Approaches to Automatic Structural Interpretation of Cryo Electron Microscopy Data

Inaugural-Dissertation

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

Vorgelegt von

Luisa Schäfer
aus Dormagen

Jülich, Oktober 2022

aus dem Institut für Biologische Informationsprozessierung
des Forschungszentrums Jülich

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

Berichterstatter:

1. Prof. Gunnar Schröder

2. Prof. Birgit Strodel

Tag der mündlichen Prüfung: 10.01.2023

Erklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der 'Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf' erstellt worden ist. Weiterhin erkläre ich, dass diese Dissertation an keiner anderen Fakultät eingereicht wurde. Ich habe bislang keine erfolglosen oder erfolgreichen Promotionsversuche unternommen.

Abstract

Due to their decisive involvement in physiological processes of living organisms, elucidation of the three-dimensional structure of proteins is essential for biomedical research and drug development. Cryogenic Electron Microscopy (cryo-EM) is a state-of-the-art method to experimentally determine a protein structure. It enables the reconstruction of a three-dimensional density map, revealing the structure of a protein with a resolution of up to a few angstrom. However, depending on map resolution, interpreting the density map in terms of an atomic model of the protein is often challenging. Computational tools are needed to facilitate this process.

This thesis presents four different methods for automatic structural interpretation of cryo-EM density maps.

(1) As contribution to the EMDB Model Metrics Challenge 2019, we have developed a procedure to optimise a given protein structure into a cryo-EM density map, regarding the conformational heterogeneity embodied in the cryo-EM data.

(2) In some situations a cryo-EM map allows for straight-forward modelling of the main-chain of the protein, but the assignment of side-chains is often ambiguous. We introduce a method to automatically sample and rank many different side-chain assignments and apply it to a density map representing the structure of IAPP fibrils.

(3) We present a routine to flexibly fit fragments from a library of backbone conformations to a given protein trace. The fitted fragments offer more detailed insights into the underlying protein structure than the plain trace. Furthermore, the fragment fitting can easily be integrated in larger frameworks for automatic protein structure modelling.

(4) If only the density map and no other structural information is available, the first step is to determine the topology of the protein. We have developed a novel integrative approach to do so. It combines the information provided by the density map with information derived from predicted inter-residue distances. We show, that incorporating distance predictions can correct errors in topology and improve traces that were built based on density information alone.

Zusammenfassung

Aufgrund ihrer entscheidenden Beteiligung an physiologischen Prozessen lebender Organismen ist die Aufklärung der dreidimensionalen Struktur von Proteinen für die biomedizinische Forschung und die Entwicklung von Arzneimitteln unerlässlich. Die kryogene Elektronenmikroskopie (Cryo-EM) ist eine Methode zur experimentellen Bestimmung einer Proteinstruktur. Sie ermöglicht die Rekonstruktion einer dreidimensionalen Dichtekarte, die die Struktur eines Proteins mit einer Auflösung von bis zu einigen wenigen Angström darstellt. Je nach Auflösung der Karte ist die Interpretation der Dichtekarte im Sinne eines atomaren Modells des Proteins jedoch oft schwierig. Hier werden computergestützte Methoden benötigt, die diesen Prozess erleichtern können.

In dieser Arbeit werden vier verschiedene Methoden zur automatischen Strukturinterpretation von Cryo-EM Dichtekarten vorgestellt.

(1) Als Beitrag zur EMDB Model Metrics Challenge 2019 haben wir ein Verfahren zur Optimierung einer gegebenen Proteinstruktur an eine Cryo-EM-Dichtekarte entwickelt. Dieses Verfahren berücksichtigt insbesondere die konformationelle Heterogenität, die in der Dichtekate enthalten ist.

(2) In manchen Situationen ermöglicht eine Cryo-EM-Karte zwar eine direkte Modellierung der Hauptkette des Proteins, aber die Zuordnung der Seitenketten kann nicht eindeutig erfolgen. Wir stellen eine Methode vor, mit der viele verschiedene Seitenkettenzuordnungen automatisch ausprobiert und in eine Rangliste eingeordnet werden können. Diese Methode wird dann auf eine Karte angewendet, die die Struktur von IAPP-Fibrillen darstellt.

(3) Wir präsentieren eine Routine zur flexiblen Anpassung von Fragmenten bekannter Rückgratkonformationen an eine gegebene Proteintopologie. Die angepassten Fragmente bieten detailliertere Einblicke in die zugrundeliegende Proteinstruktur als die einfache Topologie. Darüber hinaus kann diese Methode leicht in größere Programme zur automatischen Modellierung von Proteinstrukturen integriert werden.

(4) Wenn nur die Karte und keine anderen Strukturinformationen verfügbar sind, muss zunächst die Topologie des Proteins bestimmt werden. Wir haben dazu

einen neuen integrativen Ansatz entwickelt, der die Informationen aus der Dichtekarte mit Informationen, die aus vorhergesagten Abständen zwischen den Residuen abgeleitet werden, kombiniert. Wir zeigen, dass die Einbeziehung von Abstandsvorhersagen Fehler in der Topologie korrigieren und solche Topologien, die allein auf der Grundlage von Dichteinformationen erstellt wurden, verbessern kann.

List of Publications

Parts of the work undertaken during the course of my PhD project have been published in peer-reviewed articles or have been submitted to scientific journals:

Included in this thesis:

- **Publication I:**
  Lawson, C. L., Kryshtafovych, A., Adams, P. D., Afonine, P. V., Baker, M. L., Barad, B. A., ..., **Schäfer, L. U.**, ... & Chiu, W. (2021). Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. Nature methods, 18(2), 156-164.

- **Publication II:**
  Roeder, C., Kupreichyk, T., Gremer, L., **Schäfer, L. U.**, Pothula, K. R., Ravelli, R. B., ... & Schröder, G. F. (2020). Cryo-EM structure of islet amyloid polypeptide fibrils reveals similarities with amyloid-β fibrils. Nature Structural & Molecular Biology, 27(7), 660-667.

- **Manuscript I:**
  Schäfer, L. U. & Schröder, G. F. Predicted Distances Guide Protein Topology Tracing in Medium Resolution Density Maps
  Currently under review.

Not included:

- **Publication III:**
  Risi, C., **Schäfer, L. U.**, Belknap, B., Pepper, I., White, H. D., Schröder, G. F., & Galkin, V. E. (2021). High-resolution cryo-EM structure of the cardiac actomyosin complex. Structure, 29(1), 50-60.

Contents

1 Theoretical Background **1**
- 1.1 Basic Principles of Protein Structure . . . 1
- 1.2 Experimental Structure Determination using cryo-EM . . . 2
  - 1.2.1 General Information . . . 3
  - 1.2.2 Workflow of cryo-EM Single Particle Analysis . . . 4
  - 1.2.3 Resolution of cryo-EM Density Maps . . . 7
- 1.3 Structure Modelling based on cryo-EM Data . . . 7
  - 1.3.1 Fitting and Refinement Methods . . . 8
  - 1.3.2 De Novo Modelling Tools . . . 9
  - 1.3.3 Model Validation . . . 10

2 Objective and Outline **13**

3 Structure Optimisation by Capturing Conformational Heterogeneity in the EMDB Model Metrics Challenge **15**
- 3.1 Motivation . . . 15
- 3.2 Material and Methods . . . 16
  - 3.2.1 Material . . . 16
  - 3.2.2 Methods . . . 17
- 3.3 Results and Discussion . . . 18
  - 3.3.1 Publication 1: Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge . 18

4 Automatic Side-chain Sampling for Structure Elucidation of IAPP Fibrils **21**
- 4.1 Theoretical Background . . . 21
  - 4.1.1 Amyloid Fibrils . . . 21
  - 4.1.2 Prediction of Side-chain Conformations with Sqwrl4 . . . 23
- 4.2 Motivation . . . 23
- 4.3 Methods . . . 24
- 4.4 Results and Discussion . . . 24

4.4.1 Publication II: Cryo-EM structure of islet amyloid polypeptide fibrils reveals similarities with amyloid-β fibrils . . . 24

5 Flexible Fragment Fitting for Automatic Backbone Building 29
5.1 Theoretical Background . . . 29
5.1.1 De-Novo Protein Structure Modelling with EMfasa . . . 29
5.1.2 Deformable Elastic Network Restraints for Structure Refinement in DireX . . . 31
5.2 Motivation . . . 33
5.3 Material and Methods . . . 34
5.3.1 Fitting Procedure . . . 34
5.3.2 Benchmark Dataset . . . 36
5.3.3 Metrics to Assess Placement-Quality . . . 36
5.3.4 Adaption of the Fragment Library . . . 37
5.3.5 Tuning Fragment Rigidity with DEN-Restraints . . . 38
5.3.6 Integrating Flexible Fragment Fitting into EMfasa . . . 39
5.4 Results and Discussion . . . 39
5.4.1 Measuring Placement Quality . . . 39
5.4.2 Effects of Library Sizes . . . 41
5.4.3 Effects of Fragment Flexibility . . . 43
5.4.4 Fragment Fitting as integrated Part of EMfasa . . . 45

6 Protein Topology Tracing Guided by Predicted Distance Matrices 49
6.1 Theoretical Background . . . 49
6.1.1 Prediction of Inter-Residue Distances . . . 49
6.1.2 General Properties of Distance Matrices . . . 50
6.2 Manuscript I: Predicted Distance Maps Guide Backbone Topology Tracing in Medium Resolution Density Maps . . . 50
6.2.1 Summary . . . 50
6.2.2 Introduction . . . 51
6.2.3 Results . . . 52
6.2.4 Discussion . . . 63
6.2.5 Contribution . . . 66

7 Conclusion 67

Bibliography 69

Acknowledgements 81

List of Figures 83

List of Tables **85**

List of Abbreviations **87**

A Embedded Publication I **89**

B Embedded Publication II **111**

C Supplementary Material for Topology Tracing **133**
C.1 STAR Methods 133
C.1.1 Step 1 Trace Initialisation 133
C.1.2 Step 2 Weights Estimation 134
C.1.3 Step 3 Weights Optimisation 135
C.2 Trials, Errors and Perspectives 136
C.2.1 Guiding the Assignment 136
C.2.2 Optimising Runtime 138
C.2.3 Applying Topology Tracing 139

1. Theoretical Background

1.1. Basic Principles of Protein Structure

The functionality of living organisms is carried out through the interactions of proteins. Different proteins fulfil a huge variety of different functions, each of which is determined by the specific three-dimensional fold of the protein. While there are consequently countless different three-dimensional protein structures, they all are composed of the same building blocks. There are twenty amino acids which, when strung together in a row, build up proteins. The number and order of the amino acids in a protein, the so-called sequence, is specific for each protein.

Figure 1.1 a) shows the fundamental structure of an amino acid. The C_α atom forms the centre of the amino acid. It is bound to an hydrogen atom, an amino group (NH_2), a carboxyl group ($COOH$) and a side-chain which is specific for each of the twenty amino acid types. During protein synthesis amino acids are strung together by eliminating water and building a so-called peptide bond between the carboxyl group of one amino acid to the amino group of another amino acid (Figure 1.1 b)). At the start and the end of the chain the amino group and the carboxyl group stay intact, forming the N-terminus and the C-terminus of the protein. [1] Once an amino acid is part of a chain of two or more amino acids it is referred to as a residue. The conformation of a residue is often described with the dihedral angles Ψ and Φ. While Ψ denotes the angle around the C_α-C bond, Ψ measures the angle around the N-C_α bond. Within the chain of all residues, the atoms building the recurrent motif all residues share (the atoms that are not part of the side-chain) are called the main-chain or the backbone of the protein. The various side-chains emerge then from that common backbone (Figure 1.1 c)).[2] In a first step of computational modelling (more information about modelling of protein structures can be found in section 1.3), one simplifies the protein chain often to a C_α-trace, where each residue is represented only by its C_α-atom. This is illustrated in Figure 1.1 d).

The specific sequence of amino acids forming a protein is considered as the primary structure. Zooming out to local conformations of the protein chain, secondary structure elements become visible. In particular, there are two common dominant motifs on the secondary level of protein structure; the α-helix and the β-sheet. Both are stabilised by hydrogen bonds between the backbone atoms of

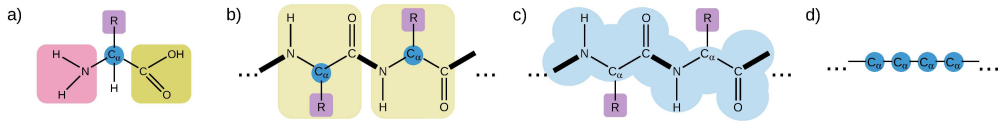


Figure 1.1.: Fundamental protein structure: a) An amino acid consists of a central C_α-atom (blue), an amino group (pink), a carboxyl group (yellow) and a specific side-chain (purple). **b)** In a protein several amino acids are strung together via a peptide bond (shown in bold). **c)** The atoms connected via the peptide bonds are referred to as main-chain or backbone (blue) of the protein from which the various side-chains (purple) emerge. **d)** In modelling a protein chain often is reduced to a C_α-trace.

the participating residues and characterised by a high regularity. In a α-helix the protein chain twists into a coil-like shape, where each turn is formed by 3.6 residues. Each residue in the helix builds a hydrogen bond to the residue located four residues ahead in the chain. The β-sheet consists of several rather straight sections of the protein chain, each single one referred to as β-strand. The strands are arranged next to each other forming a pleated sheet. Here, the stabilising hydrogen bonds are formed between adjacent strands. Irregular sections of the protein chain, which can not be assigned to a α-helix or a β-sheet are usually referred to as loop regions. The global arrangement of secondary structure elements to each other, the overall fold of a protein is considered the tertiary structure. It is mainly determined by side-chain interactions. However, the function of a protein often is not executed by a single chain protein, but from a complex built by several subunits. The quaternary structure finally describes the arrangement of multiple protein chains in such a multimeric complex. [2]

1.2. Experimental Structure Determination using cryo-EM

Now that the theoretical foundations of protein structure have been laid, the next point is how to investigate protein structures in an experiment. This section explains how protein structures can be determined using Cryogenic Electron Microscopy (cryo-EM). While subsection 1.2.1 gives a general introduction into the topic, subsection 1.2.2 provides details about the workflow of a typical cryo-EM experiment. The principle of resolution in cryo-EM and its relevance for protein

structure modelling is elucidated in subsection 1.2.3.

1.2.1. General Information

In 2017 the Nobel Prize in Chemistry was awarded to Jacques Dubochet, Joachim Frank, and Richard Henderson for "developing cryo-electron microscopy (cryo-EM) for the high-resolution structure determination of bio molecules in solution" [3–5]. This event may have been the strongest hint, that cryo-EM has become the state-of-the-art method to experimentally determine the three-dimensional structure of a protein. Common alternative techniques are X-ray Crystallography and Nuclear-Magnetic-Resonance (NMR-) Spectroscopy. However, both approaches require large amounts of sample, X-ray Crystallography is only applicable to proteins, which can be crystallised and NMR-Spectroscopy is limited to small proteins. [6]. Cryo-EM does not only overcome these drawbacks, but also entails structure determination near native conditions in an aqueous solution. Moreover, it enables researchers to investigate dynamic and heterogeneous samples [7] as well as to perform *ex vivo* or even *in situ* [8] studies, where the proteins are directly extracted from tissue or even imaged within their cellular environment, respectively. For a long time, though, cryo-EM was outperformed by X-ray crystallography and NMR spectroscopy in terms of resolution. Yet around 2013, the development of direct electron detectors as well as novel image processing algorithms initiated the so-called resolution revolution [9] enabling structure determination at a resolution that allows direct interpretation of the data in terms of an atomic model [6]. Finally, a new milestone was achieved in 2020 when Yip et al. [10] and Nakane et al. [11] published structures of Apoferritin gaining a resolution better than 1.3 Å, where individual atoms become visible.

The fundamental principle of imaging using cryo-EM is straight forward. In short, electrons are emitted from an electron gun, focused by electromagnetic lenses and accelerated by a voltage between 80 keV and 300 keV [12]. When the electron beam transmits the sample, some electrons are scattered by the Coulomb potential of the atoms in the sample, while others pass through without interacting with the sample. Scattered and unscattered electrons interfere, are detected by the electron detector and form a two dimensional projection image of the sample. Only elastically scattered electrons contribute to the image formation, inelastically scattered electrons cause radiation damage and can destroy the sample. [13] Here comes the *Cryo*-part of cryo-EM into play. To protect the sample of radiation damage it is vitirified, i.e. frozen in amorphous ice before being imaged under cryogenic conditions. To prevent interaction between electrons and air molecules, the electron microscope has to be vacuumed.

In general, one can differentiate three different modalities in cryo-EM: Tomography, crystallography and single particle analysis. Tomography is usually used to image whole cells or large molecular complexes. Here, several tilt images are merged into a three dimensional representation of the structure. If a large amount of sample is available and the protein can be crystallised, electron crystallography, also referred to as electron diffraction, can be used to achieve a high resolution reconstruction by averaging many fairly identical entities of the protein evenly arranged in a crystal. Lastly, in single particle analysis many projection images showing the protein in different orientations are aligned, averaged and combined to reconstruct the three dimensional protein structure. [12] In this thesis, only data acquired by single particle analysis is relevant. The individual steps of the workflow of single particle analysis are described in more detail in the next section.

1.2.2. Workflow of cryo-EM Single Particle Analysis

The typical workflow to conduct a cryo-EM experiment consists of four major steps, visualised in Figure 1.2. If not stated otherwise, the following descriptions are based on [14] and [15].

Sample Prepapration In a very first step the protein sample has to be purified. The pure sample is then transferred onto a grid, consisting of a metal frame and a holey carbon film. Ideally, the protein sample should be uniformly distributed within the grid holes and the proteins within the solution should be oriented randomly, revealing many different views of the protein structure. Next, the grid is frozen in liquid ethane. Performing the freezing process rapidly prevents the formation of ice crystals which would interfere with the electrons in the microscope. Instead, the ice is in an amorphous, glass-like state. It captures the proteins in their native conformation, still allowing conformational heterogeneity, and provides protection against radiation damage. The sample preparation is illustrated in Figure 1.2 a).

Data Acquisition For data collection, the grid is inserted into the microscope. After a screening process, grid holes are first selected based on protein-, or particle-, concentration as well as ice thickness and then imaged. The recorded images of the selected holes are called micrographs and contain usually several 2D projections of the particles (see Figure 1.2 b)). However, due to the radiation sensitivity of the sample, only very low electron doses, about 20-40 electrons/A^2, can be used for imaging [17]. Thus, the individual micrographs are very noisy and cannot directly resolve the atomic protein structure, but need further processing.

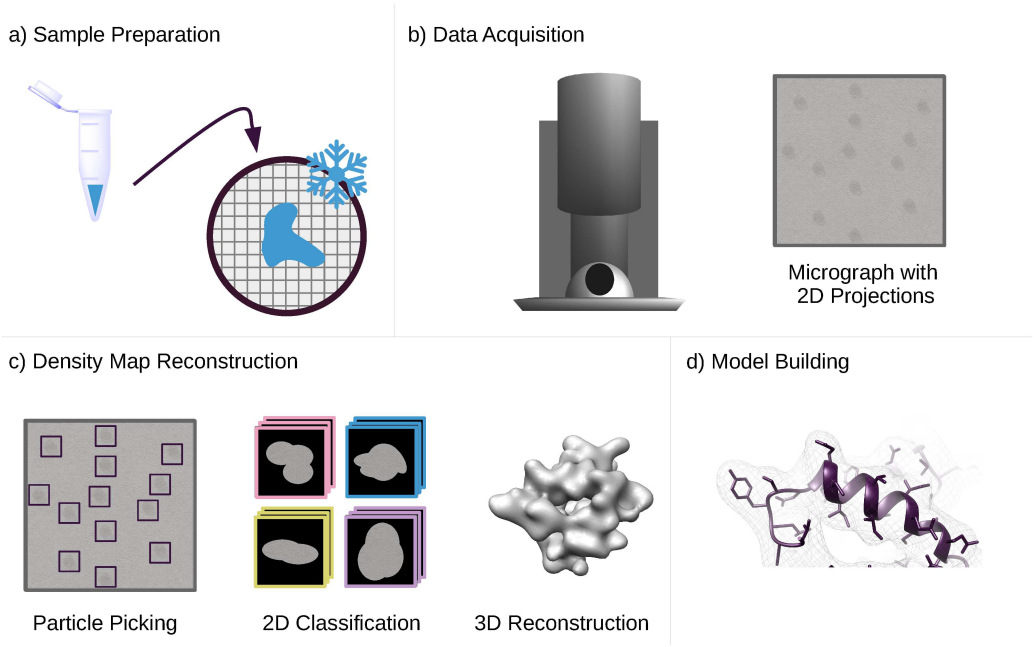


Figure 1.2.: Single particle cryo-EM workflow: A cryo-EM experiment consists of several steps. It starts with the sample preparation (**a)**) where the sample is purified, applied on a grid and rapidly frozen. The frozen grid is imaged with a microscope and several, so-called micrographs are recorded (**b)**). The noisy 2D projection images are combined into a three dimensional density map (**c)**) and finally an atomic model can be built into the map (**d)**). The figure is inspired by [16].

Density Map Reconstruction As described above, the micrographs contain several projections showing the protein of interest in various orientations. To extract the structure information from these micrographs, one needs to extract the individual particle representations from the micrographs by boxing out the corresponding areas in the micrograph. This process is referred to as particle picking and depicted in Figure 1.2 c) (left). Next, these particle images (usually in the order of 10^5) are aligned to each other and classified based on structural features, so that each class should contain a specific orientation or view of the protein (2D Classification, Figure 1.2 c) centre). Particle images within a class are averaged to improve the signal-to-noise-ratio. The Fourier-Shell-Theorem states that for each three-dimensional object, the Fourier transform of a two-dimensional projection image (of this object) is a central slice through the Fourier transform of the three-dimensional object [12]. Hence, by determining the relative orientations of the projection images, one can combine them to a three-dimensional structure (Figure 1.2 c) right). This is done by an algorithm, called projection matching, where the projection images are compared to re-projections of a three dimensional initial model (different approaches to generate an initial model are described in [18–20]), assigned to a best matching orientation, and used to update the 3D model in an iterative manner [17]. Eventually, the assigned orientations converge and the resultant three-dimensional model can be interpreted as representation of the protein structure. Although the 3D representation does not reflect the electron density but the electrostatic potential of the atoms, it is referred to as density map.

Model Building The purpose of a cryo-EM density map usually is to be interpreted in terms of an atomic model as depicted in Figure 1.2 d). In general one can distinguish two different approaches for density map interpretation: structure fitting and de-novo modelling. In the first case, if the general structure of the protein of interest is known and the cryo-EM map only represents another conformation of the protein, the known structure can be refined into the map, such that the fit of the structure to the map is maximised. DireX [21, 22] and Phenix [23] are common software packages to perform this task. In the second case, if no information about the atomic structure is available, the model has to be built into the map from scratch. This can be done manually in Coot [24]. However, manual model building is difficult and time-consuming, particularly in density maps with a resolution worse than 3 Å. Therefore, automatic modelling tools have been developed to facilitate this process. Some are the subject of this thesis. More information about existing software for automatic model refinement and model building is given in section 1.3.

1.2.3. Resolution of cryo-EM Density Maps

Every published cryo-EM map is attributed with a value describing its resolution, indicating its general quality. However, while most people have an intuitive understanding of resolution, as 'sharpness' or the degree to which details can be differentiated, it is not straight-forward to define resolution for cryo-EM maps. Usually it is computed with a consistency test between two independent 'half maps': One splits the data set of picked particle images into two random sets and reconstructs a map for each data set. Both maps are compared as a function of spatial frequency by calculating the Fourier Shell Correlation (FSC). [25] The frequency at which the FSC drops below 0.143 is then interpreted as estimate of the resolution of the cryo-EM map reconstructed based on all particle images [26, 27]. Estimating the resolution via the FSC criterion indicates that resolution is an isotropic property of the whole map. However, resolution might vary between different regions of the map, such that a flexible loop is worse resolved than a static protein core [28, 29].
Resolution has a great impact on model building. Maps obtaining resolutions better than 3 Å show clearly distinguishable side-chain densities and an unambiguous main-chain topology [30]. Thus, model building is straight-forward and can easily be performed manually. This resolution range will be referred to as high resolution or atomic resolution in the following. In the medium resolution range (3 Å to 5 Å) side-chains may not always be visible and even the identification of the correct topology may be hampered by branching or breaking. For this situation computational tools for automatic model building have been developed to support a modeller in the process of model building (see section 1.3). If the resolution of a map is worse than 5 Å (low resolution), secondary structure elements may be still identifiable [31], but de-novo model building is not possible [30, 32]. Nevertheless, refinement methods can still be used to refine a known structure into a low-resolution map.

1.3. Structure Modelling based on cryo-EM Data

The overall goal of a cryo-EM experiment is to determine the three-dimensional structure of a protein. But, while the computational steps to reconstruct a density map from the micrographs are usually performed within a single software framework, e.g. RELION [33] or cryoSPARC [34], automatic model building remains a separate task and field of research. This section gives an overview about existing approaches and software packages for automatic model building. As described above, there is a conceptional difference between tools which specialise on optimising and adapting an existing model into a cryo-EM map and other methods that

perform the modelling based on the density map alone without knowledge about a prior determined structure. The choice of which of the two approaches to follow depends on both the information available and the resolution of the map. [32] Both classes are elucidated in subsection 1.3.1 and subsection 1.3.2, respectively. Lastly, subsection 1.3.3 deals with methods to validate or to assess the quality of a built model.

