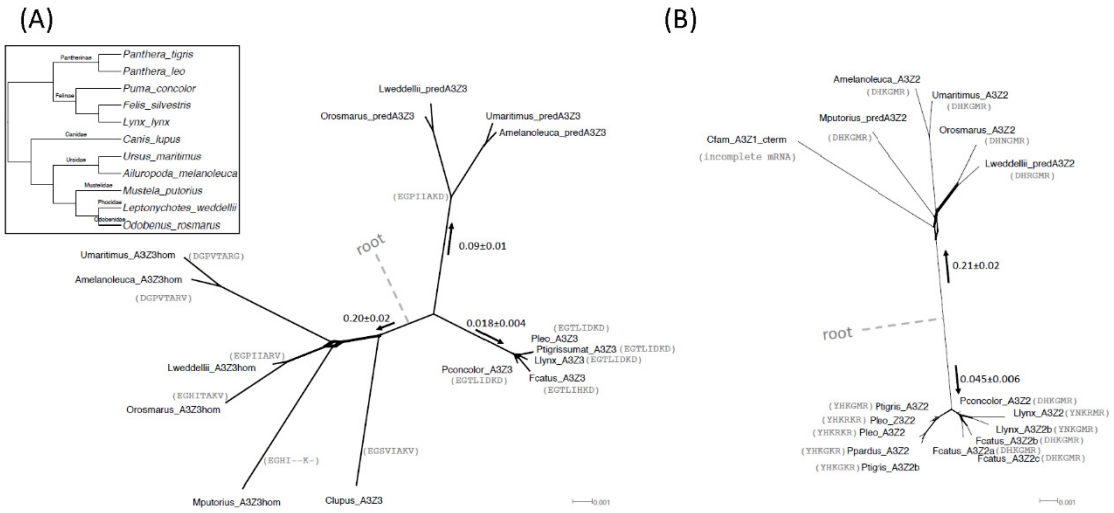
(C)


Lion FIV Vif subtype E
Domestic cat FIV Vif-34TF10
Lion FIV Vif subtype E
Domestic cat FIV Vif-34TF10
Lion FIV Vif subtype E
Domestic cat FIV Vif-34TF10
Lion FIV Vif subtype E
Domestic cat FIV Vif-34TF10
Lion FIV Vif subtype E
Domestic cat FIV Vif-34TF10

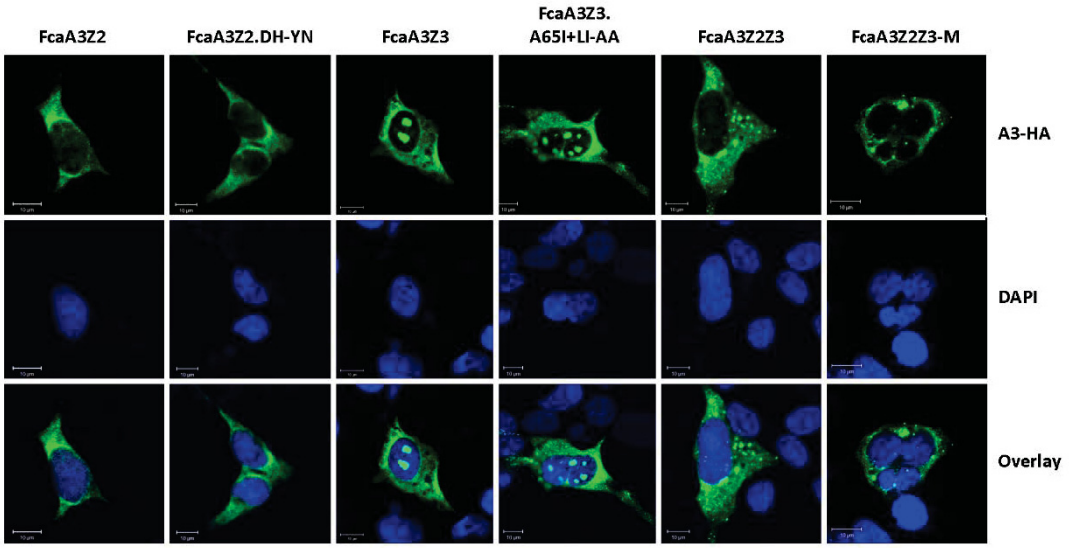
(D)


Section 1
(1) 1 10 20 30 48
HIV-1 Vif (1) MEN--RWQVMIVWQVDRMRINTWKRLVKHHMYISRKAKDWFYR-HHYE
HIV-2 Vif (1) MEEGKNWIVVPTWRVP-GRMERWHSLVKHLKYRTKDLEEVRYVPHHKV
Section 2
(49) 49 60 70 80 96
HIV-1 Vif (46) STNPKISSEVHIPL-GDAKLVITTYWGLHTGERDWHLGQGVSIEWRKK
HIV-2 Vif (48) GWAWWTCSRVIFPLEGESHLEIQAYWNLTP-EKGWLSSHSVRLTWYTE
Section 3
(97) 97 110 120 130 144
HIV-1 Vif (93) RYSTQVDPDLADQLIHLHYFDCFSESAIRNTILGRIVSPRCEYQAGHN
HIV-2 Vif (95) KFWTDVTPDCADSLIHSTYFSCFTAGEVRRAIRGEKLLSCCNYPQAHK
Section 4
(145) 145 150 160 170 180 192
HIV-1 Vif (141) -KVGSLQYLALAALIKPKQIKPPLPSVRKLTEDRWNKPQKTKGHRGSH
HIV-2 Vif (143) AQVPSLQYLALVVVQQNGRPQRKGAARKQWRRDHWRGLRVARQDYRSL
Section 5
(193) 193 200 218
HIV-1 Vif (188) TMNGH----------
HIV-2 Vif (191) KQGGSEPSAPRAHFPGVAKVLGILA-

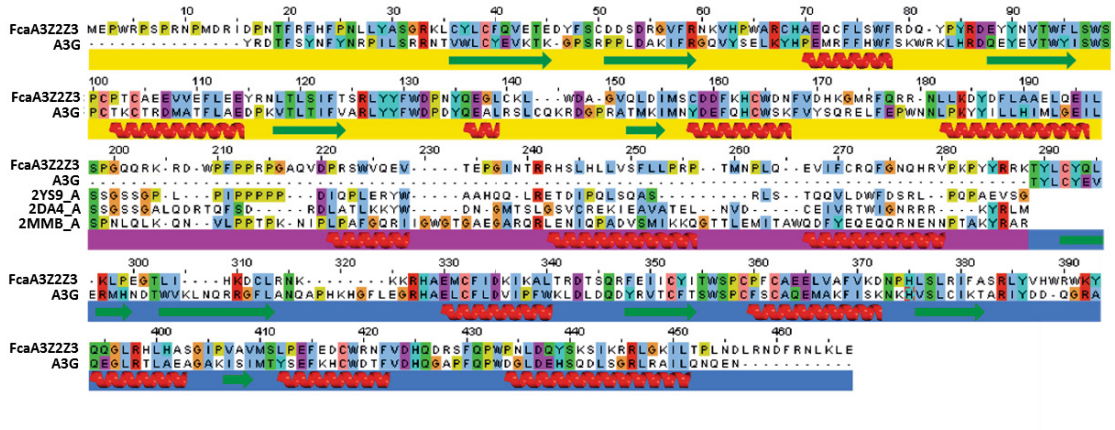
Suppl. Fig. S3



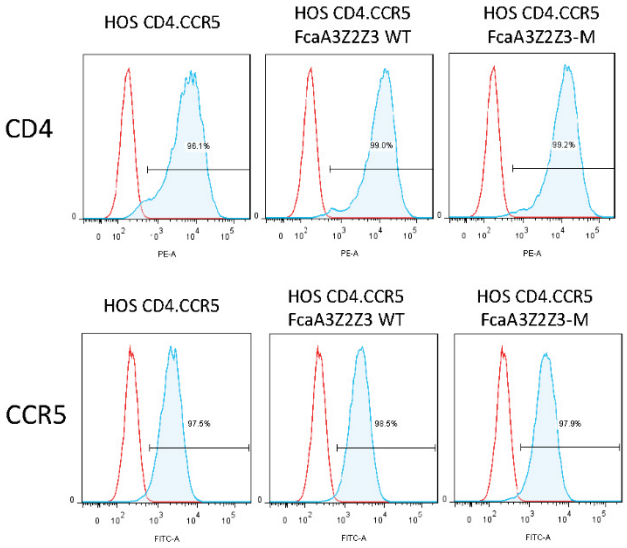
Suppl. Fig. S4



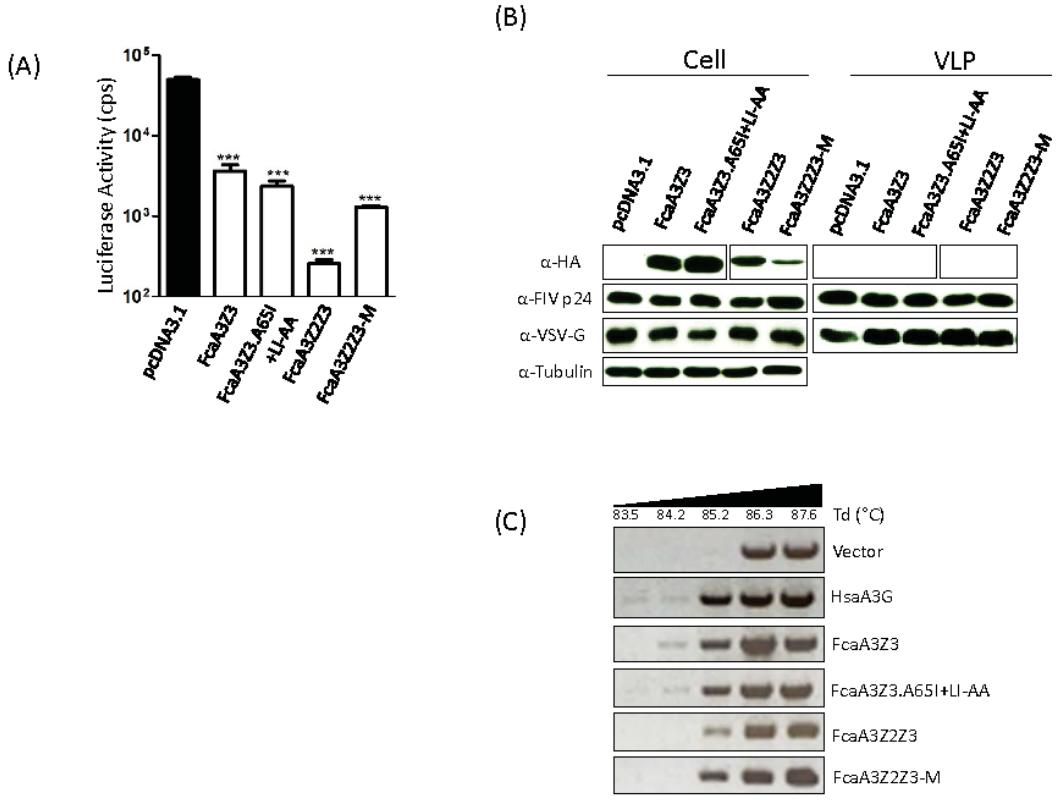
Suppl. Fig. S5



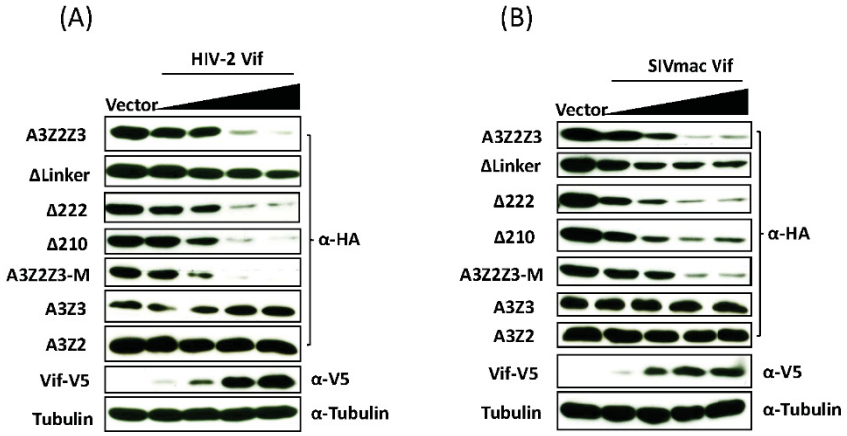
Suppl. Fig. S6



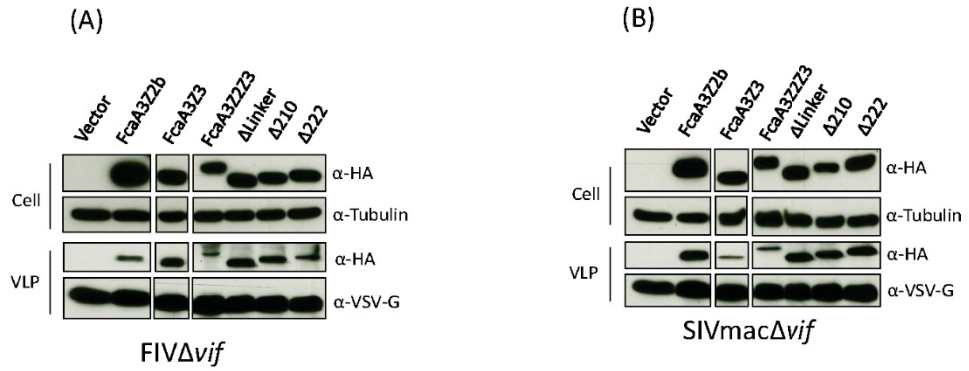
Suppl. Fig. S7



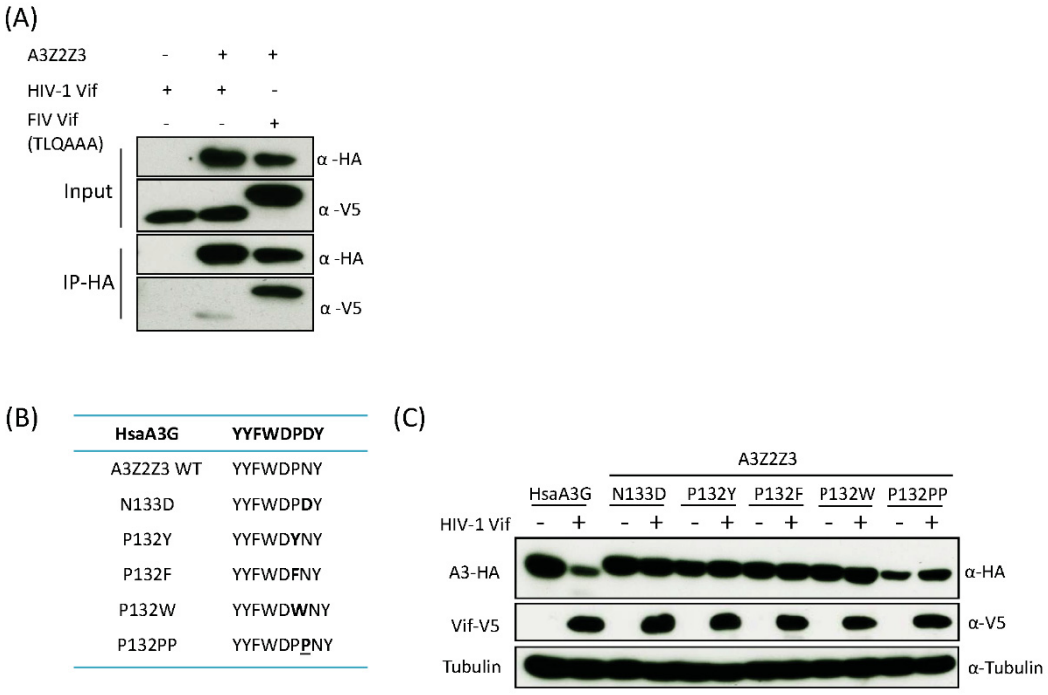
Suppl. Fig. S8



Suppl. Fig. S9



Suppl. Fig. S10



Supplemental Tables

Supplemental Table S1: Primer list used for HsaA3C/FcaA3Z2 chimeras and FcaA3Z2 mutants

Construct	Primer Name	Primer Sequence
Z2C1	fehuApo3 1-63.fw	5'tataagctttgagagaggaatggagccctggcgcccagcccaagaaacccaatggacaggatagatcctaacaccttccgtttccaatttaaaaacctatg-3'
Z2C4	hufe3C 397.fw	5'gcctctactacttctgggacccatgttaccaggagggctccgcag-3'
	hufe3C 397.rv	5'ctgcggagcccctcctggtaacatgggtcccagaagtagtagaggc-3'
Z2C5	hufe3C 493.fw	5'aaacactgttgggacaactttgtgtacaatgataatgagccattcaa-3'
	hufe3C 493.rv	5'ttgaatggctcattatcattgtacacaaagttgtccaacagtgttt-3'
Z2C30	hufe3C 485.fw	5'taaatattgttgggaaaactttgtggaccacaagggaatgcgctt-3'
	hufe3C 485.rv	5'aagcgcattcccttgtggtccacaaagttttcccaacaatattta-3'
FcaZ2bN18K	N18K.fw	5'-gatagatcctaagaccttccgtttc-3'
	N18K.rv	5'-gaaacggaaggtcttaggatctatc-3'
FcaZ2bT44R	T44R.fw	5'-cttccaagtggagagagaagactacttc-3'
	T44R.rv	5'-gaagtagtcttctctctccacttggaag-3'
FcaZ2bD165Y	D165Y.fw	5'-caactttgtgtaccacaagggaatgc-3'
	D165Y.rv	5'-gcattcccttgtggtacacaaagttg-3'
FcaZ2bH166N	H166N.fw	5'-caactttgtggacaacaagggaatgc-3'
	H166N.rv	5'-gcattcccttgttgtccacaaagttg-3'
FcaZ2b DH-YN	DH-YN.fw	5'-caactttgtgtacaacaagggaatgc-3'
	DH-YN.rv	5'-gcattcccttgttgtacacaaagttg-3'
PtiZ2Y165D	Y165D.fw	5'-caactttgtggaccacaagggaatgc-3'
	Y165D.rv	5'-gcattcccttgtggtccacaaagttg-3'
External primers	feApo3.fw	5'tataagctttgaagaggaatggagccctggcgcccag-3'
	HA-rv	5'agctcgagtcaagcgtaatctggaacatcgtatggataagcgtaatctggaacatcgtatg-3'

Supplemental Table S2: Primer list used for HsaA3H/FcaA3Z3 chimeras and FcaA3Z3 mutants

Construct	Primer Name	Primer Sequence
Z3C1	Z3C1.fw	5'ccagcaccgggtcccaaagccctactacccgaggaaggcc ctc-3'
	Z3C1.rv	5'gagggccttcctcgggtagtagggctttgggacccggtgct gg-3'
Z3C2	Z3C2.fw	5'caaagactgccttcgaaataagaaaaagtgccatgcagaa atttg-3'
	Z3C2.rv	5'caaatttctgcatggcacttttttcttatttcgaaggcagtcttt g-3'
Z3C6	Z3C6.fw	5'caagcgccgcctcagaaggccttactaccggaggaaaacc tac-3'
	Z3C6.rv	5'gtaggttttcctccggtagtaaggccttctgaggcggcgctt g-3'
Z3C7	Z3C7.fw	5'gaggctactttgaaaacaagaaaaagcgccatgcggaaat gtg-3'
	Z3C7.rv	5'cacatttccgcatggcgctttttcttgttttcaaagtagcctc- 3'
FcaZ3KL-TP	KL-TP.fw	5'-gctaccagctgacgccgcccgaaggcacc-3'
	KL-TP.rv	5'-ggtgccttcgggcggcgtcagctggtagc-3'
FcaZ3PE-QN	PE-QN.fw	5'-ccagctgaagctgcagaatggcaccctaattc-3'
	PE-QN.rv	5'-gaattagggtgccattctgcagcttcagctgg-3'
FcaZ3LI-TP	LI-TP.fw	5'-gcccgaaggcaccacacctcacaaagactgcc-3'
	LI-TP.rv	5'-ggcagtctttgtgaggtgtggtgccttcgggc-3'
FcaZ3H-T	H-T.fw	5'-cgaaggcaccctaattaccaaagactgcc-3'
	H-T.rv	5'-ggcagtctttggtaattagggtgccttcg-3'
FcaZ3DC-AA	DC-AA.fw	5'-ctaattcacaaagccgcccttcgaaataag-3'
	DC-AA.rv	5'-cttatttcgaagggcggctttgtgaattag-3'
FcaZ3LR-AA	LR-AA.fw	5'-cacaaagactgcgctgcaaataagaaaaag-3'
	LR-AA.rv	5'-ctttttcttatttgcagcgcagtctttgtg-3'
FcaZ3LI-AA	LI-AA.fw	5'-gcccgaaggcaccgcagctcacaaagactgcc-3'
	LI-AA-rv	5'-ggcagtctttgtgagctgcggtgccttcgggc-3'
External primers	FcaZ3.fw	5'-atgaattcgccaccatgaatccactacaggaag-3'
	HA-rv	5'agctcgagtcaagcgtaatctggaacatcgtatggataagc gtaatctggaacatcgtatg-3'

Supplemental Table S3: Primer list used for FcaA3Z2Z3 mutants

Construct	Primer Name	Primer Sequence
Δ210	Δ210.rv	5'ctgtagtggattcattgtgggtctttgggcccctgggcgg ggagggaagggcc-3'
Δ222	Δ222.rv	5'ctgtagtggattcattgtgggtctctctgtcacctcctgaa cccaactccttggg-3'
ΔLinker	Δlinker.fw	5'gcttcaagaaatccttagacccacaatgaatccactaca ggaag-3'
	Δlinker.rv	5'cttcctgtagtggattcattgtgggtctaaggatttcttga agc-3'
FcaZ2Z3N133D	N133D.fw	5'-ctacttctgggacccagattaccaggaggggc-3'
	N133D.rv	5'-gcccctcctggtaatctgggtcccagaagtag-3'
FcaZ2Z3P132Y	P132Y.fw	5'-ctacttctgggactacaattaccaggaggggc-3'
	P132Y.rv	5'-gcccctcctggtaattgtagtcccagaagtag-3'
FcaZ2Z3P132F	P132F.fw	5'-ctacttctgggacttcaattaccaggaggggc-3'
	P132F.rv	5'-gcccctcctggtaattgaagtcccagaagtag-3'
FcaZ2Z3P132W	P132W.fw	5'-ctacttctgggactggaattaccaggaggggc-3'
	P132W.rv	5'-gcccctcctggtaattccagtcccagaagtag-3'
FcaZ2Z3P132PP	P132PP.fw	5'-cttctgggacccaccaaattaccaggagg-3'
	P132PP.rv	5'-cctcctggtaatttggtgggtcccagaag-3'

Supplemental Table S4: Primer used to clone GST fusion constructs

Construct	Primer Name	Primer Sequence
FcaGST-Z2-HA	FcaZ2b-GST-EcoRI-F	5‘-ATAGAATTCCCATGGAGCCCTGGCGCCCC-3‘
	HA-NotI-R	5‘-ATGCGGCCGCTCAAGCGTAATCTGGAACATC-3‘
FcaGST-Z3-HA	FcaZ3-GST-EcoRI-F	5‘-ATAGAATTCCCATGAATCCACTACAGGAAG-3‘
	HA-NotI-R	5‘-ATGCGGCCGCTCAAGCGTAATCTGGAACATC-3‘
FcaGST-Linker	Fca-Linker-EcoRI-F	5‘-ATGAATTCCCAGTCCCGGCCAACAAAG-3‘
	Fca-Linker-EcoRI-R	5‘-ATGTCGACTCATGTGGGTCTGGGCAAGAG-3‘

Supplemental Table S5: The software used in TopModel for threading, alignment and model quality estimation.[a]

Threading	Alignment	Model Quality Estimation
DeltaBLAST [1]	**ClustalW*** [2]	**PROCHECK** [3]
HMMER3 [4]	**POA*** [5]	**MolProbity** [6]
HHblits [7]	**MUSCLE*** [8]	**ANOLEA** [9]
SAMT2K [10]	**ProbA*** [11]	**ProSa2003** [12]
FFAS03 [13]	**ProbCons*** [14]	**DOPE** [15]
SPARKSX [16]	**PCMA*** [17]	**GOAP** [18]
RAPTORX [19]	**DiAlign*** [20]	**ModFoldClust2** [21]
LOMETS [22]	**SAP*** [23]	**SPICKER** [24]
	TM-Align* [25]	
	MAFFT7 [26]	
	MergeAlign2 [27]	
	TCOFFEE [28]	
	PROMALS3D [29]	
	FORMATT [30]	
	MUSTANG [31]	
	3DCOMB	
	SALIGN [32]	

[a] Software marked with "*" are used within TCOFFEE.

References

1. Boratyn GM, Schaffer A, Agarwala R, Altschul SF, Lipman DJ, Madden TL: **Domain enhanced lookup time accelerated BLAST**. *Biol Direct* 2012, **7**(1):12.

2. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic acids research* 1994, **22**(22):4673-4680.

3. Laskowski RA, MacArthur MW, Moss DS, Thornton JM: **PROCHECK: a program to check the stereochemical quality of protein structures**. *Journal of applied crystallography* 1993, **26**(2):283-291.

4. Eddy SR: **Accelerated profile HMM searches**. *PLoS computational biology* 2011, **7**(10):e1002195.

5. Lee C, Grasso C, Sharlow MF: **Multiple sequence alignment using partial order graphs**. *Bioinformatics* 2002, **18**(3):452-464.

6. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC: **MolProbity: all-atom structure validation for macromolecular crystallography**. *Acta Crystallographica Section D: Biological Crystallography* 2009, **66**(1):12-21.

7. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment**. *Nature methods* 2012, **9**(2):173-175.

8. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic acids research* 2004, **32**(5):1792-1797.