1.3.1. Fitting and Refinement Methods

Fitting and refinement methods optimise an existing model into a cryo-EM map, such that the fit or overlap between the model and the map is maximised. If no structure of the protein is available, one can fall back to so-called homology modelling [35], or more general, to structure prediction methods, i.e methods which predict the three dimensional structure of a protein only based on its sequence, to generate a structure which then can be adapted to the cryo-EM map.

The least complex way of adapting a protein structure to a density map is rigid-body fitting. Here, only a translational and rotational search is performed to find the best positioning of the structure in the map. A common tool for rigid fitting is UCSF Chimera's fit-in-map method [36].

During flexible fitting or refinement the model is allowed to slightly deform, bend or relax, so that it better fits into the density map [31]. The real space refinement software DireX is specialised on medium- to low-resolution density maps and aims to only adapt those degrees of freedom for which the density map actually provides information and to keep all other degrees of freedom as close to the initial structure as possible [21]. The geometry based, efficient CONCOORD algorithm [37] is used to generate a conformational ensemble of structures. This ensemble is influenced by two forces, a force moving atoms into the density map and a counteracting force, mediated through deformable elastic network (DEN-) restraints, which prevents overfitting by biasing the ensemble towards the input structure. By balancing both forces, the structure is stepwise refined into the map. The Molecular Dynamics Flexible Fitting (MDFF) method [38] combines, as the name suggests, Molecular Dynamics Simulations with flexible fitting. To move the structure into the map, but simultaneously ensure the preservation of secondary structure elements, two external potentials are added to the MD force field. While the MD force field describes the bonded and non-bonded interactions between the atoms in the protein and therefore retains the stereochemical correctness, the potential derived from the EM data applies forces proportional to the density gradient and the second external potential acts through harmonic restraints between dihedral angles in residues participating in a helix or a sheet.

During the simulation the protein should adapt a stereochemical correct conformation that is in accordance with the cryo-EM data. Another common choice for refinement is the phenix.real_space _refine tool [23] of the PHENIX Suite [39]. Here the atomic coordinates are minimised against a target function, considering the cryo-EM data as well as stereochemical restraints, with the L-BFGS algorithm [40]. To take into account, that at lower resolutions the experimental information may be insufficient to preserve all geometry characteristics of the structure, not only standard restraints on bond-length, bond angle and dihedral angles can be applied, but additional information, like symmetry restraints, Ramachandran plot restraints or restraints on common side-chain conformations, so-called rotamers, can be included.

1.3.2. De Novo Modelling Tools

The aim of de novo modelling is to construct a model directly from a density map without the aid of a structural template [31]. Since this is a challenging task, it usually requires better resolutions than fitting methods [30, 31]. Many computational tools for automatic de novo modelling share a general rough procedure to interpret a density map: In the first step some kind of base points, often interpreted as approximate C_α positions, are identified and connected to a trace. This trace defines the global topology of the protein. The following step is to convert this C_α-trace into a full atom main-chain. Finally, the sequence is assigned to the backbone and side-chains are added accordingly. Representatives of these class of modelling tools are for example PATHWALKER [41], MAINMAST [42], the phenix.trace_and_build method [43], EMFASA [44] and also the deep learning based approach DEEPTRACER [45].
In more detail, PATHWALKER seeds as many pseudo atoms in high-density regions as there are residues in the sequence and connects them with help of a Travelling Sales Person Problem (TSP) solving algorithm [46, 47], which prefers connections traversing high density regions. Identification of secondary structure elements and removing non-protein like features help to refine the path. Completion to a full atom structure is not included in the PATHWALKER procedure. The MAINMAST approach is based on local dense points, connected by a minimal spanning tree, minimising their spatial distance. Applying a threading score which considers the matching of the volume of amino acids to the density at the local dense points converts the tree structure into a C_α-trace. To build the full-atom structure, MAINMAST employs the software PULCHRA [48]. Furthermore, the PHENIX method first identifies regions of the map representing contiguous fragments of the protein structure, models the C_α-trace of those fragments and subsequently completes the fragmented traces to full atom structures and merges them into a consensus structure. DEEPTRACER employs convolutional neural networks to pre-

dict confidence maps describing the probability of a certain voxel of the density map to contain a part of a certain secondary structure element, a part of the backbone or a C_α-atom. Those confidence maps are then used to build the C_α-trace. Here too, the sequence is mapped to the trace in a subsequent step to complete the trace into a full atom structure. Finally, EMFASA first generates a C_α-trace which determines the topology of the protein by connecting beads based on a TSP solver. Then a fragment library, consisting of sequence non-specific fragments, each seven residues long and embodying a common backbone conformation, is rigidly fitted at the C_α positions of the trace. Subsets of nicely fitting fragments are identified and then assembled into a consensus backbone. The side-chains are added with help of a profile reflecting the fit of each of the 20 amino acids at each backbone position. EMFASA will be described in more detail in section 5.1.

Wang et al. [49] follow a slightly different approach and deviate from the procedure to first build a C_α-trace and completing it afterwards. Their workflow is based on the assumption, that local similarity in sequence is accompanied by local similarity in structure. Segments of solved protein structures with local similar sequences are fitted in the map and well matching fragments are assembled to form a complete protein structure. It should be noted that, in contrast to the EMFASA approach, the fragments used in the Rosetta approach are sequence-specific and contain side-chains from the beginning.

1.3.3. Model Validation

Once a model is built, it has to be assessed and validated [30]. A validation can answer several questions: Is the geometry of the model biophysically correct? How well does the model explain the experimental data? Or, how similar is the modelled structure to a high-resolution crystal structure, which may be considered as ground truth?

The MOLPROBITY Score [50], originally developed for validation of models derived from X-Ray crystallography data, provides a common way to answer the first question. It considers several geometrical features in its analysis. In an all-atom contact analysis, the structure is checked for clashes, i.e for contacts between unbound atoms whose overlap of van-der-Waals radii exceeds 0.4 Å. The clash score gives the number of clashes per 1000 atoms. Steric hindrance also regulates the torsion angles of the backbone, such that only certain combinations of Φ and Ψ angles are allowed [51]. MolProbity compares the torsion angles of the backbone with a reference distribution of torsion angles from very high resolution structures. Residues, with an unlikely combination of the backbone torsion angles are considered as 'Ramachandran outliers'. In a similar manner, unusual side-chain conformations, so called rotamer outliers are detected. The number of clashes,

Ramachandran outliers and rotamer outliers are collected in a single MolProbity Score, which reflects the crystallographic resolution at which those values would be expected.

Another measure of model quality is, often behaving in some level antagonistic to geometric quality, the fit of the model to the experimental data. When validating the agreement of a model to a map, special attention must be paid, to the problem of overfitting, i.e the fitting to noise instead of signal. Only by using independent data, in other words data that has not been used for the optimisation of the model-to-map-fit, the concordance of model and map can be evaluated. [30]. Falkner and Schröder [52] developed a cross validation approach for cryo-EM based modelling. During cross validation the experimentally data is split into two independent data sets, one used for optimisation, the other one for validation. Cross validation approaches has been developed before for X-ray crystallography data. However, these methods randomly select about 10% of the spatial frequencies present in the data for validation. In cryo-EM though, randomly chosen spatial frequencies might correlate with each other and the two data sets would not be independent. Therefore, Falkner and Schröder spilt the spatial frequencies of the cryo-EM density map into two continuous bands of frequencies: A continuous band of higher spatial frequencies, the 'free band', is used for validation, the lower frequencies, the 'work band' are used for the DireX refinement of the model to the map. The cross correlation C_{work} between the frequencies within the work band of the cryo-EM map and of a density calculated from the model is optimised during the refinement. The C_{free} value, though, the cross correlation between the cryo-EM map and the model map for spatial frequencies in the free band, is then an independent measure of the model-to-map fit. [52]

When developing a new modelling tool, the aim is to show that the tool can help to build topologically correct and accurate models. This is usually done not (only) by assessing the model's geometrical features and its fit to the density, but by comparing it to an already known high-resolution structure of the protein of interest. In that sense, also structure comparison measures can be employed for validation. The most common measure of similarity between two structures is the Root Mean Squared Deviation (RMSD). The RMSD is the average spatial distance between n pairs of corresponding atoms of two superimposed structures:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=0}^{n} d_i^2}$$

where d_i is the distance between the atoms in the ith pair. [53] While the RMSD is a simple and straight forward metric to assess accuracy of atom placements it

also has some disadvantages. Firstly, it requires a one-to-one mapping between atoms of the the two structures to be compared and secondly, it is quite sensitive to local errors, as they would occur due to conformational variations. For example, a deviating conformation of a loop region in an otherwise perfectly match between the structures may be biophysically reasonable and in agreement with the data, but would result in a significant increase in the RMSD [53]. The second pitfall can be circumvented (e.g [45]), by including only pairs of atoms in the calculation whose distance is beneath a certain threshold (in the context of C_α-atoms this threshold is often set to around 3 Å). To complete the picture, the so-called structure overlap then indicates the percentage of atoms included in the calculation. A comparison more of the overall fold and less of error-free and accurate atom positioning is provided by the CLICK score [54]. CLICK aligns the two structures based on local structural similarities, but does not require a global matching or sequence alignment. Similarity is measured by the RMSD and structure overlap, here defined as fraction of atoms that are within a 3.5 Å to their corresponding atom, but also by the topology score. The topology score is 1 for topologically identical structures and 0 for topologically complete dissimilar structures.

2. Objective and Outline

Knowledge about protein structures is essential for our understanding of the physiological functioning of living organisms, the development of drugs against protein associated diseases as well as the development of new bio-based materials.

In the last years cryo-EM has become the state-of-the-art method for experimental protein structure determination. It enables capturing proteins near native conditions, only requires limited amount of sample and is applicable to a wide variety of different proteins.

However, analysing cryo-EM data is complicated and requires several steps. After a purified protein sample has been imaged in the microscope, the many acquired projection images can be used to reconstruct a three-dimensional density map. This map, though, is a quite complex representation of the protein structure itself and the interpretation of such a map is, particularly for lower resolutions, challenging. Computational tools are needed to extract the whole spectrum of structural information given by a cryo-EM density map, ranging from details in side chain conformations varying within different conformations of the same protein to fundamental and essential properties of a protein like the global topology of its fold.

In this thesis a top-down approach is followed to stepwise deepen the understanding of how information about protein structure is encoded in a cryo-EM density map. Four different studies are presented, each elucidating a computational method for structural interpretation of cryo-EM maps, but settled at a different point of the transformation between density map and optimised atomic model. Figure 2.1 illustrates the structure of this thesis. It leads against the typical modelling workflow, but follows the path towards a fundamental understanding of the relation between model and map from a computational perspective, characterised by less and less prior information being available.

In the first study, described in chapter 3, we have high-resolution density maps and a well fitting corresponding crystal structure at hand. We will see, how we can optimise this structure into the maps and how we can regard that a cryo-EM density map entails not only information about a single structural conformation but that a whole ensemble of structures contributes to the map.

The fold of the protein examined in the second study, presented in chapter 4, is part of a special kind of protein folds, so-called amyloid fibrils. A map with a resolution of 4.2 Å is available as well as a model of the backbone. However, the

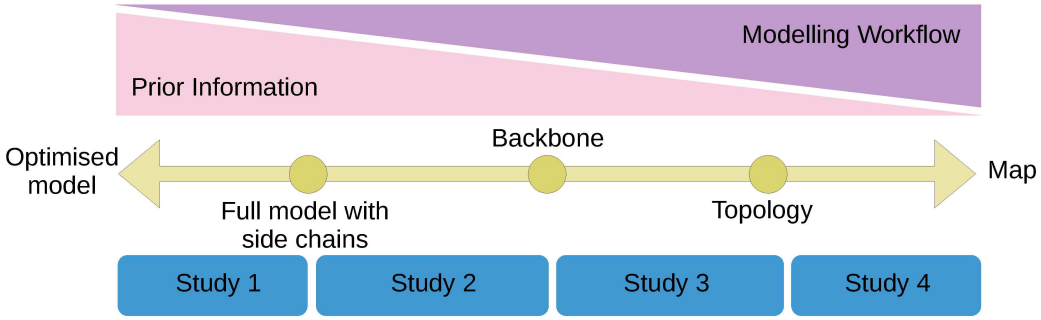


Figure 2.1.: Structure of this thesis: Atomic model and cryo-EM map can be perceived as end states of a transformation (yellow arrow). The studies (blue boxes) presented in this thesis are settled at different points of the transition. The order of the studies is not oriented at the typical modelling workflow (purple), but in a top to bottom manner, starting near the model site of the transition, and proceeding stepwise towards the map site to deepen the understanding of how structural information is encoded in a density map. This progression is characterised by a decrease in required structural input information (pink).

map does not allow for an unambiguous assignment of the side-chains. We present a method to automatically sample and rank many different side-chain assignments to understand the underlying amyloid structure on a more profound level than only given by its backbone.

The subject of the third study (chapter 5) is a method to flexibly fit fragments of known backbone conformations into a cryo-EM density map given only the map and a rough C_α trace defining the topology of the protein. Such a procedure can be integrated into modelling frameworks, like for example EMfasa to automatically build a protein backbone.

Lastly, there is no prior structural information derived from the map available in the forth study (chapter 6). Here, we present an integrative approach to determine the fundamental topology of a protein based on the map, but employ computationally predicted inter-residue distances to guide the tracing. We show that predicted inter-residue distances can serve as additional source of information in situations where the information from the density map alone is not sufficient to understand the corresponding model architecture.

In the end, chapter 7 summarises the findings of all studies and explains which conclusion can be drawn about the relation of a cryo-EM map and the underlying structural model.

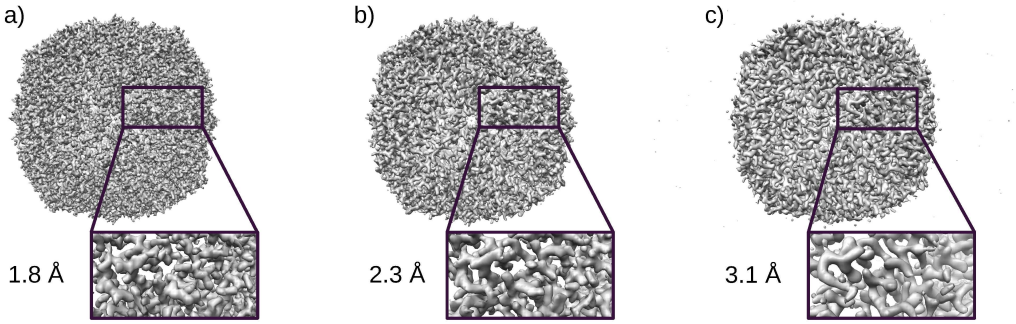
3. Structure Optimisation by Capturing Conformational Heterogeneity in the EMDB Model Metrics Challenge

This chapter describes our contribution to the EMDB Model Metrics Challenge 2019. It entails an approach to adapt a crystal structure as accurate as possible to a series of cryo-EM maps of Apoferritin with resolutions of 1.8 Å, 2.3 Å and 3.1 Å, respectively. Our idea was to not only morph the available structure into the density map, but to regard the conformational heterogeneity of the ensemble represented by the cryo-EM map.

The chapter is structured as followed: First, a short introduction is given in section 3.1, explaining the setting and the motivation of the model challenge. In the next section (section 3.2) the available material is described and our method for refining the crystal structure into the density map, while regarding the conformational heterogeneity embodied in this very map, is explained in detail. The general results of the model challenge have been published in a research article in *Nature Methods* [55]. Section 3.3 summarises the article, but also describes how well our method has performed in the challenge.

3.1. Motivation

The EMDataResource Project (EMDR) (emdataresource.org) is a global platform for archiving and retrieval of cryo-EM data, but also for community events, news and discussions. The goal of the EMDR is to establish data validation and data deposition standards through community consensus to ensure a reliable quality of published data. [56] One of the main approaches to set benchmarks are biannual community challenges, where participating community members solve a modelling task, submit their results and discuss methodologies, problems and future prospects. In 2019 the EMDR Model Metrics Challenge took place. This challenge set the focus on the evaluation of the quality of models generated by current software tools, but most of all on the comparison of different validation



[h]

Figure 3.1.: Targets of the EMDR Model Metrics Challenge 2019: Apoferritin at **a)** 1.8 Å (EMD-20026), **b)** 2.3 Å (EMD-20027) and **c)** 3.1 Å (EMD-20028) resolution. Close-ups reveal the differences in resolved atomic detail.

metrics. Scores assessing the geometry of the models, measures of model-to-map fit as well as tools to compare the models with each other or to a reference structure were considered. Challengers aimed to 'build the best quality model possible given the map data', while assessors had the task to 'decide what metrics are best for comparing models'. [57]

3.2. Material and Methods

3.2.1. Material

The targets of the Model Metrics Challenge included a series of cryo-EM maps of Apoferritin [58]. The maps differed only in the number of particles considered in the reconstruction, and thus in their final resolution [55]. Figure 3.1 shows the target maps EMD-20026, EMD-20027 and EMD-20028 yielding a resolution of 1.8 Å, 2.3 Å and 3.1 Å, respectively. In each case, a closeup illustrates the resolved atomic detail. Apoferritin is a complex of 24 identical subunits, each consisting of 183 residues folded into 5 alpha helices. Since no high-resolution cryo-EM structure was available, a 1.5 Å crystal structure [59] served as reference (PDB-ID[1] 2FHA). Maps and reference structure did not originate from the same experiment and hence, the fit between model and map was very good, but not fully optimised [55].

[1]PDB refers to the RCSB Protein Data Bank [60].

3.2.2. Methods

In view of the situation, that firstly, a high-resolution crystal structure was available and that secondly, the 1.8 Å map in particular revealed not only the general fold but specific conformational details, we decided to focus on refinement and optimisation of the crystal structure, rather than building a de-novo model.

We followed the same procedure for all three maps: In order to make use of the symmetric properties of the target we first zoned the provided map 4 Å around the asymmetrical subunit and applied the optimisation steps only on one subunit. We then rigidly aligned the crystal structure to the cropped map using Chimera. For the refinement, we split the data into two sets. Lower spatial frequencies were used for optimisation, higher spatial frequencies were left for validation. Therefore we filtered the maps to 2.0 Å (EMD-20026), 2.4 Å (EMD-20027) and 3.3 Å, respectively. The crystal structure already had a good fit to the density, but originated from another experiment and hence represented a slightly different conformation. Other than X-ray crystallography, a cryo-EM experiment mirrors not only one conformation but an average of an ensemble of differed conformations. To sample the conformational heterogeneity embodied in the cryo-EM data, we ran a MDFF simulation for 10 000 ps with default parameters and additional Berendsen pressure control [61]. For each pico second of the simulation, a frame was extracted, resulting in 10000 frames. Using DireX, we evaluated the agreement between model and map for each frame with the C_{free} value, the model-map cross-correlation based on the higher spatial frequencies not used in the MDFF refinement. Only the 1000 frames associated with the best C_{free} values were averaged. However, building an average structure of different conformations, can lead to unlikely geometrical features in the structure. That is why we performed an subsequent energy minimisation of the average structure with CNS [62, 63]. The strength of the position restraints applied during the minimisation were chosen, such that the MolProbity score was optimised. A measure to quantify conformational heterogeneity or, from another perspective local resolution, at an atomic level are atomic displacement parameters, so-called B-factors. Atoms with a position that varies among conformations are associated with a large B-factor, static atoms with a smaller B-factor. To consider those differences in the submitted model, we performed a B-factor fitting with Phenix. Finally, we aligned the optimised structure to the protomer map containing all subunits.

3.3. Results and Discussion

3.3.1. Publication 1: Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge

The results and outcomes of the Model Metrics Challenge have been published in *Nature Methods* in 2021. The article is summarised below in the first paragraph of this section and can be found in full length in Appendix A. As the publication sets another focus than this thesis, the second paragraph of this section elucidates the results relevant for the presented method in more detail.

Summary

The 2019 Model Metrics Challenge aimed to find general recommendations about new validation standards for cryo-EM model building. Four target maps were provided, three of them as a resolution series originating from the same experiment and all of them representing the state-of-the-art for cryo-EM single particle reconstructions at the time of the beginning of the challenge. Thirteen teams from Europe and the USA submitted 63 models in total, there were no instructions which software to use. All submitted models were evaluated in four tracks: model geometry, fit-to-map, comparison-to-reference-model as well as comparison-among-models. Various metrics were applied in each track. In general, most models yielded high scores in each of the tracks, most times improving the reference model. Frequent errors related to peptide bond geometry or sequence alignments. The evaluation of fit-to-map metrics revealed a poor correlation between the different metrics. Moreover, map-specific factors or user-chosen parameters as for example background noise or chosen density threshold, seemed to strongly affect the score values of many metrics. Metrics validating a models' geometry were found to be mostly orthogonal to each other, suggesting the use of multiple scores to identify all geometry issues in a model. In contrast, comparison-against reference and comparison-among-models metrics showed a strong correlation to each other. All findings were summarised in four general recommendations: Firstly, nearly all fit-to-map metrics can be used to monitor progress in the optimisation of model into a single map. Secondly, when examining local fit-to-map one should use metrics that perform the evaluation per single residue instead of metrics considering windows of several residues. Thirdly, for an archive-wide ranking of fit-to-map scores a metric which is insensitive to map background noise and not dependent on estimated input parameters is required. Examples of such metrics are the map-model FSC [27, 64], the global EM-Ringer Score [65] or the Q-score [58]. Fourthly and finally, special attention should be given to backbone

conformation errors, CaBLAM [66] could be exemplary used for that.

Performance of the described modelling method

Our submitted models scored satisfactory in all tracks. Ranked by the combined Z-score considering all metrics with an equal weight, our model for EMD-20026 took place 13 of 33, the model for EMD-20027 finished in place 9 of 31 and the model for EMD-20028 came even 5th of 31. [67] These results reveal, that the chosen method is suitable for the lower resolution in particular.

Improvements can be sought concerning the interfaces between asymmetrical subunits. Focusing the refinement only on one subunit did not take any interactions between subunits into account. This resulted in not accurately modelled residue-residue contacts between different subunits reflected in a poor PRO-Q Score [68]. Instead, refinement should be performed on the whole complex (at the expense of running time, though) or changes in side-chain conformations due to inter-subunit interactions should be taken care of in an additional step after refinement of the individual subunits.

Contribution

For this study, I participated in the challenge as modeller and contributed to the method development, preprocessed the cryo-EM maps, ran the MDFF simulations, performed the DireX validation and computed the average structures.

4. Automatic Side-chain Sampling for Structure Elucidation of IAPP Fibrils

The subject of this chapter is a method we developed in the context of a study about the atomic structure of IAPP fibrils. It describes a procedure to computationally sample and rank many different side-chain assignments for a single backbone. This can be of help in situations where a cryo-EM map allows for a more or less straight-forward building of the protein main-chain, but the resolution of the side-chains is not sufficient for an unambiguous sequence assignment. We present an approach to deal with that situation and describe how to extract as much structural information from the density map as possible.

The chapter begins with a short presentation of a few theoretical fundamentals (section 4.1) about amyloid fibrils and the prediction of side-chain conformations. Next, section 4.2 clarifies the motivation for the study. The methods for the automatic side-chain sampling are explained in section 4.3. Our findings about the structure of IAPP fibrils have been published in *Nature Structural and Molecular Biology* [69]. The first part of section 4.4.1 summarises the article, which offers a more biological perspective on the study, while the methodical and computational perspective on the results is given in the second part of section 4.4.1.

4.1. Theoretical Background

4.1.1. Amyloid Fibrils

A protein chain cannot only fold into its native three-dimensional structure as determined by its sequence. Instead, there is an alternative fold, the amyloid fold, where many entities of the same protein chain stack onto each other forming helical fibrils stabilised by well-ordered β-sheets. Each layer consists of mostly two entities of the same protein chain, or monomeric subunit, interacting with each other through side-chain interactions within the fibril interface. The structural architecture of an amyloid fibril is illustrated in Figure 4.1. While some functional amyloids have been observed mainly in bacteria, fungi and insects [70], amyloid

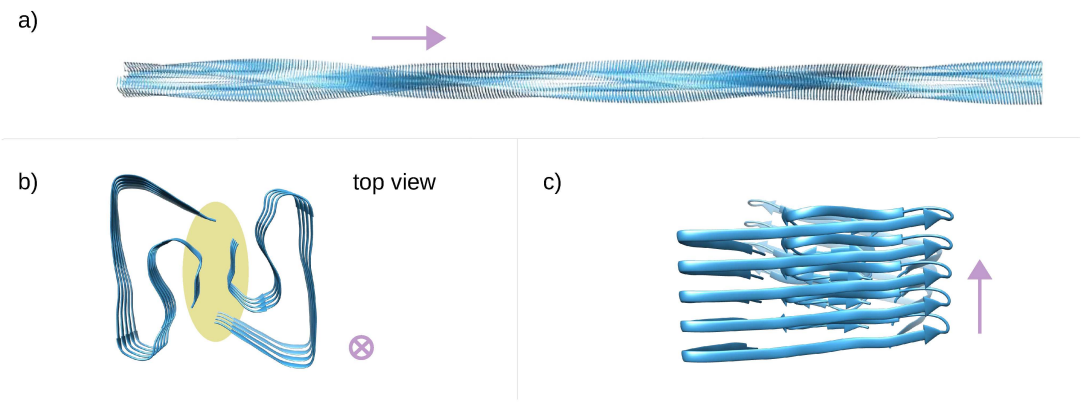


Figure 4.1.: Structural architecture of an amyloid fibril on the example of an Amyloid-β fibril [72]:
a) Macroscopic view of a fibril. **b)** Top view of a fibril showing its cross-section comprising two monomeric subunits. The fibril interface is depicted in yellow. **c)** Side view of a fibril. Layers form β strands arranged in a β sheet elongated along the fibril axis. Bright purple arrows denote the direction of the fibril axis in each panel. Figure inspired by [73].

fibrils are most infamous for their association with protein misfolding diseases [71]. For example, fibrils of Amyloid-β have been linked to Alzheimer's disease, Parkinson's disease is associated with α-synuclein fibrils and fibrils of the islet amyloid polypeptide (IAPP) have been found in patients suffering from type II diabetes [71].

Structural investigations of amyloid fibrils have a long history [74], but a major breakthrough was achieved when exact atomic details were uncovered in the first cryo-EM structures of amyloid fibrils, published in 2017 [72, 75]. Since then, cryo-EM has become the go-to-method for structure determination of amyloids and over 60 amyloid structures have been elucidated by cryo-EM so far. An overview about these structures can be found in [76]. Cryo-EM single particle analysis is exceptionally well suited to deal with a remarkable feature of amyloid samples: polymorphism, i.e. the presence of different fibril folds formed by proteins of identical sequences within one sample [77]. In cryo-EM each of the polymorphs can be processed independently. However, differences in the number of obtained images of each type and different structural features impose different prerequisites for reconstruction and may result in different resolutions of the density maps.

4.1.2. Prediction of Side-chain Conformations with Sqwrl4

SQWRL4 [78] is a program to predict side-chain conformations given a backbone and a sequence. Its workflow is described briefly in the following:
First, coordinates of the backbone atoms are read and the dihedral angles Φ and Ψ are calculated for each residue. Next, a list of possible side-chain conformations, rotamers, is generated for each residue with help of a backbone dependent rotamer library. The library contains frequencies of the rotamers as well as the mean angles, defining the conformation, and their standard deviations over a discrete Φ/Ψ grid. The resultant coordinates of the side-chain atoms are computed and each side-chain is associated with a bounding box which determines the interaction radius of the side-chain. Further, each rotamer is attributed with an energy term considering firstly its self-energy, dependent on its frequency and the interactions to the backbone atoms, and secondly a pair energy term specifying the interactions between side-chains possibly assigned to different residues. The pair energy regards van-der-Waals interactions as well as hydrogen bonding. Finding the optimal set of rotamers associated with the lowest total energy is a complex combinatorial problem. SCWRL4 approaches it by converting the data into a graph. Residues are represented by vertices, edges indicate possible interactions between residues. A combination of a dynamic programming algorithm and tree decomposition methods solves the graph and identifies the optimal assignment of side-chain conformations to the residues. [78]

4.2. Motivation

The human islet amyloid polypeptide (IAPP) is a small protein consisting of only 37 residues. Its physiological function is assumed to be involved in the glucose metabolism, the control of gastric emptying and the regulation of satiety. However, fibrillar aggregates of IAPP seem to play a decisive role in type II diabetes where they are associated with dysfunction and death of pancreatic beta-cells. Detailed knowledge about the structure of IAPP fibrils could deepen the understanding of the mechanism of amyloid formation, but also help to develop fibril growth inhibitors and soluble, non-toxic IAPP analogs. [79]

We applied cryo-EM to elucidate the structure of human IAPP fibrils grown at physiological pH. We identified three different polymorphs in our data set. While the reconstructed density of the most dominant polymorph, PM1 in the following, provided sufficient details to manually build an atomic model, the number of images of the rarest polymorph, PM3, was limited, so that the reconstruction only yielded a resolution of 8 Å. Thus, model building was not possible here. With the second polymorph, PM2, the situation was more complex. The nominal resolution

of 4.2 Å was the same as for PM1, but the assignment of the side-chains was not clear.

In this study, we present an automatic side-chain sampling and -ranking approach for the structural interpretation of the density map of PM 2. It is shown, how computational automatic modelling tools can help to deal with difficult prerequisites for model building in the case of ambiguous data. We demonstrate how they can be used to gain a more profound insight into the underlying model architecture and how they provide quantifiable results where manual model building is not reliable.

4.3. Methods

Model building for PM2 was not straight-forward. The reconstructed density map of PM2 yielded a resolution of 4.2 Å and displayed the course of the backbone clearly, but side-chains could not be assigned unambiguously. Therefore, we manually built the backbone in forward- as well as in backward direction in Coot. The backbone contained 21 residues, the remaining 16 residues were not resolved in the density. The complete sequence of 37 residues comprises 17 snippets of a length of 21 residues. Accordingly, we performed 17 sequence assignments for each backbone using Sqwrl4, resulting in 34 model in total. Each model was energy minimised with CNS. So far, the information from the density map has not been taken into account in the construction of the side-chains. Hence, we performed a refinement of all models in DireX. Spatial frequencies in the resolution range of 3.0 Å to 4.0 Å were not used in the refinement, but for the estimation of the fit-to-map in means of the real-space map correlation coefficient C_{free} (see also section 1.3.3). The C_{free} value then was used to rank the models. The best scoring model was further refined using MDFF, coordinates of the MDFF trajectory were averaged.

4.4. Results and Discussion

4.4.1. Publication II: Cryo-EM structure of islet amyloid polypeptide fibrils reveals similarities with amyloid-β fibrils

The findings of our structural investigation of IAPP fibrils have been published in *Nature Structural and Molecular Biology* in 2020 [69]. The article is summarised below in the first paragraph of section 4.4.1 and can be found in full length in Appendix B. As the publication sets another focus than this thesis, the second

paragraph of section 4.4.1 additionally elucidates the model building for PM2 in the context of this thesis.

Summary

We investigated the structure of IAPP fibrils using cryo-EM. The fibrils were formed in vitro at a physiological meaningful pH of 6.0. Various polymorphs were visible in the micrographs, three of them allowed for reconstruction of a three-dimensional density map.

The most abundant polymorph, PM1, made up for about 90 % of the fibrils. Its reconstructed density map yielded a resolution of 4.2 Å. The atomic model comprising residues 13 to 37 could be build manually in Coot and shows two S-shaped monomeric subunits per layer. The N-terminus is flexible and could not be resolved in the density. The interface of the fibril is formed by a hydrophobic cluster including the sequence motif NFGAIL[1], which has been found to be crucial for fibrillation of IAPP before [80–82]. Superimposing the atomic model of PM1 with fibril structures of Amyloid-β revealed striking similarities. These similarities might be interpreted as a hint for a molecular link between type II diabetes and Alzheimer's disease.

The second polymorph accounted for 10 % of the observed fibrils and is flatter than PM1 with a longer cross-over distance. We reconstructed a three-dimensional density map with a resolution of 4.2 Å. However, structural interpretation of the density was difficult and the assignment of the side-chains was ambiguous. Therefore, we employed automatic modelling tools to build all 34 eligible models and ranked them according to their real-space map correlation coefficient C_{free}. The most probable model according to this criterion shares some structural features with PM1. There are also two monomeric subunits per layer, but here these are not S-shaped but rather extended. The N-terminus again is flexible and cannot be resolved, which is, though, here also true for the C-terminus. Interestingly, the NFGAIL motif is again located at the centre of the fibril interface.

Lastly, we reconstructed a density map for a quite thick fibril, PM3. But the reconstruction was hampered by the limited amount of corresponding projection images in the micrographs, as only about 1 % of the fibrils could be assigned to PM3. Hence, only a resolution of 8.1 Å was achieved, so that model building was not an option. The density indicate a small interface comprising only about 3 residues.

[1] It is common to abbreviate the names of amino acids with a one letter code. NFGAIL describes the succession of amino acids asparagine, phenylalanine, glycine, alanine, isoleucine and lysine.

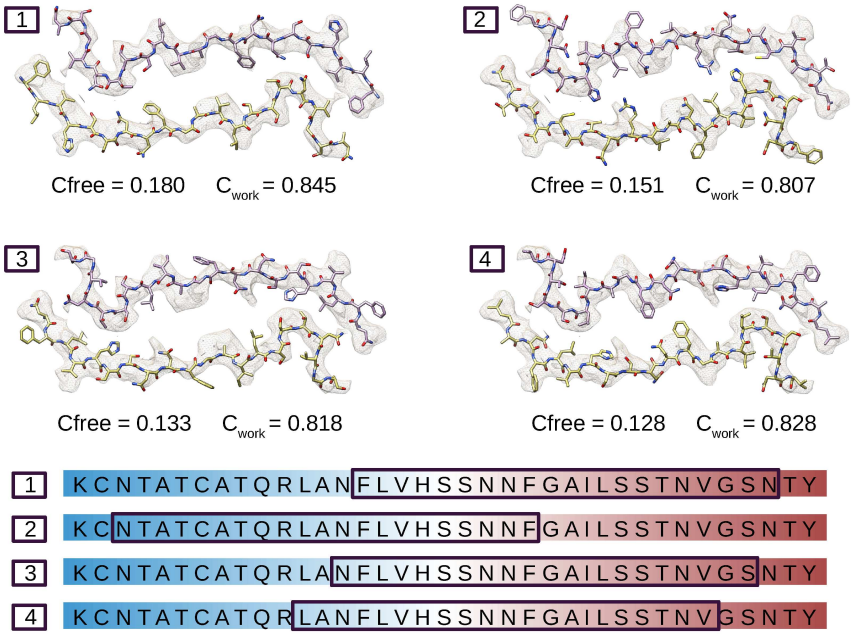


Figure 4.2.: Top four models for PM2: We built 34 models for PM2 and ranked them according to their C_{free} value. The top four ranked models are shown here, together with their C_{free}- and C_{work} values, assessing their agreement with the density map. The model with the highest C_{free} has also the highest C_{work}. On the bottom, the corresponding sequence snippets of the models are shown. The N-terminus of the IAPP sequence is coloured blue, the C-terminus red. Adapted from [69].

Structural Interpretation of PM2 based on Automatic Modelling

Structural interpretation of the density of PM2 was challenging and manual model building only was possible for the backbone, but not for an unambiguous side-chain assignment. Therefore, we sampled all possible sequence assignments using SQWRL4, ending up with 34 models of PM2 in total. To decide which of the models is the most probable, we ranked the models according to their C_{free} value, or generally speaking, according to their agreement with the map. Geometrical correctness of all models was ensured with a energy minimisation using CNS.

Figure 4.2 shows the top 4 ranked models and the corresponding sequence snippets. Model 1 has the highest C_{free} value as well as the highest C_{work} value. C_{work} describes the cross correlation between model and map in the lower spatial frequencies used in the refinement. Model 2 has also a high C_{free} value and seems to fit well in the map for the most parts. However, the Thr4 fills the side-chain density not ideally. Moreover, model 2 is unlikely to be true due to the missing disulfide bond between Cys2 and Cys7. Further, the aggregation prone sequence

motif NFGAIL is not included in model 2. Model 3 and 4 show also some imperfections in the model-to-map fit, such as not filled side-chain densities or side-chains protruding from the density. Model 4 yielded a high C_{work} value during the refinement, but the good agreement of map to model could not be confirmed to the same degree by the C_{free} value. In conclusion, model 1 seems to be the most probable model for PM2.

Thus, computational tools for automatic model building have proven helpful in various ways for the structural interpretation of the density map of PM2. Firstly, SQWRL4 made it possible to easily and quickly generate multiple models representing the various possible sequence assignments. Secondly, energy minimisation with CNS ensured a comparable geometrical quality of all models and thirdly, maybe most importantly, calculating the C_{free} value, a measure of fit-to-map completely independent of the refinement process itself, provided an objective, dispassionate and quantitative ranking of the models. The result is a sound, well-founded structural interpretation of the density in terms of a very probable, even if not completely reliable, atomic model. This would not have been possible to the same level with manual model building.

Contribution

For this study, I performed the model building and refinement for PM2. Additionally I was involved in the writing the manuscript.

5. Flexible Fragment Fitting for Automatic Backbone Building

In this chapter, we present a procedure for flexible fragment fitting. After having determined the topology of a protein in form of a rough C_α-trace, the next step in the modelling pipeline is to convert this trace into a full-atom backbone. This can be achieved by fitting fragments of known backbone conformations on the trace. We show that flexible fitting can outperform rigid fitting, examine the influence of fragment library sizes on the fitting performance and investigate which metrics are best to assess a fragment placement.

Some theoretical background information about structure modelling with EMfasa as well as about the idea of deformable elastic network restraints is given in section 5.1. The motivation for this study is explained in section 5.2. Section 5.3 gives a detailed description of the methods we used. The results regarding our flexible fitting procedure are elucidated and discussed in section 5.4.

5.1. Theoretical Background

5.1.1. De-Novo Protein Structure Modelling with EMfasa

EMfasa [44] is a fully automated protocol for de-novo model building in cryo-EM density maps. It was developed by Tatjana Braun during the course of her PhD in the group of Gunnar Schröder. Emfasa aims to rapidly interpret a medium resolution map in terms of a first atomic model. The model might not be fully optimised yet but can be refined using computationally more expensive software. The workflow of EMfasa is illustrated in Figure 5.1 and consists of six steps:

Step 1: Trace Generation The first step is to determine the topology of the protein in terms of a rough C_α-trace. To do so, the dxtraces tool which is part of the DireX Framework is employed. Twice as many beads as there are residues in the sequence are randomly placed in the map and connected based on a TSP solver [47] which is set up in way that it prefers connections between beads traversing high density regions. Then, the number of beads are halved and the

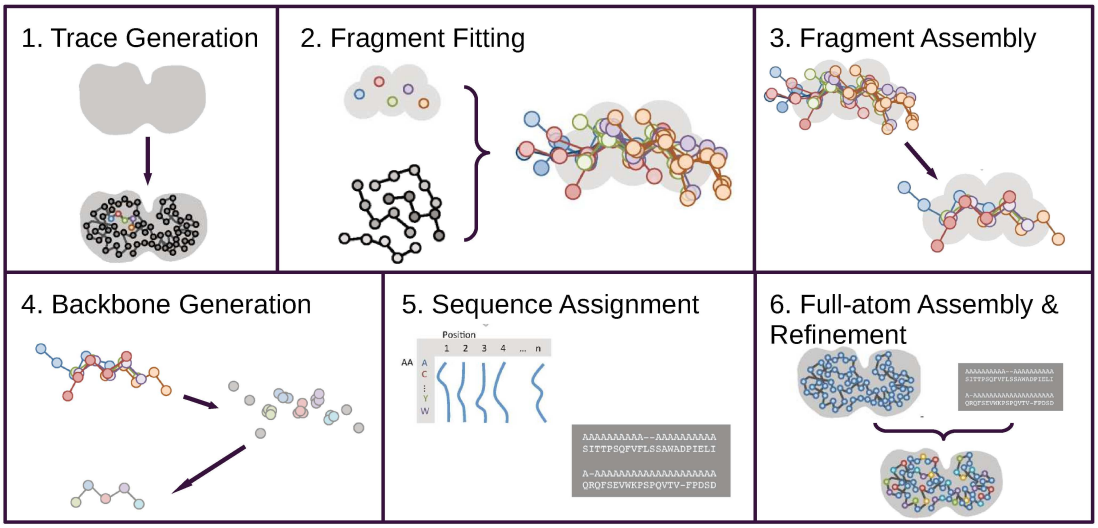


Figure 5.1.: Workflow of EMfasa: First a rough C_α-trace is generated which determines the topology of the protein. Secondly, local backbone conformations, fragments, are fitted in the map. Thirdly, a set of mutually compatible fragments is identified. Fourthly, clustering the fragments of a set gives a consensus backbone. Fifthly, a profile reflecting the fit of individual amino acids to residue positions is calculated and aligned with the sequence. Sixthly, the profile-sequence alignment and the backbone are combined to a full structure. Figure adapted from [44]

resultant trace is refined into the map using DireX. Since the direction of the trace is not known yet, both directions are considered in the following steps.

Step 2: Fragment Fitting Having the topology as well as the approximate positions of the residues, the next step is to convert the trace into a full atom backbone and to refine the residue positions. The approach to do so, is to fit local backbone conformations, fragments, from known protein structures at the bead positions into the map. The fragments originate from a library consisting of 100 non-redundant backbone 7-mer backbone fragments. The library was generated based on approximately 5000 high resolution crystal structures and the help of the ClusCo clustering algorithm [83]. During the fitting procedure, a fragment is placed onto a bead position and its position and orientation are rigidly optimised via Chimera's fitmap method. 30 global searches are performed, the top 5 placements are stored, resulting in 500 stored possible conformations per bead position.

Step 3: Fragment Assembly Local backbone conformations need not only to resemble the local map but also to match their neighbouring fragments. Therefore,

the next step is to find a set of mutually compatible fragments. The key concept of this fragment assembly is a *Monte Carlo Simulated Annealing* (MCSA) [84] sampling which identifies such a set based on the correlation of the fragments to the map, the overlap between neighbouring fragments, the occurrence of clashes between fragments assigned to residues far apart in the sequence and the agreement of the direction of a fragment with the direction of its neighbouring fragments and the direction of the trace.

Step 4: Backbone Generation In this step, the DBScan clustering algorithm [85] is employed to convert a set of mutually compatible fragments into a consensus backbone. The cluster centres represent C_α positions, the connections between residues are based on intra-fragment connections. An advantage of the DBScan algorithm is, that it does not require the number of clusters a priori. This way, it can be taken into account that it may not have been possible to identify all the residues. The cluster centres are derived by averaging the atom coordinates of fragments participating in a cluster. This averaging can lead to unphysical local geometries. Therefore the backbone is refined using the Phenix real-space-refine tool.

Step 5: Automatic Sequence Assignment Assigning side-chains to the backbone is performed in two steps. First a profile reflecting the fit of each amino acid to each residues position is generated. Then this profile is aligned to the sequence of the protein using the Needleman-Wunsch algorithm [86–88].

Step 6: Full-atom Assembly and Refinement Finally the backbone and the profile-sequence alignment are combined into a full structure using Modeller [89]. A refinement with phenix finalises the modelling process.

5.1.2. Deformable Elastic Network Restraints for Structure Refinement in DireX

The general principle of structure refinement in DireX has been elucidated in section 1.3. However, the usage of Deformable Elastic Network (DEN) restraints during the DireX routine is especially important for the following study and is here explained in more detail.

DireX employs DEN restraints to prevent overfitting. The general idea of the DEN approach is to refine only those degrees of freedom that are defined by the data and to use prior structural information for those not defined by the data [22]. The structural information is integrated in the form of distance restraints between random pairs of atoms, where the number of restraints usually equals twice the

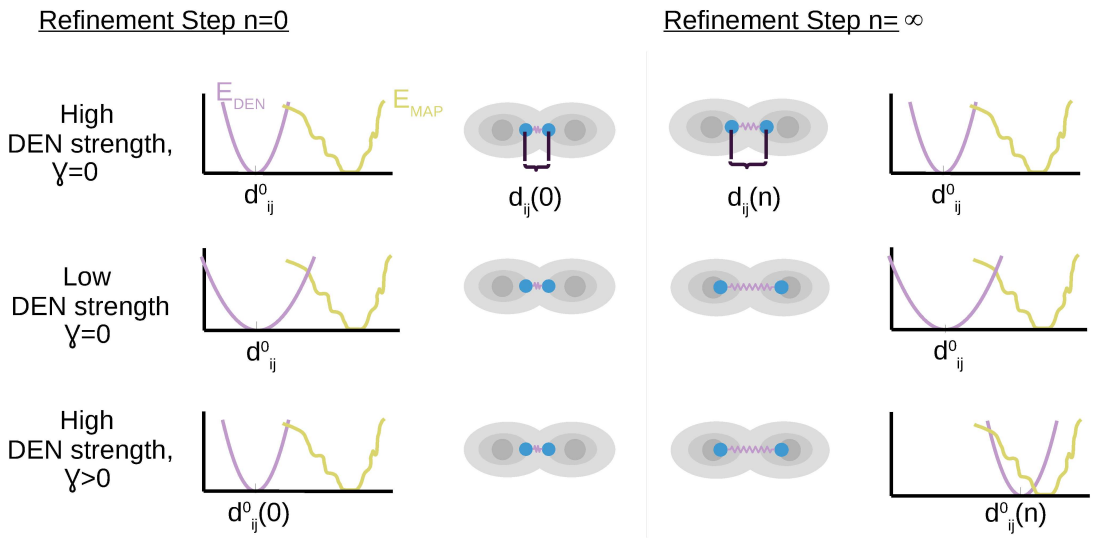


Figure 5.2.: Deformable Elastic Network restraints: DEN restraints are harmonic distance restraints between random pairs of atoms. Let´s consider two atoms i and j with a distance $d_{ij}(0)$ before the refinement of the structure: The DEN-Potential is initialised with a minimum at $d^0_{ij}(n=0) = d_{ij}(n=0)$. The DEN strength determines the steepness of the potential. During the refinement the distance between the atoms is influenced by the map potential E_{MAP} and the forces mediated through the DEN-potential E_{DEN}. It eventually converges to a distance $d_{ij}(n)$. For higher DEN-strengths $d_{ij}(n)$ is more biased towards the start value $d_{ij}(0)$, for lower DEN-strength it is more biased to resemble the density distribution in the map. The balance between E_{MAP} and E_{DEN} is also regulated by the γ parameter. For $\gamma > 0$, the DEN potential is shifted during the refinement to enable an adaption of the structural information given by the start structure to the conformation captured in the density map. Figure inspired by[22].

number of atoms. In the beginning of the refinement the restraints are mediated by a harmonic potential with the minimum at the distance $d^0_{ij}(0) = d_{ij}(0)$ where $d_{ij}(0)$ is the distance between the atoms i and j of the start structure (Figure 5.2). During the refinement the distance between the atoms i and j may change, as two opposing forces are acting on them: the force mediated through the map potential and the force applied by the DEN-potential. While the map potential pulls the atoms into high-density regions of the map, the DEN-potential damps the movement of the atoms and aims to preserve the geometry of the start structure. At refine step n the distance between atoms i and j is then $d_{ij}(n)$. Two parameters regulate the effect of the DEN-restraints: the strength of the DEN-restraints and the γ parameter. The strength is defined as the pre-factor or amplitude of the harmonic distance restraints, such that forces mediated through the DEN-restraints increase linearly with the strength. The γ parameter balances the opposing forces applied by the DEN-restraints and the map. This is achieved by regulating the adaption of the DEN-restraints during the refinement, i.e the shift of the minimum distance $d^0_{ij}(n)$. For low values of γ only slow and small changes of $d^0_{ij}(n)$ are performed and forces retaining the geometry of the start structure have a higher weight than forces adapting the geometry to the map. In contrast, for high values of γ the influence of the structural information vanishes and the map potential applies the dominant force.

For fragment fitting, γ is set to 0 such that the DEN-potential is not shifted during the refinement. Varying the strength of the DEN-restraints enables the adjustment of fragment flexibility. No or weak DEN-restraints should allow a high degree of flexibility, while strong DEN-restraints force the fragments to stay rigid and preserve their conformation during the refinement.

5.2. Motivation

In 1986 Jones and Thirup [90] discovered that backbone conformations of many different proteins often are composed of the same repeating subunits and proposed to use fragments of known protein structures for the model building in X-Ray Crystallography maps. Such approaches were then presented for example by Holm and Sander [91], Terwilliger [92] as well as by Pavelcik [93]. While Holm and Sander as well as Terwilliger make use of rigid fragments, Pavelcik allows conformational flexibility of the fragments. This way, the size of the fragment library can be reduced significantly [93].

Fragment-based methods have also been developed for model building in cryo-EM maps. Wang et al. rigidly fit sequence specific 9-mer fragments into the density map and subsequently assemble them into a full length protein structure. As described above, the procedure in EMFASA is slightly different. There, sequence

non-specific all-alanine 7-mer fragments are used in the fragment fitting step, side-chains are added later. Comparing the approach of Wang et al. and EMfasa, it becomes clear that the choice of the size of the fragments seems to be not trivial. Larger fragments are more specific, but might not be able to resemble all backbone configurations found in a backbone, such that a bigger library would be needed. Shorter fragments inherit less structure information but entail a higher adaptability to rare conformations.

Here we present an optimised fragment fitting routine to be integrated in the EMfasa framework. We use shorter fragments of 5 residues and a small fragment library comprising only ten fragments. Also, we do not perform rigid fitting of the fragments, but allow the fragments to flexibly morph into the map. We investigate which method is best suitable to assess the quality of fragment placements and we examine how fitting accuracy is influenced firstly by the size of the used fragment library and secondly by the introduction of fragment flexibility. Lastly, we illustrate the application of our method as part of protein structure modelling with EMfasa using two examples, a high-resolution map of an egg-white lysozyme acquired by electron crystallography [94] as well as a medium resolution map of Tobacco Mosaic Virus (TMV) reconstructed with the single particle workflow [95].