9. Melo F, Feytmans E: **Novel knowledge-based mean force potential at atomic level**. *Journal of molecular biology* 1997, **267**(1):207-222.

10. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R: **Combining local-structure, fold-recognition, and new fold methods for protein structure prediction**. *Proteins: Structure, Function, and Bioinformatics* 2003, **53**(S6):491-496.

11. Sierk ML, Smoot ME, Bass EJ, Pearson WR: **Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments**. *BMC bioinformatics* 2010, **11**(1):146.

12. Sippl MJ: **Recognition of errors in three-dimensional structures of proteins**. *Proteins: Structure, Function, and Genetics* 1993, **17**(4):355-362.

13. Rychlewski L, Li W, Jaroszewski L, Godzik A: **Comparison of sequence profiles. Strategies for structural predictions using sequence information**. *Protein Science* 2000, **9**(2):232-241.

14. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S: **ProbCons: Probabilistic consistency-based multiple sequence alignment**. *Genome research* 2005, **15**(2):330-340.

15. Shen My, Sali A: **Statistical potential for assessment and prediction of protein structures**. *Protein science* 2006, **15**(11):2507-2524.

16. Yang Y, Faraggi E, Zhao H, Zhou Y: **Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates**. *Bioinformatics* 2011, **27**(15):2076-2082.

17. Pei J, Sadreyev R, Grishin NV: **PCMA: fast and accurate multiple sequence alignment based on profile consistency**. *Bioinformatics* 2003, **19**(3):427-428.

18. Zhou H, Skolnick J: **GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction**. *Biophysical journal* 2011, **101**(8):2043-2052.

19. Peng J, Xu J: **RaptorX: exploiting structure information for protein alignment by statistical inference**. *Proteins: Structure, Function, and Bioinformatics* 2011, **79**(S10):161-171.

20. Al Ait L, Yamak Z, Morgenstern B: **DIALIGN at GOBICS—multiple sequence alignment using various sources of external information**. *Nucleic acids research* 2013, **41**(W1):W3-W7.

21. McGuffin LJ, Roche DB: **Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments**. *Bioinformatics* 2010, **26**(2):182-188.

22. Wu S, Zhang Y: **LOMETS: a local meta-threading-server for protein structure prediction**. *Nucleic acids research* 2007, **35**(10):3375-3382.

23. Taylor WR: **Protein structure comparison using iterated double dynamic programming**. *Protein Science* 1999, **8**(03):654-665.

24. Zhang Y, Skolnick J: **SPICKER: A clustering approach to identify near-native protein folds**. *Journal of computational chemistry* 2004, **25**(6):865-871.

25. Zhang Y, Skolnick J: **TM-align: a protein structure alignment algorithm based on the TM-score**. *Nucleic acids research* 2005, **33**(7):2302-2309.

26. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. *Molecular biology and evolution* 2013, **30**(4):772-780.

27. Collingridge PW, Kelly S: **MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments**. *BMC bioinformatics* 2012, **13**(1):117.

28. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C: **3DCoffee: combining protein sequences and structures within multiple sequence alignments**. *Journal of molecular biology* 2004, **340**(2):385-395.

29. Pei J, Kim B-H, Grishin NV: **PROMALS3D: a tool for multiple protein sequence and structure alignments**. *Nucleic acids research* 2008, **36**(7):2295-2300.

30. Daniels NM, Nadimpalli S, Cowen LJ: **Formatt: Correcting protein multiple structural alignments by incorporating sequence alignment**. *BMC bioinformatics* 2012, **13**(1):259.

31. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM: **MUSTANG: a multiple structural alignment algorithm**. *Proteins: Structure, Function, and Bioinformatics* 2006, **64**(3):559-574.

32. Madhusudhan M, Webb BM, Marti-Renom MA, Eswar N, Sali A: **Alignment of multiple protein structures based on sequence and structure features**. *Protein Engineering Design and Selection* 2009, **22**(9):569-574.

21. PUBLICATION V

Recognition Motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1

Dalibor Milic[1,4], Markus Dick[2,5], Daniel Mulnaes[2], Christopher Pfleger[2], Anna Kinnen[1], Holger Gohlke[2,3] and Georg Groth[1]

[1]Institute for Biochemical Plant Physiology and Bioeconomy Science Center (BioSC), Heinrich-Heine University, Düsseldorf, Germany

[2]Institute of Pharmaceutical and Medicinal Chemistry and Bioeconomy Science Center (BioSC), Heinrich-Heine University, Düsseldorf, Germany

[3]John von Neumann Institute for Computing (NIC), Jülich Supercomputing Center (JSC) & Institute for Complex Systems - Structural Biochemistry (ICS 6), Forschungszentrum Jülich GmbH, Jülich, Germany

[4]Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Vienna Biocenter, Vienna, Austria.

[5]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California, USA.

SCIENTIFIC REPORTS

OPEN

Recognition motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1

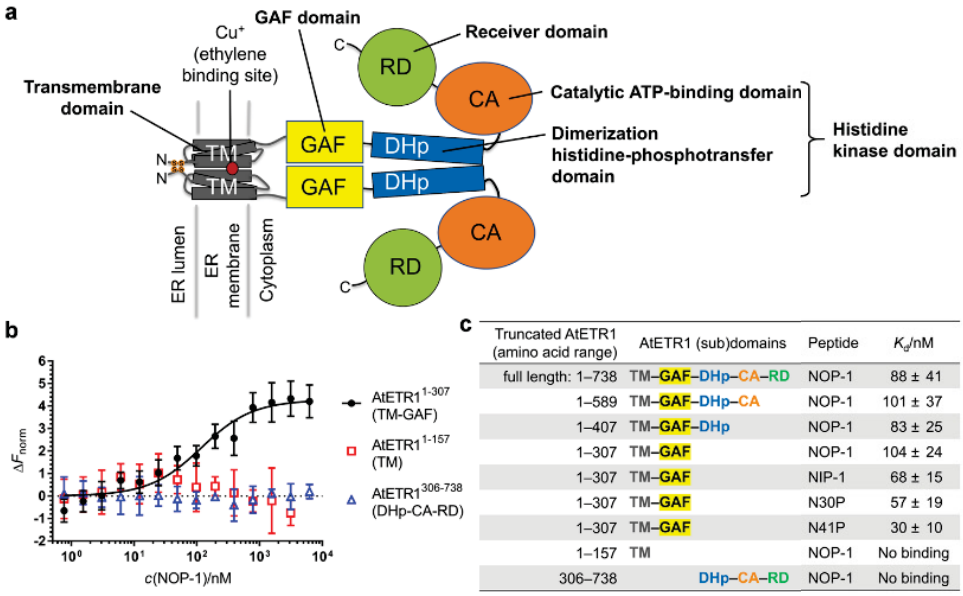
Received: 27 November 2017
Accepted: 13 February 2018
Published online: 01 March 2018

Dalibor Milić[1,4], Markus Dick[2,5], Daniel Mulnaes[2], Christopher Pfleger[2], Anna Kinnen[1], Holger Gohlke[2,3] & Georg Groth[1]

Synthetic peptides derived from ethylene-insensitive protein 2 (EIN2), a central regulator of ethylene signalling, were recently shown to delay fruit ripening by interrupting protein–protein interactions in the ethylene signalling pathway. Here, we show that the inhibitory peptide NOP-1 binds to the GAF domain of ETR1 – the prototype of the plant ethylene receptor family. Site-directed mutagenesis and computational studies reveal the peptide interaction site and a plausible molecular mechanism for the ripening inhibition.

Ripening of climacteric fruits, such as apples and tomatoes, is induced by the plant hormone ethylene. Such fruits and vegetables are usually harvested, transported, and stored in a green, unripe state, and full ripening is then induced by ethylene exposure at the final destination shortly before delivery. In order to avoid fruit damage and spoilage due to overripening, strategies have been developed to control ripening and minimize postharvest losses[1] by interfering with ethylene biosynthesis or signalling. Much of the current knowledge on signal perception and transduction of the plant hormone has been established by physiological, biochemical and genetic studies in the model plant *Arabidopsis thaliana*. Overall, more than a dozen genes have been implicated in the ethylene-signaling pathway, and their multi-stage interconnecting network has been tentatively determined using a combination of genetic and molecular approaches. In *Arabidopsis*, the ethylene signal is perceived by a family of five receptor proteins, which form homo- and heterodimers at the membrane of endoplasmic reticulum (ER) and function as negative regulators of the ethylene response[2–7]. The receptors are modular (Fig. 1a), organized similar to bacterial sensor histidine kinases and contain N-terminal transmembrane sensor domains (TM) followed by a cytosolic GAF domain (GAF), a dimerization histidine-phosphotransfer (DHp) and a catalytic ATP-binding (CA) domain forming the catalytic core, and a C-terminal response regulator domain (RD; not present in all members of the ethylene receptor family)[8,9]. Although the exact output of the receptors is still obscure, genetic studies demonstrate that in the absence of ethylene, receptors activate the Raf-like protein kinase CONSTITUTIVE TRIPLE RESPONSE 1 (CTR1), a negative regulator of the pathway[10]. Although CTR1 lacks any predicted transmembrane domains, it also resides at the ER membrane due to its physical interaction with the receptors[11]. Interaction with the receptors is considered critical for the induction of CTR1 kinase activity. Downstream of the receptors and the ER associated CTR1 kinase the membrane protein ETHYLENE INSENSITIVE 2 (EIN2) implements a positive regulatory role on ethylene signaling. The integral membrane protein was identified as the most crucial step in ethylene signaling since *ein2* is the only gene whose loss-of-function mutation confers complete ethylene insensitivity to the plant[12]. Recently, we identified inhibitory oligopeptides that delay ripening of tomatoes (*Solanum lycopersicum*) when applied onto the surface of an unripe fruit before or after its harvesting[13–15]. Their amino acid sequences are based on a highly conserved nuclear localization signal (NLS) found at the C-terminus of EIN2[16]. Molecular and genetic studies revealed that the C-terminal cytoplasmic part of EIN2 (EIN2-CEND) gets cleaved in the presence of ethylene by a so far unknown mechanism and has a

[1]Institute of Biochemical Plant Physiology and Bioeconomy Science Center (BioSC), Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [2]Institute of Pharmaceutical and Medicinal Chemistry and Bioeconomy Science Center (BioSC), Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [3]John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC) & Institute for Complex Systems - Structural Biochemistry (ICS 6), Forschungszentrum Jülich GmbH, Jülich, Germany. [4]Present address: Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Vienna Biocenter, Vienna, Austria. [5]Present address: Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California, USA. Dalibor Milić and Markus Dick contributed equally to this work. Correspondence and requests for materials should be addressed to H.G. (email: gohlke@hhu.de) or G.G. (email: georg.groth@hhu.de)



c

Truncated AtETR1 (amino acid range)	AtETR1 (sub)domains	Peptide	K_d/nM
full length: 1–738	TM–GAF–DHp–CA–RD	NOP-1	88 ± 41
1–589	TM–GAF–DHp–CA	NOP-1	101 ± 37
1–407	TM–GAF–DHp	NOP-1	83 ± 25
1–307	TM–GAF	NOP-1	104 ± 24
1–307	TM–GAF	NIP-1	68 ± 15
1–307	TM–GAF	N30P	57 ± 19
1–307	TM–GAF	N41P	30 ± 10
1–157	TM	NOP-1	No binding
306–738	DHp–CA–RD	NOP-1	No binding

Figure 1. Identification of the AtETR1 domain interacting with inhibitory octapeptide NOP-1. (**a**) Modular organization of the AtETR1 structure. The receptor forms a covalent dimer *via* two disulfide bridges at the N-terminus. The ethylene binding site (Cu^+ ion) is situated at the interface of two α-helical transmembrane (TM) domains immersed in a membrane of the endoplasmic reticulum (ER). The highly flexible cytoplasmic part of AtETR1 is composed of four domains: a GAF, a dimerization histidine-phosphotransfer (DHp), a catalytic ATP-binding (CA), and a receiver (RD) domain. DHp and CA domains are parts of a histidine kinase functional unit. (**b**) Binding of NOP-1 to the truncated AtETR1 constructs studied by microscale thermophoresis (MST). ΔF_{norm} is a relative normalized fluorescence measured for a fluorescently labelled protein at constant concentration (25 nM) in the presence of NOP-1 at different concentrations, *c*(NOP-1). $AtETR1^{1–307}$ (TM–GAF) still binds NOP-1, while $AtETR1^{1–157}$ (TM) and $AtETR1^{306–738}$ (DHp–CA–RD) show no binding. Mean values and standard deviations of ΔF_{norm} are plotted. (**c**) Dissociation constants (K_d) determined in MST binding experiments with the truncated AtETR1 constructs. All corresponding binding curves are presented in Supplementary Fig. S2.

crucial role in regulating expression of ethylene response genes[17–21]. Recent work in our laboratory showed that the synthetic inhibitory peptides derived from the NLS motif at the EIN2 C-terminus bind directly to ethylene receptors[14,15] and disrupt their interactions with EIN2-CEND[13,14].

In this report, we demonstrate that the inhibitory peptides bind to the GAF domain of ethylene receptor 1 (ETR1). Furthermore, the results of our experimental and computational biophysical studies not only indicate the peptide interaction site but also suggest a probable molecular mechanism of the ripening inhibition.

Results and Discussion

To understand the structural basis of interactions between ethylene receptors and inhibitory peptides, we heterologously expressed and purified C-terminally truncated constructs of ETR1 from the plant model organism *A. thaliana* (AtETR1), which were successively lacking protein domain modules starting from the C-terminus (Fig. 1a and Supplementary Fig. S1). Our goal was to identify AtETR1 domain(s) crucial for the interaction with the archetypal inhibitory octapeptide NOP-1 (LKRYKRRL-NH_2)[13–15], the sequence of which matches exactly the NLS sequence found in EIN2 of most plant species[14], including *A. thaliana* and tomato. Therefore, we used microscale thermophoresis to characterize binding of NOP-1 to the fluorescently labelled full-length AtETR1 and each of its four C-terminally truncated constructs (Fig. 1b,c and Supplementary Fig. S2). Out of these, $AtETR1^{1–157}$, containing the transmembrane (TM) domain only, showed no binding to the inhibitory peptide. All other C-terminally truncated constructs bound NOP-1 with binding affinities very similar to those of the full-length protein (dissociation constant $K_d = 88 + 41$ nM; Fig. 1c). To further explore the role of the histidine kinase (DHp and CA) or receiver domains (RD) in binding of NOP-1, we prepared $AtETR1^{306–738}$ containing only these domains. To our surprise, we observed no binding of NOP-1 to $AtETR1^{306–738}$ (Fig. 1b), thus ruling out our initial hypothesis that the NOP-1 binding site corresponds to a canonical phosphorylation site in the ETR1 histidine kinase or receiver domain[13]. Taken together, these results pinpointed the GAF domain as the ETR1 structural unit that interacts with NOP-1. Moreover, the three extended peptides NIP-1 (AFPKGKENLASVLKRYKRRL-NH_2)[13], N30P (GRTGTAAGDVAFPKGKENLASVLKRYKRRL-NH_2), and N41P (KDVEMAISSRKGRTGTAAGDVAFPKGKENLASVLKRYKRRL-NH_2) – all of which were derived from the AtEIN2 sequence and contain the NLS motif with additional 12, 22, or 33 upstream amino acid residues, respectively – also showed binding to $AtETR1^{1–307}$ (Fig. 1c and Supplementary Fig. S2). Their binding affinities improved with increasing peptide length, highlighting the importance of the NLS-core motif in this

interaction along with the positive correlation of sequence length on folding and/or stability of the biologicals (Supplementary Fig. S3).

Previous *in vivo* studies by various labs[7,22,23] have demonstrated a crucial role of the GAF domain for noncovalent homo- and hetero-oligomerization of ethylene receptors. Even before these discoveries, several researchers proposed that non-covalent interactions between the receptors and formation of higher-order oligomers might have functional implications in ethylene signalling and could explain the high sensitivity and broad concentration range of ethylene response[24–27].

To further understand the nature of peptide–GAF domain interactions, we first focused on predicting possible common structural motifs of peptides NOP-1, NIP-1, N30P, and N41P. We used 50 μs long molecular dynamics (MD) simulations, with three independent replicates for each system, in implicit solvent to perform *ab initio* folding simulations, motivated by recent successful studies[28,29]. In neither case did we see tertiary structure formation, and, except for specific regions (amino acids 7–9, 11–15 that tend to form α-helices), the major secondary structural elements were random coils (Supplementary Fig. S3a); these predictions were confirmed by CD spectroscopy (Supplementary Fig. S3b,c). Hence, it was not possible to identify a common structural motif. Nevertheless, such a result is not completely unexpected, considering the short length and high number of positive charges of the peptides, and the fact that the peptide sequences are part of the C-terminal domain of AtEIN2, which is predicted to be mainly disordered (60% disordered regions according to DISOPRED[30]).

As no experimental structure of the ETR1 GAF domain has been reported so far, we used our in-house software package TopModel[31] to build a structural model based on available templates (Supplementary Fig. S4 and Supplementary Table S1) applying the sequence of AtETR1$^{118–305}$ as the target (PDB ID and chain identifier of the templates given, with sequence identity indicated in parentheses: 3P01_A (18%), 3TRC_A (15%), 3CI6_A (13%), 3W2Z_A (12%), and 1YKD_B (15%)). A structural alignment between the GAF domain model and the templates used is shown in Supplementary Fig. S5. The final model built by TopModel (Fig. 2a) was assessed with our in-house model quality assessment program TopScore (D. Mulnaes, H. Gohlke, unpublished results; see Materials and Methods section for details) to be 71% correct, with the majority of inaccuracies being located in the flexible loop regions (residues in AtETR1 228–247 and 257–272: 47% and 52% inaccuracies, respectively).

Previous findings suggest that ethylene receptors form a dimer in their simplest functional state that is also mediated by their GAF domains[32]. We therefore built a dimer model of the AtETR1 GAF domain using our in-house protein–protein docking software TopDock (D. Mulnaes, H. Gohlke, unpublished results). TopDock predicts protein–protein contacts based on a structure-based homology search that is independent of sequence. TopDock identified five different homologous interfaces (PDB ID and chain identifiers given: 3G6O_AB, 3IBJ_AB, 3K2N_AB, 3P01_AB, and 3TRC_AB) all of which indicate that the dimer interface consists of the N- and C-terminal helices of the GAF domain (Supplementary Fig. S6). TopDock-predicted residue–residue contacts from each homologous interface were used for restrained docking of the GAF domains with HADDOCK[33]. The docking solutions were pooled and clustered by TopDock, and ranked according to HADDOCK energy, cluster size, distance to cluster centroid, and fulfilment of predicted contacts to select a docking solution (Fig. 2a). Each monomeric subunit of our final model contains a central, antiparallel, seven-fold β-sheet, flanked by one short α-helix (amino acids 213–220) and three, parallel-oriented α-helices that cover the N- and C-terminal regions (amino acids 118–173 and 290–305). Both N-terminal α-helices form the dimeric interface resulting in a six-helix bundle in the homodimeric structure (Fig. 2a). MD simulations of the protein of 500 ns length in the absence of any peptide ligand revealed overall moderate structural variations within both monomers (Supplementary Fig. S7), when the unstructured loop regions (residues 222–290) were omitted.

To identify interaction sites on the GAF dimer to which NOP-1 binds, we performed 15 independent MD simulations of 2 μs length each of free NOP-1 diffusion around the dimer, motivated by our own experience[31] and that of others[34,35] in related studies. To prevent any bias, NOP-1 was randomly placed in the simulation box also containing the ETR1 GAF dimer and explicit solvent (Fig. 2b). Over the simulation times, the locations of NOP-1 at the GAF dimer converge to three binding regions (Fig. 2b): (I) in the upper loop region (residues 283–286), (II) nearby the central β-sheets (residues 190–205), and (III) at the helices of the dimeric interface (residues 152–170). The propensity of hydrogen bond and salt bridge formation between a protein residue and NOP-1, averaged over the entire MD simulation data, confirmed preferred NOP-1/GAF dimer interactions with the three sites (Fig. 2c).

To validate the predictions of the interaction sites, we mutated the residues with the highest frequency of hydrogen bond formation (region I: E177, E178, E246, D283; region II: E190, E204; region III: E152, E169; Fig. 3a) to alanine and probed for NOP-1/GAF dimer interactions *in vitro*. AtETR1$^{1–307}$ variants II (E190A, E204A) and III (E152A, E169A) showed no binding of NOP-1 in the MST experiments (Fig. 3b). In contrast, AtETR1$^{1–307}$ variant I (E177A, E178A, E246A, D283A) interacted with NOP-1 with a similar affinity ($K_d = 128 \pm 65$ nM) as the unmutated AtETR1$^{1–307}$ ($K_d = 104 \pm 24$ nM), but with a smaller change in the relative normalized fluorescence (ΔF_{norm}). This is probably due to an increased net electric charge of the variant I and the related change in its hydration sphere, which ultimately influence both temperature-induced fluorescence jump and thermophoresis, and yet do not prevent NOP-1 from binding to the fluorescently labelled protein. Altogether, these results eliminate region I as a NOP-1 interaction site, however they do not clarify the roles of regions II and III in the NOP-1 binding.

To obtain more insights, we performed intrinsic fluorescence quenching experiments. Initially, we mutated two tryptophan residues in the AtETR1 GAF domain (W265 and W288) to phenylalanine to reduce background noise by natural tryptophan residues. The third tryptophan (W182) is located in the interior of the GAF domain and might be important for its structural integrity; hence, we left it unchanged resulting in the AtETR1$^{1–307}$-W265F-W288F construct. This variant was used as reference for individually introducing a tryptophan fluorescence reporter in close proximity of each predicted binding region (Fig. 3a). We then monitored intrinsic tryptophan fluorescence of four Trp-mutants (plus reference variant) in the presence of NOP-1 and found the largest quenching effect in the case of AtETR1$^{1–307}$-M148W-W265F-W288F – a variant with a tryptophan reporter

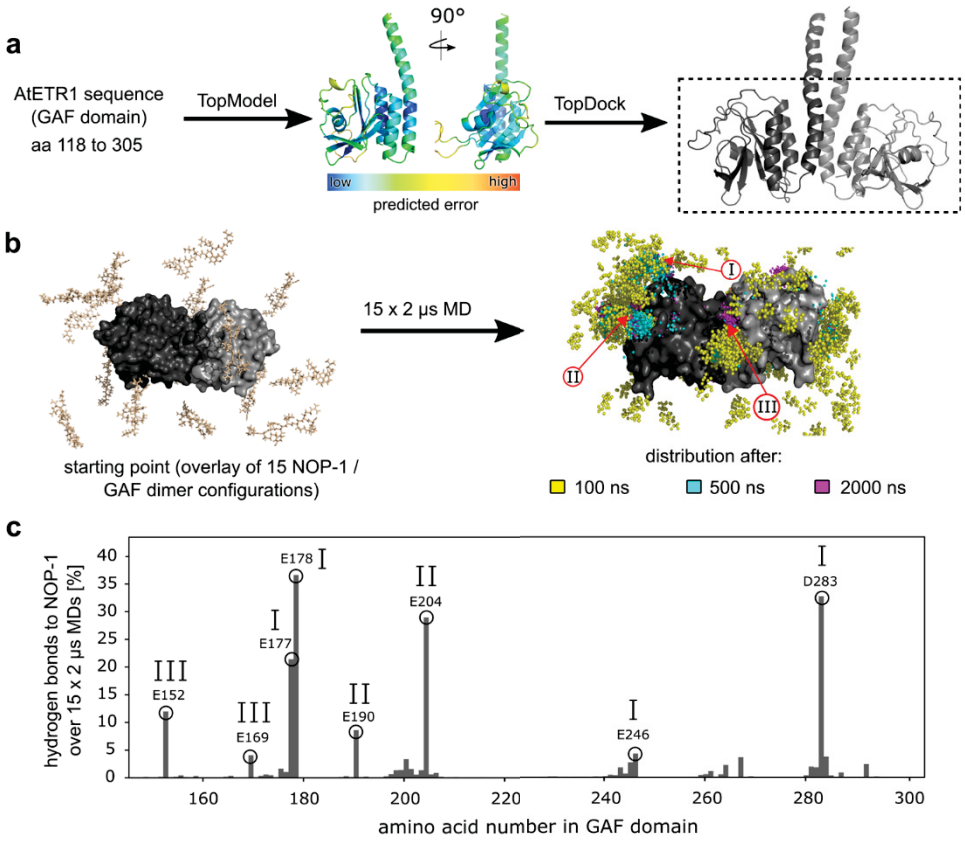


Figure 2. Molecular modelling of NOP-1 interactions with the GAF domain of AtETR1. (**a**) Model building of the GAF domain (dimeric form). Amino acids 118 to 305 of AtETR1 were used as a target sequence to build a homology model using TopModel[31]. The colouring of the monomeric structures represents the residue-wise uncertainty of the predicted model computed by TopScore. Next, protein–protein docking guided by positional restraints was performed to determine the interface between both monomeric subunits. As is known from experimental data (see Fig. 1b), amino acids 118 to 141 do not interact with NOP-1 and are not needed for the dimer formation. Thus, only the part of the protein shown in the dashed black box was used for further studies. (**b**) Starting from different initial NOP-1 positions (left, NOP-1 structures are coloured in beige, while the GAF domains are labelled in dark and light grey) 15 MD simulations of 2 µs length were performed. The cumulative distribution of the peptide after 100 ns (yellow), 500 ns (turquois), and 2000 ns (pink) over the 15 MD simulations is shown as points (representing the centre of mass of NOP-1) superimposed onto the average structure of the GAF dimer. The three main binding sites are highlighted by red arrows and labelled with Roman numerals (I to III). (**c**) The overall percentage of hydrogen bond and salt bridge formation with NOP-1 is shown for each residue of the GAF domain over the 15 MD simulations; results obtained for either domain in the GAF dimer were averaged. All residues chosen for mutation to alanine are labelled. The Roman numerals represent the corresponding binding sites as in panel b.