5.3. Material and Methods

5.3.1. Fitting Procedure

We implemented a new fragment fitting procedure as part of DireX. The workflow is illustrated in Figure 5.3.

First, the C_α trace, the density map as well as a fragment library are read as input. Then, DEN-restraints are initialised for each fragment.

In the next steps, the fitting is performed in a loop over all beads: For each bead position-restraints are defined, which bias the fragments to be fitted along the trace. The principle of the position-restraints is illustrated in Figure 5.4 a). Two harmonic position-restraints are applied on each fragment. For a fragment comprising f C_α-atoms (f should be an odd number) and for a bead associated with a position i in the trace, the first C_α-atom is restrained to the coordinates of bead $j = i - (\frac{f-1}{2})$. Accordingly the last C_α-atom of the fragment is restrained to the bead $k = i + (\frac{f-1}{2})$. With this set-up 30 independent fitting circles are performed: The central C_α of each fragment is placed onto the bead and randomly rotated (all fragments are fitted simultaneously and undergo the same rotation). Next, the fragments are refined into the map following the usual DireX routine, allowing translations, rotations but also slight conformational changes. The DEN-restraints regulate the degree of the conformational changes, while the position

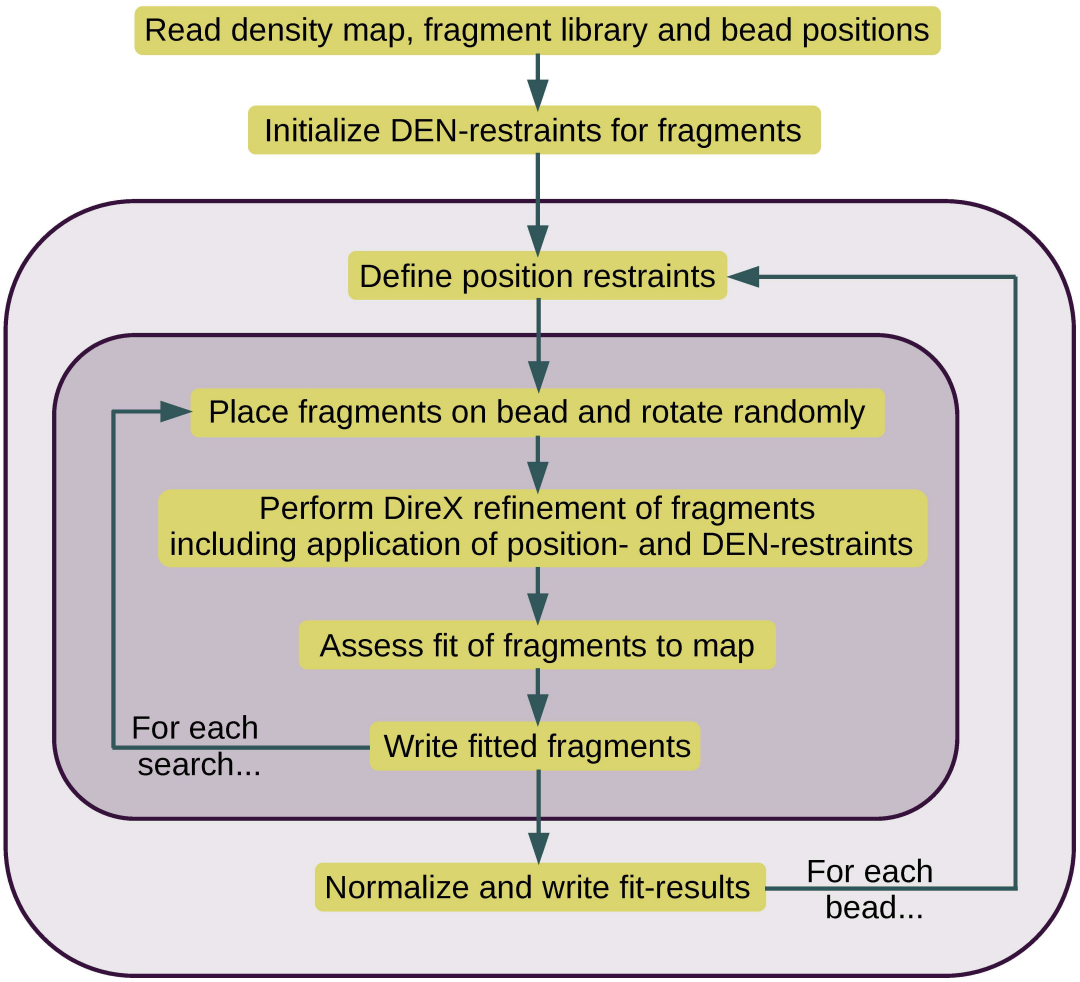


Figure 5.3.: Fragment Fitting with DireX: Diagram showing the implementation of a fragment fitting procedure with DireX. A trace defining the topology of the protein, a fragment library and the density map are required as input. The fitting is performed in a loop over all beads in the trace and over 30 independent searches per bead.

restraints bias the fragments towards orientations along the direction of the trace. After the fitting, the fit-results are assessed (see also sections 5.3.3 and 5.4.1) and the procedure is repeated for a new rotation of the fragments. When all 30 searches for a bead have been conducted, all fit results are normalised, such that the best fragment placement is associated with a score of 1, and all placements (for m fragments, there are $30 \times m$ placements per bead) are ranked and stored.

Afterwards the next bead is chosen as fitting position, new position restraints are defined and 30 searches are carried out. The procedure is repeated for each bead.

The result is a list of $30 \times m$ ranked fragment placements for each bead.

5.3.2. Benchmark Dataset

We analysed the performance of our fitting method using three exemplary density maps.

The first one is a high-resolution map of human ferritin, determined with X-ray crystallography with a resolution of 1.9 Å [96]. The corresponding atomic model is mainly comprised of α-helices and has been deposited with the PDB ID 2FHA.

The second example is a high-resolution map of an egg white lysozyme acquired by electron crystallography. The resolution of the map is 1.8 Å. The secondary structure of the lysozyme is quite diverse and comprises α-helices, β-sheets and extensive loop regions. The structure can be found by its PDB ID 6S2N.

As third example we use a a map of the asymmetrical subunit of TMV which has been reconstructed with a resolution of 3.4 Å using cryo-EM single particle analysis. The deposited corresponding atomic structure (PDB 4UDV) comprises 153 residues. The fold of TMV is dominated by α-helices.

5.3.3. Metrics to Assess Placement-Quality

A crucial component of a fragment fitting procedure is the assessment of the quality of placement. The evaluation of placement quality is the foundation for the selection of fragment placements used for further modelling steps and has therefore direct impact on the quality of the final model.

While fit-to-map metrics like DireX' C_{free} value or Chimera's model-to-map correlation seem to be an obvious choice at first glance, the situation for backbone fragment fitting is in fact more complex. Fit-to-map measures do not differentiate between density caused by backbone atoms and side-chain densities. In consequence, a backbone fragment associated with a good fit-to-map does not necessarily resemble the underlying backbone conformation but may lie partly within a density region which should be occupied by a side-chain. Having this in mind,

the average map-value per atom could be an easily accessible and more suitable measure, as high-density regions usually follow the main-chain of a protein [43].

A good measure of fragment placements associates fragment placements resulting in a low RMSD to the ground truth fragment with a good score, and, vice versa, placements with a high RMSD to the true backbone conformation should get a worse score. While false negatives, i.e. good placements with a low score, hamper effectiveness, further modelling can still be successful if other good placements have been found. On the other hand, false positives, unfavourable placements with a high score, have direct impact on the quality of the final model and should be particularly avoided.

We tested if the average map value per atom (AMVA), the average map value weighted with its reciprocal standard deviation (AMVA-STD) as well as a score which we call the i-o score fulfil this behaviour. The i-o score assumes that the inner backbone atoms that form the $N - C - C$ chain should be placed in higher density regions than the outer backbone atoms, i.e the adjacent O and H atoms (see Figure 5.4 b)). It is defined as :

$$\text{i-o score} = (< \rho_{\text{inner}} > - < \rho_{\text{outer}} >) * \text{AMVA}$$

where $< \rho_{\text{inner}} >$ and $< \rho_{\text{outer}} >$ denote the average mapvalues of the atoms corresponding to the inner or the outer backbone, respectively.

As test case we performed fragment fitting on the map of ferritin. The coordinates of the C_α-atoms of the deposited structure served as input trace. A fragment library of only one fragment was used, no DEN-restraints were applied to enable fully flexible fitting. For each fitted fragment, we calculated the AMVA, the ANVA-STD and the i-o score. Further we calculated for each fitted fragment its RMSD to the corresponding fragment in the deposited structure.

5.3.4. Adaption of the Fragment Library

EMFasa uses a sequence non-specific fragment library consisting of 100 7-mers. In order to increase accuracy, we decided to resort to a library of 150 5-mer fragments previously generated in the group, analogously to the EMfasa library. We then employed ClusCo to re-cluster this library into libraries consisting of 100, 50 and 20 and 10 fragments using default parameters. To decide which library size offers the best compromise between run-time and accuracy, we tested all of them on two test data sets. Data set 1 consisted of 11 beads in a helix region of the ferritin map, data set 2 comprised 11 beads in a loop region of this very map. We performed fragment fitting on both data sets with all libraries, allowing full flexibility of the fragemnts (DEN strength 0). Finally, we calculated the RMSD of each fitted fragment to the corresponding true fragment of the deposited TMV structure.

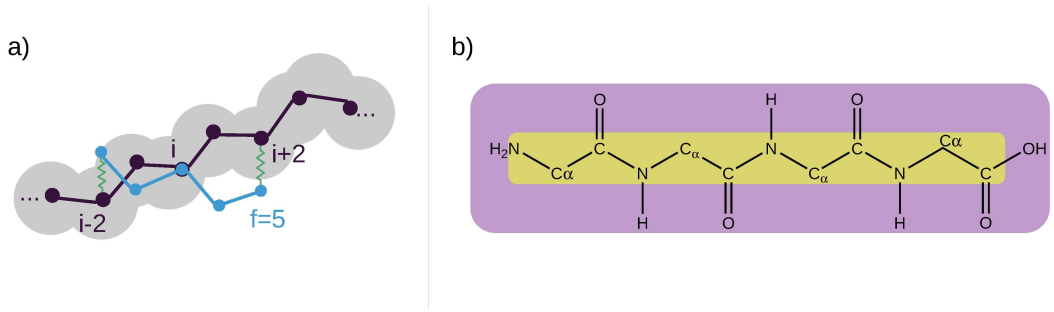


Figure 5.4.: a) Position-restraints: During the fragment fitting procedure, position-restraints are applied such that the fragments are placed in direction of the trace and are not too strongly influenced by side-chain densities emerging from the main-chain. When a fragment consisting of $f = 5$ C_α-atoms is fitted on a bead i, the first C_α is restrained to the coordinates of bead $i-2$ and the last C_α is restrained to the coordinates of bead $i+2$.
b) i-o score: For the calculation of the i-o score, one differentiates between the inner backbone consisiting of the $N-C-C$ chain (yellow) and the outer backbone comprising the other backbone atoms (purple).

5.3.5. Tuning Fragment Rigidity with DEN-Restraints

Fitting fragments flexibly instead of rigidly has the advantage that only a small fragment library is needed to accurately resemble the structural geometry given by the map. On the other hand, overfitting needs to be prevented and as much as possible structural information should be derived from the geometry of the fragments. Usually DEN-restraints are employed to overcome overfitting. In our fragment fitting procedure, we reinterpreted the purpose of the DEN-restraints and used them to regulate flexibility or rigidity of the fragments.

We performed several runs of fragment fitting on the high-resolution structure of the egg-white lysozyme varying the strength of the DEN-restraints from 0.0 to 0.2 but setting the γ parameter to 0. The resulting data was used for different analyses:

In a first step we aimed to investigate the relation between the strength of the DEN-restraints and the rigidity of the fragments. For each run and each fitted fragment, we aligned the fitted fragment on its corresponding input fragment from the fragment library and calculated the RMSD between them. This gives a measure to which the fragment has been deflected or deformed during the fitting procedure. Then, all fragment deformations corresponding to a DEN-strength were averaged.

Further, we examined how varying fragment flexibility effects fitting accuracy. We measured fitting accuracy by the means of the mean minimal RMSD per run. Meaning, for each run of fragment fitting we determined the RMSD between each

fitted fragment and its corresponding true fragment. We identified the best fit at each fitting position or bead as the minimal RMSD between a fitted fragment and the true fragment of the deposited structure at that position. To get a general measure for the whole run, we then averaged those minimal RMSDs over all beads, resulting in the mean minimal RMSD. For this analysis, we additionally filtered the map of the lysozyme to 3 Å and 5 Å and repeated the fitting and evaluation.

The effect of fragment flexibility on fitting accuracy may also depend on the characteristics of the local underlying backbone conformation. We therefore also had a look on the individual minimal RMSDs at each fitting position, and subsequently averaged only over beads participating in similar secondary structure geometries, resulting in three separate measurement series for helical residues, residues in sheet regions and residues in loop regions.

5.3.6. Integrating Flexible Fragment Fitting into EMfasa

The final purpose of the here presented fragment fitting method is to be integrated into the EMfasa framework for automatic structure modelling. Therefore, we went through the complete EMfasa workflow to build the structures corresponding to the TMV as well as to the lysozyme map. But, instead of performing the usual EMfasa fragment fitting, we employed our new flexible fitting method with DireX. Also, we needed to change the trace generation, as it was not possible to build a trace with the correct topology using dXtraces. Alternatively, we used MAINMAST to determine a rough C_α-trace. A trace built by MAINMAST is a refined minimal spanning tree between many local dense points. In the usual MAINMAST procedure the sequence is threaded onto this trace to finalise the structure and to convert the tree into a backbone comprising the correct number of residues. However, to stay as close as possible to the EMfasa workflow, we decided to not use sequence information in the tracing step and to skip the sequence threading. Instead, we directly resampled the minimal spanning tree to the correct number of beads. All other steps of EMfasa were performed conventionally, applying default parameters.

5.4. Results and Discussion

5.4.1. Measuring Placement Quality

The overall aim of a fragment fitting procedure is to provide a selection of fragments which resemble the underlying backbone conformation as accurate as possible. Such a selection requires a metric that measures how well a fitted fragment matches the true backbone geometry. A promising candidate is the AMVA, or metrics

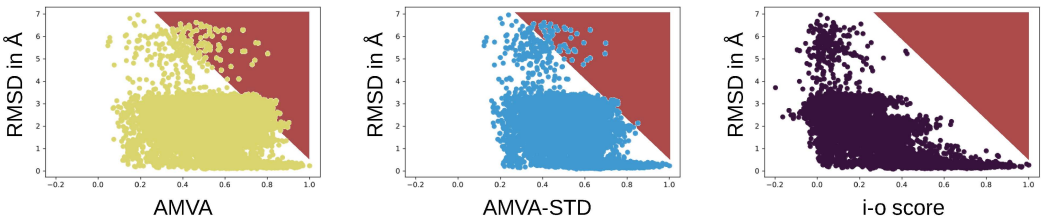


Figure 5.5.: Metrics for assessment of placement quality: Left: Average Map-value per atom (AMVA), centre: average map value per atom weighted with its standard deviation and, right: i-o score. All scores has been normalised for better comparability, such that the best scored fragment is associated with a score of 1. A suitable metric must not yield high scores for poor placements (identified by a high RMSD to the ground truth), this "forbidden" area is illustrated with a red triangle in each of the plots. While AMVA as well as AMVA-STD show data points in the forbidden area, the i-o score does not.

derived from the AMVA, like the AMVA-STD or the i-o score. The AMVA is based on the observation that the main-chain of a protein usually lies in high-density regions of the map. A possible drawback of this method could arise from the averaging, though. A fragment located partly in very high density regions, but also party in low density regions would yield a relatively high score but its orientation is probably not ideal. To suppress such situations, the AMVA-STD prefers fragment placements with evenly distributed, high map values. The i-o score offers a more detailed look on the distribution of map values along the fitted fragment. Here, placements yield a high score if the AMVA is high and the inner backbone, the N and C atoms, have significantly higher map values than the surrounding H and O atoms, indicating a centred placement of the fragment in the high density region.

We investigated the ability to identify good fragment placements for all three measures. Figure 5.5 shows the results of this investigation.

One can see the dependence of the assigned scores on the ‘true’ quality, calculated as RMSD between the fitted fragments to the corresponding true fragment of the deposited structure. An ideal metric would give data points on the diagonal of such a plot (top left to bottom right), where fragments with a high RMSD are associated with a low score and fragment placements with a low RMSD yield a high score. None of the tested metrics shows this behaviour.

Instead, fragments with low RMSDs are associated with a wide variety of scores, meaning that many good placements are not identified as such. In fact, this is true, albeit to a lesser extent, for medium high RMSDs up to 3.5 Å, too. Nevertheless,

there are differences between the metrics. The AMVA is for fragments with a score worse than ≈ 0.9 no reliable measure of the placement quality, fragments scoring better have indeed a low RMSD to the ground truth. For the AMVA-STD this border can be slightly shifted to ≈ 0.8, but other than this the two metrics behave similar. The correlation between the i-o score and the RMSD appears more distinct, while there is still a variety of scores associated for fragments with a similar RMSD, very high scores are only yielded by fragments with a RMSD lower than about 0.5 Å, a medium i-o score down to 0.5 can be mainly only achieved by placements with a RMSD of 3.5 Å or better, while worse placed fragments only get scored up to ≈ 0.4.

As described before, a crucial requirement for a suitable metric is that no poor placed fragments are identified as good ones. Data points corresponding to these situations would be located somewhere within the red marked areas in Figure 5.5. While the AMVA and the AMVA-STD give data points in the 'forbidden' area and therefore do not fulfil the requirement, the i-o score does.

It might be noticed that the i-o score give negative values for some placements. Although this is unusual for metrics in general, it make sense with regard to the definition of the i-o score. A negative i-o score means that the outer backbone atoms are, on average, located in higher density regions than the inner backbone atoms. This would be a quite unlikely distribution for backbone conformations in density maps, and therefore a low score for such situations is particularly justified.

Based on the here described observations we implemented the i-o score in our fragment fitting procedure for the assessment of placement quality.

5.4.2. Effects of Library Sizes

The size of the fragment library has direct impact on the performance of the fragment fitting procedure. While larger libraries improve accuracy, they negatively affect run-time issues. For flexible fitting, though, smaller libraries could achieve similar accuracy to larger libraries in rigid fitting.

Therefore, we investigated the accuracy of fragment libraries consisting of 10, 20, 50, 100 and 150 fragments as shown in Figure 5.6. The distribution of RMSDs between fitted fragments and corresponding true fragment is illustrated as violin plot for each library. A violin plot displays the kernel density estimation of the underlying distribution with whiskers marking the minimal and the maximal observed value. In each violin plot dashed lines depict the quartiles of the distribution.

The performance of the libraries was examined on two independent data sets, in a helix region (Figure 5.6 left) as well as in a loop region (Figure 5.6 right). While helix regions should be fairly easy to cover, since all residues in a helix adapt well-defined, nearly identical conformations, loop regions can form a wide variety of conformations and are therefore more difficult to model. In terms of fragment

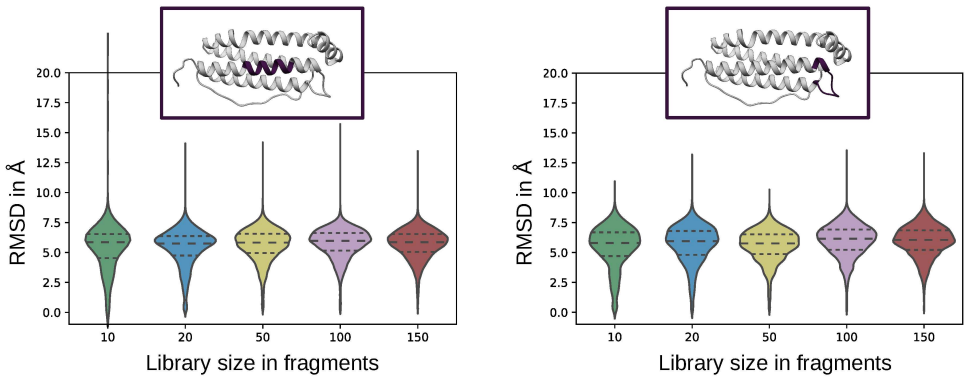


Figure 5.6.: Different library sizes for fragment fitting: Fragment fitting was performed on two test data sets comprising 11 beads each in a map of human ferritin, representing a helix-region (left) and a loop region (right) respectively. The corresponding region is depicted dark purple in the deposited protein structure of ferritin at the top of the plots. Different library sizes were used for fitting and the resulting distributions of RMSDs between fitted fragments and corresponding true fragments are illustrated as violin plots. Dashed lines show the quartiles of the distributions.

fitting, only one helical fragment should be needed to find a suitable fragment to fit in a helix region, but multiple fragments might be needed to cover the various loop conformations. In consequence, smaller libraries might be expected to perform worse in loop regions compared to larger libraries, but might yield comparable accuracy in helix regions.

On first glance, all distributions appear similar, no obvious differences are visible, neither with regard to library size nor with regard to the different test regions. No trends between accuracy and library size are recognisable for both data sets. Interestingly, the smaller libraries comprising 10 and 20 fragments have a slightly broader tail towards lower RMSDs, indicating that relatively more fragments were fitted with higher accuracy. This is also visible when looking at the 25 % quartile lines. Moreover, the smallest library with only 10 fragments yielded the lowest minimal RMSD in both data sets, meaning that the best fitted fragment was placed during the fitting procedure with the library consisting of 10 fragments. On the other hand, this library produced also the worst placed fragment, resulting in a huge upper whisker of the corresponding violin plot for the helix data set. However, as long as good placements are found, individual worse placements are not crucial.

Why were the best fitting results achieved using the smallest library? The following consideration may help to clarify this: The smaller libraries are no subsets

of the larger libraries, but may include fragments that are not part of a larger library. That is, because all libraries were generated by clustering a large library of 150 fragments. A smaller library comes from clustering the original library in fewer, but larger clusters, while a larger library results from more, but smaller clusters. The fragments in the smaller library are the cluster centres of the few, large clusters and therefore potentially different than the fragments in a larger library representing different, namely smaller clusters. Either the library consisting of only 10 fragments consists of fragments which fit particularly well in the map, but are not found in larger libraries. However, small conformational differences between fragments should be compensated for by the fragment flexibility. Or, the difference in fitting performance is random and only due to a too small number of fitting repeats. In that case, the analysis should be repeated with a larger number of fitting repeats.

In conclusion though, no negative impacts of small libraries on accuracy could be confirmed for our flexible fitting method. This is even true for fitting in loop regions. Hence, the library consisting of 10 fragments serves as default library.

5.4.3. Effects of Fragment Flexibility

Fragment Fitting with DireX offers the possibility to tune the flexibility of the fragments via the strength of DEN-restraints. But how exactly does fragment flexibility influence the fitting procedure?

We first wanted to understand the relation between the strength of the DEN-restraints and fragment flexibility during the fitting process. Figure 5.7 a) shows the dependence of the deformation of the fragments on the DEN strength. If no DEN-restraints are applied (DEN strength is 0.0) the fragments are deformed by an RMSD of about 1.7 Å on average during the fitting. But, with increasing DEN strength this value decreases rapidly, until it converges to ≈ 0.1 Å for DEN strength ≥ 0.13. So, flexibility of fragments can indeed be regulated via DEN-restraints, enabling flexible fitting as well as nearly rigid fitting. For low DEN-restraints, the degree of flexibility reacts quite sensitively to changes in DEN strength, but for higher values variations in DEN strength do not result in significant changes in fragment flexibility.

The relation between DEN strength and fitting accuracy measured by the mean minimal RMSD to the true fragment, is illustrated in Figure 5.7 b). Showing a mirrored behaviour compared to Figure 5.7 a) the mean minimal RMSD increases with increasing DEN-strength. The slope of the curve is quite steep in the beginning, and, analogously to Figure 5.7 a) the values converge with DEN strengths ≥ 0.13. While the best fitted fragment only differs by 0.25 Å from its true fragment, the RMSD between best fitted fragment and true fragment is ≈ 1 Å for large DEN-strengths. Interestingly, there is nearly no difference to different map

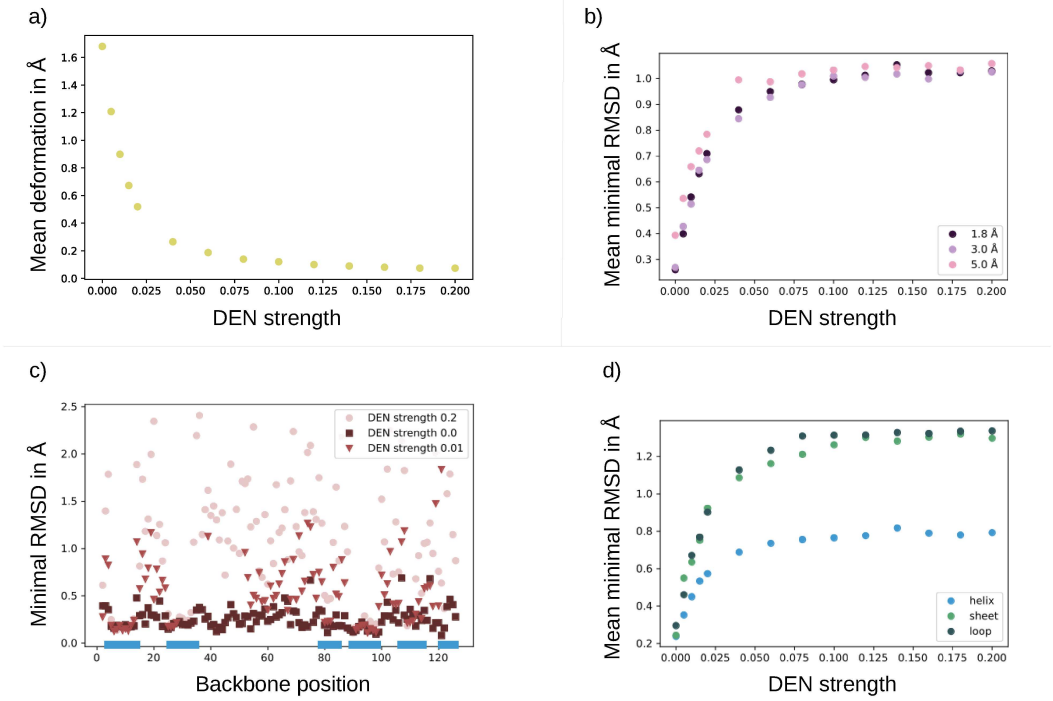


Figure 5.7.: Effects of DEN-restraints on fitting accuracy: a) With increasing strength of the DEN-restraints the deformation of the fragments decreases, i.e increasing of DEN-restraints results in a higher rigidity of the fragments. Deformation is measured as the RMSD between the fitted fragments and the input fragment, after alignment of the two fragments. b) With increasing rigidity of the fragments, the minimal RMSD, averaged over all beads, increases. The fitting becomes less accurate. This is true nearly independent of resolution of the map. c) Accuracy of fitting, measured as the minimal RMSD between fitted and true fragment, varies along the backbone. This effect is more significant for increasing rigidity of fragments. Blue bars at the bottom indicate location of helices in the backbone. d) Fitting is more accurate in helical regions of the protein compared to sheet or loop regions. This effect is less significant for low DEN-strengths.

resolutions. While it could be expected, that fragment flexibility is particularly beneficial in high-resolution maps, but can cause problems in lower-resolution maps due to overfitting, this assumption can not be confirmed based on the present data. On the contrary, full fragment flexibility results in maximal fitting accuracy for all tested resolutions. Moreover, higher resolution does not yield significantly higher accuracy compared to lower resolutions.

Figure 5.7 c) contemplates variations in fitting accuracy along the protein trace. For fully flexible fitting (DEN strength 0.0) no distinct pattern in accuracy can be observed, the RMSD between best fitted fragment and corresponding true fragment varies only slightly between ≈ 0.2 Å and 0.5 Å along the backbone with few outliers in the C-terminal region of the trace. The picture is completely different for higher DEN strengths. High accuracy, meaning low minimal RMSD, is achieved also for higher DEN strengths but only in single sections of the trace. In other regions accuracy is significantly worse with minimal RMSDs up to 1.2 Å (outliers up to 1.8 Å) for a DEN strengths of 0.1 and even up to 2.5 Å for a DEN strength of 0.2. Interestingly, the regions of high accuracy seem to overlay with helical regions of the trace (depicted by blue bars in Figure 5.7 c)). Thus, fitting in helical regions seems to be more accurate than fitting in other regions for higher DEN strengths. Figure 5.7 d) confirms this impression. Averaging the minimal RMSDs over beads participating in helix-, sheet and loop regions, respectively, shows that accuracy is significantly higher in helical regions for higher DEN strengths than in sheet and loop regions. There is no difference between sheet and loop regions. Considering that backbone conformations forming β-strands are well defined and therefore less diverse than conformations in loop regions, this is surprising and may indicate that a suitable fragment is missing in the fragment library. For low DEN strengths there is no significant difference in accuracy for different secondary structures, which is in line with the observations made during the analysis regarding library sizes.

In conclusion, we recommend to perform fragment fitting without the usage of DEN-restraints. Allowing full flexibility of fragments during the fitting procedure improves accuracy significantly. This is also true for lower map resolutions.

5.4.4. Fragment Fitting as integrated Part of EMfasa

The typical application of a fragment fitting procedure is to be part of a structure modelling workflow. We employed a customised MAINMAST version to build a C_α trace, performed our flexible fragment fitting, and fed the fitted fragments into the EMfasa pipeline, where the fragments were assembled to mutually compatible subsets and clustered to form a full atom backbone. Side-chains were added with help of a profile sequence alignment and the whole structure was finalised and refined. The results are shown in Figure 5.8.

The final EMfasa structure for the medium resolution map of TMV yielded a

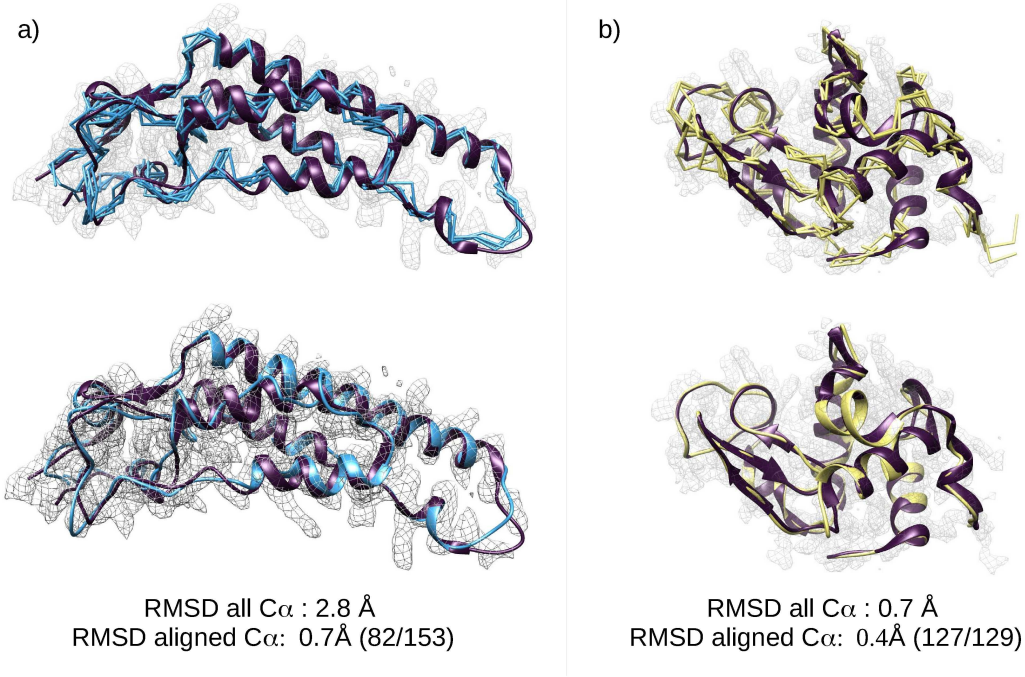


Figure 5.8.: EMfasa with flexible fragment fitting: a) Results for the 3.4 Å map of TMV. Upper panel shows the deposited PDB structure in dark purple, overlaid with the best fitted fragments for each bead position in blue. Lower panel presents an overlay of the true structure (purple) with the final structure modelled with EMFASA. The RMSD was calculated between all C_α-atoms, and separately for all aligned pairs of C_α, where a pair of C_α-atoms is considered to be aligned if there spatial distance is < 2 Å. The number of aligned C_α-atoms compared to the total number of C_α-atoms is given in parentheses. b) Same arrangement as in a), but for the 1.8 Å map of the egg white lysozyme. Here, fitted fragments and modelled structure are depicted in yellow.

C_α RMSD of 2.8 Å. Just over the half of the C_α-atoms of the EMFASA structure were in a 2 Å neighbourhood of the corresponding C_α-atom of the true structure. Among these, the RMSD is only 0.7 Å. This shows, that the RMSD is not due to a total shift in the backbone, but that there are regions of the map which could be modelled quite accurately, and those where modelling was more challenging. Roughly speaking, helices have been reconstructed more accurately than loop regions. The structure of TMV has been also modelled in [44]. Compared to the here presented structure, the structure built with the original version of EMFASA yielded a C_α RMSD of 2.1 Å and 129 out of 153 matched C_αs. However, it should be noted at this point that the difference in accuracy already occurs before the fragment fitting namely with the trace generation. The C_α-trace in [44] yielded 139 of 153 matched C_αs compared to only 37 aligned C_αs in the MAINMAST trace, which was the basis for the modelling here. So, applying flexible fragment fitting and the subsequently following EMFASA steps could improve the accuracy, while in [44] the accuracy of the C_{alpha} trace was higher than the accuracy of the final structure. Consequently, the difference in RMSDs of the final structures can not be attributed to a poorer performance of the fragment fitting but to a more challenging initial situation.

The density map of the lysozyme has a very high resolution, which facilitates the model building significantly. The RMSD between the modelled structure and the deposited true structure is only 0.7 Å. Only two C_α-atoms were not in a 2 Å radius of their corresponding C_α-atom in the true structure. Even loop regions could be modelled accurately. Although the number of tested maps is not sufficient for a reliable statement, the example of the 1.8 Å map of the egg white lysozyme gives a hint, that EMFASA with flexible fragment fitting can build highly accurate models for high-resolution maps.

6. Protein Topology Tracing Guided by Predicted Distance Matrices

This chapter presents a method to automatically determine the topology of a protein based on a cryo-EM map. It represents the very first step in a modelling pipeline. The key feature of our method, dXtopology, is the integration of predicted inter-residue distances to supplement the information provided by the density map.

As in the previous chapters, some theoretical background information is given in the beginning (section 6.1), here some fundamentals about prediction of inter-residue distances are explained. The motivation, methods, results and discussion are not given in individual sections, but in form of a manuscript. This manuscript has been also submitted to Structure in October 2022.

A more technical description of the methods as well as a section about trials, errors and perspectives corresponding to this chapter can be found in appendix C.

6.1. Theoretical Background

6.1.1. Prediction of Inter-Residue Distances

The principal idea of predicting inter-residue distances is based on co-evolutionary analysis. Proteins, emerging from the same protein family share many structural features but show variations in sequence. Residues which are in close spatial proximity to each other tend to co-evolve, that means that if one residue is mutated, the other one will probably undergo a mutation, too. So-called correlated mutations can therefore indicate contacts between residues. [97]
Thus, the common approach nowadays to predict inter-residue distances is to construct a multiple sequence alignment (MSA) containing evolutionary related sequences for a target protein and feed the MSA as input to a neural network which then extracts the geometrical information.

TrRosetta [98, 99] is one representative of such a method. A MSA is generated and subsequently processed by a deep neural network, which then predicts inter-residue distances and orientations.

The predicted distances between C_β-atoms are given in form of a histogram,

where the distance range between 2 Å and 20 Å is binned into 36 equally spaced segments, 0.5 Å each, plus one bin indicating that residues are not in contact. For each pair of residues and each bin, the neural network predicts the probability of the distance between the pair being in a particular bin.

6.1.2. General Properties of Distance Matrices

Distance distributions in proteins are often arranged in a symmetric distance matrix D, where each residue in the sequence corresponds to a row and a column of the matrix. The distance between residue i and j is then stored in D_{ij} (and also in Dji). Consequently, the main diagonal of a distance matrix is always filled with 0s. Also, secondary structure elements can be identified by characteristic patterns in the distance matrix. α-helices appear as thickenings of the main diagonal, as a C_β-atom which is part of helix is in close proximity to the residues surrounding it in sequence space. In β-sheets, residues which are far away in sequence can be in close spatial proximity to each other when they are part of neighbouring strands. In a distance matrix, this conformation forms lines perpendicular to the main diagonal for anti-parallel strands and lines parallel to the main diagonal for parallel strands.

6.2. Manuscript I: Predicted Distance Maps Guide Backbone Topology Tracing in Medium Resolution Density Maps

6.2.1. Summary

Cryo-electron microscopy in principle can reach atomic resolution, however, in many cases only medium resolutions between 3 Å to 5 Å are achieved. In this resolution range, building atomic models can be difficult and poses a time consuming challenge. A first step in map interpretation typically is to determine the trace of the protein chain through the map. Finding this correct trace, i.e. the topology of the protein chain, is crucial for all following modelling steps. Here we present a novel approach, DXTOPOLOGY, to determine the topology of the protein in medium resolution density maps which is inspired by the recent success of machine learning based structure prediction programs. DXTOPOLOGY combines backbone tracing with inter-residue distances predicted with the program TRROSETTA. The aim is to connect pseudo-atoms that were placed into the density map such that their distance patterns best resemble the predicted inter-residue distances. We show that using information about inter-residue distances can correct errors in

the topology and improve traces which were built using the density map alone. Our tool provides a quick initial representation of the protein backbone including the sequence assignment. Moreover, the procedure can be easily incorporated into existing frameworks and serves as a basis for further automatic model building.

6.2.2. Introduction

Over the past few years cryo-electron microscopy (cryo-EM) has become a major technique for protein structure determination [100–102]. The number of high-resolution structures, gaining a resolution better than 3 Å has increased dramatically. Nevertheless, the majority of density maps deposited to the EMDataBank still have medium resolutions between 3 Å to 5 Å [103]. Interpretation of medium-resolution density maps often is not straightforward, though. The limited resolution and the consequently lower signal-to-noise ratio leads to ambiguities in the density. Those ambiguities, like branches or breaks in the density hamper the determination of the correct topology of the protein chain and the localisation of atom positions. In consequence, de novo atomic model building based on medium-resolution density maps is often difficult, time consuming and highly dependent on the expertise of the modeller. Computational tools for automatic backbone tracing have been developed to facilitate the interpretation process. For example, Chen et al. developed a fully automated program for backbone tracing named Pathwalker [41]. During the Pathwalker procedure pseudo-atoms are placed into the density map and connected by solving the Travelling Sales Person Problem (TSP). Another method for automatic main-chain modelling in the medium resolution range is MAINMAST [42]. In MAINMAST, the geometry of the backbone is represented by a tree structure, given by the minimum spanning tree, connecting local dense points so that their total spatial distance is minimised. The Rosetta approach [49] assumes that local similarity in sequence implies local similarity in structure. Segments of solved protein structures with local similar sequences are fitted into the density map and well matching fragments are assembled to form a complete protein structure. Further, the widely used Phenix framework addresses the backbone tracing problem with the phenix.trace_and_build tool [43]. Besides the methods mentioned above, several deep learning approaches for automatic protein structure modelling have been published recently [45, 104, 105]. One of them is DeepTracer, which can be used via a web server and builds all-atom structures of protein complexes. While all of these tools can help to overcome the difficulties of model building in medium-resolution density maps, there are still cases where the information gained from the density map is not sufficient to reconstruct the correct topology of the protein chain. Typical examples of topology errors are a false identification of termini, cross connections between β-strands in a β-sheet or an incorrect arrangement of secondary structure elements.

Beyond experimental data, information about the structure of a protein can also be predicted from its sequence. Deep Mind's machine learning based program AlphaFold [106] has revolutionised the field of structure prediction and the question of how to predict the native structure of a protein based on its sequence, may, at least for single-chain structures, be considered solved [107, 108]. Similar results to those of AlphaFold can also be produced with RosettaFold [109]. A key feature in both methods are two-dimensional pair representations. These are $N_{\text{res}} \times N_{\text{res}}$ matrices (where N_{res} is the number of residues), storing in each entry a measure of distance between the two corresponding residues. In this sense a pair representation can be understood as a predicted distance matrix.

In this work, we present a novel integrative approach, dxTopology, for the interpretation of medium-resolution density maps: We combine conventional backbone tracing with predicted inter-residue distances to determine the topology of a protein. More precisely, the aim is to build a backbone trace such that the resulting distance matrix best resembles the predicted inter-residue distances produced by the program trRosetta [98, 99]. In a first step, pseudo-atoms are placed into the density map and connected using a TSP solver. The resulting initial trace, in the following referred to as the conventional trace, might have many correct subtraces, but potentially a few wrong connections.

We detect those errors by comparing the distance matrix of the conventional trace with the predicted distance matrix. This comparison leads to an approximate assignment of where the pseudo-atoms are located along the sequence. The approximate sequence assignment then allows for reordering the correctly connected subtraces and yields an integrative trace, based on density information as well as predicted inter-residue distances.

We show that using predictions of inter-residue distances can correct errors in the topology and improve conventional traces which were built using the density map alone. dxTopology gives an initial representation of the backbone including the sequence assignment. We further show how our tool can be used to check traces that were built with other tools for potential topology errors. Moreover, dxTopology can be easily incorporated into existing frameworks and provides a solid basis for further modelling steps.

6.2.3. Results

Workflow

The workflow of dxTopology consists of three steps (see Figure 6.1).

Step 1 Trace Initialisation: First, pseudo-atoms, also referred to as "beads", are placed into high-density regions of the map using the dxbeadgen tool from the refinement software DireX [21, 22]. Hereby, the number of beads, N_{bead}, is

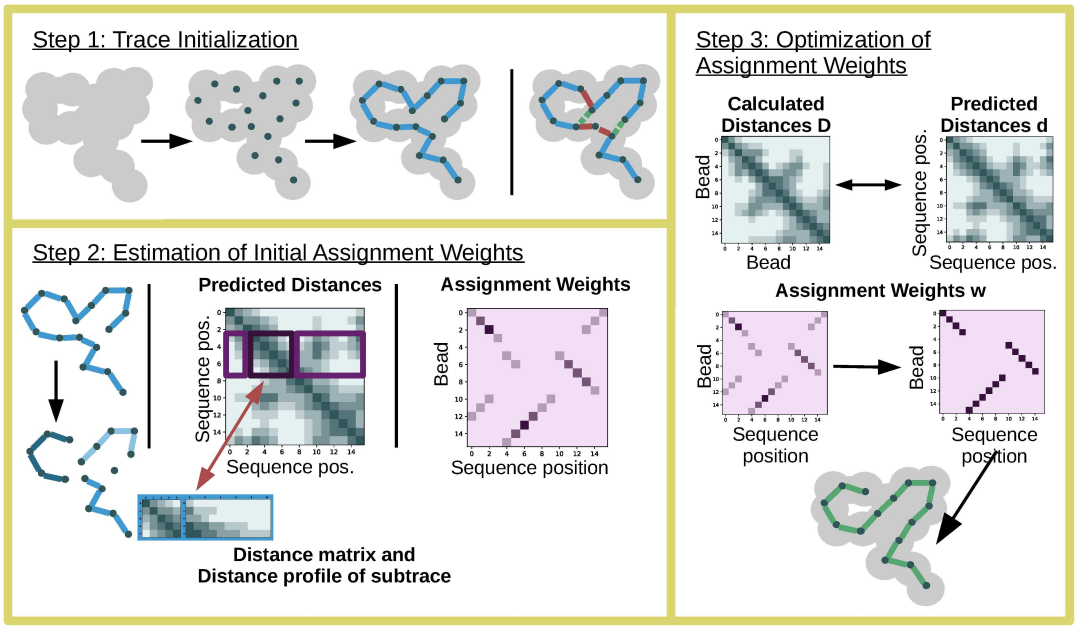


Figure 6.1.: Workflow Overview: In step 1, beads are placed in high-density regions and connected by solving the TSP Problem. The resulting conventional trace has many correct subtraces, but a few wrong connections. In step 2, distance matrices and distance profiles are calculated for many possible subtraces. Inter-residue distances are predicted by trRosetta. Distance patterns of the predicted distance matrix are compared to the distance matrix / distance profile pairs of the subtraces. Best matches between distance patterns of subtraces and distance patterns in the predicted distance matrix are stored in an assignment matrix, representing possible assignments of beads to sequence positions. In step 3, this initial assignment is optimised by minimising the difference between the predicted distance matrix and the distance matrix calculated on the basis of the assigned beads. Applying the optimised assignment weights yields an improved integrative trace.

chosen to be equal to the number of residues, N_{res}, in the sequence.

The beads are connected to form a trace by using the Lin-Kernighan heuristic (LKH) [47] which solves the TSP problem. The cost matrix for the LKH algorithm was generated with the PATHWALKER program, which takes into account that connections through high-density regions should be preferred. We have modified the calculation of this cost matrix in PATHWALKER to adapt it to our purposes (see STAR Methods). The obtained trace is referred to as the *conventional* trace in the following. This conventional trace often is correct in some parts but may have a few wrong connections. Even though, these wrong connections may only occur isolated and locally, they might lead to a completely wrong global topology and therefore need to be corrected as described in the following steps.

Step 2 Estimation of Initial Assignment Weights: To help finding the correct topology, additional information is obtained from the structure prediction tool trROSETTA, which predicts the distances between the residues of a protein from its amino acid sequence. The correct trace can be expected to have inter-residue distances that are similar to the corresponding predicted distances. To find errors and improve the trace, we compare the distances computed from the conventional trace with the predicted distances. It is important to note that this comparison requires to know which bead corresponds to which amino acid in the sequence. However, the conventional trace mainly defines the connectivity but does not yield a reliable assignment of beads to sequence positions, since small errors in the connectivity can lead to large errors in the assignment. The bead distance matrix and the predicted distance matrix are therefore not easily comparable. To solve this problem we compare only smaller subtraces which have a higher probability of having the correct connectivity. We assume that the conventional trace has correct subtraces, but that these subtraces may need to be reordered and connected differently to eventually obtain the correct topology of the protein, which is described in the following.

Given N_{bead} beads in total and a subtrace, consisting of n ($n \leq N_{\text{bead}}$) beads, all distances between these n beads within the subtrace are calculated and arranged in a distance matrix, which we refer to as the "inner" distance matrix. These inner distances describe the shape of the subtrace, but not its position within the protein structure. The position of the subtrace is rather encoded in the distances between the n beads within the subtrace to the $N_{\text{bead}} - n$ beads that are outside the subtrace; we refer to these distances as the "outer" distances. However, we cannot arrange these outer distances in a distance matrix, because we do not know the global topology and suspect that some connections outside the subtrace might be wrong. Instead we arrange the outer distances in a distance profile, which just contains the distribution of distances, and is independent of the assignment. The distance profile is a $n \times (N_{\text{bead}} - n)$ matrix, where the i-th row contains the sorted

outer distances of bead i. Analogously, we can associate each subset of adjacent n residues with a distance matrix and a distance profile, given by submatrices of the predicted distance matrix.

To compare the subtrace with a residue subset, we calculate the root-mean-square deviation between the distance matrices and between the distance profiles. Collecting all best matches of subtraces to residue subsets yields a bead-to-sequence assignment matrix **w** (see STAR Methods, appendix C). The entries w_{ki} of the assignment matrix describe the likelihood that bead k corresponds to residue i.

Beads that are connected correctly in the conventional (initial) trace appear as diagonal patterns in this matrix. If the subtrace is also at the correct sequence position within the protein, the diagonal pattern would be on the main diagonal. Diagonals parallel to the main diagonal are shifted within the sequence, but have the correct direction. Diagonals orthogonal to the main diagonal represent subtraces whose direction is wrong and needs to be reversed.

Step 3 Optimization of Assignment Weights: At this point, there are possibly still many side-maxima in the assignment. Especially shorter subtraces might be ambiguously assigned to several sequence positions. Therefore, we implemented a gradient descent optimisation of the weights. The ideal assignment of beads to sequence position, i.e the ideal arrangement of subtraces, minimises the difference between the predicted inter-residue distances and the distances between the beads. Therefore we minimise the following scoring function:

$$S = -\sum_{ijkl} \frac{w_{ki} w_{lj}}{1 + (d_{ij} - D_{kl})^2}$$

Here w_{ki} denotes entries of the assignment matrix, while D_{kl} are distances between beads and d_{ij} are the predicted inter-residue distances, respectively. Thus, i and j iterate over positions in the amino acid sequence, while k and l are indices numbering the beads.

Over the course of the optimisation, the weights converge and the final assignment is used to reorder the beads into a new trace. Badly placed beads or wrong assignments may lead to outliers in this trace. There are two types of outliers: 1) a bead can be considered an outlier in Cartesian space. This case arises if a bead has an unphysically large distance to its neighbours. 2) A bead can be an outlier in sequence space. If a single bead is assigned to a sequence position that is far away from the sequence positions of neighboring beads in the conventional trace, this assignment is considered unreliable. Both types of outliers are removed from the trace. The result is an integrative trace, whose estimated atom positions and local connectivity patterns are derived from the density map, but whose global topology is based on predicted inter-residue distances.

Matrix representations of topology

A key feature of DXTOPOLOGY is the matrix representation of traces and their topologies. As described above, our approach considers three matrices: the matrix of distances between the beads, a matrix of predicted distances between residues, and the bead-to-sequence assignment matrix. An example of those matrix representations is shown in Figure 6.2 for the case of the Bordetella bacteriophage cementing protein (EMD-5764, PDB 3J4U:H) [110]; the interpretation of these matrices is explained in the following.