(M148W) located in binding region III (Fig. 3c and Supplementary Fig. S8). When placing the Trp reporter at a more distant position (T161W) to the proposed binding motif at site III, no significant quenching was observed, emphasizing that the NOP-1 inhibitory peptide binds in close proximity to acidic residues E152 and E169 in region III. In addition, the electrostatic potentials mapped onto the molecular surfaces of the GAF dimer and NOP-1 show a strong complementarity at site III, which supports a potential binding motif of NOP-1 at this site (Fig. 3d).

To probe a potential influence of NOP-1 binding on the structural stability of the GAF dimer, we used an ensemble-based perturbation approach[37] integrated into a method for analysing biomolecular rigidity and flexibility[38]. Initially, we clustered snapshots from the 15 MD simulations of free NOP-1 diffusion, in which NOP-1 binds to binding site III of the GAF domain on chain A (Fig. 4a,b), in order to combine similar configurations of bound NOP-1. Comparing the GAF dimer with and without bound NOP-1 for clusters 1–4 (which cover ~60% of all snapshots) revealed an increase in structural stability upon NOP-1 binding for about 60% of the residues (Fig. 4c). The largest $\Delta G_{i,CNA}$ were found for the loop region (A175–A180) and residues in the neighbouring helix (L167–L174) of the NOP-1-binding domain (Fig. 4c,d), with a maximal $\Delta G_{i,CNA} = 0.5\,\text{kcal mol}^{-1}$ for residue L176. Notably, even residues up to 20 Å away from the binding site III were influenced by NOP-1 binding, with E273 being the most distant one located in the other domain (Fig. 4c,d). The affected residues form a narrow pathway running across the dimer interface and extending into the other domain. Root mean square fluctuations (RMSF), a measure for atomic mobility, averaged over all MD simulations of the GAF dimer with NOP-1, are

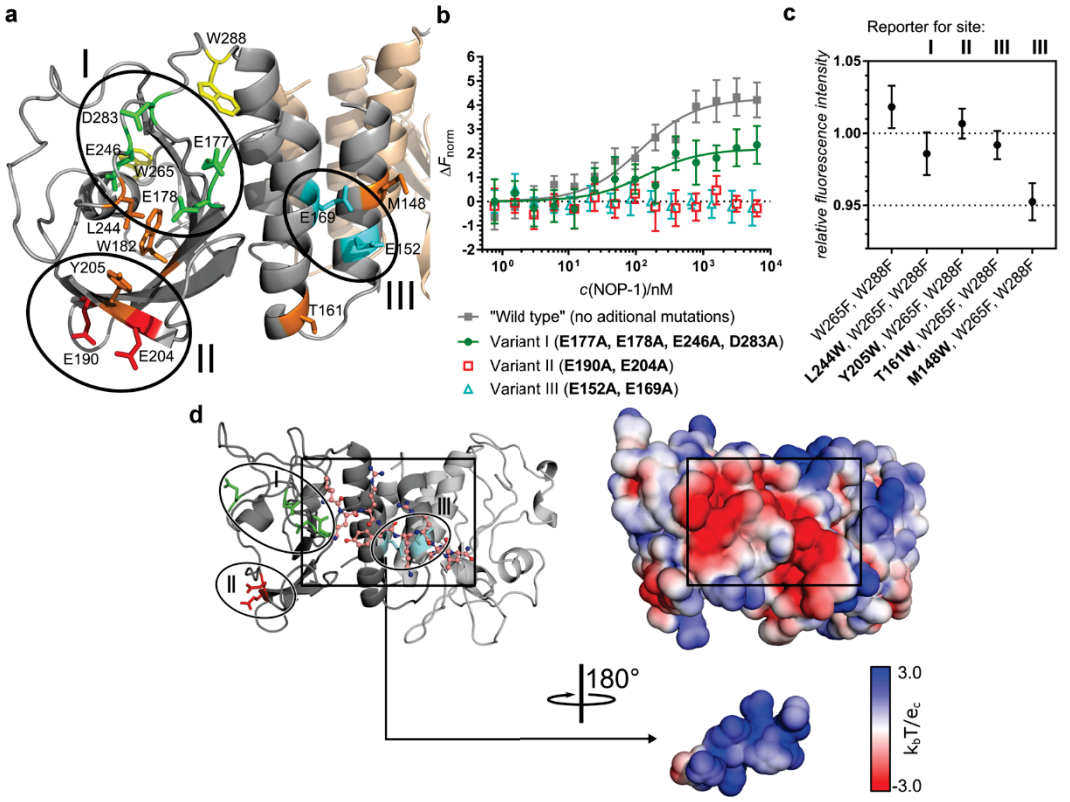


Figure 3. Evaluation of the predicted binding regions in the GAF domain of AtETR1. (**a**) Model of AtETR1 GAF domain with the highlighted acidic residues potentially involved in binding of NOP-1 (region I – *red*, region II – *green*, region III - *cyan*). Two tryptophans (W265 and W288) mutated to phenylalanine for the intrinsic fluorescence quenching experiments are shown in *yellow*. The remaining tryptophan (W182) and the four residues separately exchanged for tryptophan (fluorescence reporter) are highlighted in *orange*. (**b**) Binding of NOP-1 to the fluorescently labelled AtETR1$^{1-307}$ and its three variants monitored via microscale thermophoresis (relative normalized fluorescence, ΔF_{norm}). (**c**) Tryptophan fluorescence quenching of the AtETR1$^{1-307}$ Trp-variants by NOP-1. Fluorescence intensity of each Trp-variant in the presence of 10-fold excess of NOP-1 is given relative to fluorescence intensity of each protein measured without NOP-1. The complete titration data are presented in Supplementary Fig. S8. Mean values and standard deviations of independent triplicate measurements are shown in panels (**b**) and (**c**). (**d**) NOP-1 (within the black box) bound to the GAF domain at binding site III, taken from the merged clusters Cl 1–4 (Fig. 4). Circles indicate the three potential binding sites of the peptide as in panel (**a**). The colour scale of the electrostatic potentials ranges from −3.0 (red) to + 3.0 (blue) k_BT/e; the potentials were computed with the Adaptive Poisson-Boltzmann Solver (APBS)[36]. The view of NOP-1 is rotated by 180°, depicting the binding interface with the GAF dimer.

smaller by up to ~2 Å compared to MD simulations of the GAF dimer alone in regions distant to binding site III (residues 201–207 and 267–276; Supplementary Fig. S9); these regions coincide with those of higher structural stability identified by the rigidity and flexibility analysis (Fig. 4c,d). Thus, both independent approaches mutually corroborate each other. As the GAF dimer is rotationally symmetric, such an influence will also be felt *vice versa* if NOP-1 binds to the other domain. As a consequence, we speculate that due to the increased structural stability of the GAF dimer, the transmission of a signal, arising from ethylene binding to the TM domain of AtETR1, to domains C-terminal of the GAF domain is hampered (Fig. 4d). The structural stabilization does not contradict the observed Trp fluorescence quenching of the M148W mutant upon NOP-1 binding. We believe a positive charge of NOP-1 in close vicinity of W148 outweighs the positive effect that packing stabilization might have on the fluorescence intensity and results in the overall fluorescence quenching.

In summary, we have shown that the archetypical ripening inhibitory peptide NOP-1 interacts with the GAF domain of the plant ethylene receptor AtETR1 at helices of the dimeric interface. As a result, signal transmission from the TM domain of AtETR1 to the histidine kinase or receiver domains may be hampered, which may explain how NOP-1 inhibits ripening. While currently a full understanding of the AtETR1 signal transduction is hindered by the lack of a complete atomistic structure, our speculation is supported in that for a related histidine kinase[39] such signal transmission involved TM helix movements that are predicted in computational models to modulate the structural dynamics of the cytoplasmic domains. The predominant predicted binding mode involves primarily residues at the C-terminus of NOP-1, which may explain why the extension of NOP-1 at the N-terminus resulting in NIP-1, N30P, and N41P did not interfere with binding. Hence, this peptide part may be used to further optimize binding, stability, and applicability.

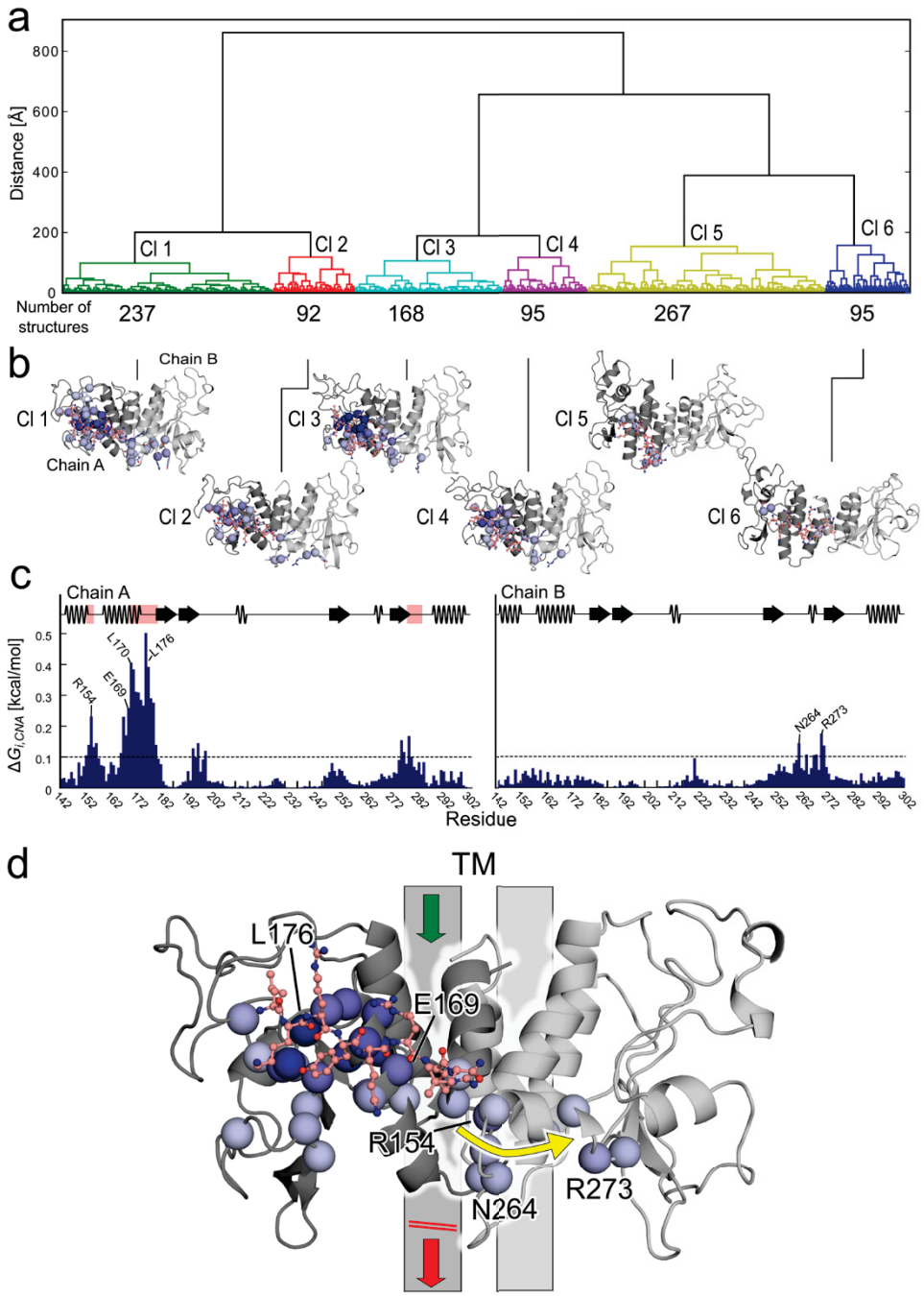


Figure 4. Influence of NOP-1 binding on the structural stability of the AtETR1 GAF dimer. (**a**) The dendrogram shows the clustering of 954 NOP-1 configurations bound at site III of the GAF dimer model (see Fig. 2). Hierarchical clustering was performed using the all-atom RMSD of NOP-1 as distance metric and Ward's minimum variance algorithm. The dendrogram was cut at a distance threshold $\delta(c_1, c_2) = 160$ Å resulting in six clusters (Cl 1–6). $\delta(c_1, c_2)$ is the square-root of the change in total sum of squares resulting from the fusion of clusters c_1 and c_2.[40,41] (**b**) CNA was applied on each cluster separately, and residues with $\Delta G_{i,CNA}$ above a threshold of 0.1 kcal mol^{-1} are depicted as spheres on the GAF dimer of each cluster centroid[45]. Blue colors reflect predicted $\Delta G_{i,CNA}$ values, with darker colors indicating larger values. (**c**) The histogram shows the per-residue $\Delta G_{i,CNA}$ of the merged clusters Cl 1–4. The dashed line at 0.1 kcal mol^{-1} indicates the threshold above which residues are considered perturbed, and pink colors highlight the region where NOP-1 binds. (**d**) Same information as shown in (**c**) for the merged clusters Cl 1–4 with NOP-1 bound at site III (salmon). The yellow arrow indicates how the perturbation upon removal of NOP-1 influences residues in chain B. The grey bars indicate connections to the transmembrane (TM) domain and dimerization domain. Due to the increased structural stability of the GAF dimer upon NOP-1 binding, we speculate that the transmission of a signal, arising from ethylene binding to the TM domain of AtETR1, to domains C-terminal of the GAF domain is hampered.

Materials and Methods

Inhibitory peptides. C-terminally amidated peptides NOP-1 (LKRYKRRL-NH_2), NIP-1 (AFPKGKENLASV LKRYKRRL-NH_2), N30P (GRTGTAAGDVAFPKGKENLASVLKRYKRRL-NH_2) and N41P (KDVEMAISSRKGRT GTAAGDVAFPKGKENLASVLKRYKRRL-NH_2) were purchased from GenScript as lyophilized trifluoacetate (TFA) salts with > 98% HPLC purity and stored at −20 °C. After dissolving a white peptide powder in a buffer of choice, peptide concentration in the resulting solution was determined spectroscopically from absorbance at 280 nm and the calculated molar attenuation coefficient (ProtParam)[42].

Molecular cloning. All truncated AtETR1 constructs and $AtETR1^{1–307}$ mutants were prepared in pTEV-16b vector backbone[43], a modified version of pET-16b (Novagen, Darmstadt, Germany) containing the N-terminal decahistidine-tag followed by a linker (SSGH) and a tobacco etch virus (TEV) protease cleavage site (ENLYFQG; instead of a Factor Xa cleavage site in pET-16b). The new constructs were made by using a two-fragment PCR approach[44] starting from the expression plasmid pTEV-16b-AtETR1 that contains the full-length *Arabidopsis thaliana* ethylene receptor 1 (AtETR1) cDNA. In short, the mutagenesis PCR primers were designed in either PCRdesign or AAscan program[45] with a 21-nucleotides overlap for a mutagenesis primer pair. Each fragment was amplified in a PCR with Phusion or Q5 high-fidelity DNA polymerase (both from New England BioLabs) or purchased from Integrated DNA Technologies as a gBlocks gene fragment. A pair of fragments was combined into the target plasmid in Gibson assembly[46], as described in our earlier report[44]. A detailed overview of the molecular cloning as well as the sequences of primers and gene fragments are given in Supplementary Tables S2–S4. The target constructs were verified by sequencing at SEQLAB Sequence Laboratories Göttingen or at the Biological-Medical Research Centre (BMFZ) of the Heinrich Heine University Düsseldorf.

Expression and purification of AtETR1, its C-terminally truncated constructs and $AtETR1^{1–307}$ mutants. For production of AtETR1 and its variants containing the transmembrane domain, we slightly modified our previous protocol[27]. In brief, the chemically competent *E. coli* C43 (DE3) (Lucigen Corporation) cells were transformed with the corresponding pTEV-16b expression plasmid. Transformants were precultured overnight in 2YT medium [16 g L^{-1} peptone, 10 g L^{-1} yeast extract and 5 g L^{-1} NaCl] with 100 µg mL^{-1} ampicillin at 30 °C. Typically, 30 mL preculture was diluted in 500 mL 2YT medium containing 100 µg mL^{-1} ampicillin in a 1-L baffled flask. Cultures were incubated at 30 °C while shaking at 180 rpm. The cells were grown to an optical density at 600 nm (OD_{600}) between 0.8 and 1.0 and induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). After incubation for additional 5 h, cells were spun down at 7,500 *g* for 15 min at 4 °C, flash-frozen in liquid nitrogen and stored at −20 °C. If not stated otherwise, all further purification steps were done on ice or at 4 °C. Cell pellets thawed on ice were resuspended by vortexing in ice-cold lysis buffer 1 [pH 8.0, 140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, 100 g L^{-1} glycerol, 20 mg L^{-1} phenylmethylsulfonyl fluoride (PMSF) and 10 mg L^{-1} DNase I (PanReac AppliChem); 5 mL lysis buffer per 1 g cells] and broken with Constants Cell Disruption System (Constant Systems) at 2.4 kbar and 5 °C. Cell debris and inclusion bodies were removed by centrifugation at 14,000 *g* for 30 min. The supernatant was centrifuged further at 40,000 *g* for 30 min, the resulting pellet was washed with the lysis buffer and centrifuged again at 34,000 *g* for 60 min to isolate cell membranes. Membrane pellets were used immediately in further purification or flash-frozen in liquid nitrogen and stored at −80 °C. To isolate the His-tagged proteins, membranes were resuspended with a paint brush in the solubilization buffer [50 mM Tris/HCl, pH 8.0 at 4 °C, 200 mM NaCl, 12 g L^{-1} fos-choline-16 (*n*-hexadecyl-phosphocholine; Glycon Biochemicals), 20 mg L^{-1} PMSF; 10 mL per 1 g membranes] and incubated for 1 h at 4 °C while mixing. Insoluble part was spun down at 200,000 *g* for 30 min and the supernatant was loaded to a 5-mL Ni-NTA HisTrap FF column (GE Healthcare Life Sciences) equilibrated with buffer A1 [50 mM Tris/HCl, pH 8.0 at 4 °C, 200 mM NaCl, 0.15 g L^{-1} fos-choline-16, 20 mg L^{-1} PMSF]. The protein-loaded column was washed with 25 mL buffer A1, followed by 100 mL buffer ATP1 [buffer A1 with additional 50 mM KCl, 20 mM $MgCl_2$ and 10 mM adenosine triphosphate (ATP)] to remove copurified chaperone DnaK, 50 mL buffer A and, finally, 50 mL wash buffer [buffer A1 with 50 mM imidazole]. His-tagged proteins were eluted with 25 mL elution buffer 1 [buffer A1 with 250 mM imidazole] and concentrated in a 100-kDa-MWCO Amicon Ultra-15 concentrator (EDM Millipore) to a final volume 2.5 mL. Buffer was exchanged for storage buffer 1 [50 mM Tris/HCl, pH 8.0 at 20 °C, 300 mM NaCl, 0.15 g L^{-1} fos-choline-16, 50 g L^{-1} glycerol] on a desalting PD-10 column (GE Healthcare Life Sciences) and the sample was centrifuged at 200,000 *g* for 30 min. Protein concentration in the supernatant was determined from absorbance measured at 280 nm and a corresponding molar attenuation coefficient computed using the ProtParam tool[42]. Glycerol was added to purified protein samples to final concentration 200 g L^{-1}. The samples with glycerol were distributed into 50-µL aliquots in 200-µL PCR tubes, flash-frozen in liquid nitrogen and stored at −80 °C. Purified proteins were analysed in SDS-PAGE followed by colloidal Coomassie staining[47] or western blotting to PVDF membrane (Amersham, GE Healthcare Life Sciences) and immunodetection with anti-His-HRP monoclonal antibody (Miltenyi Biotech).

Expression and purification of $AtETR1^{306–738}$. $AtETR1^{306–738}$ was expressed in chemically competent *E. coli* BL21 (DE3) Gold cells (Stratagene) additionally transformed with pBB540 and pBB542 plasmids[48] (a kind gift from Bernd Bukau, Heidelberg University), carrying the genes for chaperones GrpE, ClpB, DnaK, DnaJ, GroEL and GroES. Typically, 500 mL terrific broth (TB) medium (12 g L^{-1} tryptone, 24 g L^{-1} yeast extract, 5 g L^{-1} glycerol, 2.31 g L^{-1} KH_2PO_4 and 12.54 g L^{-1} K_2HPO_4) with 100 µg mL^{-1} ampicillin, 34 µg mL^{-1} chloramphenicol and 50 µg mL^{-1} spectinomycin in a 1-L baffled flask was inoculated with 1 mL overnight preculture and incubated at 37 °C while shaking at 160 rpm. The bacteria were grown to OD_{600} between 1.1 and 1.3, when they were cooled down on ice (5 min incubation), induced with 0.4 mM IPTG and further grown for 18 h at 20 °C. Cells were spun down (15 min, 7,500 *g*), flash-frozen in liquid nitrogen and stored at −20 °C. As already observed for some other AtETR1 constructs without the transmembrane domain (AtETR1-ΔTM)[32], purified $AtETR1^{306–738}$

precipitated at higher protein concentrations (>1 mg mL^{-1}) in our preliminary purification trials. To circumvent this, we used 0.15 g L^{-1} fos-choline-16 in our purification buffers (the same detergent concentration as for the other AtETR1 constructs with the transmembrane domain described in this work). If not stated otherwise, all purification steps were performed at 4 °C or on ice. The frozen cell pellet was thawed on ice, resuspended in lysis buffer 2 [5 mL buffer per 1 g wet cell pellet; 50 mM Tris/HCl, pH 8.5 at 4 °C, 250 mM NaCl, 20 mM imidazole, 2.5 mM dithiothreitol (DTT), cOmplete EDTA-free protease inhibitor cocktail (Roche) and 10 mg L^{-1} DNase I] and lysed in Constants Cell Disruption System at 2.4 kbar and 5 °C. Insoluble cell debris was separated by centrifugation at 200,000 *g* for 30 min, the supernatant was filtered through 0.22-µm syringe filter and loaded on a 5-mL HisTrap HP column (GE Healthcare Life Sciences) equilibrated with buffer A2 (50 mM Tris/HCl, pH 8.5 at 4 °C, 250 mM NaCl, 2.5 mM DTT, 0.15 g L^{-1} fos-choline-16, cOmplete EDTA-free protease inhibitor cocktail). The column was washed with 50 mL buffer A2, followed by 100 mL buffer ATP2 [50 mM Tris/HCl, pH 8.5 at 4 °C, 250 mM NaCl, 2.5 mM DTT, 0.15 g L^{-1} fos-choline-16, 50 mM KCl, 20 mM $MgCl_2$ and 10 mM ATP], 50 mL buffer A2 and 75 mL buffer A2 with 100 mM imidazole. Finally, $AtETR1^{306–738}$ was eluted with 50 mL elution buffer 2 (buffer A2 with 250 mM imidazole) and analysed in SDS-PAGE. The fractions containing the target protein were poured, concentrated (10-kDa-MWCO Amicon Ultra-15 concentrator, EDM Millipore) and imidazole removed by buffer exchange on a PD-10 column for storage buffer 2 [50 mM Tris/HCl, pH 8.5 at at 4 °C, 250 mM NaCl, 0.15 g L^{-1} fos-choline-16, 50 g L^{-1} glycerol, 2.5 mM DTT, cOmplete EDTA-free protease inhibitor cocktail]. The protein sample was centrifuged at 200,000 *g* for 30 min to remove potential aggregates. Finally, glycerol concentration in the supernatant was adjusted to 200 g L^{-1}, the sample divided into 50-µL aliquots in 200-µL PCR-tubes, flash-frozen in liquid nitrogen and stored at −80 °C.

Circular dichroism spectroscopy. Peptides and purified protein constructs were characterized in circular dichroism (CD) spectroscopy. For that, peptides were directly dissolved in degassed ultrapure Milli-Q water (Millipore) or degassed and filtered (0.22-µm filter) CD buffer (10 mM KH_2PO_4/K_2HPO_4, pH 8.0 at 20 °C) and subsequently diluted to 0.10 mg mL^{-1}. Original buffer of protein samples was exchanged for the CD buffer on a desalting PD MiniTrap G-25 column (GE Healthcare Life Sciences). Protein and fos-choline-16 concentrations were determined by using a Direct Detect infrared spectrometer (EMD Millipore) and the samples diluted to final protein concentration 0.10–0.20 mg mL^{-1}. Fos-choline-16 was added to each blank buffer solution to match detergent concentration in the final protein samples. CD spectra were recorded at room temperature on a J-715 spectropolarimeter (JASCO) using a 1-mm-path-length cylindrical quartz cuvette (Hellma). Each spectrum represents an average of 10 continuous scans (100 nm min^{-1}) with response time 0.25 s and bandwidth 1.0 nm. CD spectra of the peptides were analysed using the K2D2 web server[49] (Supplementary Fig. S3b,c). Secondary structure content of the protein constructs was calculated in programs CDSSTR[50], CONTIN[51] and SELCON3[52,53] from CDPro software package[54] using the reference protein set SMP50 (Supplementary Fig. S10 and S11).

Fluorescent labelling. For the microscale thermophoresis binding experiments, the proteins were labelled with thiol-reactive Alexa Fluor™ 488 C_5 maleimide fluorescent dye (ThermoFisher Scientific). For that, buffer of a concentrated freshly purified protein sample was exchanged on a desalting PD MiniTrap G-25 column resulting in 800 µL protein sample in labelling buffer [50 mM K_2HPO_4/KH_2PO_4, 300 mM NaCl and 0.15 g L^{-1} fos-choline-16]. 10 mg mL^{-1} Alexa Fluor™ 488 C_5 maleimide dimethyl sulfoxide (DMSO) solution was added to the protein sample in 3:1 dye:protein molar ratio and incubated in dark for 30 min at 20 °C while mixing slightly. Buffer was exchange for the storage buffer 2 ($AtETR1^{306–738}$) or storage buffer 1 (all other protein constructs) and the sample centrifuged for 30 min at 200,000 *g* and 4 °C. Spectroscopically determined degrees of labelling in the supernatants ranged from 140% to 300% for different AtETR1 constructs. After adjusting glycerol concentration to 200 g L^{-1}, the labelled protein samples were divided into 20-µL aliquots in 200-µL PCR tubes, flash-frozen in liquid nitrogen and stored at −80 °C.