There are $N_{\text{res}} = 140$ residues in the sequence of the cementing protein. Hence, $N_{\text{bead}} = 140$ beads were placed into the density map and connected to obtain the conventional trace, as described above in step 1 of the workflow. The topology of this trace is represented by a $N_{\text{bead}} \times N_{\text{bead}}$ distance matrix of pairwise distances between the beads, as shown in Figure 6.2a). The goal of step 2 is to obtain an initial estimate for the assignment of the beads to the amino acid sequence. For this, distance patterns in the bead distance matrix are compared to distance patterns in the predicted distance matrix. The $N_{\text{res}} \times N_{\text{res}}$ matrix of the predicted inter-residue distances as obtained by TRROSETTA is shown in Figure 6.2b).

The estimated assignment weights obtained from the distance comparison are stored in a $N_{\text{bead}} \times N_{\text{res}}$ assignment matrix (Figure 6.2e) upper panel). Those weights are optimised by minimising the difference between Figure 6.2a) and Figure 6.2b) in step 3 of the workflow. The result is the assignment matrix depicted in the lower panel of Figure 6.2e). Reordering the beads according to this assignment matrix and removing outliers gives the integrative trace, represented by its distance matrix in Figure 6.2d). The distance matrix of the deposited PDB structure 3J4U:H is shown for comparison in Figure 6.2c). The predicted distances, the PDB structure and the integrative trace share similar global distance patterns. Examples are the two β-strands, recognisable as two long diagonals perpendicular to the main diagonal (marked in red in Figure 6.2b),c),d)).

In general both, the predicted distance matrix as well as distance matrix of the integrative trace encode the same correct global topology as the PDB derived distance matrix. While the distance matrix of the conventional trace has similar local features, the global pattern is different and encodes a different global topology. The errors in the topology can be understood by looking at the weights in the assignment matrix Figure 6.2e) (lower panel). Subtraces, consisting of correctly connected beads, appear as diagonals, wrong connections as breaks between the diagonals. If the conventional trace had the correct topology including correct N- and C-terminus, the weights would be distributed along the main diagonal. Here, the conventional trace consists of five correct subtraces, recognisable as five diagonal lines. Two of the subtraces should be flipped, as can be seen from their orientation perpendicular to the main diagonal. Wrong connections were found

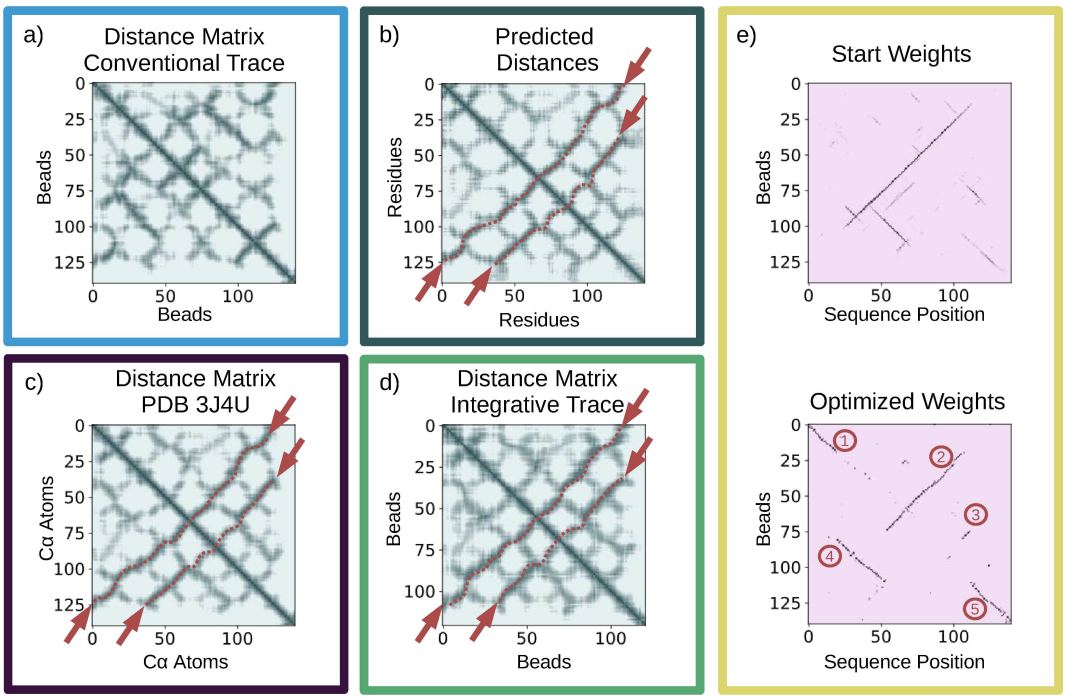


Figure 6.2.: Matrix representations of Bordetella bacteriophage cementing protein (EMD-5764, PDB 3J4U:H): **a)** Distance matrix of the conventional trace. **b)** Predicted inter-residue distances. **c)** Distance matrix of the PDB structure. **d)** Distance matrix of the integrative trace. **e)** Assignment weights before and after optimisation. **b)**, **c)** and **d)** show similar distance patterns, indicating a concordant topology, whereas **A** exhibits a clearly different appearance. Dotted red lines and red arrows in **b)**, **c)** and **d)** denote β-sheet structures. Assignment weights in **e)** encode the topology deviations of the conventional trace from the other traces. Even before the optimisation it is obvious that the middle part of the trace runs in the wrong direction, recognisable by the off-diagonal. After optimisation, correct subtraces can be identified as diagonals, wrong connections as breaks between them. Red numbers indicate the order of the subtraces in the conventional trace.

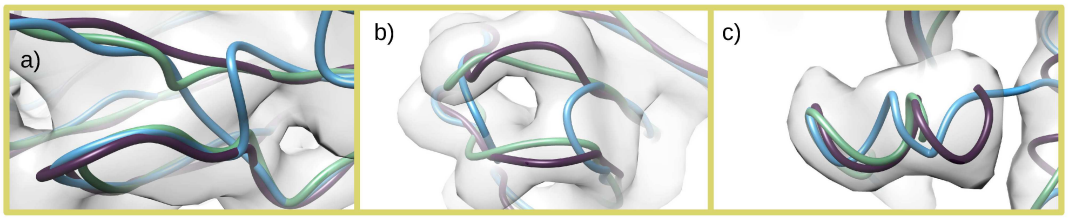


Figure 6.3.: Correcting topology errors: Shown is the PDB structure (dark purple), the conventional trace (blue) and the integrative trace (green). **a)** and **b)** are showing detailed views of the Bordetella bacteriophage cementing protein (EMD-5764, PDB 3J4U:H), **c)** shows details of the T20S proteasome α-subunit (EMD-5623, PDB: 3J9I:K). The conventional trace contains wrong connections. Those topology errors are fixed in the integrative trace.

around sequence positions 20, 50, 105 and 110. Rearranging the subtraces, such that the order of the subtraces, denoted by the red circled numbers in Figure 6.2e) changes to 1, 4, 2 (flipped), 3 (flipped), and 5, results in a trace with the correct topology.

Information about predicted inter-residue distances corrects topology errors

A more visual impression of topology errors and how they can be corrected is given in Figure 6.3. Figure 6.3a) shows a close-up of two β-strands in the Bordetella bacteriophage cementing protein (EMD-5764, PDB 3J4U:H). As β-strands run very close to each other (about 4.7 Å), there is a high probability that a TSP solver will mistakenly create a cross-connection between two neighboring β-strands. Taking into account global distance distributions given by predicted inter-residue distances can prevent those cross-connections, such that the complete β-strands run in parallel to each other in the integrative trace. In Figure 6.3b), large side-chain densities biased the conventional trace to form wrong connections. The pathway followed by these wrong connections may appear as a possible alternative on the local level, but it results in a false global topology. Rearranging the involved subtraces and adapting the connectivity pattern between them corrected the error in the integrative trace. Figure 6.3c) shows the terminus of the T20S proteasome α-subunit protein (EMD-5623, PDB: 3J9I:K) [111], which was not identified correctly by the conventional trace, since the terminus is surrounded by other density regions. The additional information provided by the predicted inter-residue distances facilitated the identification of termini, so that the terminus was correctly detected in the integrated trace.

We tested our method on seven experimental density maps in a resolution range from 3.2 Å to 4.8 Å [110–114] (see Table 6.1) . An overview of the results can

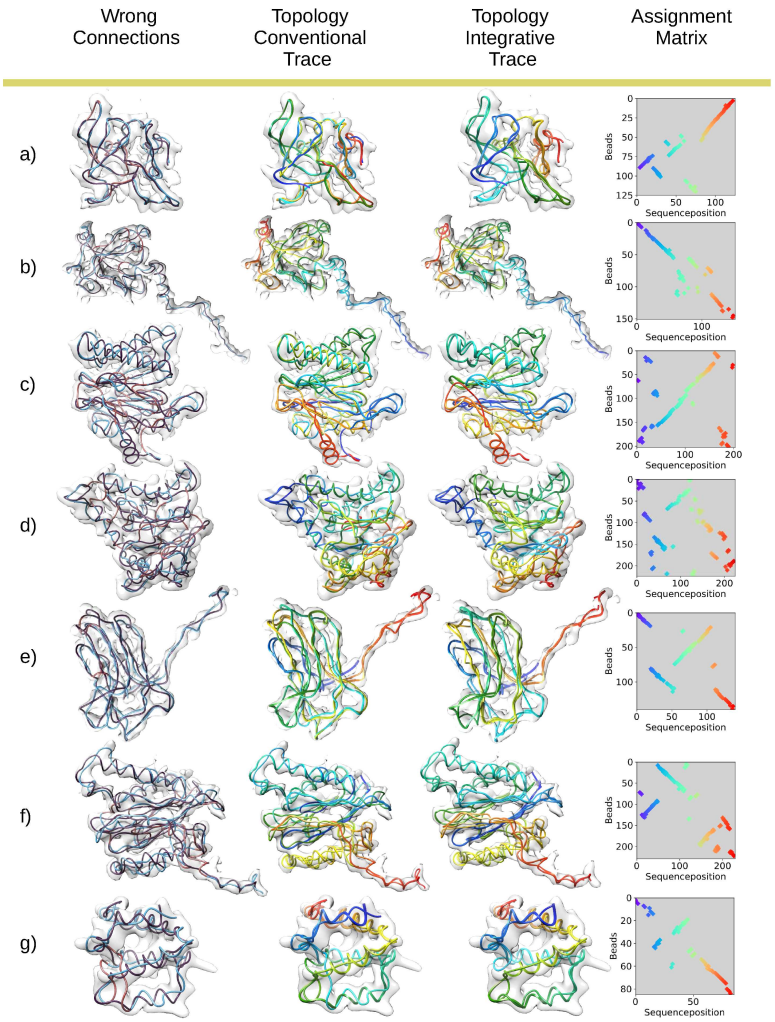


Figure 6.4.: Tracing results: First column shows an overlay of the PDB structure (dark purple) and the conventional trace (blue). Wrong connections in the conventional trace are marked red. Second column shows the same overlay, but both traces are colour-coded from the N-terminus (blue) to the C-terminus (red). Different topologies become recognisable as corresponding parts of the two traces showing different colours. An overlay of the PDB structure and the integrative trace can be seen in the third column. Traces are again colour-coded from the N-terminus (blue) to the C-terminus (red). Matching colours indicate matching topologies. Column 4 shows the assignment matrix, also colour-coded from N- to C-terminus (blue to red). Wrong connections correspond to breaks in diagonals, resulting deviations in the sequence assignment are recognisable as rotations or shifts relative to the main diagonal. **a)** EMD-2566, PDB: 3J6B:I, **b)** EMD-2566, PDB: 3J6B:M, **c)** EMD-5623, PDB: 3J9I:F, **d)** EMD-5623, PDB: 3J9I:K, **e)** EMD-5764, PDB 3J4U:H, **f)** EMD-3535, PDB: 5MPA and **g)** EMD-3241, PDB: 5FNA.

Protein	EMDB ID	PDB ID	Resolution in Å	N_{res}
Yeast mitochondrial large ribosomal subunit	EMD-2566	3J6B:I	3.2	125
Yeast mitochondrial large ribosomal subunit	EMD-2566	3J6B:M	3.2	151
T20S proteasome β-subunit	EMD-5623	3J9I:F	3.3	203
T20S proteasome α-subunit	EMD-5623	3J9I:K	3.3	224
Bordetella bacteriophage cementing protein	EMD-5764	3J4U	3.5	140
26S proteasome	EMD-3535	5MPA	4.5	229
Caspase-1 CARD	EMD-3535	5FNA	4.8	85

Table 6.1.: List of test proteins. The method was applied to seven experimental density maps with resolutions between 3.2 Å and 4.8 Å. Deposited PDB structures serve for validation of the reconstructed trace.

be seen in Figure 6.4. The first column shows an overlay of the conventional trace and the PDB structure. Wrong connections in the conventional trace are marked in red. We found wrong connections in all seven examples. However, the number of false connections varied significantly. For example, Figure 6.4 a), e) and g) exhibit only some isolated deviations of the conventional trace from the PDB structure, but in Figure 6.4d) the conventional trace runs only partially alongside the corresponding PDB structure. This is also apparent, when looking at the assignment matrices in the fourth column of Figure 6.4. While there are relatively long diagonals with only a few breaks in a), e) and g), the pattern of the matrix in d) appears more scattered. Local false connections lead to errors in the global topology. This is illustrated in the second column of Figure 6.4, where the same overlay as in column 1 is shown, but with another colour code. Both traces are coloured blue to red following the chain from the N-terminus to the C-terminus. Deviations in topology appear as deviations in colour between two corresponding strands of the traces. For example for PDB ID 3J9I:F, the upper helix in the front (column 2 in Figure 6.4c)) is cyan in the PDB structure but yellow in the conventional trace. This means that the helix should be located closer to the N-terminus, compared to its assigned position in the conventional trace. Similarly for PDB ID 5MPA, the lower loop in Figure 6.4f) should be blue, however, it is green in the conventional trace showing that it has been assigned to a position that is too close to the C-terminus. The third column shows the corresponding overlay for the integrated trace. Here, all topology errors of the conventional trace have been corrected. Consequently, all corresponding parts of the traces appear in the same colour.

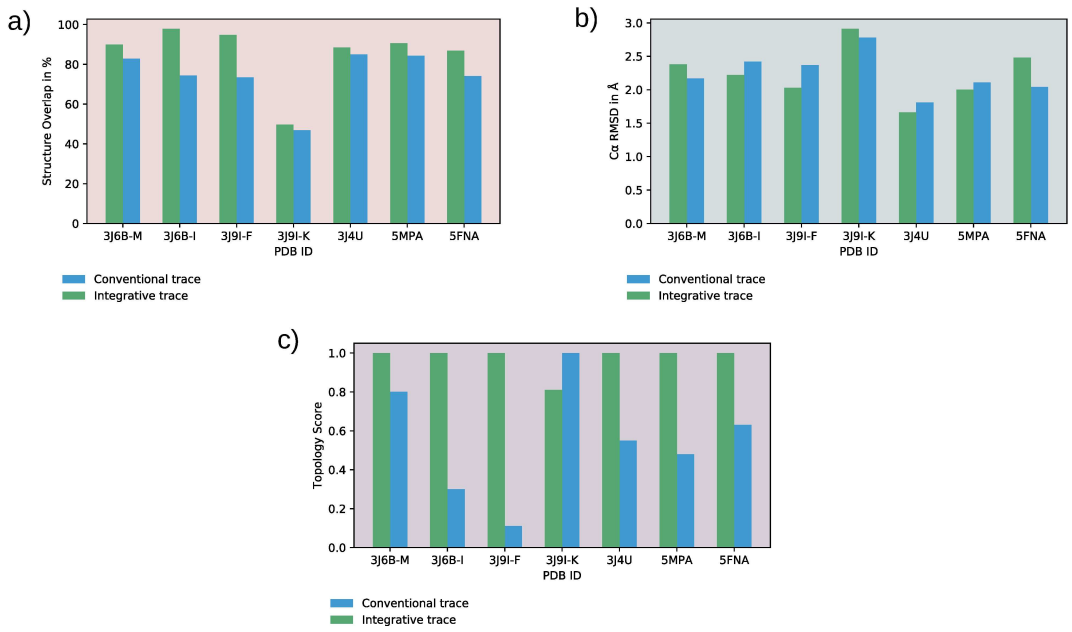


Figure 6.5.: Similarity to the PDB structure assessed with CLICK: Bars denote scores obtained by comparing the PDB structure with the conventional trace (blue) or with the integrative trace (green). **a)** Structure overlap, **b)** RMSD between matched C_α atoms, and **c)** topology score.

Integrative traces outperform conventional traces with regard to topology scores and structure overlap

To quantify the quality of the integrative traces we used the CLICK method [54]. CLICK superimposes two structures based on local fold similarities. A sequence assignment is not used in the alignment. After superposition, we consider three similarity scores that are calculated by CLICK: The structure overlap describes the percentage of C_α atoms of protein A which are in a 3.5 Å radius of their matched C_α atoms in protein B. Among these correctly registered atoms the root mean square deviation (RMSD) gives an additional measure of accuracy. Most relevant in our case is the topology score, which compares global fold patterns. A topology score of 1 means that there is a perfect match between the topology of protein A and the topology of protein B. A protein pair with a topology score of 0 is topologically completely dissimilar. If the superposition is not successful, the calculated measures are unreliable. The scores for our seven test proteins superimposed to the corresponding PDB structure are shown in Figure 6.5.

Figure 6.5a) illustrates the structure overlap. Blue bars correspond to the com-

parison of the conventional trace to the PDB structure, green bars to the comparison of the integrative trace to the PDB structure. It can be seen that the structure overlap could be increased for all the traces by including information derived from predicted inter-residue distances. Six of seven examples have a structure overlap of higher than 80 % indicating a successful structural alignment and a high structural similarity. For one case, the T20S proteasome α-subunit, the alignment by the CLICK algorithm failed, resulting in a low structure overlap for both traces.

The RMSD was calculated based on the correctly matched C_α atoms. The results are shown in Figure 6.5b). All six examples, for which the alignment was successful, yield an RMSD below 2.5 Å. Further, in four of the six examples the integrative trace yielded a lower RMSD to the PDB structure, than the conventional trace. But, for the yeast mitochondrial large ribosomal subunit chain M (PDB 3J6B:M) as well as for the caspase 1 CARD (PDB 5FNA) the RMSD was increased in the integrated trace compared to the conventional trace. It should be noted that the RMSD is calculated over different sets of matched C_α atoms and are therefore not directly comparable. Considering the concurrent increase in structure overlap for the integrative trace, the set of atoms used for RMSD calculation is larger. Therefore, the increase in RMSD is caused by beads that could not be matched in the conventional trace, but are considered in the RMSD calculation for the integrated trace. For the T20S α-subunit, the RMSD calculation was unreliable due to the failed structure alignment, as mentioned above.

Figure 6.5c) shows the topology scores of the conventional trace and the integrated trace. The scores show that including information about predicted inter-residue distances improved the topology score for all properly aligned examples. Moreover, for all of these six cases, the integrative trace yields a topology score of 1, indicating a perfect matching topology with the corresponding PDB structure. The topology of the T20S proteasome α-subunit was examined manually and it could be revealed, that the topology of the integrative trace is identical to the topology of the PDB structure.

Accuracy of predicted distance matrices

In the presented method predicted inter-residue distances play a key role. In the context of backbone tracing based on cryo-EM density maps, they provide an additional source of information for the solution of the problem, which is independent of map resolution. Therefore, considering predicted inter-residue distances can be particularly helpful for maps of lower resolution. However, it should be noted, that predicted inter-residue distances cannot be treated as ground truth information, but they may contain errors and inaccuracies themselves. To assess the quality of the distance matrices used in this study we compared the predicted distance matrix to the distance matrix of the corresponding deposited PDB structure. The

PDB ID	Mean deviation (all distances)	Fraction of outliers	Mean deviation (outliers)	Maximum deviation (outliers)
3J6B:I	(2.6 ± 3.0)Å	9%	(5.8 ± 2.9)Å	6.3 Å
3J6B:M	(2.6 ± 3.4)Å	11%	(6.6 ± 3.7)Å	15.3 Å
3J9I:F	(2.3 ± 2.7)Å	4%	(5.3 ± 2.6)Å	7.8 Å
3J9I:K	(2.3 ± 2.7)Å	5%	(5.4 ± 1.6)Å	10.3Å
3J4U	(2.6 ± 3.1)Å	8%	(6.0 ± 3.0)Å	14.3Å
5MPA	(2.3 ± 2.8)Å	5%	(5.7 ± 3.0)Å	8.3Å
5FNA	(3.1 ± 3.2)Å	14%	(5.8 ± 2.9)Å	10.3Å

Table 6.2.: Accuracy of distances predicted by trRosetta.

results are summarised in Table 6.2. Note that for the comparison it was necessary to match the distribution of distances between C_α atoms of the PDB structures (in Å) to the predicted distances, which are originally given in dimensionless bins. The matching of the distribution was done analogously to matching the distribution of bead distances to the predicted distances as described in *Calculation of inter-bead distances* of the STAR Methods.

The quality of the prediction was assessed using different metrics. The mean deviation provides a general measure of similarity between PDB distances and predicted distances. A prediction is considered correct, if the difference between the distances was < 6 bins, which corresponds to 4.75 Å (see STAR Methods). A distance is considered an outlier, if the deviation ≥ 6 bins and the fraction of outliers was calculated. Furthermore, Table 6.2 gives the maximum and the mean deviation including standard deviation of the outliers only. While the quality of distance matrices varies between the test cases, all predictions were sufficient to identify and correct topology errors in the traces. For all the tested proteins, more than 85 % of the predictions were correct, ensuring a sufficient representation of the folding patterns, so that the remaining deviations could be tolerated.

6.2.4. Discussion

We developed a tool that incorporates information about predicted inter-residue distances into a backbone tracing routine to correct errors in the global topology. By comparing the topology of a conventionally built trace with the topology given by predicted inter-residue distances we can identify correct subtraces as well as false connections. We interpret the comparison of topologies as an assignment problem and solve it by minimising the difference between distance matrices of predicted distances and distances between beads. This eventually enables rearranging of the correct subtraces and eliminating false connections. Resulting integrated traces

outperform traces built conventionally regarding structure overlap and topology score.

Both traces, the conventional and the integrative trace, considered here are C_α traces and no complete backbones yet. Building the complete backbone structure based on the C_α trace could be performed by a fragment fitting procedure similar to the approach described in [49]. Applying such an approach could mitigate some deviations in the bead placement and improve the RMSD. A backbone built this way would profit from the topology determined by our method and the atom placements based on the fragment fitting. We work currently on such a framework that combines the topology tracing with a backbone fragment fitting procedure.

It should be further noted, that, while the global topology can be determined correctly, not all structural features can necessarily be identified. For example, the integrative trace of the 26s proteasome (PDB: 5MPA) does not contain the loop consisting of residues Leu127 to Ser138. In general there are two possible reasons, why a part of a structure is not included in the integrated trace. The first reason is that there are no beads placed in the corresponding density regions. This may be the case for very flexible parts of the protein where density often is not well resolved. The integrative trace is based on the same beads as the conventional trace and cannot pass through density regions that are not traversed by the conventional trace. A second reason may be a cluster of false connections between the beads located in the density region of interest. The integrative trace is based on rearranged subtraces. The assignment of beads that are not part of a subtrace is much more difficult, because they do not get assigned an initial weight in step 2 of the workflow. In consequence, a cluster of such isolated beads may not be assigned correctly and also may not be included in the integrative trace.

The placement of beads is a decisive factor not only for the conventional trace but also for the integrative trace. Poorly placed beads can be sorted out as outliers when transforming the assignment weights back into a trace, but, nevertheless, they can cause problems, if there are too many of them. If there are cliques of poorly placed beads, it becomes difficult to distribute them along the trace and the optimisation of the weights gets distorted, because the assignment aims for a one-to-one matching of beads to residues. The combination of the correct assignment of the well-placed beads and the assignment of the poorly placed beads to the remaining residues might not be the minimum of the scoring function anymore. In that case, the algorithm cannot find the correct assignment anymore.

Observed fluctuations in accuracy of predicted distances did not show strong influences on the final tracing results. However, it may have been advantageous that examples with worse quality of predicted distance maps, e.g. 3J6B:M and 5FNA, are associated with rather small proteins consisting of less than 200 residues. For those proteins the topology problem is less complex than for bigger proteins. Look-

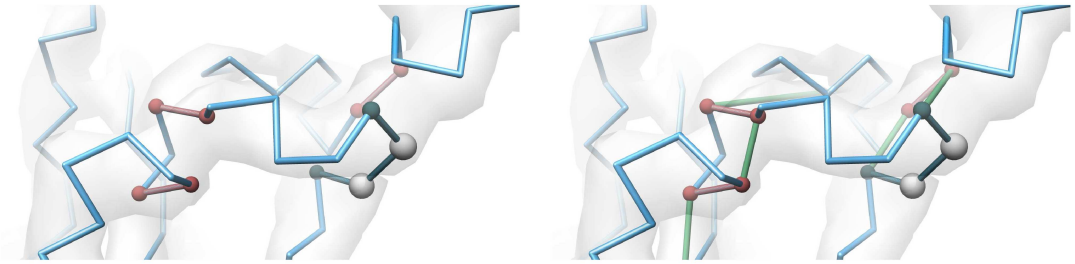


Figure 6.6.: Visualisation of topology problems with Chimera: dxTopology gives a .bild-file which can be opened in Chimera and visualises possible topology problems. Shown is a part of the conventional trace of the 26s proteasome (EMD-3535) in blue. The .bild file colours connections that may be wrong in red, beads that could not be assigned in grey, and connections to beads that could not be assigned in dark grey. Alternatives to the wrong connections are depicted in green.

ing at larger examples, 3J9I:K or 5MPA, the predicted distances were more accurate, supporting the successful determination of the protein topology. Moreover, topology is defined by distance patterns, describing the relationship between various distances, not single values of distances. Therefore the approach is rather robust against local errors in distance prediction. A certain degree of deviation is due to a systematic difference: while predicted inter-residue distances relate to C_β atoms, beads are placeholders for C_α atoms. Considering the typical bond length between C_α and C_β atoms of 1.5 Å and a fixed distance between C_α atoms, the distance between C_β atoms can vary by 6 Å between a conformation where side-chains are pointing towards each other and a conformation where side-chains are pointing away from each other. The observation, that those variations seem not to hamper the tracing process confirms the robustness of our approach.

dxTopology cannot only be used to build a new C_α trace in a cryo-EM map, but also to assess the topology of manually built traces or traces modelled with other software. In that case, only step 2 and 3 of the workflow are performed and the topology of the input and resulting integrative trace are compared. The difference between the traces shows possibly wrong connections and suggests alternative connections, which can easily be visualised by Chimera [36] (see Figure 6.6). This way a modeller can inspect a trace manually with regard to topology and may use the suggestions as guidance for trace improvements.

It could be considered that the distance information could also come from other sources, such as NMR and cross-linking experiments, which could be incorporated in a similar way.

In conclusion, we present a novel approach to improve backbone tracing in

medium-resolution density maps by incorporating information from predicted inter-residue distances. These predictions are in particular helpful to determine the correct topology of a protein trace. The benefit is especially large for challenging map resolutions, since the informative value of distance predictions is completely independent of map resolution.

6.2.5. Contribution

For this study, I developed the method under the supervision of Gunnar Schröder, performed all calculations, made all figures and wrote the manuscript.

7. Conclusion

In this thesis four different approaches for the automatic structural interpretation of cryo-EM density maps were presented. Each of them focuses on modelling a different feature of protein structure based on the information provided by the cryo-EM data. The amount of required prior structural information derived from the map decreased from study to study:

In chapter 3 a method was introduced that adapts a protein structure, originally determined with another experiment, to a cryo-EM density map, while regarding the conformational heterogeneity embodied in the map. Here, the rational is, that a cryo-EM map is based on not a single conformation of a protein but on a whole ensemble captured under near-native conditions in an aqueous solution. The atomic structure associated with such a map, should therefore be an average of all conformations present in the imaged ensemble. Conformational heterogeneity in structure space result in varying local resolution in map space. Thus, the averaging of coordinates is augmented with B-factor fitting, assigning each residue with an atomic displacement parameter, describing the variations of the residues position within different conformations. Our presented method yielded pleasing results in the model metrics challenge 2019. It can be applied for structural interpretation of cryo-EM maps when a structure from another experiment is available.

Chapter 4 presented an approach for automatic side-chain sampling. Using the example of a cryo-EM data set of IAPP fibrils it is shown how computational tools can open up profound insights about side-chain assignments where the density map does not allow for reliable manual modelling. Having a manually built backbone at hand, we automatically performed all possible side-chain assignments and ranked the resulting models according to their model-to-map fit. Computational approaches outperform manual model building in terms of runtime and provide insights that are independent of the personal judgement of the modeller. Our approach is helpful in situations where the data is ambiguous and provides a quantitative structural analysis of the density map.

Chapter 5 described a procedure for flexible fragment fitting with DireX. Fragment fitting enables one to interpret a given protein trace in terms of backbone conformations found in other proteins. Our investigations led to the conclusion,

that introducing a flexible fitting routine, instead of a rigid fitting routine allows to use smaller fragment libraries and benefits accuracy. This is particularly true for loop and sheet regions. In these regions, information derived from the map seem to be more important than information from known reference conformations. Further, metrics for assessing fragment placements should reliably detect poor placements, these requirements seem to be met by the i-o score, but not by the average map value per atom. In the future, the presented method can be integrated in the modelling framework EMfasa.

Lastly, subject of chapter 6 was a method to automatically determine the topology or the trace of a protein based on a cryo-EM map. Hereby, the tracing is guided by predicted inter-residue distances. Other prior structural information is not used. It was shown that integrating information from predicted inter-residue distances improves traces that were built based on density information alone. The method provides a quick initial representation of the protein trace including the sequence assignment. It can also be used to assess traces modelled with other tools and suggests alternatives were it suspect a connection between atoms to be wrong. In the future, we would like to provide a Chimera plug-in to assess and, if necessary, correct connections on the fly with a simple GUI. Moreover, the tool might also be helpful to identify single chains or domains in a multi chain protein complex, details about the workflow of such an approach need still to be worked out.

Together, these four studies cover the whole transformation between map and model. They contribute to our holistic understanding of how an atomic protein structure is represented by its density map and how we can extract this information using computational tools. Their application helps to solve protein structures using cryo-EM. Therefore, this work can be of help in the context of biomedical research, drug development or biotechnology.

Bibliography

[1] Carl Ivar Branden and John Tooze. *Introduction to protein structure*. Garland Publishing, 1999.

[2] Eric D Scheeff and J Lynn Fink. 'Fundamentals of protein structure'. In: *Methods of Biochemical Analysis* 44 (2003), pp. 15–40.

[3] Richard Henderson. 'From electron crystallography to single particle cryoEM (Nobel Lecture)'. In: *Angewandte Chemie International Edition* 57.34 (2018), pp. 10804–10825.

[4] Jacques Dubochet. 'On the Development of Electron Cryo-Microscopy (Nobel Lecture)'. In: *Angewandte Chemie International Edition* 57.34 (2018), pp. 10842–10846.

[5] Joachim Frank. 'Single-Particle Reconstruction of Biological Molecules—Story in a Sample (Nobel Lecture)'. In: *Angewandte Chemie International Edition* 57.34 (2018), pp. 10826–10841.

[6] Eva Nogales. 'The development of cryo-EM into a mainstream structural biology technique'. In: *Nature methods* 13.1 (2016), pp. 24–27.

[7] Kazuyoshi Murata and Matthias Wolf. 'Cryo-electron microscopy for structural analysis of dynamic biological macromolecules'. In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1862.2 (2018), pp. 324–334.

[8] Stefan Pfeffer and Julia Mahamid. 'Unravelling molecular complexity in structural cell biology'. In: *Current opinion in structural biology* 52 (2018), pp. 111–118.

[9] Werner Kühlbrandt. 'The resolution revolution'. In: *Science* 343.6178 (2014), pp. 1443–1444.

[10] Ka Man Yip et al. 'Atomic-resolution protein structure determination by cryo-EM'. In: *Nature* 587.7832 (2020), pp. 157–161.

[11] Takanori Nakane et al. 'Single-particle cryo-EM at atomic resolution'. In: *Nature* 587.7832 (2020), pp. 152–156.

[12] Jacqueline LS Milne et al.
‘Cryo-electron microscopy–a primer for the non-microscopist’.
In: *The FEBS journal* 280.1 (2013), pp. 28–45.

[13] Niels Volkmann and Dorit Hanein. ‘ELECTRON MICROSCOPY IN THE CONTEXT OF STRUCTURAL SYSTEMS BIOLOGY’.
In: *Structural Bioinformatics* 44 (2009), p. 143.

[14] Allison Doerr. ‘Single-particle cryo-electron microscopy’.
In: *Nature methods* 13.1 (2016), pp. 23–23.

[15] Pamela A Thuman-Commike. ‘Single particle macromolecular structure determination via electron microscopy’.
In: *Febs Letters* 505.2 (2001), pp. 199–205.

[16] Mara Zielinski, Christine Röder and Gunnar F Schröder. ‘Challenges in sample preparation and structure determination of amyloids by cryo-EM’.
In: *Journal of Biological Chemistry* 297.2 (2021).

[17] Eva Nogales and Sjors HW Scheres. ‘Cryo-EM: a unique tool for the visualization of macromolecular complexity’.
In: *Molecular cell* 58.4 (2015), pp. 677–689.

[18] Michael Radermacher. ‘Three-dimensional reconstruction of single particles from random and nonrandom tilt series’.
In: *Journal of electron microscopy technique* 9.4 (1988), pp. 359–394.

[19] Marin Van Heel. ‘Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction’.
In: *Ultramicroscopy* 21.2 (1987), pp. 111–123.

[20] Hans Elmlund, Dominika Elmlund and Samy Bengio.
‘PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy’. In: *Structure* 21.8 (2013), pp. 1299–1306.

[21] Gunnar F Schröder, Axel T Brunger and Michael Levitt.
‘Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution’.
In: *Structure* 15.12 (2007), pp. 1630–1641.

[22] Zhe Wang and Gunnar F Schröder. ‘Real-space refinement with DireX: From global fitting to side-chain improvements’.
In: *Biopolymers* 97.9 (2012), pp. 687–697.

[23] Pavel V Afonine et al.
‘Real-space refinement in PHENIX for cryo-EM and crystallography’.
In: *Acta Crystallographica Section D: Structural Biology* 74.6 (2018), pp. 531–544.

[24] Paul Emsley and Kevin Cowtan.
'Coot: model-building tools for molecular graphics'.
In: *Acta crystallographica section D: biological crystallography* 60.12 (2004), pp. 2126–2132.

[25] Marin Van Heel and Michael Schatz.
'Fourier shell correlation threshold criteria'.
In: *Journal of structural biology* 151.3 (2005), pp. 250–262.

[26] Sjors HW Scheres and Shaoxia Chen.
'Prevention of overfitting in cryo-EM structure determination'.
In: *Nature methods* 9.9 (2012), pp. 853–854.

[27] Peter B Rosenthal and Richard Henderson.
'Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy'.
In: *Journal of molecular biology* 333.4 (2003), pp. 721–745.

[28] Satinder Kaur et al. 'Local computational methods to improve the interpretability and analysis of cryo-EM maps'.
In: *Nature communications* 12.1 (2021), pp. 1–12.

[29] Giovanni Cardone, J Bernard Heymann and Alasdair C Steven.
'One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions'.
In: *Journal of structural biology* 184.2 (2013), pp. 226–236.

[30] F DiMaio and W Chiu. 'Tools for model building and optimization into near-atomic resolution electron cryo-microscopy density maps'.
In: *Methods in enzymology* 579 (2016), pp. 255–276.

[31] Matthew L Baker et al.
'Analyses of subnanometer resolution cryo-EM density maps'.
In: *Methods in enzymology*. Vol. 483. Elsevier, 2010, pp. 1–29.

[32] Sony Malhotra et al. 'Modelling structures in cryo-EM maps'.
In: *Current Opinion in Structural Biology* 58 (2019), pp. 105–114.

[33] Sjors HW Scheres. 'RELION: implementation of a Bayesian approach to cryo-EM structure determination'.
In: *Journal of structural biology* 180.3 (2012), pp. 519–530.

[34] Ali Punjani et al. 'cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination'.
In: *Nature methods* 14.3 (2017), pp. 290–296.

[35] Melissa R Pitman and R Ian Menz.
'Methods for protein homology modelling'.
In: *Applied mycology and biotechnology*. Vol. 6. Elsevier, 2006, pp. 37–59.

[36] Eric F Pettersen et al. 'UCSF Chimera—a visualization system for exploratory research and analysis'.
In: *Journal of computational chemistry* 25.13 (2004), pp. 1605–1612.

[37] BL De Groot et al.
'Prediction of protein conformational freedom from distance constraints'.
In: *Proteins: Structure, Function, and Bioinformatics* 29.2 (1997), pp. 240–251.

[38] Leonardo G Trabuco et al. 'Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics'.
In: *Single-Particle Cryo-Electron Microscopy: The Path Toward Atomic Resolution: Selected Papers of Joachim Frank with Commentaries*. World Scientific, 2008, pp. 433–443.

[39] Paul D Adams et al. 'The Phenix software for automated determination of macromolecular structures'. In: *Methods* 55.1 (2011), pp. 94–106.

[40] Dong C Liu and Jorge Nocedal.
'On the limited memory BFGS method for large scale optimization'.
In: *Mathematical programming* 45.1 (1989), pp. 503–528.

[41] Muyuan Chen et al.
'De Novo modeling in cryo-EM density maps with Pathwalking'.
In: *Journal of structural biology* 196.3 (2016), pp. 289–298.

[42] Genki Terashi and Daisuke Kihara.
'De novo main-chain modeling for EM maps using MAINMAST'.
In: *Nature communications* 9.1 (2018), pp. 1–11.

[43] Thomas C Terwilliger et al. 'Cryo-EM map interpretation and protein model-building using iterative map segmentation'.
In: *Protein science* 29.1 (2020), pp. 87–99.

[44] Tatjana Braun. 'Protein Structure Modelling using Evolutionary Information and Cryo-EM Data'. PhD thesis. 2017.

[45] Jonas Pfab, Nhut Minh Phan and Dong Si.
'DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes'.
In: *Proceedings of the National Academy of Sciences* 118.2 (2021).

[46] Michael Jünger, Gerhard Reinelt and Giovanni Rinaldi.
'The traveling salesman problem'. In: *Handbooks in operations research and management science* 7 (1995), pp. 225–330.

[47] Keld Helsgaun. 'An effective implementation of the Lin–Kernighan traveling salesman heuristic'.
In: *European journal of operational research* 126.1 (2000), pp. 106–130.

[48] Piotr Rotkiewicz and Jeffrey Skolnick. 'Fast procedure for reconstruction of full-atom protein models from reduced representations'.
In: *Journal of computational chemistry* 29.9 (2008), pp. 1460–1465.

[49] Ray Yu-Ruei Wang et al. 'De novo protein structure determination from near-atomic-resolution cryo-EM maps'.
In: *Nature methods* 12.4 (2015), pp. 335–338.

[50] Vincent B Chen et al. 'MolProbity: all-atom structure validation for macromolecular crystallography'. In: *Acta Crystallographica Section D: Biological Crystallography* 66.1 (2010), pp. 12–21.

[51] C Ramakrishnan and GN Ramachandran.
'Stereochemical criteria for polypeptide and protein chain conformations: II. Allowed conformations for a pair of peptide units'.
In: *Biophysical journal* 5.6 (1965), pp. 909–933.

[52] Benjamin Falkner and Gunnar F Schröder.
'Cross-validation in cryo-EM–based structural modeling'. In: *Proceedings of the National Academy of Sciences* 110.22 (2013), pp. 8930–8935.

[53] Irina Kufareva and Ruben Abagyan.
'Methods of protein structure comparison'. In: *Homology modeling*. Springer, 2011, pp. 231–257.

[54] MN Nguyen, Kuan Pern Tan and Mallur S Madhusudhan.
'CLICK—topology-independent comparison of biomolecular 3D structures'. In: *Nucleic acids research* 39.suppl_2 (2011), W24–W28.

[55] Catherine L Lawson et al. 'Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge'.
In: *Nature methods* 18.2 (2021), pp. 156–164.

[56] EMDataResource. *Unified Data Resource for 3DEM*.
URL: `http://emdataresource.org/mission.html`.
(accessed: 15.07.2022).

[57] EMDataResource. *2019 Model Metrics Challenge*.
URL: `https://challenges.emdataresource.org/?q=model-metrics-challenge-2019`. (accessed: 15.07.2022).

[58] Grigore Pintilie et al.
'Measurement of atom resolvability in cryo-EM maps with Q-scores'.
In: *Nature methods* 17.3 (2020), pp. 328–334.

[59] Taro Masuda et al.
'The universal mechanism for iron translocation to the ferroxidase site in ferritin, which is mediated by the well conserved transit site'.
In: *Biochemical and biophysical research communications* 400.1 (2010), pp. 94–99.

[60] Helen M Berman et al. 'The protein data bank'.
In: *Nucleic acids research* 28.1 (2000), pp. 235–242.

[61] Herman JC Berendsen et al.
'Molecular dynamics with coupling to an external bath'.
In: *The Journal of chemical physics* 81.8 (1984), pp. 3684–3690.

[62] Axel T Brünger et al. 'Crystallography & NMR system: A new software suite for macromolecular structure determination'.
In: *Acta Crystallographica Section D: Biological Crystallography* 54.5 (1998), pp. 905–921.

[63] Axel T Brunger. 'Version 1.2 of the Crystallography and NMR system'.
In: *Nature protocols* 2.11 (2007), pp. 2728–2733.

[64] Pavel V Afonine et al. 'New tools for the analysis and validation of cryo-EM maps and atomic models'. In: *Acta Crystallographica Section D: Structural Biology* 74.9 (2018), pp. 814–840.

[65] Benjamin A Barad et al. 'EMRinger: side chain–directed model and map validation for 3D cryo-electron microscopy'.
In: *Nature methods* 12.10 (2015), pp. 943–946.

[66] Christopher J Williams et al. 'MolProbity: More and better reference data for improved all-atom structure validation'.
In: *Protein Science* 27.1 (2018), pp. 293–315.

[67] EMDataResource. *Model ranks per target*. URL: `http://model-compare.emdataresource.org/2019/cgi-bin/em_model_ranks.cgi`. (accessed: 19.07.2022).

[68] Björn Wallner and Arne Elofsson.
'Can correct protein models be identified?'
In: *Protein science* 12.5 (2003), pp. 1073–1086.

[69] Christine Roeder et al. 'Cryo-EM structure of islet amyloid polypeptide fibrils reveals similarities with amyloid-β fibrils'.
In: *Nature Structural & Molecular Biology* 27.7 (2020), pp. 660–667.

[70] Douglas M Fowler et al. ‘Functional amyloid–from bacteria to humans’. In: *Trends in biochemical sciences* 32.5 (2007), pp. 217–224.

[71] Fabrizio Chiti and Christopher M Dobson. ‘Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade’. In: *Annu. Rev. Biochem* 86.1 (2017), pp. 27–68.

[72] Lothar Gremer et al. ‘Fibril structure of amyloid-β (1–42) by cryo–electron microscopy’. In: *Science* 358.6359 (2017), pp. 116–119.

[73] Christine Röder. ‘A Glimpse into the Polymorphic Landscape of Amyloids-Structural Investigation of Amyloid Fibrils by Cryo-Electron Microscopy’. PhD thesis. 2021.

[74] Matthew G Iadanza et al. ‘A new era for understanding amyloid structures and disease’. In: *Nature Reviews Molecular Cell Biology* 19.12 (2018), pp. 755–773.

[75] Anthony WP Fitzpatrick et al. ‘Cryo-EM structures of tau filaments from Alzheimer’s disease’. In: *Nature* 547.7662 (2017), pp. 185–190.

[76] Dieter Willbold et al. ‘Amyloid-type protein aggregation and prion-like properties of amyloids’. In: *Chemical reviews* 121.13 (2021), pp. 8285–8307.

[77] Rodrigo Gallardo, Neil A Ranson and Sheena E Radford. ‘Amyloid structures: much more than just a cross-β fold’. In: *Current opinion in structural biology* 60 (2020), pp. 7–16.

[78] Georgii G Krivov, Maxim V Shapovalov and Roland L Dunbrack Jr. ‘Improved prediction of protein side-chain conformations with SCWRL4’. In: *Proteins: Structure, Function, and Bioinformatics* 77.4 (2009), pp. 778–795.

[79] Rehana Akter et al. ‘Islet amyloid polypeptide: structure, function, and pathophysiology’. In: *Journal of diabetes research* 2016 (2016).

[80] Per Westermark et al. ‘Islet amyloid polypeptide: pinpointing amino acid residues linked to amyloid fibril formation.’ In: *Proceedings of the National Academy of Sciences* 87.13 (1990), pp. 5036–5040.

[81] Christer Betsholtz et al.
'Sequence divergence in a specific region of islet amyloid polypeptide (IAPP) explains differences in islet amyloid formation between species'.
In: *FEBS letters* 251.1-2 (1989), pp. 261–264.

[82] Konstantinos Tenidis et al.
'Identification of a penta-and hexapeptide of islet amyloid polypeptide (IAPP) with amyloidogenic and cytotoxic properties'.
In: *Journal of molecular biology* 295.4 (2000), pp. 1055–1071.

[83] Michal Jamroz and Andrzej Kolinski.
'ClusCo: clustering and comparison of protein models'.
In: *Bmc Bioinformatics* 14.1 (2013), pp. 1–6.

[84] Scott Kirkpatrick, C Daniel Gelatt Jr and Mario P Vecchi.
'Optimization by simulated annealing'.
In: *science* 220.4598 (1983), pp. 671–680.

[85] Martin Ester et al. 'A density-based algorithm for discovering clusters in large spatial databases with noise.' In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[86] Saul B Needleman and Christian D Wunsch. 'A general method applicable to the search for similarities in the amino acid sequence of two proteins'.
In: *Journal of molecular biology* 48.3 (1970), pp. 443–453.

[87] Osamu Gotoh. 'An improved algorithm for matching biological sequences'.
In: *Journal of molecular biology* 162.3 (1982), pp. 705–708.

[88] Jacek Blazewicz et al. 'Protein alignment algorithms with an efficient backtracking routine on multiple GPUs'.
In: *BMC bioinformatics* 12.1 (2011), pp. 1–17.

[89] Andrej Šali and Tom L Blundell.
'Comparative protein modelling by satisfaction of spatial restraints'.
In: *Journal of molecular biology* 234.3 (1993), pp. 779–815.

[90] T Alwyn Jones and Soren Thirup. 'Using known substructures in protein model building and crystallography.'
In: *The EMBO journal* 5.4 (1986), pp. 819–822.

[91] Liisa Holm and Chris Sander. 'Database algorithm for generating protein backbone and side-chain co-ordinates from a Cα trace: application to model building and detection of co-ordinate errors'.
In: *Journal of molecular biology* 218.1 (1991), pp. 183–194.

[92] Thomas C Terwilliger. 'Automated main-chain model building by template matching and iterative fragment extension'. In: *Acta Crystallographica Section D: Biological Crystallography* 59.1 (2003), pp. 38–44.

[93] Frantisek Pavelcik. 'Automatic model building based on flexible fragment formalism. The case of high-resolution protein structures'. In: *Acta Crystallographica Section A: Foundations of Crystallography* 59.5 (2003), pp. 487–494.

[94] Robert Bücker et al. 'Serial protein crystallography in an electron microscope'. In: *Nature communications* 11.1 (2020), pp. 1–8.

[95] Simon A Fromm et al. 'Seeing tobacco mosaic virus through direct electron detectors'. In: *Journal of structural biology* 189.2 (2015), pp. 87–97.

[96] Paul D Hempstead et al. 'Comparison of the three-dimensional structures of recombinant human H and horse L ferritins at high resolution'. In: *Journal of molecular biology* 268.2 (1997), pp. 424–448.

[97] Debora S Marks et al. 'Protein 3D structure computed from evolutionary sequence variation'. In: *PloS one* 6.12 (2011), e28766.

[98] Jianyi Yang et al. 'Improved protein structure prediction using predicted interresidue orientations'. In: *Proceedings of the National Academy of Sciences* 117.3 (2020), pp. 1496–1503.

[99] Zongyang Du et al. 'The trRosetta server for fast and accurate protein structure prediction'. In: *Nature protocols* 16.12 (2021), pp. 5634–5651.

[100] Ewen Callaway. 'Revolutionary cryo-EM is taking over structural biology.' In: *Nature* 578.7794 (2020), pp. 201–202.

[101] Xiao-Chen Bai, Greg McMullan and Sjors HW Scheres. 'How cryo-EM is revolutionizing structural biology'. In: *Trends in biochemical sciences* 40.1 (2015), pp. 49–57.

[102] Yifan Cheng. 'Single-particle cryo-EM—How did it get here and where will it go'. In: *Science* 361.6405 (2018), pp. 876–880.

[103] Catherine L Lawson et al. 'EMDataBank unified data resource for 3DEM'. In: *Nucleic acids research* 44.D1 (2016), pp. D396–D403.

[104] Jiahua He and Sheng-You Huang. 'Full-length de novo protein structure determination from cryo-EM maps using deep learning'. In: *Bioinformatics* 37.20 (2021), pp. 3480–3490.

[105] Sheng Chen et al. 'SEGEM: a Fast and Accurate Automated Protein Backbone Structure Modeling Method for Cryo-EM'. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2021, pp. 24–31.

[106] John Jumper et al. 'Highly accurate protein structure prediction with AlphaFold'. In: *Nature* 596.7873 (2021), pp. 583–589.

[107] R Service and R Service. 'The game has changed.'AI triumphs at solving protein structures'. In: *Science* 370 (2020), pp. 1144–1145.

[108] Ewen Callaway. ''It will change everything': DeepMind's AI makes gigantic leap in solving protein structures'. In: *Nature* 588.7837 (2020), pp. 203–205.

[109] Minkyung Baek et al. 'Accurate prediction of protein structures and interactions using a three-track neural network'. In: *Science* 373.6557 (2021), pp. 871–876.