Microscale thermophoresis (MST). Each inhibitory peptide was dissolved in the binding buffer [50 mM Tris/HCl, pH 8.0 at 20 °C, 300 mM NaCl, 0.15 g L^{-1} fos-choline-16] and serially diluted for MST measurements. Alexa-Fluor™-488-labelled AtETR1 constructs were diluted with the binding buffer to concentration 50 nM and mixed in a 1:1 volume ratio with each member of the peptide dilution series, resulting in 25 nM fluorescently labelled protein in the final 20-µL mixture. The protein–peptide mixtures were centrifuged at 14,000 *g* for 2 min before filling-up standard treated Monolith NT.115 MST glass capillaries (NanoTemper Technologies). Binding interactions were characterized in Monolith NT.115 Blue/Green (NanoTemper Technology) at 23–25 °C without temperature control. Power of the blue LED (excitation wavelength ca 470 nm) was adjusted depending on a degree of fluorescent labelling of each particular construct and fluorescence. Fluorescence in each capillary (emission wavelength 520 nm) was measured for 5 s without heating, then 30 s heating with 80% infrared laser (MST) power followed by 5 s without heating and 25 s delay before measurement of the next capillary. All measurements were run in at least three independent replicates. Data were evaluated from temperature jump (fluorescence signal between 0.5 s and 1.5 s after applying the laser normalized with the fluorescence signal in the last second before applying the laser) and fitted with nonlinear regression to the one-binding-site model[55–57] in GraphPad Prism version 7.00 for Windows (GraphPad Software, La Jolla California USA). As a negative control, a protein sample was diluted in the denaturation buffer [50 mM Tris/HCl, pH 8.0 at 20 °C, 300 mM NaCl, 0.15 g L^{-1} fos-choline-16, 40 g L^{-1} sodium dodecyl sulfate (SDS) and 40 mM DTT] and the MST measurements were carried out as described above.

Model building. The model structure of the GAF domain (amino acid 142 to 305 of AtETR1) was predicted using our in-house automated structure prediction pipeline TopModel[31,58]. TopModel is a multi-template meta-approach in which 20 different state-of-the-art threaders (see Supplementary Table S1) are used to detect

homologous templates. For each template the Topmodel-Score[59] to the native structure, a measure of structural similarity, is predicted using deep neural networks. These networks use alignment features, PSIPRED[60] secondary structure agreement, threading scores from individual threaders, model quality predicted by TopScore (D. Mulnaes, H. Gohlke, unpublished results; see also below for details), and structural consensus as input. Based on the neural network predictions, false positive templates are removed, consensus alignments are calculated, and the templates are ranked according to predicted TopModel-Scores. To sample different alignments, TopModel makes an ensemble of multiple sequence alignments (MSAs) using all combinations of the top five templates and eight different sequence and structure alignment programs (see Supplementary Table S1). These MSAs are used to generate 3D models of the GAF domain using Modeller9[61] and the template structures. Loops without template were refined using the DOPE potential[62] and secondary structure restraints based on PSIPRED predictions. The generated models were ranked with TopScore, and the highest ranked model for each template combination was selected for model combination and refinement. The selected models are refined with ModRefiner[63] and scored with TopScore. Based on TopScore predictions, regions with errors are removed and the remaining regions used as templates to construct meta-models. Two iterations of this refinement and model combination is performed, after which the best scoring model according to TopScore is selected as the final model of the GAF domain.

The correctness of the model is measured by TopScore as the predicted global and local lDDT score compared to the native structure. The lDDT score compares all intra-molecular heavy-atom distances within two structures and, thus, is superposition-free. Two models are considered completely different if all distances deviate by more than 4 Å, and completely identical if all distances deviate by less than 0.5 Å. Since the native structure is unknown in our case, the score is predicted by a deep neural network which uses multiple sources of information as input. These include knowledge-based angle, distance and contact potentials, residue stereochemistry, atom clashes, model clustering, and agreement between features predicted from the sequence and measured in the model, such as secondary structure, solvent accessibility, and residue contacts. The deep neural network was trained on a large data-set of 660 protein targets totaling over 133,000 models and over $19{\cdot}10^6$ residues.

Molecular dynamics (MD) simulations. The model structure of the GAF domain (amino acid 142 to 305 of AtETR1) and the linear forms ($\phi = \psi = 180°$) of NOP-1, NIP-1, N30P, and N41P with a C-terminal amino (NHE)-cap served as input structures for MD simulations. For receptor–peptide interaction studies, NOP-1 was randomly placed next to the GAF dimer with a minimum distance of 8 Å using the software package PackMol[64]; fifteen representative systems were generated that way. The solutes were placed in a truncated octahedral box of TIP3P[65] water leaving a distance of at least 11 Å between the protein and the solvation box boundaries, and Na^+ and Cl^- ions were added to reach a final salt concentration of 0.15 M. MD simulations were performed with the ff14SB force field[66]. Hydrogen mass repartitioning was used, allowing a time step of 4 fs[67]. Further parameters for system preparation, thermalization, and production runs are described in Minges *et al.*[68]. In short, each system was prepared performing a conjugate gradient minimization, followed by rising the temperature from 0 K to 300 K (over 100 ps) and adjusting the system density under NPT conditions. Production NVT-MD simulations were performed at 300 K utilizing the Berendsen thermostat[69], and conformations were saved every 100 ps.

For peptide folding simulations, three independent replicates (initiated by slightly different thermalization temperatures) of 50 µs simulation length were performed for each system. All simulations were performed in implicit solvent using the ff14SBonlysc force field in combination with mbondi3 radii and the GB-Neck2 model[70] as described by Nguyen *et al.*[28]. In short, after minimization and thermalization, MD simulations were performed with a time step of 4 fs using hydrogen mass repartitioning[67], temperature control at 300 K with a Langevin thermostat[71], and a long-range distance cut-off of 999 Å. Conformations were saved every 1 ns.

The trajectories were analysed with respect to secondary structure formation, distribution of NOP-1 around the GAF dimer, and RMSF using *cpptraj*[72]. The DSSP method of Kabsch and Sander[73] was utilized to calculate secondary structure types of each residue of NOP-1, NIP-1, N30P, and N41P. Values were averaged over all trajectories. For calculating the distribution of NOP-1 around the GAF dimer along the 15 MD simulations of free NOP-1 diffusion, the snapshots were superimposed onto the starting structure of the GAF dimer, a cubic grid with bin size 3×250 Å^2 was placed in the simulation box, and the presence of the centre of mass of NOP-1 within a grid bin was assessed after 100, 500, and 2000 ns of simulation time over all snapshots. The number of hydrogen bonds (and salt bridges) formed between NOP-1 and each residue of the GAF dimer over all trajectories was determined using VMD[74], where NOP-1 was chosen as donor and the receptor as acceptor molecule. Prior to computing C_α atom RMSF, snapshots of either the 15 MD simulations of free NOP-1 diffusion or the three MD simulations of the *apo* GAF dimer were superimposed onto the starting structure of the GAF dimer.

Tryptophan fluorescence. Steady-state intrinsic fluorescence of the freshly prepared AtETR1$^{1-307}$ Trp-mutants was measured on a LS-55 fluorescence spectrometer (PerkinElmer) using an excitation wavelength 295 nm. In the last protein purification step, the elution buffer 1 was exchanged for the binding buffer on a desalting PD MiniTrap G-25 column. To monitor binding of NOP-1 by fluorescence quenching, each protein sample was diluted with the same buffer to final concentration 1 µM and titrated with a concentrated stock solution of NOP-1 in the binding buffer at room temperature (22 °C) while stirring slowly in a 4-mm Quartz SUPRASIL Macro/Semi-micro cell with a small magnet (PerkinElmer). At the same time, intensity of an emission maximum at 344 nm was recorded as an average of 5 measurements. Fluorescence readings were corrected for the dilution effect. The inner filter effect of NOP-1 was negligible and could be ignored.

Constraint Network Analysis. To detect changes in biomolecular rigidity and flexibility upon NOP-1 binding, we analysed ensembles of snapshots in the biomolecule's bound and unbound states in terms of a perturbation approach[37]. First, an ensemble of network topologies is saved every 2 ns from the 15×2 µs of independent, unbiased MD simulations of free NOP-1 diffusion around the GAF dimer (see above). From this ensemble of

150,000 conformations, those conformations were extracted that have a hydrogen bond between NOP-1 and the residues E152 or E169, indicative of NOP-1 binding to site III of the GAF dimer; this yielded 954 snapshots for the ground state. The perturbed state is obtained by removing the covalent and non-covalent interactions associated with NOP-1 from each network topology of the ground state. In order to further group similar binding modes of NOP-1, we clustered NOP-1 conformations based on a pairwise all-atom RMDS according to Ward's method as implemented in SciPy[75]. This resulted in six clusters (see Fig. 4a). Second, altered biomolecular stability due to removal of NOP-1 is quantified in terms of a per-residue decomposition $\Delta G_{i,CNA}$ of the perturbation free energy. $\Delta G_{i,CNA}$ was computed based on rigidity analyses performed with the CNA software package[38] on the ensembles of network topologies of the ground and perturbed states. Network topologies (containing nodes (atoms) and constraints (covalent and non-covalent interactions)) were constructed with the FIRST (Floppy Inclusions and Rigid Substructure Topography) software (version 6.2)[76] to which CNA is a front and back end. The strength of hydrogen bonds (including salt bridges) were assigned by the energy E_{HB} computed by FIRST[77]. Hydrophobic interactions between carbon or sulfur atoms were taken into account if the distance between these atoms was less than the sum of their van der Waals radii (C: 1.7 Å, S: 1.8 Å) plus $D_{cut} = 0.25$ Å[78]. Non-covalent interactions between NOP-1 and the GAF domain were identified using knowledge-based DrugScore pair potentials[79].

When CNA was applied on each cluster 1–6 (see above) separately, the clusters 5 and 6 revealed only minor and local altered structural stability of the GAF dimer upon NOP-1 removal (see Fig. 4b) and, thus, were excluded from further analyses. Clusters 1–4 were merged for subsequent analyses. This resulted in a final ensemble of 592 snapshots used as input for CNA. Upon perturbation, the network topologies lose on average 7.5 (=1.3% of all) hydrogen bond constraints and 2.2 (=1.6% of all) hydrophobic tether constraints. About 60% of the residues in the GAF domain show altered stability characteristic, with 9% of the residues having $\Delta G_{i,CNA}$ values > 0.1 kcal mol^{-1} upon removal of NOP-1.

Electrostatic surface potential. The electrostatic surface potential for the GAF dimer and NOP-1 was calculated using the Adaptive Poisson-Boltzmann Solver (APBS)[36]. The complex structure of the GAF dimer and NOP-1 were first split into their single components. For the APBS calculations, default parameters were used, the temperature of the system was set to 300 K, and the concentration of 1:1 counterions to 0.15 M.

Data availability statement. The data generated and analysed during the current study are either included in this published article and its Supplementary Information file or available from the corresponding authors on reasonable request.

References

1. Payasi, A. & Sanwal, G. G. Ripening of climacteric fruits and their control. *J. Food Biochem.* **34**, 679–710, https://doi.org/10.1111/j.1745-4514.2009.00307.x (2010).
2. Bleecker, A. B., Estelle, M. A., Somerville, C. & Kende, H. Insensitivity to Ethylene Conferred by a Dominant Mutation in *Arabidopsis thaliana*. *Science* **241**, 1086–1089, https://doi.org/10.1126/science.241.4869.1086 (1988).
3. Chang, C., Kwok, S., Bleecker, A. & Meyerowitz, E. *Arabidopsis* ethylene-response gene ETR1: similarity of product to two-component regulators. *Science* **262**, 539–544, https://doi.org/10.1126/science.8211181 (1993).
4. Hua, J., Chang, C., Sun, Q. & Meyerowitz, E. Ethylene insensitivity conferred by *Arabidopsis* ERS gene. *Science* **269**, 1712–1714, https://doi.org/10.1126/science.7569898 (1995).
5. Hua, J. & Meyerowitz, E. M. Ethylene Responses Are Negatively Regulated by a Receptor Gene Family in *Arabidopsis thaliana*. *Cell* **94**, 261–271, https://doi.org/10.1016/S0092-8674(00)81425-7 (1998).
6. Hua, J. *et al.* EIN4 and ERS2 Are Members of the Putative Ethylene Receptor Gene Family in Arabidopsis. *The Plant Cell* **10**, 1321–1332, https://doi.org/10.1105/tpc.10.8.1321 (1998).
7. Grefen, C. *et al.* Subcellular Localization and *In Vivo* Interactions of the *Arabidopsis thaliana* Ethylene Receptor Family Members. *Mol. Plant* **1**, 308–320, https://doi.org/10.1093/mp/ssm015 (2008).
8. Bleecker, A. B. & Kende, H. Ethylene: A Gaseous Signal Molecule in Plants. *Annu. Rev. Cell Dev. Biol.* **16**, 1–18, https://doi.org/10.1146/annurev.cellbio.16.1.1 (2000).
9. Stepanova, A. N. & Ecker, J. R. Ethylene signaling: from mutants to molecules. *Curr. Opin. Plant. Biol.* **3**, 353–360, https://doi.org/10.1016/s1369-5266(00)00096-0 (2000).
10. Kieber, J. J., Rothenberg, M., Roman, G., Feldmann, K. A. & Ecker, J. R. CTR1, a negative regulator of the ethylene response pathway in Arabidopsis, encodes a member of the Raf family of protein kinases. *Cell* **72**, 427–441, https://doi.org/10.1016/0092-8674(93)90119-B (1993).
11. Gao, Z. *et al.* Localization of the Raf-like Kinase CTR1 to the Endoplasmic Reticulum of *Arabidopsis* through Participation in Ethylene Receptor Signaling Complexes. *J. Biol. Chem.* **278**, 34725–34732, https://doi.org/10.1074/jbc.M305548200 (2003).
12. Alonso, J. M., Hirayama, T., Roman, G., Nourizadeh, S. & Ecker, J. R. EIN2, a Bifunctional Transducer of Ethylene and Stress Responses in *Arabidopsis*. *Science* **284**, 2148–2152, https://doi.org/10.1126/science.284.5423.2148 (1999).
13. Bisson, M. M. A. & Groth, G. Targeting Plant Ethylene Responses by Controlling Essential Protein–Protein Interactions in the Ethylene Pathway. *Mol. Plant* **8**, 1165–1174, https://doi.org/10.1016/j.molp.2015.03.014 (2015).
14. Bisson, M. M. A. *et al.* Peptides interfering with protein-protein interactions in the ethylene signaling pathway delay tomato fruit ripening. *Sci. Rep.* **6**, 30634, https://doi.org/10.1038/srep30634 (2016).
15. Kessenbrock, M. *et al.* Novel Protein-Protein Inhibitor Based Approach to Control Plant Ethylene Responses: Synthetic Peptides for Ripening Control. *Front. Plant Sci.* **8**, https://doi.org/10.3389/fpls.2017.01528 (2017).
16. Bisson, M. M. A. & Groth, G. New paradigm in ethylene signaling: EIN2, the central regulator of the signaling pathway, interacts directly with the upstream receptors. *Plant Signal. Behav.* **6**, 164–166, https://doi.org/10.4161/psb.6.1.14034 (2011).
17. Qiao, H., Chang, K. N., Yazaki, J. & Ecker, J. R. Interplay between ethylene, ETP1/ETP2 F-box proteins, and degradation of EIN2 triggers ethylene responses in. *Arabidopsis*. *Genes Dev.* **23**, 512–521, https://doi.org/10.1101/gad.1765709 (2009).
18. Qiao, H. *et al.* Processing and Subcellular Trafficking of ER-Tethered EIN2 Control Response to Ethylene Gas. *Science* **338**, 390–393, https://doi.org/10.1126/science.1225974 (2012).
19. Wen, X. *et al.* Activation of ethylene signaling is mediated by nuclear translocation of the cleaved EIN2 carboxyl terminus. *Cell Res.* **22**, 1613–1616, https://doi.org/10.1038/cr.2012.145 (2012).
20. Li, W. *et al.* EIN2-directed translational regulation of ethylene signaling in *Arabidopsis*. *Cell* **163**, 670–683, https://doi.org/10.1016/j.cell.2015.09.037 (2015).
21. Merchante, C. *et al.* Gene-Specific Translation Regulation Mediated by the Hormone-Signaling Molecule EIN2. *Cell* **163**, 684–697, https://doi.org/10.1016/j.cell.2015.09.036 (2015).

22. Xie, F., Liu, Q. & Wen, C.-K. Receptor Signal Output Mediated by the ETR1 N Terminus Is Primarily Subfamily I Receptor Dependent. *Plant Physiol.* **142**, 492–508, https://doi.org/10.1104/pp.106.082628 (2006).
23. Gao, Z. *et al.* Heteromeric Interactions among Ethylene Receptors Mediate Signaling in Arabidopsis. *J. Biol. Chem.* **283**, 23801–23810, https://doi.org/10.1074/jbc.M800641200 (2008).
24. Gamble, R. L., Qu, X. & Schaller, G. E. Mutational Analysis of the Ethylene Receptor ETR1. Role of the Histidine Kinase Domain in Dominant Ethylene Insensitivity. *Plant Physiol.* **128**, 1428–1438, https://doi.org/10.1104/pp.010777 (2002).
25. Binder, B. M. & Bleecker, A. B. A Model for Ethylene Receptor Function and 1-Methylcyclopropene Action *Acta Hortic.*, 177-187, https://doi.org/10.17660/ActaHortic.2003.628.21 (2003).
26. Binder, B. M. *et al.* *Arabidopsis* Seedling Growth Response and Recovery to Ethylene. A Kinetic Analysis. *Plant Physiol.* **136**, 2913–2920, https://doi.org/10.1104/pp.104.050369 (2004).
27. Binder, B. M., Mortimore, L. A., Stepanova, A. N., Ecker, J. R. & Bleecker, A. B. Short-Term Growth Responses to Ethylene in *Arabidopsis* Seedlings Are EIN3/EIL1 Independent. *Plant Physiol.* **136**, 2921–2927, https://doi.org/10.1104/pp.104.050393 (2004).
28. Nguyen, H., Maier, J., Huang, H., Perrone, V. & Simmerling, C. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J. Am. Chem. Soc.* **136**, 13959–13962, https://doi.org/10.1021/ja5032776 (2014).
29. Maffucci, I. & Contini, A. An Updated Test of AMBER Force Fields and Implicit Solvent Models in Predicting the Secondary Structure of Helical, beta-Hairpin, and Intrinsically Disordered Peptides. *J. Chem. Theory Comput.* **12**, 714–727, https://doi.org/10.1021/acs.jctc.5b01211 (2016).
30. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645, https://doi.org/10.1016/j.jmb.2004.02.002 (2004).
31. Gohlke, H. *et al.* Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion. *J. Chem. Inf. Model.* **53**, 2493–2498, https://doi.org/10.1021/ci400370y (2013).
32. Mayerhofer, H. *et al.* Structural Model of the Cytosolic Domain of the Plant Ethylene Receptor 1 (ETR1). *J. Biol. Chem.* **290**, 2644–2658, https://doi.org/10.1074/jbc.M114.587667 (2015).
33. Dominguez, C., Boelens, R. & Bonvin, A. M. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737, https://doi.org/10.1021/ja026939x (2003).
34. Ahmad, M., Gu, W. & Helms, V. Mechanism of fast peptide recognition by SH3 domains. *Angew. Chem. Int. Ed. Engl.* **47**, 7626–7630, https://doi.org/10.1002/anie.200801856 (2008).
35. Zwier, M. C. *et al.* Efficient Atomistic Simulation of Pathways and Calculation of Rate Constants for a Protein-Peptide Binding Process: Application to the MDM2 Protein and an Intrinsically Disordered p53 Peptide. *J. Phys. Chem. Lett.* **7**, 3440–3445, https://doi.org/10.1021/acs.jpclett.6b01502 (2016).
36. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **27**, 112–128, https://doi.org/10.1002/pro.3280 (2018).
37. Pfleger, C. *et al.* Ensemble- and rigidity theory-based perturbation approach to analyze dynamic allostery. *J. Chem. Theory Comput.*. https://doi.org/10.1021/acs.jctc.7b00529 (2017).
38. Pfleger, C., Rathi, P. C., Klein, D. L., Radestock, S. & Gohlke, H. Constraint Network Analysis (CNA): A Python Software Package for Efficiently Linking Biomacromolecular Structure, Flexibility, (Thermo-)Stability, and Function. *J. Chem. Inf. Model.* **53**, 1007–1015, https://doi.org/10.1021/ci400044m (2013).
39. Lemmin, T., Soto, C. S., Clinthorne, G., DeGrado, W. F. & Dal Peraro, M. Assembly of the transmembrane domain of E. coli PhoQ histidine kinase: implications for signal transduction from molecular simulations. *PLoS Comput. Biol.* **9**, e1002878, https://doi.org/10.1371/journal.pcbi.1002878 (2013).
40. Kaufman, L. & Rousseeuw, P. J. Finding Groups in Data: An Introduction to Cluster Analysis. Vol. 344 (John Wiley & Sons, 2009).
41. Legendre, P. & Legendre, L. F. J. Numerical ecology. Vol. 24 (Elsevier, 2012).
42. Gasteiger, E. *et al.* In *The Proteomics Protocols Handbook* (ed John M. Walker) 571-607 (Humana Press, 2005).
43. Classen, E. & Groth, G. Cloning, expression and purification of orthologous membrane proteins: a general protocol for preparation of the histidine sensor kinase ETR1 from different species. *Mol. Membr. Biol.* **29**, 26–35, https://doi.org/10.3109/09687688.2012.667576 (2012).
44. Heydenreich, F. M. *et al.* High-throughput mutagenesis using a two-fragment PCR approach. *Sci. Rep.* **7**, 6787, https://doi.org/10.1038/s41598-017-07010-4 (2017).
45. Sun, D. *et al.* AAscan, PCRdesign and MutantChecker: A Suite of Programs for Primer Design and Sequence Analysis for High-Throughput Scanning Mutagenesis. *PLoS One* **8**, e78878, https://doi.org/10.1371/journal.pone.0078878 (2013).
46. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345, https://doi.org/10.1038/nmeth.1318 (2009).
47. Kang, D.-H., Gho, Y.-S., Suh, M.-K. & Kang, C.-H. Highly Sensitive and Fast Protein Detection with Coomassie Brilliant Blue in Sodium Dodecyl Sulfate-Polyacrylamide GelElectrophoresis. *Bull. Korean Chem. Soc.* **23**, 1511–1512, https://doi.org/10.5012/bkcs.2002.23.11.1511 (2002).
48. de Marco, A., Deuerling, E., Mogk, A., Tomoyasu, T. & Bukau, B. Chaperone-based procedure to increase yields of soluble recombinant proteins produced in E. coli. *BMC Biotechnol.* **7**, 32, https://doi.org/10.1186/1472-6750-7-32 (2007).
49. Perez-Iratxeta, C. & Andrade-Navarro, M. A. K2D2: estimation of protein secondary structure from circular dichroism spectra. *BMC Struct Biol* **8**, 25, https://doi.org/10.1186/1472-6807-8-25 (2008).
50. Johnson, W. C. Analyzing protein circular dichroism spectra for accurate secondary structures. *Proteins: Struct., Funct., Bioinf.* **35**, 307–312, https://doi.org10.1002/(SICI)1097-0134(19990515)35:3<307::AID-PROT4>3.0.CO;2-3 (1999).
51. Provencher, S. W. & Gloeckner, J. Estimation of globular protein secondary structure from circular dichroism. *Biochemistry* **20**, 33–37, https://doi.org/10.1021/bi00504a006 (1981).
52. Sreerama, N. & Woody, R. W. A Self-Consistent Method for the Analysis of Protein Secondary Structure from Circular Dichroism. *Anal. Biochem.* **209**, 32–44, https://doi.org/10.1006/abio.1993.1079 (1993).
53. Sreerama, N., Venyaminov, S. Y. U. & Woody, R. W. Estimation of the number of α-helical and β-strand segments in proteins using circular dichroism spectroscopy. *Protein Sci.* **8**, 370–380, https://doi.org/10.1110/ps.8.2.370 (1999).
54. Sreerama, N. & Woody, R. W. Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set. *Anal. Biochem.* **287**, 252–260, https://doi.org/10.1006/abio.2000.4880 (2000).
55. Wienken, C. J., Baaske, P., Rothbauer, U., Braun, D. & Duhr, S. Protein-binding assays in biological liquids using microscale thermophoresis. *Nat. Commun.* **1**, 100, https://doi.org/10.1038/ncomms1093 (2010).
56. Seidel, S. A. I. *et al.* Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions. *Methods* **59**, 301–315, https://doi.org/10.1016/j.ymeth.2012.12.005 (2013).
57. Jerabek-Willemsen, M., Wienken, C. J., Braun, D., Baaske, P. & Duhr, S. Molecular Interaction Studies Using Microscale Thermophoresis. *Assay Drug Dev. Technol.* **9**, 342–353, https://doi.org/10.1089/adt.2011.0380 (2011).
58. Widderich, N. *et al.* Molecular dynamics simulations and structure-guided mutagenesis provide insight into the architecture of the catalytic core of the ectoine hydroxylase. *J. Mol. Biol.* **426**, 586–600, https://doi.org/10.1016/j.jmb.2013.10.028 (2014).
59. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309, https://doi.org/10.1093/nar/gki524 (2005).

60. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405, https://doi.org/10.1093/bioinformatics/16.4.404 (2000).
61. Webb, B. & Sali, A. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics* **5.6**, 1–5.6. 32, https://doi.org/10.1002/0471250953.bi0506s15 (2014).
62. Shen, M. Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524, https://doi.org/10.1110/ps.062416606 (2006).
63. Xu, D. & Zhang, Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys. J.* **101**, 2525–2534, https://doi.org/10.1016/j.bpj.2011.10.024 (2011).
64. Martinez, L., Andrade, R., Birgin, E. G. & Martinez, J. M. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **30**, 2157–2164, https://doi.org/10.1002/jcc.21224 (2009).
65. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **79**, 926–935, https://doi.org/10.1063/1.445869 (1983).
66. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713, https://doi.org/10.1021/acs.jctc.5b00255 (2015).
67. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **11**, 1864–1874, https://doi.org/10.1021/ct5010406 (2015).
68. Minges, A. *et al.* Structural intermediates and directionality of the swiveling motion of Pyruvate Phosphate Dikinase. *Sci. Rep.* **7**, 45389, https://doi.org/10.1038/srep45389 (2017).
69. Berendsen, H. J. C., Postma, J. P. M., Vangunsteren, W. F., Dinola, A. & Haak, J. R. Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **81**, 3684–3690, https://doi.org/10.1063/1.448118 (1984).
70. Nguyen, H., Roe, D. R. & Simmerling, C. Improved Generalized Born Solvent Model Parameters for Protein Simulations. *J. Chem. Theory Comput.* **9**, 2020–2034, https://doi.org/10.1021/ct3010485 (2013).
71. Larini, L., Mannella, R. & Leporini, D. Langevin stabilization of molecular-dynamics simulations of polymers by means of quasisymplectic algorithms. *J. Chem. Phys.* **126**, 104101, https://doi.org/10.1063/1.2464095 (2007).
72. Roe, D. R. & Cheatham, T. E. 3rd PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084–3095, https://doi.org/10.1021/ct400341p (2013).
73. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637, https://doi.org/10.1002/bip.360221211 (1983).
74. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**(33-38), 27–38, https://doi.org/10.1016/0263-7855(96)00018-5 (1996).
75. Jones, E. *SciPy: Open Source Scientific Tools for Python*,http://www.scipy.org/ (2001).
76. Jacobs, D. J., Rader, A. J., Kuhn, L. A. & Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins* **44**, 150–165, https://doi.org/10.1002/prot.1081 (2001).
77. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337, https://doi.org/10.1002/pro.5560060622 (1997).
78. Rader, A. J., Hespenheide, B. M., Kuhn, L. A. & Thorpe, M. F. Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. USA* **99**, 3540–3545, https://doi.org/10.1073/pnas.062492699 (2002).
79. Gohlke, H., Hendlich, M. & Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **295**, 337–356, https://doi.org/10.1006/jmbi.1999.3371 (2000).

Acknowledgements

This work was supported by a grant from the Ministry of Innovation, Science and Research within the framework of the NRW Strategieprojekt BioSC (No. 313/323-400-002 13) by the boost fund 'RIPE' granted to H.G. and G.G. and by the Deutsche Forschungsgemeinschaft, CRC 1208, project B06 (G.G.) and project A03 (H.G.). We are grateful for computational support and infrastructure provided by the "Zentrum für Informations- und Medientechnologie" (ZIM) at the Heinrich Heine University Düsseldorf and the computing time provided by the John von Neumann Institute for Computing (NIC) to H.G. on the supercomputer JURECA at Jülich Supercomputing Centre (JSC) (user ID: HKF7). Financial support by DFG for funds (INST 208/704-1 FUGG) to purchase the hybrid computer cluster used in this study is gratefully acknowledged.

Author Contributions

G.G. and H.G. designed research; D.M., M.D., D.M., C.P., A.K. performed research; all authors analysed data; all authors wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-21952-3.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information

Recognition motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1

Dalibor Milić[1,†,‡], Markus Dick[2,†,§], Daniel Mulnaes[2], Christopher Pfleger[2], Anna Kinnen[1], Holger Gohlke[2,3,*] and Georg Groth[1,*]

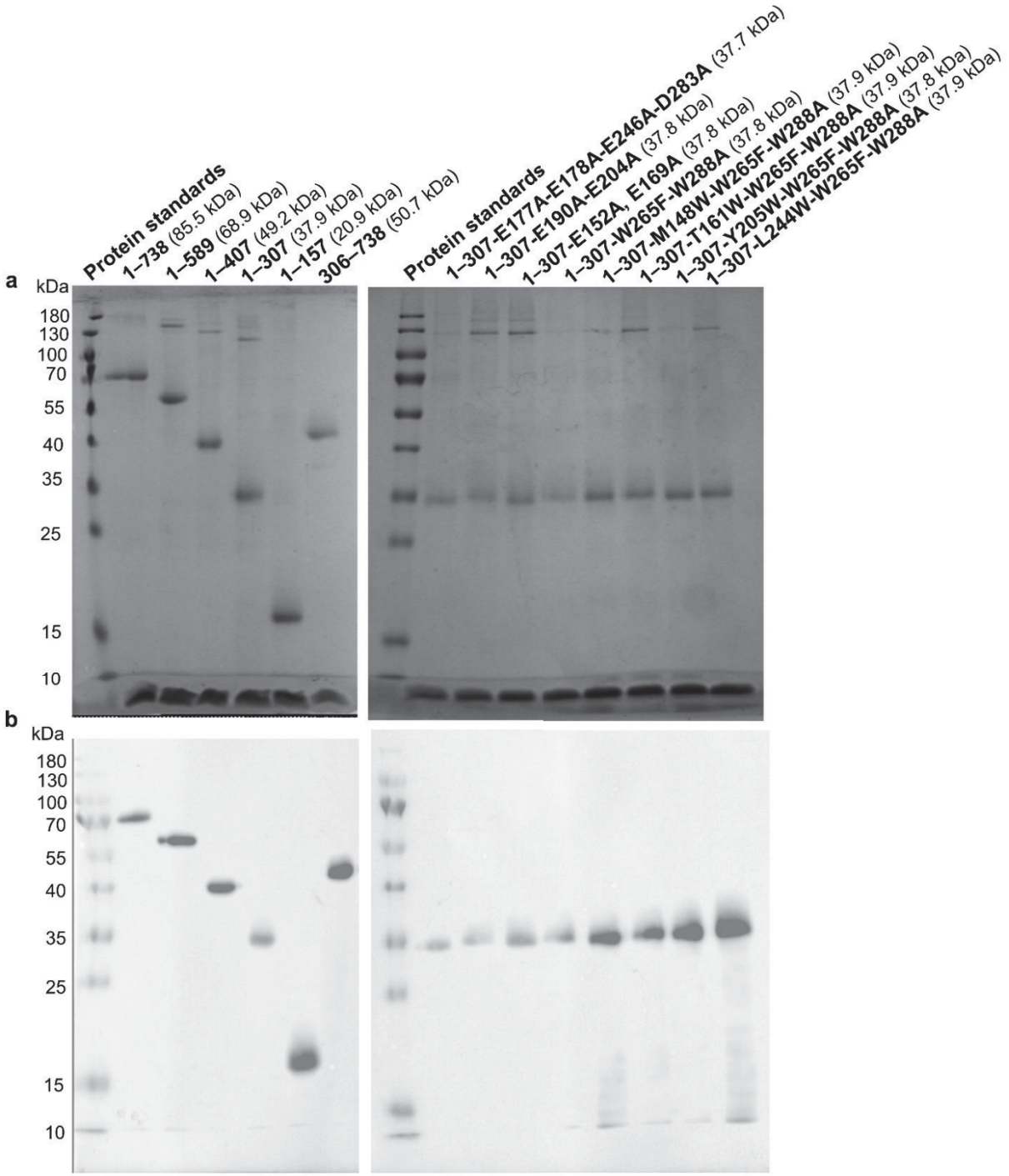
[1]Institute of Biochemical Plant Physiology and Bioeconomy Science Center (BioSC), Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[2]Institute of Pharmaceutical and Medicinal Chemistry and Bioeconomy Science Center (BioSC), Heinrich Heine University Düsseldorf, Düsseldorf, Germany

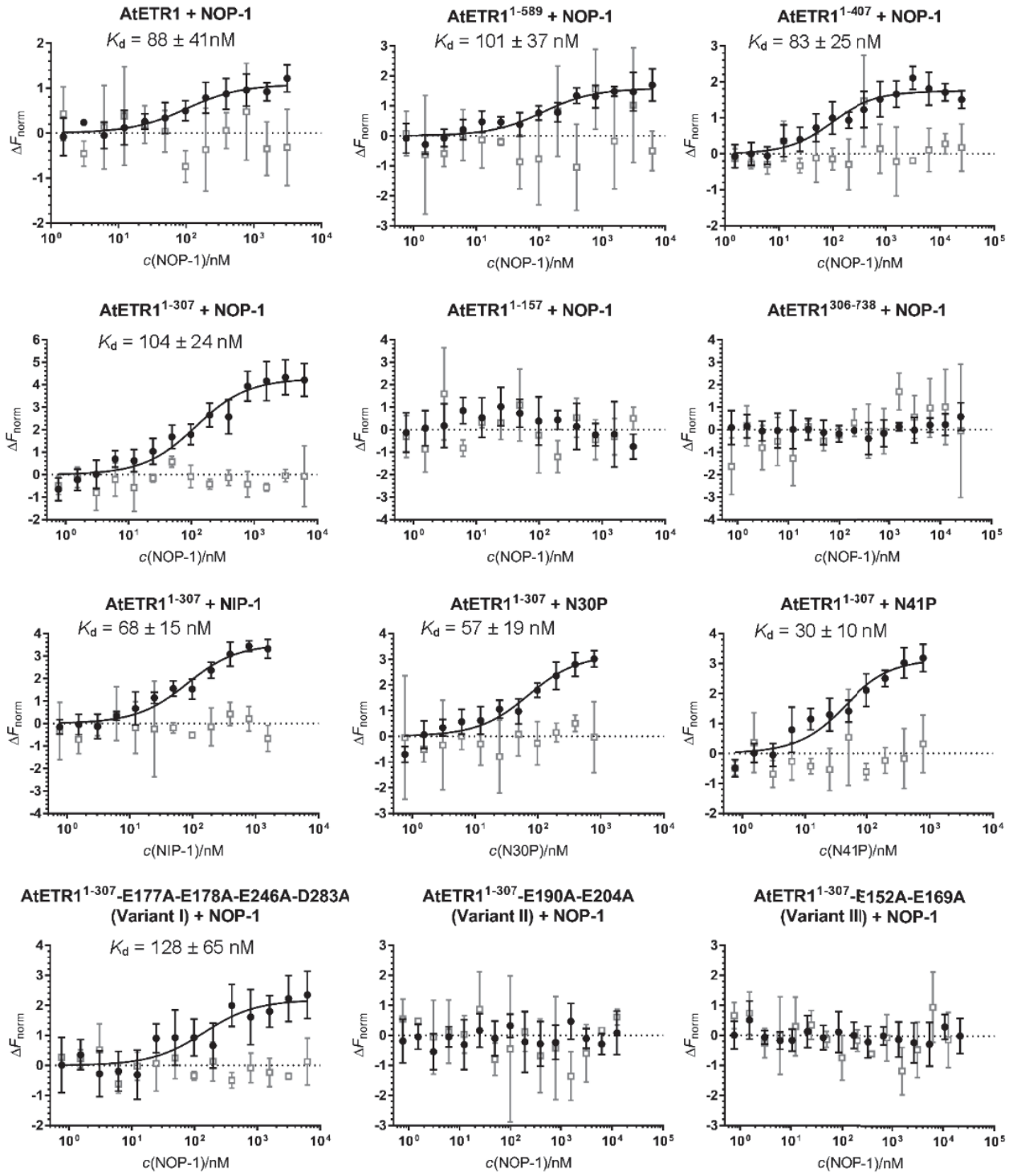
[3]John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC) & Institute for Complex Systems - Structural Biochemistry (ICS 6), Forschungszentrum Jülich GmbH, Jülich, Germany

[†]D.M. and M.D. contributed equally to this work.
[‡]D.M. present address: Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Vienna Biocenter, Vienna, Austria
[§]M.D. present address: Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California, USA

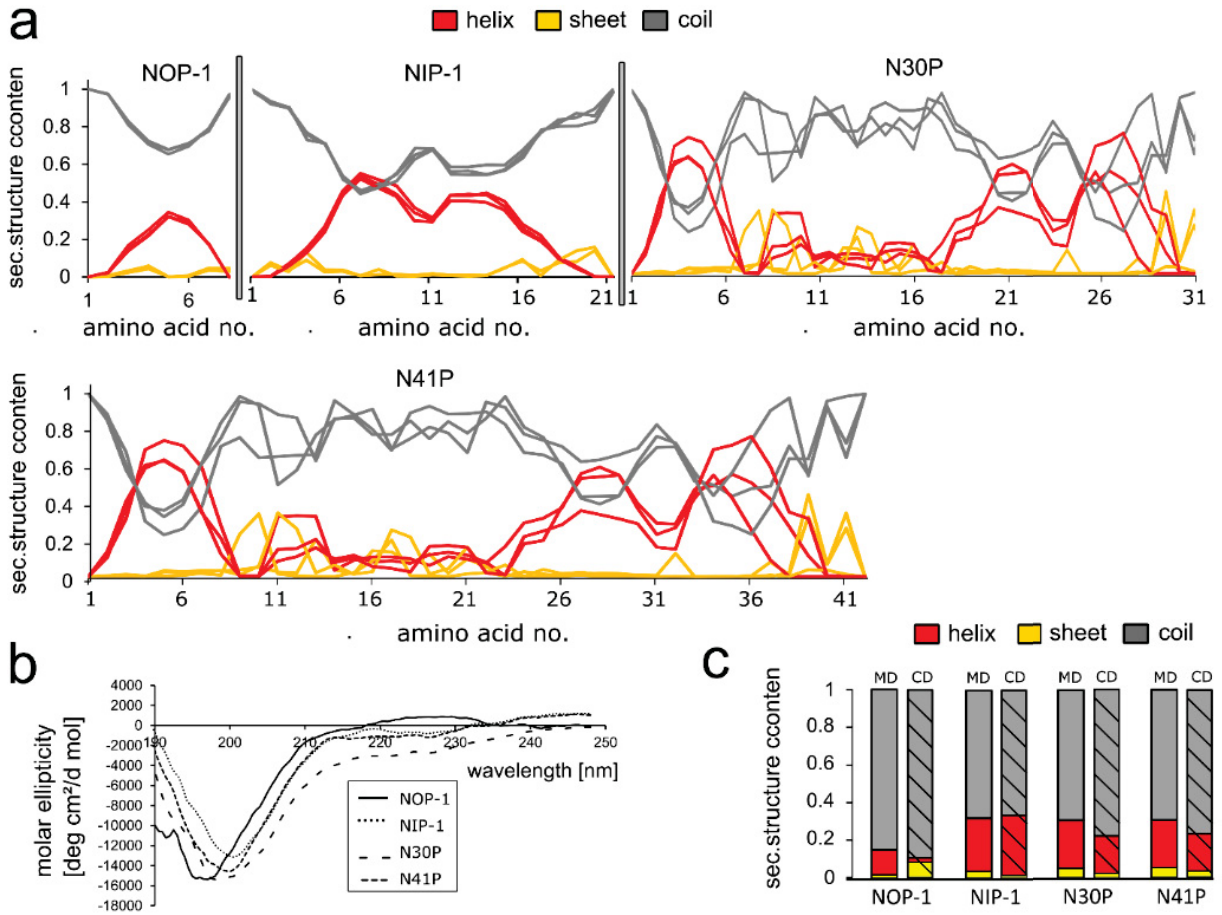
*Correspondence should be addressed to G.G., e-mail: georg.groth@hhu.de or H.G., e-mail: gohlke@hhu.de, h.gohlke@fz-juelich.de



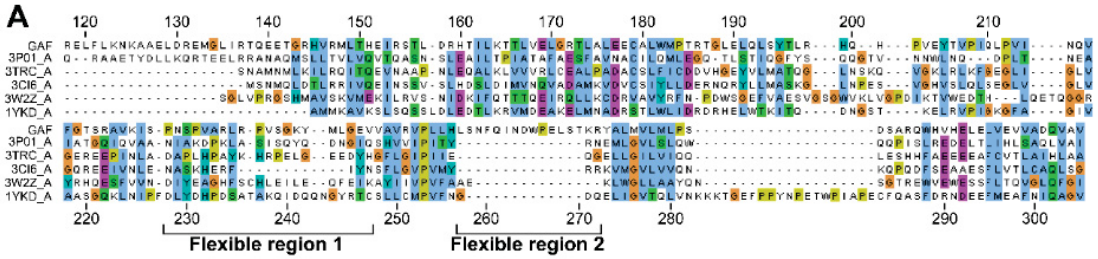
Supplementary Figure S1 | SDS-PAGE of the truncated His-tagged AtETR1 constructs. (**a**) Coomassie-stained gels. (**b**) Western blotting with the anti-His antibody. Molecular weight of each construct is given in parentheses. The gel bands below the dye front are artefacts due to fos-choline-16 – a detergent used in purification.



Supplementary Figure S2 | Microscale thermophoresis (MST) interaction studies of the fluorescently labelled AtETR1 mutants with peptides NOP-1, NIP-1, N30P, and N41P. Relative normalized fluorescence (ΔF_{norm}; ●) was fitted to the one-binding-site model, and the corresponding K_d value is given for each binding curve. Data for the chemically denatured proteins (□) are given for comparison.



Supplementary Figure S3 | Secondary structure of the four EIN2-derived peptides. (**a**) Residue-wise secondary structure prediction based on three MD simulations for each system in implicit solvent. The secondary structure content was calculated by DSSP [1] as an average over all snapshots in 50 µs of MD simulations. (**b**) Far UV-CD scan for NOP-1, NIP-1, N30P and N41P. (**c**) Comparison of secondary structure contents predicted by MD simulations (mean over all residues and three MD simulations) and computed from CD data (using the K2D2 web server[2]).



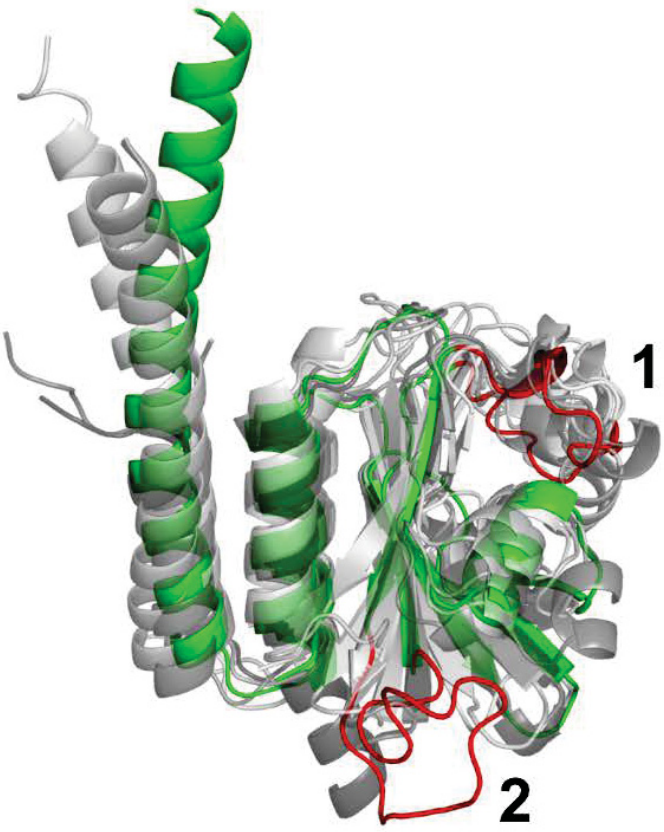
Supplementary Figure S4 | Sequence alignment between the AtETR1 GAF domain (residues 118–304) and five template sequences used to predict the structure. The flexible regions 1 (residues 228–247) and 2 (residues 257–272) with the highest degree of inaccuracy in the final GAF domain model are indicated in brackets. The alignment was calculated using the in-house meta-alignment tool TopAligner as a consensus between alignments calculated by eight different state-of-the-art multiple alignment programs. The alignment programs used are given in **Supplementary Table 1**.

Supplementary Table S1 | TopModel methods used for threading, alignment and model quality assessment of the GAF domain model

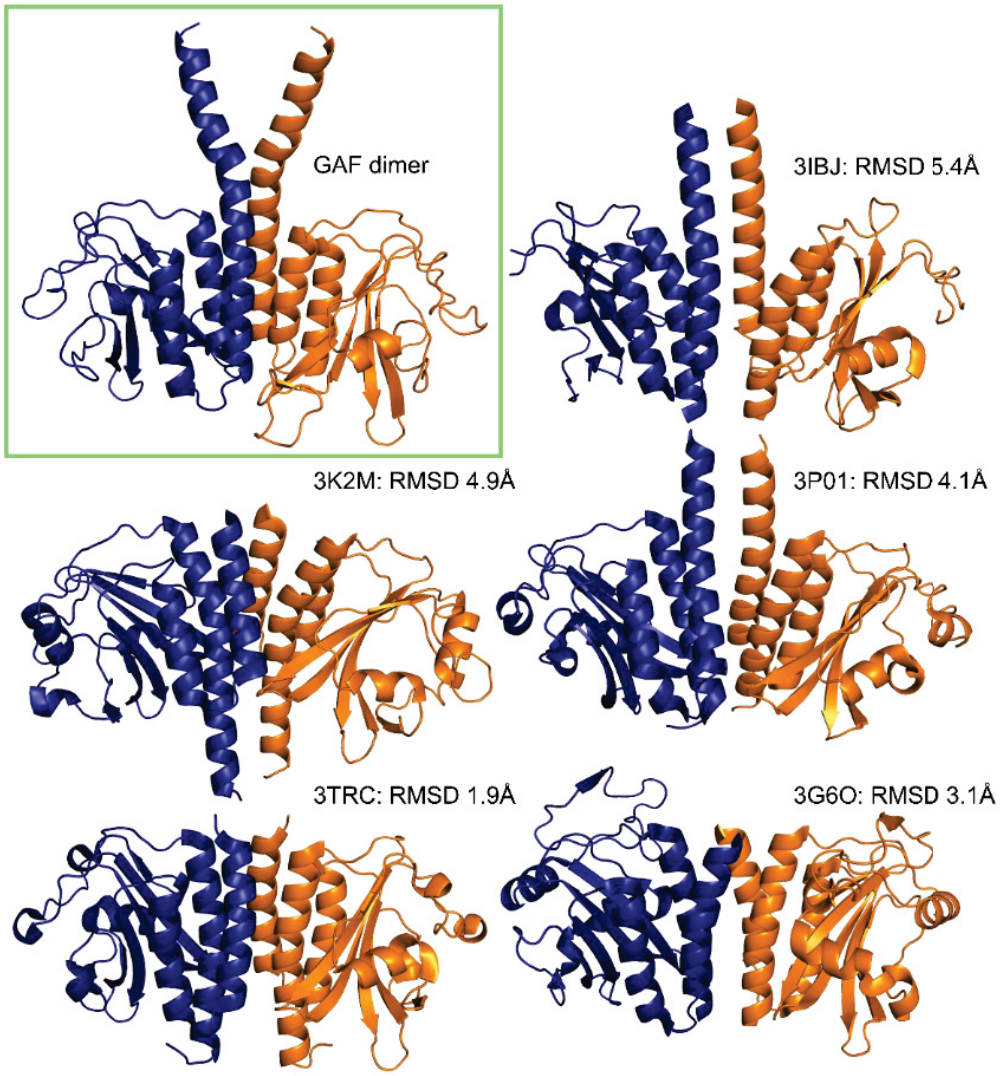
Threading	Alignment	Quality Assessment
DeltaBLAST [3]	TCOFFEE [4] [1]	PROCHECK [5]
HMMER3 [6]	MAFFT7 [7]	MolProbity [8]
HHBlits [9]	MergeAlign2 [10]	ANOLEA [11]
HHSearch [12]	SAlign [13]	ProSA2003 [14]
FFAS03 [15]	PROMALS3D [16]	DOPE [17]
SPARKSX [18]	FORMATT [19]	GOAP [20]
RAPTORX [21]	MUSTANG [22]	ModFOLDClust2 [23]
LOMETS [24] [2]	3DCOMB [25]	PCONS [26]
pGenThreader [27]		SPICKER [28]
pDomThreader [27]		QMEAN6 [29]
FASTA [30]		PROQ2 [31]
SAMT2K [32]		SELECTPRO [33]

[1] The following programs are used within the TCOFFEE suite as the default methods for calculating alignments: ClustalW [34], POA [35], MUSCLE [36], ProbA [37], PCMA [38], ProbCons [39], DiAlign [40], SAP [41], and TM-Align [42].

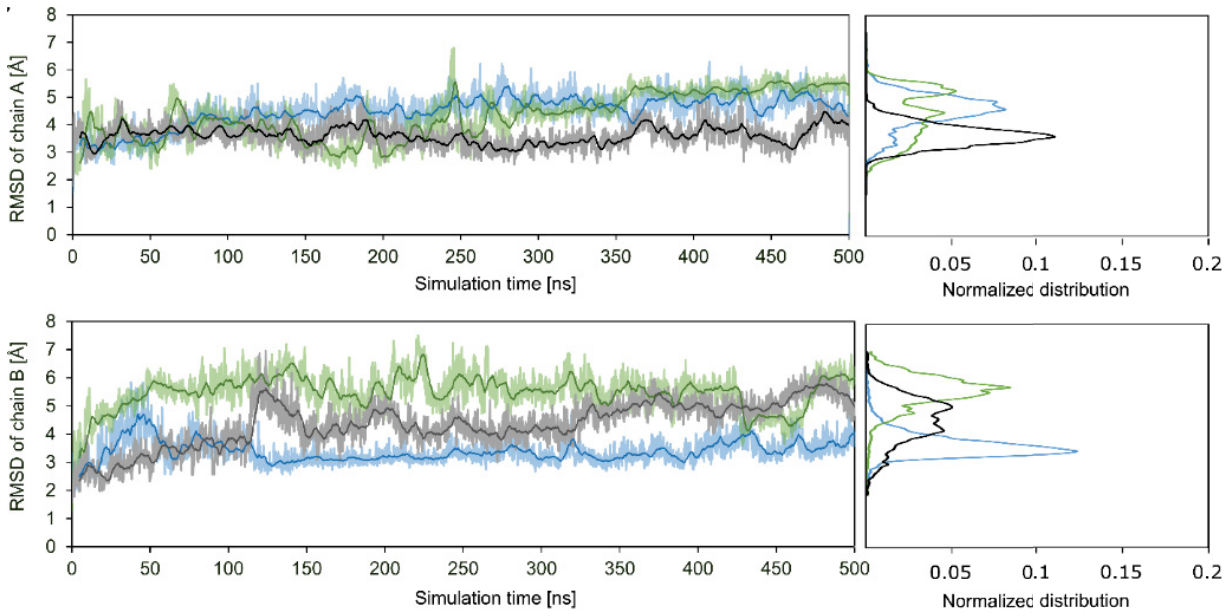
[2] The LOMETS software includes the algorithms PPAS, wPPAS, dPPAS, wdPPAS, PPAS2, dPPAS2, Env-PPAS, MUSTER, and wMUSTER.



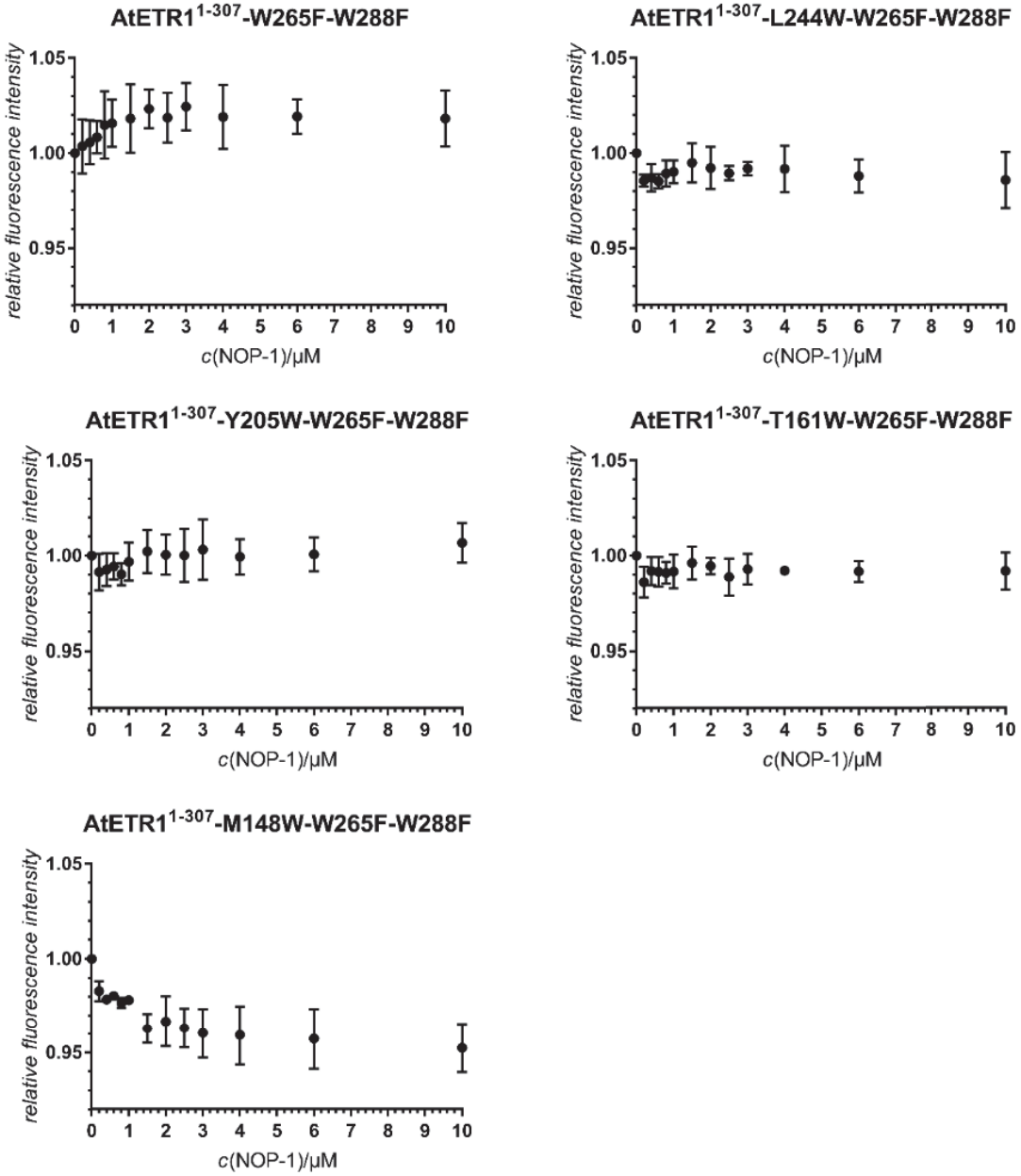
Supplementary Figure S5 | Structural alignment between the final GAF domain model and the five templates used to predict the structure. The flexible regions 1 (residues 228–247) and 2 (residues 257–272) with the highest degree of inaccuracy are indicated in red.



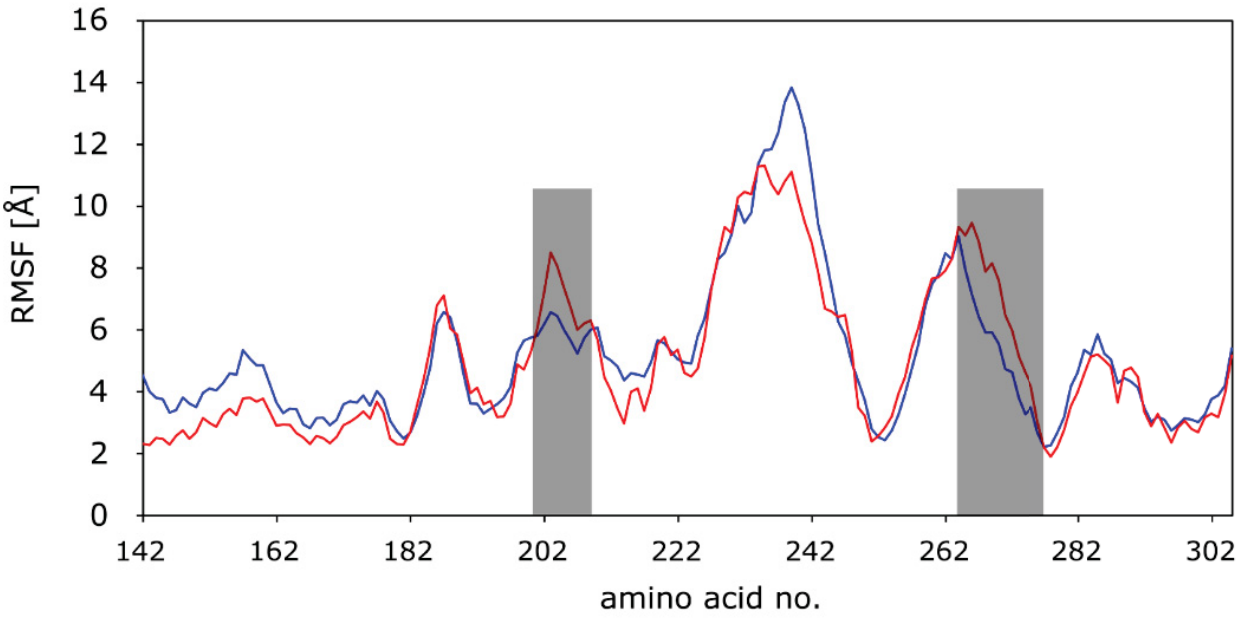
Supplementary Figure S6 | The homologous interfaces identified by TopDock and used to calculate protein–protein contacts for guided protein–protein docking. For each interface, the C_α RMSD to the GAF docking solution is shown. All five interfaces show similar folds and interaction patterns despite low sequence identity to the GAF domain.



Supplementary Figure S7 | Root mean square deviations (RMSD) along three independent, unbiased MD simulations of the dimeric GAF domain of AtETR1 (amino acids 118 to 305) without ligand. RMSD values were calculated separately for chains A (top) and B (bottom) with respect to the starting structure, excluding unstructured loop regions (residues 222–290). The histograms to the right depict the frequency distributions of RMSD values within the single trajectories.



Supplementary Figure S8 | Binding of NOP-1 to AtETR1$^{1–307}$ Trp variants monitored by intrinsic tryptophan fluorescence. Excitation and emission wavelengths were 295 nm and 344 nm, respectively. Protein concentration was 1.0 µM.



Supplementary Figure S9 | Root mean square fluctuations (RMSF) of the dimeric GAF domain of AtETR1 with (blue) and without (red) NOP-1 derived from MD simulations. The average RMSF value of chains A and B is depicted. Regions of residues with lower RMSF in the presence of NOP-1 are marked by grey boxes. These regions coincide with those identified by the rigidity and flexibility analysis (CNA; Fig. 4c,d in the main text) as having an increased structural stability upon NOP-1 binding.

Supplementary Table S2 | Molecular cloning scheme

Target plasmid	Template plasmid	Primer pair or *gene fragment**	
		Fragment 1	Fragment 2
pTEV-16b-AtETR1$^{1-589}$	pTEV-16b-AtETR1	ColE1-F AtETR1-589-R	AtETR1-STOP-F ColE1-R
pTEV-16b-AtETR1$^{1-407}$	pTEV-16b-AtETR1	ColE1-F AtETR1-407-R	AtETR1-STOP-F ColE1-R
pTEV-16b-AtETR1$^{1-307}$	pTEV-16b-AtETR1	ColE1-F AtETR1-307-R	AtETR1-STOP-F ColE1-R
pTEV-16b-AtETR1$^{1-157}$	pTEV-16b-AtETR1	ColE1-F AtETR1-157-R	AtETR1-STOP-F ColE1-R
pTEV-16b-AtETR1$^{306-738}$	pTEV-16b-AtETR1	ColE1-F pET16b-TEV-R	AtETR1-306-F ColE1-R
pTEV-16b-AtETR1$^{1-307}$- E177A-E178A-E246A-D283A	pTEV-16b-AtETR1$^{1-307}$	1_F_20170417 1_R_20170502	*A_20170417* *(gene fragment)*
pTEV-16b-AtETR1$^{1-307}$- E152A-E169A	pTEV-16b-AtETR1$^{1-307}$	2_F_20170417 2_R_20170502	*B_20170417* *(gene fragment)*
pTEV-16b-AtETR1$^{1-307}$- E190A-E204A	pTEV-16b-AtETR1$^{1-307}$	2_F_20170417 2_R_20170417	*C_20170417* *(gene fragment)*
pTEV-16b-AtETR1$^{1-307}$- W265F-W288F	pTEV-16b-AtETR1$^{1-307}$	F_for_20170614 F_rev_20170614	*W265F-W288F* *(gene fragment)*
pTEV-16b-AtETR1$^{1-307}$- L244W-W265F-W288F	pTEV-16b-AtETR1$^{1-307}$- W265F-W288F	ColE1-F 244_R	244_R ColE1-R
pTEV-16b-AtETR1$^{1-307}$- Y205W-W265F-W288F	pTEV-16b-AtETR1$^{1-307}$- W265F-W288F	ColE1-F 205_R	205_F ColE1-R
pTEV-16b-AtETR1$^{1-307}$- M148W-W265F-W288F	pTEV-16b-AtETR1$^{1-307}$- W265F-W288F	ColE1-F 148_F	148_R ColE1-R
pTEV-16b-AtETR1$^{1-307}$- T161W-W265F-W288F	pTEV-16b-AtETR1$^{1-307}$- W265F-W288F	ColE1-F 161_R	161_F ColE1-R

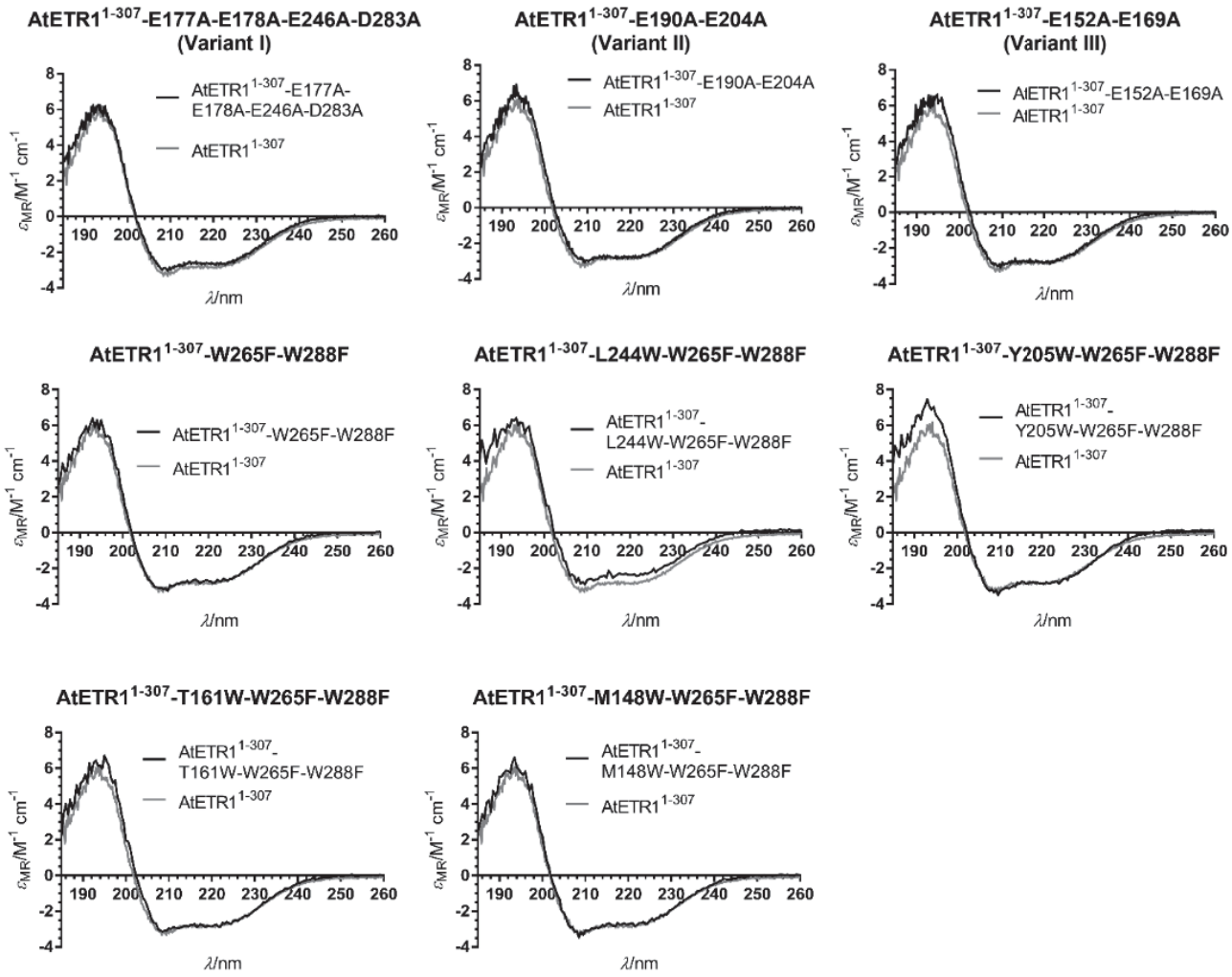
* Each target plasmid was assembled in a Gibson reaction from the two fragments. Fragments were either purchased as a synthetic double-stranded DNA (gBlocks gene fragments from Integrated DNA Technologies) or amplified in a PCR using an indicated pair of primers.

Supplementary Table S3 | Primers used in the molecular cloning

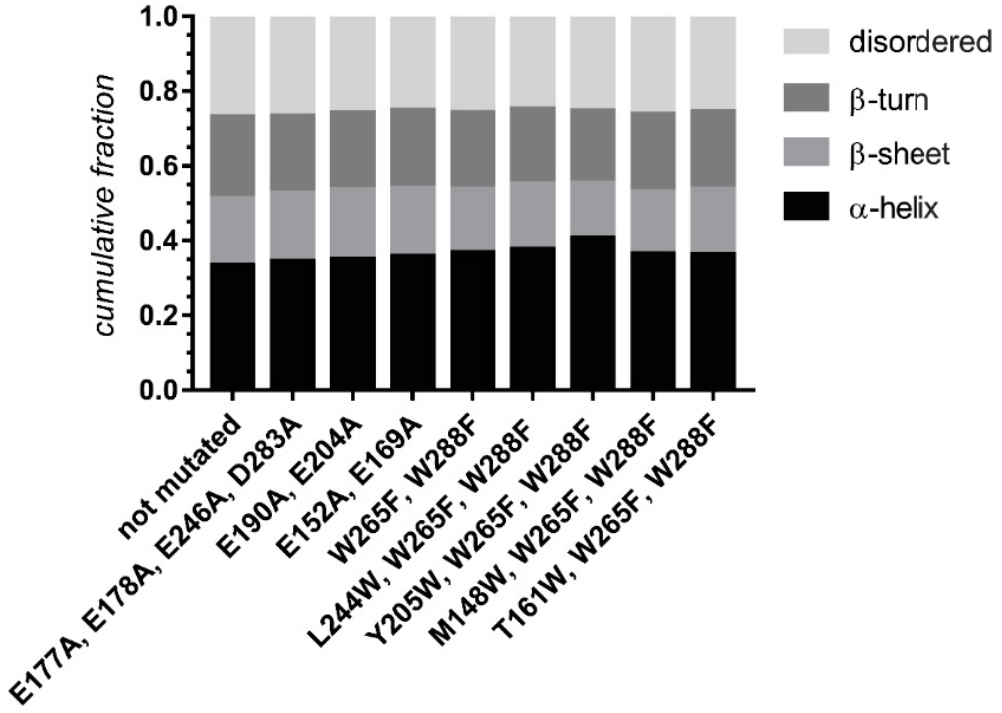
Primer	Sequence (5' → 3')
ColE1-F	GGAGCGAACGACCTACACCGAACTGAGATACCTACAGCG
ColE1-R	CGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCC
AtETR1-STOP-F	TAAGGATCCGGCTGCTAACAAAGCCCGAAAG
AtETR1-589-R	TTGTTAGCAGCCGGATCCTTATTCGTTTGAACGTTCTGAGATCCCAAGTTTAAC
AtETR1-407-R	TTGTTAGCAGCCGGATCCTTAATCTTCTAACCTTGAAAGATCTAAGACATCATTCATC
AtETR1-307-R	TTGTTAGCAGCCGGATCCTTAGAGAGCTACAGCCACCTGATCAGC
AtETR1-157-R	TTGTTAGCAGCCGGATCCTTATAAAGTGCTTCTAATCTCATGAGTCAACATTCTCAC
pET16b-TEV-R	ATGTCCCTGAAAATACAGGTTTTCATGGCCGCTG
AtETR1-306-F	AACCTGTATTTTCAGGGACATGCTCTCTCACATGCTGCGATCCTAG
1_F_20170417	AGTGCAAGGCAATGGCATGTCCATGAG
1_R_20170502	CAAAGCTAATGTCCTACCAAGCTCAACAAG
2_F_20170417	TATACGGTTCCTATTCAATTACCGGTGATTAACC
2_R_20170502	ATGAGTCAACATTCTCACATGCCTTCCGGTTTC
2_R_20170417	GAAACCGGAAGGCATGTGAGAATGTTGACTCAT
2_F_20170417	TATACGGTTCCTATTCAATTACCGGTGATTAACC
F_for_20170614	GTCATTAATCTGAAAATTAGAAAGGTGGAGAAGCGGAAC
F_rev_20170614	TAAGGATCCGGCTGCTAACAAAGCCCGAAAG
244_F	ATGTGGGGGGAGGTGGTCGCTGTG
244_R	CCTCCCCCCACATATATTTCCCAGAAACAGG
205_F	CCCGTGGAGTGGACGGTTCCTATTCAATTAC
205_R	GGAACCGTCCACTCCACGGGATGTTG
148_F	GGCATGTGAGATGGTTGACTCATGAGATTAG
148_R	CCATCTCACATGCCTTCCGGTTTCTTCC
161_F	TGGATTTTAAAGACTACACTTGTTGAGCTTGGTAGGAC
161_R	GTCTTTAAAATCCAATGTCTATCTAAAGTGCTTCTAATCTCATG

Supplementary Table S4 | Gene fragments used in the molecular cloning

Gene fragment	Sequence (5'→3')
A_20170417	CTTGGTAGGACATTAGCTTTGGCGGCGTGTGCATTGTGGATGCCTACTAGAACTGGGTTA GAGCTACAGCTTTCTTATACACTTCGTCATCAACATCCCGTGGAGTATACGGTTCCTATTCA ATTACCGGTGATTAACCAAGTGTTTGGTACTAGTAGGGCTGTAAAAATATCTCCTAATTCT CCTGTGGCTAGGTTGAGACCTGTTTCTGGGAAATATATGCTAGGGGCGGTGGTCGCTGT GAGGGTTCCGCTTCTCCACCTTTCTAATTTTCAGATTAATGACTGGCCTGAGCTTTCAACAA AGAGATATGCTTTGATGGTTTTGATGCTTCCTTCAGCTAGTGCAAGGCAATGGCATGTC
B_20170417	CATGTGAGAATGTTGACTCATGCGATTAGAAGCACTTTAGATAGACATACTATTTTAAAG ACTACACTTGTTGCGCTTGGTAGGACATTAGCTTTGGAGGAGTGTGCATTGTGGATGCCT ACTAGAACTGGGTTAGAGCTACAGCTTTCTTATACACTTCGTCATCAACATCCCGTGGAGT ATACGGTTCCTATTCAATTA
C_20170417	CATGTGAGAATGTTGACTCATGAGATTAGAAGCACTTTAGATAGACATACTATTTTAAAG ACTACACTTGTTGAGCTTGGTAGGACATTAGCTTTGGAGGAGTGTGCATTGTGGATGCCT ACTAGAACTGGGTTAGCGCTACAGCTTTCTTATACACTTCGTCATCAACATCCCGTGGCGT ATACGGTTCCTATTCAATTA
W265F-W288F	TCTAATTTTCAGATTAATGACTTTCCTGAGCTTTCAACAAAGAGATATGCTTTGATGGTTTT GATGCTTCCTTCAGATAGTGCAAGGCAATTCCATGTCCATGAGTTGGAACTCGTTGAAGT CGTCGCTGATCAGGTGGCTGTAGCTCTCTAAGGATCCGGCTGCTAACAA



Supplementary Figure S10 | CD spectra of the AtETR1$^{1–307}$ variants. The spectrum of each variant (black curve) is compared with that of the unchanged AtETR1$^{1-307}$ (grey curve).



Supplementary Figure S11 | Secondary structure content of the not mutated AtETR1$^{1–307}$ and its variants. The plotted fractions are averaged values determined from the CD spectra by three different methods (as described in the Materials and Methods).

Supplementary References

1 Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637, doi:10.1002/bip.360221211 (1983).

2 Perez-Iratxeta, C. & Andrade-Navarro, M. A. K2D2: estimation of protein secondary structure from circular dichroism spectra. *BMC Struct Biol* **8**, 25, doi:10.1186/1472-6807-8-25 (2008).

3 Boratyn, G. M. *et al.* Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7**, 12, doi:10.1186/1745-6150-7-12 (2012).

4 O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G. & Notredame, C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.* **340**, 385-395, doi:10.1016/j.jmp.2004.04.056 (2004).

5 Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283-291, doi:Doi 10.1107/S0021889892009944 (1993).

6 Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

7 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

8 Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **66**, 12-21, doi:10.1107/S0907444909042073 (2009).

9 Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173-175, doi:10.1038/Nmeth.1818 (2012).

10 Collingridge, P. W. & Kelly, S. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC Bioinformatics* **13**, 117, doi:10.1186/1471-2105-13-117 (2012).

11 Melo, F. & Feytmans, E. Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.* **267**, 207-222, doi:DOI 10.1006/jmbi.1996.0868 (1997).

12 Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244-W248, doi:10.1093/nar/gki408 (2005).

13 Madhusudhan, M., Webb, B. M., Marti-Renom, M. A., Eswar, N. & Sali, A. Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des. Sel.* **22**, 569-574, doi:10.1093/protein/gzp040 (2009).

14 Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Genetics* **17**, 355-362, doi:10.1002/prot.340170404 (1993).

15 Rychlewski, L., Li, W., Jaroszewski, L. & Godzik, A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9**, 232-241, doi:10.1110/ps.9.2.232 (2000).

16 Pei, J., Kim, B.-H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295-2300, doi:10.1093/nar/gkn072 (2008).

17 Shen, M. y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507-2524, doi:10.1110/ps.062416606 (2006).

18 Yang, Y., Faraggi, E., Zhao, H. & Zhou, Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional

structural properties of query and corresponding native properties of templates. *Bioinformatics* **27**, 2076-2082, doi:10.1093/bioinformatics/btr350 (2011).

19 Daniels, N. M., Nadimpalli, S. & Cowen, L. J. Formatt: Correcting protein multiple structural alignments by incorporating sequence alignment. *BMC Bioinformatics* **13**, 259, doi:10.1186/1471-2105-13-259 (2012).

20 Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**, 2714-2726, doi:10.1110/ps.0217002 (2002).

21 Peng, J. & Xu, J. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics* **79**, 161-171, doi:10.1002/prot.23175 (2011).

22 Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. & Lesk, A. M. MUSTANG: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics* **64**, 559-574, doi:10.1002/prot.20921 (2006).

23 McGuffin, L. J. & Roche, D. B. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics* **26**, 182-188, doi:10.1093/bioinformatics/btp629 (2010).

24 Wu, S. & Zhang, Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* **35**, 3375-3382, doi:10.1093/nar/gkm251 (2007).

25 Wang, S., Peng, J. & Xu, J. Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. *Bioinformatics* **27**, 2537-2545, doi:10.1093/bioinformatics/btr432 (2011).

26 Wallner, B. & Elofsson, A. Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci.* **15**, 900-913, doi:10.1110/ps.051799606 (2006).

27 Lobley, A., Sadowski, M. I. & Jones, D. T. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* **25**, 1761-1767, doi:10.1093/bioinformatics/btp302 (2009).

28 Zhang, Y. & Skolnick, J. SPICKER: A clustering approach to identify near-native protein folds. *J. Comput. Chem.* **25**, 865-871, doi:10.1002/jcc.20011 (2004).

29 Benkert, P., Schwede, T. & Tosatto, S. C. QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. *BMC Struct. Biol.* **9**, 35, doi:10.1186/1472-6807-9-35 (2009).

30 Pearson, W. R. Finding protein and nucleotide similarities with FASTA. *Current protocols in bioinformatics* **Chapter 3**, 3.9. 1-3.9. 25, doi:10.1002/0471250953.bi0309s04 (2004).

31 Ray, A., Lindahl, E. & Wallner, B. Improved model quality assessment using ProQ2. *BMC Bioinformatics* **13**, 224, doi:10.1186/1471-2105-13-224 (2012).

32 Karplus, K. SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res.* **37**, gkp403, doi:10.1093/nar/gkp403 (2009).

33 Randall, A. & Baldi, P. SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *BMC Struct. Biol.* **8**, 52, doi:10.1186/1472-6807-8-52 (2008).

34 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680, doi:10.1093/nar/22.22.4673 (1994).

35 Lee, C., Grasso, C. & Sharlow, M. F. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**, 452-464, doi:10.1093/bioinformatics/18.3.452 (2002).

36 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).
37 Sierk, M. L., Smoot, M. E., Bass, E. J. & Pearson, W. R. Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. *BMC Bioinformatics* **11**, 146, doi:10.1186/1471-2105-11-146 (2010).
38 Pei, J., Sadreyev, R. & Grishin, N. V. PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics* **19**, 427-428, doi:10.1093/bioinformatics/btg008 (2003).
39 Do, C. B., Mahabhashyam, M. S., Brudno, M. & Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15**, 330-340, doi:10.1101/gr.2821705 (2005).
40 Al Ait, L., Yamak, Z. & Morgenstern, B. DIALIGN at GOBICS—multiple sequence alignment using various sources of external information. *Nucleic Acids Res.* **41**, W3-W7, doi:10.1093/nar/gkt283 (2013).
41 Taylor, W. R. Protein structure comparison using iterated double dynamic programming. *Protein Sci.* **8**, 654-665, doi:10.1110/ps.8.3.654 (1999).
42 Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302-2309, doi:10.1093/nar/gki524 (2005).

22. CURRICULUM VITAE

Name: Daniel Mulnaes

Date of birth: 26/12/1987 in Gladsaxe (Denmark)

Education

2018-2020 **Post Doc., Research assistant**

Group of Computational Pharmaceutical Chemistry and Molecular Bioinformatics at the Heinrich-Heine University, Düsseldorf.

Project: *Hybrid approach to predict super-tertiary and quaternary structure of proteins and protein complexes in cells*

Supervisor: Prof. Dr. Holger Gohlke

2013-2018 **PhD student and research assistant**

Group of Computational Pharmaceutical Chemistry and Molecular Bioinformatics at the Heinrich-Heine University, Düsseldorf.

Project: *TopSuite: A meta-suite for protein structure prediction using deep neural networks*

Supervisor: Prof. Dr. Holger Gohlke

2010-2012 **Master of Science in Bioinformatics**

Structural Bioinformatics group, Bio-Centre, University of Copenhagen.

Project: *Modelling pH dependent conformational fluctuations in proteins*

Final Grade: 12/12 (100%)

2007-2010 **Bachelor of Science in Biophysics**

Bio-membrane research centre, August Krogh Institute, University of Copenhagen.

Project: *Menkes Disease. A study and review of 15 missense mutations in ATP7A causing Menkes Disease*

Final Grade: 12/12 (100%)

2004-2007 **High school degree**

Greve Gymnasium, Greve (Denmark).

Final grade average: 10.7/11 (97%)

1995-2004 **Private School**

Greve Privat Skole, Hundige (Denmark).

Final grade average 10.7/11 (97%)

Work experience and internships

2012-2012 **Software Developer**

Novozymes A/S, Brudelysvej 32, Bagsvaerd (Denmark)

Supervisor: Dr. Jens Erik Nielsen

Project: *Implementation of high-throughput molecular dynamics simulation and analysis pipeline for industrial biotechnology*.

Teaching Experience

2013-2019 Teaching Assistant

Teaching and supervising pharmacy students within the Inorganic Chemistry practical course

Heinrich Heine University, Düsseldorf

Additional skills

Computer skills Operating Systems: Linux, Microsoft Windows

Programming skills Python, R, bash, Tensorflow, KERAS

Structure Evaluation PROCHECK, MolProbity, ANOLEA; ProSA, DOPE, GOAP, SPICKER, ModFoldClust, PCONS, PROQ, PROQ3D, SVMQA, QMEAN, SELECTpro

Sequence Alignment TCOFFEE, PROMALS3D, SAlign, FORMATT, 3DCOMB, MUSTANG, TMAlign, SAP, DiAlign, POA, ProbA, ProbCons, MUSCLE, ClustalW, FAST, PCMA, MAFFT7, kAlign, CDHit

Threading	LOMETS, RAPTOR-X, SPARKS-X, HHSUITE, BLAST, SAMT2K, FFAS03, HMMER3, FASTA, pGenThreader, pDomThreader
Structure Prediction	Modeller, ROSETTA, CONFOLD, I-TASSER, CNS
Feature Prediction	PSICOV, EVFOLD, CCMPred, MetaPSICOV, NebCon, NNCON, SVMCON, DeepCov, DNCON2, PCONSC4, COLORS, SPOT, CMAPpro, BETApro, FragHMMent, SPIDER3, MuFoldSS, DeepCNF_SS, DeepCNF_D, PSIPRED, SOLVPRED, SSPro5, SCRATCH, AcconPred, NetSurfP, SANN, SCAMPI, PHILIUS, PHOBIUS, POLYPHOBIUS, OCTOPUS, SPOCTOPUS, TMHMM, HMMTOP, TOPCONS, MEMSAT3, PROTEUS, BOCTOPUS, BETAWARE, LIPS, RYTHM, COILS2, SIGNALP, GlobPlot, TRUST, DeepCoil, T-REKS, HHrep, ScoobyDomain, PPRODO, DOMPRO, DROP, DOBO, DOMCUT, DeepDom, ConDo, FIEFDom, ThreaDom, DOMPRED, InterProScan5
Protein Docking	HADDOCK, FRODOCK, JET
Ligand Docking	Autodock Vina, IONCOM, CAVER, OpenBabel, PocketAnalyzer, FunFold3, P2Rank, CONCAVITY, SiteHound, eFindSite, fPocket, LigSiteCSC
Others:	Microsoft Office, Open Office, Inkscape.

Advanced Training

Interdiciplinary Graduate and Research Academy (iGRAD), Düsseldorf

- Good Scientific Practise for Doctoral Researchers
- Presenting (in) Science - How to own the stage on (international) conferences
- Optimizing Writing Strategies for Publishing Research in English (for CEPLAS)

CLIB Graduate Cluster, Düsseldorf

- Introduction to Project- and Innovation Management and Patent Law

Publications

I) Daniel Mulnaes, and Holger Gohlke. TopScore: Using deep neural networks and large diverse datasets for accurate protein model quality assessment. *Journal of Chemical Theory and Computation,* 2018, 14, 6117-6126.

II) Daniel Mulnaes, Nicola Porta, Rebecca Clemens, Irina Apanasenko, Jens Reiners, Lothar Gremer, Philipp Neudecker, Sander Smits, Holger Gohlke. TopModel: A deep neural network and model quality driven meta-approach to template-based protein structure prediction. *Journal of Chemical Theory and Computation,* 2019, Submitted.

III) Holger Gohlke, Ulrike Hergert, Tatu Meyer, Daniel Mulnaes, Manfred K. Grieshaber, Sander H.J. Smits and Lutz Schmitt. Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion. *Journal of Chemical Information and Modelling*, 2013, 53, 2493–2498.

IV) Zeli Zhang, Qinyong Gu, Ananda Ayyappan Jaguva Vasudevan, Anika Hain, Björn-Philipp Kloke, Sascha Hasheminasab, Daniel Mulnaes, Kei Sato, Klaus Cichutek, Dieter Häussinger, Ignatio G. Bravo, Sander H.J. Smits, Holger Gohlke and Carsten Münk. Determinants of FIV and HIV Vif sensitivity of feline APOBEC3 restriction factors. *Retrovirology*, 2016, 13, 46.

V) Dalibor Milić, Markus Dick, Daniel Mulnaes, Christopher Pfleger, Anna Kinnen, Holger Gohlke and Georg Groth. Recognition motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1. *Nature Scientific Reports,* 2018, 8, 3890.

VI) Nils Widderich, Marco Pittelkow, Astrid Höppner, Daniel Mulnaes, Wolfgang Buckel, Holger Gohlke, Sander H.J. Smits, Erhard Bremer. Molecular dynamics simulations and structure-guided mutagenesis provide insight into the architecture of the catalytic core of the ectoine hydroxylase. *Journal of Chemical Information and Modelling*, 2013, 53, 2493–2498.

VII) Sakshi Khosa, Benedikt Frieg, Daniel Mulnaes, Diana Kleinschrodt, Astrid Höppner, Holger Gohlke, Sander H.J. Smits. Structural basis of lantibiotic recognition by the nisin resistance protein from Streptococcus agalactiae. *Nature Scientific Reports* 2016, 6, 18679.

VIII) Prakash Chandra Rathi, Daniel Mulnaes and Holger Gohlke. VisualCNA: a GUI for interactive Constraint Network Analysis and protein engineering for improving thermostability. *Bioinformatics,* 2015, 31, 2394–2396

23. REFERENCES

1. Rathi, P. C.; HöFfken, H. W.; Gohlke, H., Quality matters: Extension of clusters of residues with good hydrophobic contacts stabilize (hyper) thermophilic proteins. *J. Chem. Inf. Model.* **2014,** *54* (2), 355-361.
2. Widderich, N.; Pittelkow, M.; Höppner, A.; Mulnaes, D.; Buckel, W.; Gohlke, H.; Smits, S. H.; Bremer, E., Molecular dynamics simulations and structure-guided mutagenesis provide insight into the architecture of the catalytic core of the ectoine hydroxylase. *J. Mol. Biol.* **2014,** *426* (3), 586-600.
3. Ingles-Prieto, A.; Ibarra-Molero, B.; Delgado-Delgado, A.; Perez-Jimenez, R.; Fernandez, J. M.; Gaucher, E. A.; Sanchez-Ruiz, J. M.; Gavira, J. A., Conservation of protein structure over four billion years. *Structure* **2013,** *21* (9), 1690-1697.
4. Gohlke, H.; Hergert, U.; Meyer, T.; Mulnaes, D.; Grieshaber, M. K.; Smits, S. H.; Schmitt, L., Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion. *J. Chem. Inf. Model.* **2013,** *53* (10), 2493-2498.
5. Yang, J.; Roy, A.; Zhang, Y., Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, btt447.
6. Janin, J., Assessing predictions of protein–protein interaction: the CAPRI experiment. *Protein Science* **2005,** *14* (2), 278-283.
7. Cheng, Y.; Grigorieff, N.; Penczek, P. A.; Walz, T., A primer to single-particle cryo-electron microscopy. *Cell* **2015,** *161* (3), 438-449.
8. Cressey, D.; Callaway, E., Cryo-electron microscopy wins chemistry Nobel. *Nature News* **2017,** *550* (7675), 167.
9. Aehle, W.; Sobek, H.; Amory, A.; Vetter, R.; Wilke, D.; Schomburg, D., Rational protein engineering and industrial application: Structure prediction by homology and rational design of protein-variants with improved 'washing performance'—the alkaline protease from Bacillus alcalophilus. *Journal of Biotechnology* **1993,** *28* (1), 31-40.
10. Cavasotto, C. N.; Phatak, S. S., Homology modeling in drug discovery: current trends and applications. *Drug discovery today* **2009,** *14* (13), 676-683.
11. Roy, A.; Yang, J.; Zhang, Y., COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **2012**, gks372.
12. Roche, D. B.; Buenavista, M. T.; Mcguffin, L. J., The FunFOLD2 server for the prediction of protein–ligand interactions. *Nucleic Acids Res.* **2013,** *41* (W1), W303-W307.
13. Zhang, Y., Protein structure prediction: when is it useful? *Current Opinion in Structural Biology* **2009,** *19* (2), 145-155.
14. Petsko, G. A., The grail problem. *Genome Biol.* **2000,** *1* (1), comment002. 001.
15. Jones, D. T.; Singh, T.; Kosciolek, T.; Tetchner, S., MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **2014,** *31* (7), 999-1006.
16. Wallner, B.; Larsson, P.; Elofsson, A., Pcons. net: protein structure prediction meta server. *Nucleic Acids Res.* **2007,** *35* (suppl_2), W369-W374.
17. Wu, S.; Zhang, Y., LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* **2007,** *35* (10), 3375-3382.
18. Pawlowski, M.; Gajda, M. J.; Matlak, R.; Bujnicki, J. M., MetaMQAP: a meta-server for the quality assessment of protein models. *BMC Bioinformatics* **2008,** *9* (1), 403.

19. Necci, M.; Piovesan, D.; Dosztányi, Z.; Tosatto, S. C., MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* **2017,** *33* (9), 1402-1404.
20. Bernsel, A.; Viklund, H.; Hennerdal, A.; Elofsson, A., TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* **2009,** *37* (suppl_2), W465-W468.
21. De Vries, S. J.; Bonvin, A. M., CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One* **2011,** *6* (3), e17695.
22. Collingridge, P. W.; Kelly, S., MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC Bioinformatics* **2012,** *13* (1), 117.
23. Lundström, J.; Rychlewski, L.; Bujnicki, J.; Elofsson, A., Pcons: A neural-network–based consensus predictor that improves fold recognition. *Protein Science* **2001,** *10* (11), 2354-2362.
24. Zhang, Y., I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **2008,** *9* (1), 40.
25. Wang, Z.; Eickholt, J.; Cheng, J., MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics* **2010,** *26* (7), 882-888.
26. Weik, M.; Colletier, J.-P., Temperature-dependent macromolecular X-ray crystallography. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2010,** *66* (4), 437-446.
27. Durbin, S.; Feher, G., Protein crystallization. *Annu. Rev. Phys. Chem.* **1996,** *47* (1), 171-204.
28. Hendrickson, W. A., Analysis of protein structure from diffraction measurement at multiple wavelengths. *Trans. Am. Crystallogr. Assoc* **1985,** *21* (11).
29. De La Fortelle, E.; Bricogne, G., [27] Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. In *Methods Enzymol.*, Elsevier: 1997; Vol. 276, pp 472-494.
30. Ealick, S. E., Advances in multiple wavelength anomalous diffraction crystallography. *Curr. Opin. Chem. Biol.* **2000,** *4* (5), 495-499.
31. Navaza, J., AMoRe: an automated package for molecular replacement. *Acta Crystallogr. Sect. A: Found. Crystallogr.* **1994,** *50* (2), 157-163.
32. Mcpherson, A.; Gavira, J. A., Introduction to protein crystallization. *Acta Crystallogr F Struct Biol Commun* **2013,** *70* (Pt 1), 2-20.
33. Dale, G. E.; Oefner, C.; D'arcy, A., The protein as a variable in protein crystallization. *J. Struct. Biol.* **2003,** *142* (1), 88-97.
34. Derewenda, Z. S.; Vekilov, P. G., Entropy and surface engineering in protein crystallization. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2006,** *62* (1), 116-124.
35. Derewenda, Z. S., The use of recombinant methods and molecular engineering in protein crystallization. *Methods* **2004,** *34* (3), 354-363.
36. Wüthrich, K., Protein structure determination in solution by NMR spectroscopy. *Journal of Biological Chemistry* **1990,** *265* (36), 22059-22062.
37. Brünger, A. T.; Adams, P. D.; Clore, G. M.; Delano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J.-S.; Kuszewski, J.; Nilges, M.; Pannu, N. S., Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **1998,** *54* (5), 905-921.
38. Wüthrich, K., The way to NMR structures of proteins. *Nat. Struct. Mol. Biol.* **2001,** *8* (11), 923.
39. Kuszewski, J.; Schwieters, C. D.; Garrett, D. S.; Byrd, R. A.; Tjandra, N.; Clore, G. M., Completely automated, highly error-tolerant macromolecular structure

determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. *J. Am. Chem. Soc.* **2004,** *126* (20), 6258-6273.
40. Pervushin, K.; Riek, R.; Wider, G.; Wüthrich, K., Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proceedings of the National Academy of Sciences* **1997,** *94* (23), 12366-12371.
41. Kuszewski, J.; Clore, G. M., Sources of and solutions to problems in the refinement of protein NMR structures against torsion angle potentials of mean force. *Journal of Magnetic Resonance* **2000,** *146* (2), 249-254.
42. Henderson, R.; Baldwin, J. M.; Ceska, T.; Zemlin, F.; Beckmann, E. A.; Downing, K. H., Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **1990,** *213* (4), 899-929.
43. Campbell, M. G.; Cheng, A.; Brilot, A. F.; Moeller, A.; Lyumkis, D.; Veesler, D.; Pan, J.; Harrison, S. C.; Potter, C. S.; Carragher, B., Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. *Structure* **2012,** *20* (11), 1823-1828.
44. Tang, G.; Peng, L.; Baldwin, P. R.; Mann, D. S.; Jiang, W.; Rees, I.; Ludtke, S. J., EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **2007,** *157* (1), 38-46.
45. Scheres, S. H., RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **2012,** *180* (3), 519-530.
46. Topf, M.; Lasker, K.; Webb, B.; Wolfson, H.; Chiu, W.; Sali, A., Protein structure fitting and refinement guided by cryo-EM density. *Structure* **2008,** *16* (2), 295-307.
47. Matthies, D.; Bae, C.; Toombes, G. E.; Fox, T.; Bartesaghi, A.; Subramaniam, S.; Swartz, K. J., Single-particle cryo-EM structure of a voltage-activated potassium channel in lipid nanodiscs. *Elife* **2018,** *7*, e37558.
48. Hura, G. L.; Hodge, C. D.; Rosenberg, D.; Guzenko, D.; Duarte, J. M.; Monastyrskyy, B.; Grudinin, S.; Kryshtafovych, A.; Tainer, J. A.; Fidelis, K., Small angle X-ray scattering-assisted protein structure prediction in CASP13 and emergence of solution structure differences. *Proteins: Struct. Funct. Bioinform.* **2019**.
49. Svergun, D. I.; Koch, M. H., Small-angle scattering studies of biological macromolecules in solution. *Rep. Prog. Phys.* **2003,** *66* (10), 1735.
50. Suchanek, M.; Radzikowska, A.; Thiele, C., Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells. *Nature Methods* **2005,** *2* (4), 261.
51. Fancy, D. A.; Kodadek, T., Chemistry for the analysis of protein–protein interactions: rapid and efficient cross-linking triggered by long wavelength light. *Proceedings of the National Academy of Sciences* **1999,** *96* (11), 6020-6024.
52. Schneider, M.; Belsom, A.; Rappsilber, J., Protein tertiary structure by crosslinking/mass spectrometry. *Trends Biochem. Sci* **2018,** *43* (3), 157-169.
53. Fajardo, J. E.; Shrestha, R.; Gil, N.; Belsom, A.; Crivelli, S. N.; Czaplewski, C.; Fidelis, K.; Grudinin, S.; Karasikov, M.; Karczyńska, A. S., Assessment of chemical-crosslink-assisted protein structure modeling in CASP13. *Proteins: Struct. Funct. Bioinform.* **2019**.
54. Kalinin, S.; Peulen, T.; Sindbert, S.; Rothwell, P. J.; Berger, S.; Restle, T.; Goody, R. S.; Gohlke, H.; Seidel, C. A., A toolkit and benchmark study for FRET-restrained high-precision structural modeling. *Nature Methods* **2012,** *9* (12), 1218.
55. Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A., Critical assessment of methods of protein structure prediction (CASP)—round x. *Proteins: Struct. Funct. Bioinform.* **2014,** *82* (S2), 1-6.

56. Croll, T. I.; Sammito, M. D.; Kryshtafovych, A.; Read, R. J., Evaluation of template-based modeling in CASP13. *Proteins: Struct. Funct. Bioinform.* **2019**.
57. Kinch, L.; Yong Shi, S.; Cong, Q.; Cheng, H.; Liao, Y.; Grishin, N. V., CASP9 assessment of free modeling target predictions. *Proteins: Struct. Funct. Bioinform.* **2011,** *79* (S10), 59-73.
58. Kryshtafovych, A.; Barbato, A.; Fidelis, K.; Monastyrskyy, B.; Schwede, T.; Tramontano, A., Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins: Struct. Funct. Bioinform.* **2014,** *82* (S2), 112-126.
59. Schaarschmidt, J.; Monastyrskyy, B.; Kryshtafovych, A.; Bonvin, A. M., Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Struct. Funct. Bioinform.* **2018,** *86* (S1), 51-66.
60. Nugent, T.; Cozzetto, D.; Jones, D. T., Evaluation of predictions in the CASP10 model refinement category. *Proteins: Struct. Funct. Bioinform.* **2014,** *82*, 98-111.
61. Gallo Cassarino, T.; Bordoli, L.; Schwede, T., Assessment of ligand binding site predictions in CASP10. *Proteins: Struct. Funct. Bioinform.* **2014,** *82*, 154-163.
62. Lensink, M. F.; Brysbaert, G.; Nadzirin, N.; Velankar, S.; Chaleil, R. A.; Gerguri, T.; Bates, P. A.; Laine, E.; Carbone, A.; Grudinin, S., Blind prediction of homo-and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins: Struct. Funct. Bioinform.* **2019**.
63. Sanchez, R.; Sali, A., Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proceedings of the National Academy of Sciences* **1998,** *95* (23), 13597-13602.
64. Boratyn, G. M.; Schaffer, A.; Agarwala, R.; Altschul, S. F.; Lipman, D. J.; Madden, T. L., Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **2012,** *7* (1), 12.
65. Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997,** *25* (17), 3389-3402.
66. Pearson, W. R., Finding protein and nucleotide similarities with FASTA. *Curr. Protoc. Bioinf.* **2004**, 3.9. 1-3.9. 25.
67. Peng, J.; Xu, J., RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Struct. Funct. Bioinform.* **2011,** *79* (S10), 161-171.
68. Dayhoff, M. O., A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure* **1972,** *5*, 89-99.
69. Henikoff, S.; Henikoff, J. G., Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **1992,** *89* (22), 10915-10919.
70. Needleman, S. B.; Wunsch, C. D., A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970,** *48* (3), 443-453.
71. Smith, T. F.; Waterman, M. S., Identification of common molecular subsequences. *J. Mol. Biol.* **1981,** *147* (1), 195-197.
72. Gotoh, O., An improved algorithm for matching biological sequences. *J. Mol. Biol.* **1982,** *162* (3), 705-708.
73. Thompson, J. D.; Higgins, D. G.; Gibson, T. J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994,** *22* (22), 4673-4680.
74. Katoh, K.; Standley, D. M., MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **2013,** *30* (4), 772-780.

75. O'sullivan, O.; Suhre, K.; Abergel, C.; Higgins, D. G.; Notredame, C., 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.* **2004,** *340* (2), 385-395.
76. Daniels, N. M.; Nadimpalli, S.; Cowen, L. J., Formatt: Correcting protein multiple structural alignments by incorporating sequence alignment. *BMC Bioinformatics* **2012,** *13* (1), 259.
77. Wang, S.; Peng, J.; Xu, J., Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. *Bioinformatics* **2011,** *27* (18), 2537-2545.
78. Panchenko, A. R., Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res.* **2003,** *31* (2), 683-689.
79. Rychlewski, L.; Li, W.; Jaroszewski, L.; Godzik, A., Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science* **2000,** *9* (2), 232-241.
80. Lobley, A.; Sadowski, M. I.; Jones, D. T., pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* **2009,** *25* (14), 1761-1767.
81. Jones, D. T., Protein secondary structure prediction based on position-specific s coring matrices. *J. Mol. Biol.* **1999,** *292* (2), 195-202.
82. Boratyn, G. M.; Schäffer, A. A.; Agarwala, R.; Altschul, S. F.; Lipman, D. J.; Madden, T. L., Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **2012,** *7* (1), 12.
83. Rice, D. W.; Eisenberg, D., A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* **1997,** *267* (4), 1026-1038.
84. Biegert, A.; Söding, J., Sequence context-specific profiles for homology searching. *Proceedings of the National Academy of Sciences* **2009,** *106* (10), 3770-3775.
85. Marchler-Bauer, A.; Derbyshire, M. K.; Gonzales, N. R.; Lu, S.; Chitsaz, F.; Geer, L. Y.; Geer, R. C.; He, J.; Gwadz, M.; Hurwitz, D. I., CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **2014,** *43* (D1), D222-D226.
86. Söding, J., Protein homology detection by HMM–HMM comparison. *Bioinformatics* **2005,** *21* (7), 951-960.
87. Eddy, S. R., Accelerated profile HMM searches. *PLoS Computational Biology* **2011,** *7* (10), e1002195.
88. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J., HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **2012,** *9* (2), 173-175.
89. Madera, M., Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics* **2008,** *24* (22), 2630-2631.
90. Karplus, K., SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res.* **2009**, gkp403.
91. Yoon, B.-J., Hidden Markov models and their applications in biological sequence analysis. *Curr. Genomics* **2009,** *10* (6), 402-415.
92. Söding, J.; Biegert, A.; Lupas, A. N., The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **2005,** *33* (suppl_2), W244-W248.
93. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The protein data bank. *Nucleic Acids Res.* **2000,** *28* (1), 235-242.
94. Consortium, U., UniProt: a hub for protein information. *Nucleic Acids Res.* **2014,** *43* (D1), D204-D212.

95. Jaroszewski, L.; Rychlewski, L.; Li, Z.; Li, W.; Godzik, A., FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Res.* **2005,** *33* (suppl 2), W284-W288.
96. Petersen, B.; Petersen, T. N.; Andersen, P.; Nielsen, M.; Lundegaard, C., A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* **2009,** *9* (1), 51.
97. Magnan, C. N.; Baldi, P., SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* **2014,** *30* (18), 2592-2597.
98. Wang, S.; Peng, J.; Ma, J.; Xu, J., Protein secondary structure prediction using deep convolutional neural fields. *Sci. Rep.* **2016,** *6*, 18962.
99. Mcguffin, L. J.; Bryson, K.; Jones, D. T., The PSIPRED protein structure prediction server. *Bioinformatics* **2000,** *16* (4), 404-405.
100. Fang, C.; Shang, Y.; Xu, D., MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction. *Proteins: Struct. Funct. Bioinform.* **2018,** *86* (5), 592-598.
101. Heffernan, R.; Yang, Y.; Paliwal, K.; Zhou, Y., Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics* **2017,** *33* (18), 2842-2849.
102. Heffernan, R.; Paliwal, K.; Lyons, J.; Singh, J.; Yang, Y.; Zhou, Y., Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *JCoCh* **2018**.
103. Ma, J.; Wang, S., AcconPred: Predicting solvent accessibility and contact number simultaneously by a multitask learning framework under the conditional neural fields model. *BioMed research international* **2015,** *2015*.
104. Joo, K.; Lee, S. J.; Lee, J., Sann: solvent accessibility prediction of proteins by nearest neighbor method. *Proteins: Struct. Funct. Bioinform.* **2012,** *80* (7), 1791-1797.
105. Xu, D.; Jaroszewski, L.; Li, Z.; Godzik, A., FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics* **2013**, btt578.
106. Chakravarty, S.; Varadarajan, R., Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* **1999,** *7* (7), 723-732.
107. Wu, S.; Zhang, Y., MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Struct. Funct. Bioinform.* **2008,** *72* (2), 547-556.
108. Ward, J. J.; Mcguffin, L. J.; Bryson, K.; Buxton, B. F.; Jones, D. T., The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **2004,** *20* (13), 2138-2139.
109. Yang, Y.; Faraggi, E.; Zhao, H.; Zhou, Y., Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **2011,** *27* (15), 2076-2082.
110. Zhou, H.; Zhou, Y., Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins: Struct. Funct. Bioinform.* **2005,** *58* (2), 321-328.
111. Fernandez-Fuentes, N.; Rai, B. K.; Madrid-Aliste, C. J.; Fajardo, J. E.; Fiser, A., Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinformatics* **2007,** *23* (19), 2558-2565.

112. Jones, D. T.; Miller, R. T.; Thornton, J. M., Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins: Struct. Funct. Bioinform.* **1995,** *23* (3), 387-397.
113. Rychlewski, L.; Fischer, D., LiveBench-8: The large-scale, continuous assessment of automated protein structure prediction. *Protein Science* **2005,** *14* (1), 240-245.
114. Wang, S.; Ma, J.; Peng, J.; Xu, J., Protein structure alignment beyond spatial proximity. *Sci. Rep.* **2013,** *3*, 1448.
115. Konagurthu, A. S.; Whisstock, J. C.; Stuckey, P. J.; Lesk, A. M., MUSTANG: a multiple structural alignment algorithm. *Proteins: Struct. Funct. Bioinform.* **2006,** *64* (3), 559-574.
116. Webb, B.; Sali, A., Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinf.* **2014**, 5.6. 1-5.6. 32.
117. Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D., Protein structure prediction using Rosetta. In *Methods Enzymol.*, Elsevier: 2004; Vol. 383, pp 66-93.
118. Zhang, W.; Yang, J.; He, B.; Walker, S. E.; Zhang, H.; Govindarajoo, B.; Virtanen, J.; Xue, Z.; Shen, H. B.; Zhang, Y., Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11. *Proteins: Struct. Funct. Bioinform.* **2016,** *84* (S1), 76-86.
119. Adhikari, B.; Bhattacharya, D.; Cao, R.; Cheng, J., CONFOLD: residue-residue contact-guided ab initio protein folding. *Proteins: Struct. Funct. Bioinform.* **2015,** *83* (8), 1436-1449.
120. Moult, J., A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology* **2005,** *15* (3), 285-289.
121. Floudas, C.; Fung, H.; Mcallister, S.; Mönnigmann, M.; Rajgaria, R., Advances in protein structure prediction and de novo protein design: A review. *Chem. Eng. Sci.* **2006,** *61* (3), 966-988.
122. Mulnaes, D.; Gohlke, H., TopScore: Using Deep Neural Networks and Large Diverse Data Sets for Accurate Protein Model Quality Assessment. *J. Chem. Theory Comput.* **2018**.
123. Brown, P. J.; Fuller, W. A., *Statistical Analysis of Measurement Error Models and Applications: Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference Held June 10-16, 1989, with Support from the National Science Foundation and the US Army Research Office*. American Mathematical Soc.: 1990; Vol. 112.
124. Wallner, B.; Elofsson, A., Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Science* **2006,** *15* (4), 900-913.
125. Zhang, Y.; Skolnick, J., TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005,** *33* (7), 2302-2309.
126. Zemla, A., LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **2003,** *31* (13), 3370-3374.
127. Siew, N.; Elofsson, A.; Rychlewski, L.; Fischer, D., MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* **2000,** *16* (9), 776-785.
128. Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G., Optimal protein-folding codes from spin-glass theory. *Proceedings of the National Academy of Sciences* **1992,** *89* (11), 4918-4922.
129. Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T., lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **2013,** *29* (21), 2722-2728.

130. Olechnovič, K.; Kulberkytė, E.; Venclovas, Č., CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins: Struct. Funct. Bioinform.* **2013,** *81* (1), 149-162.
131. Xu, D.; Zhang, Y., Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys. J.* **2011,** *101* (10), 2525-2534.
132. Bhattacharya, D.; Nowotny, J.; Cao, R.; Cheng, J., 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic Acids Res.* **2016,** *44* (W1), W406-W409.
133. Zhang, J.; Liang, Y.; Zhang, Y., Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* **2011,** *19* (12), 1784-1795.
134. Park, H.; Ovchinnikov, S.; Kim, D. E.; Dimaio, F.; Baker, D., Protein homology model refinement by large-scale energy optimization. *Proceedings of the National Academy of Sciences* **2018,** *115* (12), 3054-3059.
135. Buchan, D. W.; Jones, D. T., EigenTHREADER: analogous protein fold recognition by efficient contact map threading. *Bioinformatics* **2017,** *33* (17), 2684-2690.
136. Jones, D. T.; Buchan, D. W.; Cozzetto, D.; Pontil, M., PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **2011,** *28* (2), 184-190.
137. Seemayer, S.; Gruber, M.; Söding, J., CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **2014,** *30* (21), 3128-3130.
138. Zhang, H.; Gao, Y.; Deng, M.; Wang, C.; Zhu, J.; Li, S. C.; Zheng, W.-M.; Bu, D., Improving residue–residue contact prediction via low-rank and sparse decomposition of residue correlation matrix. *Biochemical and Biophysical Research Communications* **2016,** *472* (1), 217-222.
139. Kaján, L.; Hopf, T. A.; Kalaš, M.; Marks, D. S.; Rost, B., FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* **2014,** *15* (1), 85.
140. Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J., Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Computational Biology* **2017,** *13* (1), e1005324.
141. He, B.; Mortuza, S.; Wang, Y.; Shen, H.-B.; Zhang, Y., NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics* **2017,** *33* (15), 2296-2306.
142. Adhikari, B.; Hou, J.; Cheng, J., DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* **2017,** *34* (9), 1466-1472.
143. Randall, A.; Baldi, P., SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *BMC Struct. Biol.* **2008,** *8* (1), 52.
144. Mabrouk, M.; Putz, I.; Werner, T.; Schneider, M.; Neeb, M.; Bartels, P.; Brock, O., RBO Aleph: leveraging novel information sources for protein structure prediction. *Nucleic Acids Res.* **2015,** *43* (W1), W343-W348.
145. Tegge, A. N.; Wang, Z.; Eickholt, J.; Cheng, J., NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res.* **2009,** *37* (suppl_2), W515-W518.
146. Cheng, J.; Baldi, P., Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics* **2007,** *8* (1), 113.

147. Björkholm, P.; Daniluk, P.; Kryshtafovych, A.; Fidelis, K.; Andersson, R.; Hvidsten, T. R., Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue–residue contacts. *Bioinformatics* **2009,** *25* (10), 1264-1270.
148. Liu, Y.; Palmedo, P.; Ye, Q.; Berger, B.; Peng, J., Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Systems* **2018,** *6* (1), 65-74. e63.
149. Skwark, M. J.; Raimondi, D.; Michel, M.; Elofsson, A., Improved contact predictions using the recognition of protein like contact patterns. *PLoS Computational Biology* **2014,** *10* (11), e1003889.
150. Jones, D. T.; Kandathil, S. M., High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* **2018,** *1*, 8.
151. Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y.; Valencia, A., Accurate Prediction of Protein Contact Maps by Coupling Residual Two-Dimensional Bidirectional Long Short-Term Memory with Convolutional Neural Networks. *Bioinformatics* **2018**.
152. Breiman, L., Random forests. *MLear* **2001,** *45* (1), 5-32.
153. Geurts, P.; Ernst, D.; Wehenkel, L., Extremely randomized trees. *MLear* **2006,** *63* (1), 3-42.
154. Michel, M.; Skwark, M. J.; Menéndez Hurtado, D.; Ekeberg, M.; Elofsson, A., Predicting accurate contacts in thousands of Pfam domain families using PconsC3. *Bioinformatics* **2017,** *33* (18), 2859-2866.
155. Chen, X.-W.; Jeong, J. C., Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics* **2009,** *25* (5), 585-591.
156. Ebina, T.; Toh, H.; Kuroda, Y., DROP: an SVM domain linker predictor trained with optimal features selected by random forest. *Bioinformatics* **2010,** *27* (4), 487-494.
157. Manavalan, B.; Lee, J., SVMQA: support–vector-machine-based protein single-model quality assessment. *Bioinformatics* **2017,** *33* (16), 2496-2503.
158. Wang, Z.; Tegge, A. N.; Cheng, J., Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins: Struct. Funct. Bioinform.* **2009,** *75* (3), 638-647.
159. Deng, L.; Yu, D., Deep learning: methods and applications. *Foundations and Trends in Signal Processing* **2014,** *7* (3–4), 197-387.
160. Shrestha, A.; Mahmood, A., Review of Deep Learning Algorithms and Architectures. *IEEE Access* **2019,** *7*, 53040-53065.
161. Mamoshina, P.; Vieira, A.; Putin, E.; Zhavoronkov, A., Applications of deep learning in biomedicine. *Mol. Pharm.* **2016,** *13* (5), 1445-1454.
162. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. In *The application of two-level attention models in deep convolutional neural network for fine-grained image classification*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015; pp 842-850.
163. Michel, M.; Menéndez Hurtado, D.; Elofsson, A., PconsC4: fast, accurate and hassle-free contact predictions. *Bioinformatics* **2018**.
164. Li, M.; Zuo, W.; Gu, S.; Zhao, D.; Zhang, D. In *Learning convolutional networks for content-weighted image compression*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018; pp 3214-3223.
165. Li, F.; Qiao, H.; Zhang, B., Discriminatively boosted image clustering with fully convolutional auto-encoders. *Pattern Recognition* **2018,** *83*, 161-173.
166. Graves, A.; Mohamed, A.-R.; Hinton, G. In *Speech recognition with deep recurrent neural networks*, 2013 IEEE international conference on acoustics, speech and signal processing, IEEE: 2013; pp 6645-6649.

167. Irsoy, O.; Cardie, C. In *Opinion mining with deep recurrent neural networks*, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014; pp 720-728.
168. Hawkins, J.; Boden, M., The applicability of recurrent neural networks for biological sequence analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **2005,** *2* (3), 243-253.
169. Faraggi, E.; Zhang, T.; Yang, Y.; Kurgan, L.; Zhou, Y., SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *JCoCh* **2012,** *33* (3), 259-267.
170. Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Yang, Y.; Zhou, Y., Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep.* **2015,** *5*, 11476.
171. Denton, E. L.; Chintala, S.; Fergus, R. In *Deep generative image models using a*, Adv. Neural Inf. Process. Syst., 2015; pp 1486-1494.
172. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z. In *Photo-realistic single image super-resolution using a generative adversarial network*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017; pp 4681-4690.
173. Hochreiter, S., The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **1998,** *6* (02), 107-116.
174. He, K.; Zhang, X.; Ren, S.; Sun, J. In *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; pp 770-778.
175. Zhang, B.; Li, J.; Lu, Q., Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics* **2018,** *19* (1), 293.
176. Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. R., Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* **2012**.
177. Steiger, J. H., Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **1980,** *87* (2), 245.
178. Ghent, A. W., A method for exact testing of 2X2, 2X3, 3X3, and other contingency tables, employing binomial coefficients. *Am. Midl. Nat.* **1972**, 15-27.
179. Schneidman-Duhovny, D.; Hammel, M.; Tainer, J. A.; Sali, A., Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **2013,** *105* (4), 962-974.
180. Schneidman-Duhovny, D.; Hammel, M.; Tainer, J. A.; Sali, A., FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* **2016,** *44* (W1), W424-429.
181. Svergun, D. I.; Petoukhov, M. V.; Koch, M. H., Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **2001,** *80* (6), 2946-2953.
182. Kozin, M. B.; Svergun, D. I., Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.* **2001,** *34*, 33-41.
183. Harcet, M.; Perina, D.; Pleše, B., Opine dehydrogenases in marine invertebrates. *Biochem. Genet.* **2013,** *51* (9-10), 666-676.
184. Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E., How does a drug molecule find its target binding site? *J. Am. Chem. Soc.* **2011,** *133* (24), 9181-9183.

185. Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhani, D. W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D. E., Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences* **2011,** *108* (32), 13118-13123.
186. Giorgino, T.; Buch, I.; De Fabritiis, G., Visualizing the induced binding of SH2-phosphopeptide. *J. Chem. Theory Comput.* **2012,** *8* (4), 1171-1175.
187. Buch, I.; Giorgino, T.; De Fabritiis, G., Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **2011,** *108* (25), 10184-10189.
188. Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O., Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Natur* **2012,** *482* (7386), 552.
189. Endo, N.; Kan-No, N.; Nagahisa, E., Purification, characterization, and cDNA cloning of opine dehydrogenases from the polychaete rockworm Marphysa sanguinea. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **2007,** *147* (2), 293-307.
190. Hartmann, K., Clinical aspects of feline immunodeficiency and feline leukemia virus infection. *Vet. Immunol. Immunopathol.* **2011,** *143* (3-4), 190-201.
191. De Rozières, S.; Mathiason, C. K.; Rolston, M. R.; Chatterji, U.; Hoover, E. A.; Elder, J. H., Characterization of a highly pathogenic molecular clone of feline immunodeficiency virus clade C. *J. Virol.* **2004,** *78* (17), 8971-8982.
192. Diehl, L. J.; Mathiason-Dubard, C. K.; O'neil, L. L.; Obert, L. A.; Hoover, E. A., Induction of accelerated feline immunodeficiency virus disease by acute-phase virus passage. *J. Virol.* **1995,** *69* (10), 6149-6157.
193. Obert, L. A.; Hoover, E. A., Feline immunodeficiency virus clade C mucosal transmission and disease courses. *AIDS Res. Hum. Retroviruses* **2000,** *16* (7), 677-688.
194. Lehman, T. L.; O'halloran, K. P.; Hoover, E. A.; Avery, P. R., Utilizing the FIV model to understand dendritic cell dysfunction and the potential role of dendritic cell immunization in HIV infection. *Vet. Immunol. Immunopathol.* **2010,** *134* (1-2), 75-81.
195. Yamamoto, J. K.; Sanou, M. P.; Abbott, J. R.; Coleman, J. K., Feline immunodeficiency virus model for designing HIV/AIDS vaccines. *Curr. HIV Res.* **2010,** *8* (1), 14-25.
196. Elder, J. H.; Lin, Y.-C.; Fink, E.; Grant, C. K., Feline immunodeficiency virus (FIV) as a model for study of lentivirus infections: parallels with HIV. *Curr. HIV Res.* **2010,** *8* (1), 73-80.
197. Larue, R. S.; Jónsson, S. R.; Silverstein, K. A.; Lajoie, M.; Bertrand, D.; El-Mabrouk, N.; Hötzel, I.; Andrésdóttir, V.; Smith, T. P.; Harris, R. S., The artiodactyl APOBEC3 innate immune repertoire shows evidence for a multi-functional domain organization that existed in the ancestor of placental mammals. *BMC Mol. Biol.* **2008,** *9* (1), 104.
198. Larue, R. S.; Andrésdóttir, V.; Blanchard, Y.; Conticello, S. G.; Derse, D.; Emerman, M.; Greene, W. C.; Jónsson, S. R.; Landau, N. R.; Löchelt, M., Guidelines for naming nonprimate APOBEC3 genes and proteins. *J. Virol.* **2009,** *83* (2), 494-497.
199. Münk, C.; Willemsen, A.; Bravo, I. G., An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. *BMC Evol. Biol.* **2012,** *12* (1), 71.
200. Münk, C.; Beck, T.; Zielonka, J.; Hotz-Wagenblatt, A.; Chareza, S.; Battenberg, M.; Thielebein, J.; Cichutek, K.; Bravo, I. G.; O'brien, S. J., Functions, structure, and read-through alternative splicing of feline APOBEC3 genes. *Genome Biol.* **2008,** *9* (3), R48.

201. Sheehy, A. M.; Gaddis, N. C.; Choi, J. D.; Malim, M. H., Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Natur* **2002,** *418* (6898), 646.
202. Derse, D.; Hill, S. A.; Princler, G.; Lloyd, P.; Heidecker, G., Resistance of human T cell leukemia virus type 1 to APOBEC3G restriction is mediated by elements in nucleocapsid. *Proceedings of the National Academy of Sciences* **2007,** *104* (8), 2915-2920.
203. Löchelt, M.; Romen, F.; Bastone, P.; Muckenfuss, H.; Kirchner, N.; Kim, Y.-B.; Truyen, U.; Rösler, U.; Battenberg, M.; Saib, A., The antiretroviral activity of APOBEC3 is inhibited by the foamy virus accessory Bet protein. *Proceedings of the National Academy of Sciences* **2005,** *102* (22), 7982-7987.
204. Stavrou, S.; Nitta, T.; Kotla, S.; Ha, D.; Nagashima, K.; Rein, A. R.; Fan, H.; Ross, S. R., Murine leukemia virus glycosylated Gag blocks apolipoprotein B editing complex 3 and cytosolic sensor access to the reverse transcription complex. *Proceedings of the National Academy of Sciences* **2013,** *110* (22), 9078-9083.
205. Gerpe, M. C. R.; Renner, T. M.; Bélanger, K.; Lam, C.; Aydin, H.; Langlois, M.-A., N-linked glycosylation protects gammaretroviruses against deamination by APOBEC3 proteins. *J. Virol.* **2015,** *89* (4), 2342-2357.
206. Kolokithas, A.; Rosenke, K.; Malik, F.; Hendrick, D.; Swanson, L.; Santiago, M. L.; Portis, J. L.; Hasenkrug, K. J.; Evans, L. H., The glycosylated Gag protein of a murine leukemia virus inhibits the antiretroviral function of APOBEC3. *J. Virol.* **2010,** *84* (20), 10933-10936.
207. Willems, L.; Gillet, N., APOBEC3 interference during replication of viral genomes. *Viruses* **2015,** *7* (6), 2999-3018.
208. Holmes, R. K.; Koning, F. A.; Bishop, K. N.; Malim, M. H., APOBEC3F can inhibit the accumulation of hiv-1 reverse transcription products in the absence of hypermutation comparisons with APOBEC3G. *Journal of Biological Chemistry* **2007,** *282* (4), 2587-2595.
209. Iwatani, Y.; Chan, D. S.; Wang, F.; Maynard, K. S.; Sugiura, W.; Gronenborn, A. M.; Rouzina, I.; Williams, M. C.; Musier-Forsyth, K.; Levin, J. G., Deaminase-independent inhibition of HIV-1 reverse transcription by APOBEC3G. *Nucleic Acids Res.* **2007,** *35* (21), 7096-7108.
210. Gillick, K.; Pollpeter, D.; Phalora, P.; Kim, E.-Y.; Wolinsky, S. M.; Malim, M. H., Suppression of HIV-1 infection by APOBEC3 proteins in primary human CD4+ T cells is associated with inhibition of processive reverse transcription as well as excessive cytidine deamination. *J. Virol.* **2013,** *87* (3), 1508-1517.
211. Wang, X.; Ao, Z.; Chen, L.; Kobinger, G.; Peng, J.; Yao, X., The cellular antiviral protein APOBEC3G interacts with HIV-1 reverse transcriptase and inhibits its function during viral replication. *J. Virol.* **2012,** *86* (7), 3777-3786.
212. Mbisa, J. L.; Barr, R.; Thomas, J. A.; Vandegraaff, N.; Dorweiler, I. J.; Svarovskaia, E. S.; Brown, W. L.; Mansky, L. M.; Gorelick, R. J.; Harris, R. S., Human immunodeficiency virus type 1 cDNAs produced in the presence of APOBEC3G exhibit defects in plus-strand DNA transfer and integration. *J. Virol.* **2007,** *81* (13), 7099-7110.
213. Mbisa, J. L.; Bu, W.; Pathak, V. K., APOBEC3F and APOBEC3G inhibit HIV-1 DNA integration by different mechanisms. *J. Virol.* **2010,** *84* (10), 5250-5259.
214. Zielonka, J.; Marino, D.; Hofmann, H.; Yuhki, N.; Löchelt, M.; Münk, C., Vif of feline immunodeficiency virus from domestic cats protects against APOBEC3 restriction factors from many felids. *J. Virol.* **2010,** *84* (14), 7312-7324.

215. Münk, C.; Zielonka, J.; Constabel, H.; Kloke, B.-P.; Rengstl, B.; Battenberg, M.; Bonci, F.; Pistello, M.; Löchelt, M.; Cichutek, K., Multiple restrictions of human immunodeficiency virus type 1 in feline cells. *J. Virol.* **2007,** *81* (13), 7048-7060.
216. Stern, M. A.; Hu, C.; Saenz, D. T.; Fadel, H. J.; Sims, O.; Peretz, M.; Poeschla, E. M., Productive replication of Vif-chimeric HIV-1 in feline cells. *J. Virol.* **2010,** *84* (14), 7378-7395.
217. Larue, R. S.; Lengyel, J.; Jónsson, S. R.; Andrésdóttir, V.; Harris, R. S., Lentiviral Vif degrades the APOBEC3Z3/APOBEC3H protein of its mammalian host and is capable of cross-species activity. *J. Virol.* **2010,** *84* (16), 8193-8201.
218. Yoshikawa, R.; Takeuchi, J. S.; Yamada, E.; Nakano, Y.; Ren, F.; Tanaka, H.; Münk, C.; Harris, R. S.; Miyazawa, T.; Koyanagi, Y., Vif determines the requirement for CBF-β in APOBEC3 degradation. *The Journal of General Virology* **2015,** *96* (Pt 4), 887.
219. Bisson, M. M.; Groth, G., Targeting plant ethylene responses by controlling essential protein–protein interactions in the ethylene pathway. *Molecular plant* **2015,** *8* (8), 1165-1174.
220. Bisson, M. M.; Kessenbrock, M.; Müller, L.; Hofmann, A.; Schmitz, F.; Cristescu, S. M.; Groth, G., Peptides interfering with protein-protein interactions in the ethylene signaling pathway delay tomato fruit ripening. *Sci. Rep.* **2016,** *6*, 30634.
221. Kessenbrock, M.; Klein, S. M.; Müller, L.; Hunsche, M.; Noga, G.; Groth, G., Novel protein-protein inhibitor based approach to control plant ethylene responses: synthetic peptides for ripening control. *Frontiers in Plant Science* **2017,** *8*, 1528.
222. Mayerhofer, H.; Panneerselvam, S.; Kaljunen, H.; Tuukkanen, A.; Mertens, H. D.; Mueller-Dieckmann, J., Structural model of the cytosolic domain of the plant ethylene receptor 1 (ETR1). *Journal of Biological Chemistry* **2015,** *290* (5), 2644-2658.
223. Mezulis, S.; Sternberg, M. J.; Kelley, L. A., PhyreStorm: A web server for fast structural searches against the PDB. *J. Mol. Biol.* **2016,** *428* (4), 702-708.
224. Dominguez, C.; Boelens, R.; Bonvin, A. M., HADDOCK: a protein− protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **2003,** *125* (7), 1731-1737.
225. Pfleger, C.; Minges, A.; Boehm, M.; Mcclendon, C. L.; Torella, R.; Gohlke, H., Ensemble-and rigidity theory-based perturbation approach to analyze dynamic allostery. *J. Chem. Theory Comput.* **2017,** *13* (12), 6343-6357.
226. Pfleger, C.; Rathi, P. C.; Klein, D. L.; Radestock, S.; Gohlke, H., Constraint Network Analysis (CNA): a Python software package for efficiently linking biomacromolecular structure, flexibility,(thermo-) stability, and function. ACS Publications: 2013.