[110] Xing Zhang et al. 'A new topology of the HK97-like fold revealed in Bordetella bacteriophage by cryoEM at 3.5 Å resolution'. In: *Elife* 2 (2013), e01299.

[111] Xueming Li et al. 'Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM'. In: *Nature methods* 10.6 (2013), pp. 584–590.

[112] Alvin Lu et al. 'Molecular basis of caspase-1 polymerization and its inhibition by a new capping mechanism'. In: *Nature structural & molecular biology* 23.5 (2016), pp. 416–425.

[113] Alexey Amunts et al. 'Structure of the yeast mitochondrial large ribosomal subunit'. In: *Science* 343.6178 (2014), pp. 1485–1489.

[114] Marc Wehmer et al. 'Structural insights into the functional cycle of the ATPase module of the 26S proteasome'. In: *Proceedings of the National Academy of Sciences* 114.6 (2017), pp. 1305–1310.

[115] Guang Tang et al. 'EMAN2: an extensible image processing suite for electron microscopy'. In: *Journal of structural biology* 157.1 (2007), pp. 38–46.

[116] Liam J McGuffin, Kevin Bryson and David T Jones.
'The PSIPRED protein structure prediction server'.
In: *Bioinformatics* 16.4 (2000), pp. 404–405.

[117] Philipp Mostosi et al. 'Haruspex: a neural network for the automatic identification of oligonucleotides and protein secondary structure in cryo-electron microscopy maps'. In: *Angewandte Chemie International Edition* 59.35 (2020), pp. 14788–14795.

[118] Diederik P Kingma and Jimmy Ba.
'Adam: A method for stochastic optimization'.
In: *arXiv preprint arXiv:1412.6980* (2014).

Acknowledgements

Thank you to ...

... my supervisor Gunnar Schröder for giving me the freedom and trust to follow my own ideas, but supporting and guiding me whenever it was needed. I learned so much from you and had a lot of fun during our scientific discussions.

... all the coauthors of the publications I participated in for great teamwork.

... my colleagues and friends of the schroderlab, for making my office life so incomparably awesome. Mara, Christine, James, Simon, Benedikt and Karunakar, thank you so much for all your help and support and laughter we shared in the office, during coffee- and frisbee-breaks and during all the evenings in Ehrenfeld, Aachen and Düsseldorf.

... Johanna and Greta, for keeping me company in the home office during the pandemic. You made difficult times so much easier and nicer.

... Moritz, because you were always there for me, always believed in me, did everything to support me and always encouraged me.

... and my family, Mechthild, Gert and Julia, for always supporting me, no matter what, and for being my safe haven in life, where I can always find refuge and support. You guys just mean so much to me.

List of Figures

1.1 Fundamental protein structure 2
1.2 Single particle cryo-EM workflow 5

2.1 Structure of the thesis 14

3.1 Targets of the EMDR Model Metrics Challenge 2019 16

4.1 Structural architecture of an amyloid fibril 22
4.2 Top four models for PM2 26

5.1 Workflow of EMfasa 30
5.2 Deformable Elastic Network restraints 32
5.3 Diagram of the workflow for flexible fragment fitting with DireX . 35
5.4 Principle of position restraints and definition of the i-o score for flexible fragment fitting 38
5.5 Metrics for assessment of placement quality 40
5.6 Different library sizes for fragment fitting 42
5.7 Effects of DEN-restraints on fitting accuracy 44
5.8 EMfasa with flexible fragment fitting 46

6.1 Workflow topology tracing 53
6.2 Matrix representations 57
6.3 Corrcting topology errors 58
6.4 Tracing results 59
6.5 Similarity to the PDB structure assessed with CLICK 61
6.6 Visualisation of topology problems 65

List of Tables

6.1 List of test proteins 60
6.2 Accuracy of distances predicted by trRosetta. 63

List of Abbreviations

cryo-EM	Cryogenic Electron Microscopy
NMR	Nuclear Magnetic Resonance
FSC	Fourier Shell Correlation
DEN	Deformable Elastic Network
MDFF	Molecular Dynamics Flexible Fitting
TSP	Travelling Sales Person Problem
RMSD	Root Mean Square Deviation
EMDR	EMDataResource Project
PDB	Protein Data Bank
IAPP	Islet Amyloid Polypeptide
PM	Polymorph
MCSA	Monte Carlo Simulated Annealing
TMV	Tobacco Mosaic Virus
AMVA	Average Map Value per Atom
AMVA-STD	Average Map Value per Atom weighted with its reciprocal standard deviation
MSA	Multiple Sequence Alignment

A. Embedded Publication I

ANALYSIS
https://doi.org/10.1038/s41592-020-01051-w

nature methods

Check for updates

OPEN

Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge

Catherine L. Lawson[1]✉, Andriy Kryshtafovych[2], Paul D. Adams[3,4], Pavel V. Afonine[3], Matthew L. Baker[5], Benjamin A. Barad[6], Paul Bond[7], Tom Burnley[8], Renzhi Cao[9], Jianlin Cheng[10], Grzegorz Chojnowski[11], Kevin Cowtan[7], Ken A. Dill[12], Frank DiMaio[13], Daniel P. Farrell[13], James S. Fraser[14], Mark A. Herzik Jr[15], Soon Wen Hoh[7], Jie Hou[16], Li-Wei Hung[17], Maxim Igaev[18], Agnel P. Joseph[8], Daisuke Kihara[19,20], Dilip Kumar[21], Sumit Mittal[22,23], Bohdan Monastyrskyy[2], Mateusz Olek[7], Colin M. Palmer[8], Ardan Patwardhan[24], Alberto Perez[25], Jonas Pfab[26], Grigore D. Pintilie[27], Jane S. Richardson[28], Peter B. Rosenthal[29], Daipayan Sarkar[19,22], Luisa U. Schäfer[30], Michael F. Schmid[31], Gunnar F. Schröder[30,32], Mrinal Shekhar[22,33], Dong Si[26], Abishek Singharoy[22], Genki Terashi[18], Thomas C. Terwilliger[34], Andrea Vaiana[18], Liguo Wang[35], Zhe Wang[24], Stephanie A. Wankowicz[14,36], Christopher J. Williams[28], Martyn Winn[8], Tianqi Wu[37], Xiaodi Yu[38], Kaiming Zhang[27], Helen M. Berman[39,40] and Wah Chiu[27,31]✉

This paper describes outcomes of the 2019 Cryo-EM Model Challenge. The goals were to (1) assess the quality of models that can be produced from cryogenic electron microscopy (cryo-EM) maps using current modeling software, (2) evaluate reproducibility of modeling results from different software developers and users and (3) compare performance of current metrics used for model evaluation, particularly Fit-to-Map metrics, with focus on near-atomic resolution. Our findings demonstrate the relatively high accuracy and reproducibility of cryo-EM models derived by 13 participating teams from four benchmark maps, including three forming a resolution series (1.8 to 3.1 Å). The results permit specific recommendations to be made about validating near-atomic cryo-EM structures both in the context of individual experiments and structure data archives such as the Protein Data Bank. We recommend the adoption of multiple scoring parameters to provide full and objective annotation and assessment of the model, reflective of the observed cryo-EM map density.

Cryo-EM has emerged as a key method to visualize and model biologically important macromolecules and cellular machines. Researchers can now routinely achieve resolutions better than 4 Å, yielding new mechanistic insights into cellular processes and providing support for drug discovery[1].

The recent explosion of cryo-EM structures raises important questions. What are the limits of interpretability given the quality of maps and resulting models? How can model accuracy and reliability be quantified under the simultaneous constraints of map density and chemical rules?

The EMDataResource Project (EMDR) (emdataresource.org) aims to derive validation methods and standards for cryo-EM structures through community consensus[2]. EMDR has convened an EM Validation Task Force[3] analogous to those for X-ray crystallography[4] and NMR[5] and has sponsored challenges, workshops and conferences to engage cryo-EM experts, modelers and end-users[2,6]. During this period, cryo-EM has evolved rapidly (Fig. 1).

This paper describes outcomes of EMDR's most recent challenge, the 2019 Model 'Metrics' Challenge. Map targets representing the state-of-the-art in cryo-EM single particle reconstruction were selected in the near-atomic resolution regime (1.8–3.1 Å) with a twist: three form a resolution series from the same specimen/imaging experiment. Careful evaluation of submitted models by participating teams leads us to several specific recommendations for validating near-atomic cryo-EM structures, directed toward both individual researchers and the Protein Data Bank (PDB) structure data archive[7].

Results

Challenge design. Challenge targets (Fig. 2) consisted of a three-map human heavy-chain apoferritin (APOF) resolution series (a 500-kDa octahedral complex of 24 α-helix-rich subunits), with maps differing only in the number of particles used in reconstruction[8], plus a single map of horse liver alcohol dehydrogenase (ADH) (an 80-kDa α/β homodimer with NAD and Zn ligands)[9].

A key criterion for target selection was availability of high-quality, experimentally determined model coordinates to serve as references (Fig. 3a). A 1.5 Å X-ray structure[10] served as the APOF reference

A full list of affiliations appears at the end of the paper.

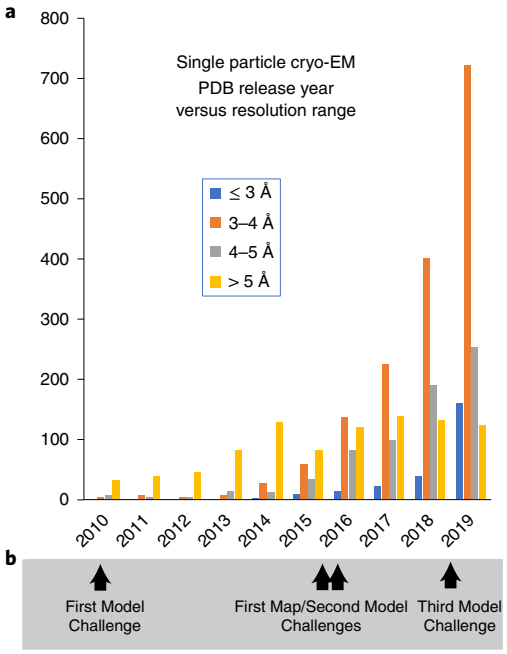


Fig. 1 | Single particle cryo-EM models in the Protein Data Bank. a, Plot of reported resolution versus PDB release year. Models derived from single particle cryo-EM maps have increased dramatically since the 'resolution revolution' circa 2014. Higher-resolution structures (blue bars) are also trending upward. **b**, EMDataResource challenge activities timeline.

since no cryo-EM model was available. The X-ray model provides an excellent although not a fully optimized fit to each map, owing to method/sample differences. For ADH, the structure deposited by the original cryo-EM study authors served as the reference[9].

Thirteen teams from the USA and Europe submitted 63 models in total, using whatever modeling software they preferred, yielding 15–17 submissions per target (Fig. 3b and Table 1). Most (51) were created ab initio, sometimes supported by additional manual steps, while others (12) were optimizations of publicly available models. The estimated human effort per model was 7 h on average, with a wide range (0–80 h).

Submitted models were evaluated as in the previous challenge[11] with multiple metrics in each of four tracks: Fit-to-Map, Coordinates-only, Comparison-to-Reference and Comparison-among-Models (Fig. 3c). The metrics include many in common use as well as several recently introduced.

Metrics to evaluate global Fit-to-Map included Map-Model Fourier shell correlation (FSC)[12], FSC average[13], Atom Inclusion[14], EMRinger[15], density-based correlation scores from TEMPy[16–18], Phenix[19] and the recently introduced *Q*-score to assess atom resolvability[8].

Metrics to evaluate overall Coordinates-only quality included Clashscore, Rotamer outliers and Ramachandran outliers from MolProbity[20], as well as standard geometry measures (for example, bond, chirality, planarity) from Phenix[21]. PDB currently uses all of these validation measures based on community recommendations[3–5]. New to this challenge round was CaBLAM, which evaluates protein backbone conformation using virtual dihedral angles[22].

Metrics assessing similarity of model to reference included Global Distance Test[23], Local Difference Distance Test[24], CaRMSD[25] and Contact Area Difference[26]. Davis-QA was used to measure similarity among submitted models[27]. These measures are widely used in critical assessment of protein structure prediction (CASP) competitions[27].

Several metrics were also evaluated per residue. These were Fit-to-Map: EMRinger[15], *Q*-score[8], Atom Inclusion[14], SMOC[18] and CCbox[19]; and for Coordinates-only: Clashes, Ramachandran outliers[20] and CaBLAM[22].

Evaluated metrics are tabulated with brief definitions in Table 2 and extended descriptions are provided in Methods.

An evaluation system website with interactive tables, plots and tools (Fig. 3d) was established to organize and enable analysis of the challenge results and make the results accessible to all participants (model-compare.emdataresource.org).

Overall and local quality of models. Most submitted models scored well, landing in 'acceptable' regions in each of the evaluation tracks, and in many cases performing better than the associated reference structure that served as a control (Supplementary Fig. 1). Teams that submitted ab initio models reported that additional manual adjustment was beneficial, particularly for the two lower resolution targets.

Evaluation exposed four fairly frequent issues: mis-assignment of peptide-bond geometry, misorientation of peptides, local sequence misalignment and failure to model associated ligands. Two-thirds of submitted models had one or more peptide-bond geometry errors (Extended Data Fig. 1).

At resolutions near 3 Å or in weak local density, the carbonyl O protrusion disappears into the tube of backbone density (Fig. 2), and *trans* peptide bonds are more readily modeled in the wrong orientation. If peptide torsion ϕ (C,N,C_α,C), ψ (N,C_α,C,N) values are explicitly refined, adjacent sidechains can be pushed further in the wrong direction. Such cases are not flagged as Ramachandran outliers but they are recognized by CaBLAM[28] (Extended Data Fig. 2).

Sequence misthreadings misplace residues over very large distances. The misalignment can be recognized by local Fit-to-Map criteria, with ends flagged by CaBLAM, bad geometry, *cis*-nonPro peptides and clashes (Extended Data Fig. 3).

ADH contains tightly bound ligands: an NADH cofactor as well as two zinc ions per subunit, with one zinc in the active site and the other in a spatially separate site coordinated by four cysteine residues[9]. Models lacking these ligands had considerable local modeling errors, sometimes even mistracing the backbone (Extended Data Fig. 4).

Although there was evidence for ordered water in higher-resolution APOF maps[8], only two groups elected to model water. Submissions were also split roughly 50/50 for (1) inclusion of predicted H-atom positions and (2) refinement of isotropic B factors. Although near-atomic cryo-EM maps do not have a sufficient level of detail to directly identify H-atom positions, inclusion of predicted positions can still be useful for identifying steric properties such as H-bonds or clashes[20]. Where provided, refined B factors modestly improved Fit-to-Map scores (Extended Data Fig. 5).

Evaluating metrics: Fit-to-Map. Score distributions of Fit-to-Map metrics (Table 2) were systematically compared (Fig. 4a–d). For APOF, single subunits were evaluated against masked subunit maps, whereas for ADH, dimeric models were evaluated against the full sharpened cryo-EM map (Fig. 2d). To control for the varied impact of H-atom inclusion or isotropic *B*-factor refinement on different metrics, all evaluated scores were produced with H atoms removed and all B factors were set to zero.

Score distributions were first evaluated for all 63 models across all four challenge targets. A wide diversity in performance was observed, with poor correlations between most metrics (Fig. 4a). This means that a model that scored well relative to all 62 others using one metric may have a much poorer ranking using another. Hierarchical analysis identified three distinct clusters of similarly performing metrics (Fig. 4a, labels c1–c3).

The unexpected sparse correlations and clustering can be understood by considering per-target score distribution ranges, which

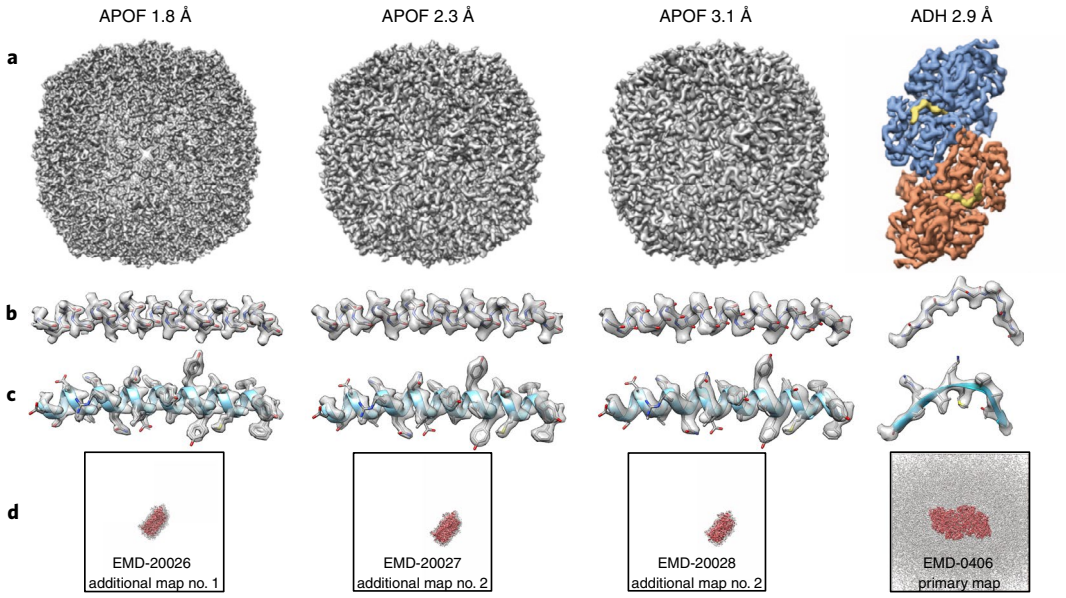


Fig. 2 | Challenge targets: cryo-EM maps at near-atomic resolution. Shown from left to right are α-helix-rich APOF at 1.8, 2.3 and 3.1 Å (EMDB entries EMD-20026, EMD-20027 and EMD-20028) and ADH at 2.9 Å (EMDB entry EMD-0406). **a**, Full maps for each target. **b**,**c**, Representative secondary structural elements (APOF, residues 14–42; ADH, residues 34–45) with masked density for protein backbone atoms only (**b**), and for all protein atoms (**c**). Visible map features transition from near-atomic to secondary -structure dominated over the 1.8–3.1 Å resolution range. **d**, EMDB maps used in model Fit-to-Map analysis (APOF targets, masked single subunits; ADH, unmasked sharpened map). The molecular boundary is shown in red at the EMDB recommended contour level, background noise is represented in gray at one-third of the EMDB recommended contour level and the full map extent is indicated by the black outline.

differ substantially from each other. The three clusters identify sets of metrics that share similar trends (Fig. 4c).

Cluster 1 metrics (Fig. 4c, top row) share the trend of decreasing score values with increasing map resolution. The cluster consists of six real-space correlation measures, three from TEMPy[16–18] and three from Phenix[19]. Each evaluates a model's fit in a similar way: by correlating calculated model-map density with experimental map density. In most cases (five out of six), correlation is performed after model-based masking of the experimental map. This observed trend is contrary to the expectation that a Fit-to-Map score should increase as resolution improves. The trend arises at least in part because map resolution is an explicit input parameter for this class of metrics. For a fixed map/model pair, changing the input resolution value will change the score. As map resolution increases, the level of detail that a model-map must faithfully replicate to achieve a high correlation score must also increase.

Cluster 2 metrics (Fig. 4c, middle row) share the inverse trend: score values improve with increasing map target resolution. Cluster 2 metrics consist of Phenix Map-Model FSC = 0.5 (ref. [19]), *Q*-score[8] and EMRinger[15]. The observed trend is expected: by definition, each metric assesses a model's fit to the experimental map in a manner that is intrinsically sensitive to map resolution. In contrast with cluster 1, cluster 2 metrics do not require map resolution to be supplied as an input parameter.

Cluster 3 metrics (Fig. 4c, bottom row) share a different overall trend: score values are substantially lower for ADH relative to APOF map targets. These measures include three unmasked correlation functions from TEMPy[16–18], Refmac FSCavg[13], Electron Microscopy Data Bank (EMDB) Atom Inclusion[14] and TEMPy ENV[16]. All of these measures consider the full experimental map without masking, so can be sensitive to background noise, which is substantial in the unmasked ADH map and minimal in the masked APOF maps (Fig. 2d).

Score distributions were also evaluated for how similarly they performed per target, and in this case most metrics were strongly correlated with each other (Fig. 4b). This means that for any single target, a model that scored well relative to all others using one metric also fared well using nearly every other metric. This situation is illustrated by comparing scores for two different metrics, CCbox from cluster 1 and *Q*-score from cluster 2 (Fig. 4d). The plot's four diagonal lines demonstrate that the scores are tightly correlated with each other within each map target. But, as described above, the two metrics have different sensitivities to map-specific factors. It is these different sensitivities that give rise to the separated, parallel spacings of the four diagonal lines, indicating score ranges on different relative scales.

One Fit-to-Map metric showed poor per-target correlation with all others: TEMPy ENV (Fig. 4b). ENV evaluates atom positions relative to a density threshold that is based on sample molecular weight. At near-atomic resolution this threshold is overly generous. TEMPy Mutual Information and EMRinger also diverged from others (Fig. 4b). Mutual information scores reflected strong influence of ADH background noise. In contrast, masked MI_OV correlated well with other measures. EMRinger yielded distinct distributions owing to its focus on backbone placement[15].

Collectively these results reveal that multiple factors such as using experimental map resolution as an input parameter, presence of background noise and density threshold selection can strongly affect Fit-to-Map score values, depending on the chosen metric. These are not desirable features for archive-wide validation of deposited cryo-EM structures.

Evaluating metrics: Coordinates-only and versus Reference. Metrics to assess model quality based on Coordinates-only (Table 2), as well as Comparison-to-Reference and Comparison-among-Models (Table 2) were also evaluated and compared (Fig. 4e,f).

Most Coordinates-only metrics were poorly correlated with each other (Fig. 4e), with the exception of bond, bond angle and chirality root mean squared deviation (r.m.s.d.), which form a

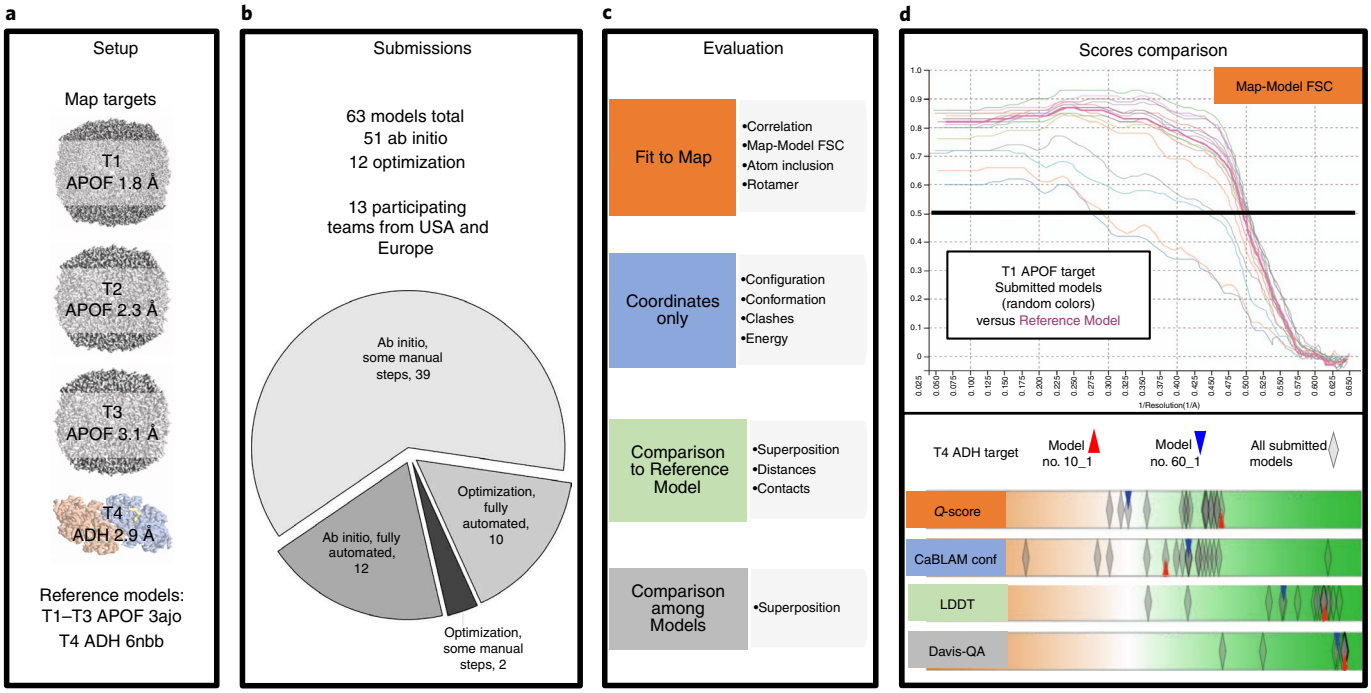


Fig. 3 | Challenge pipeline. a–c, Overview of the challenge setup (**a**), submissions (**b**) and evaluation (**c**) strategy. **d**, Scores comparison. Multiple interactive tabular and graphical displays enable comparative evaluations (model-compare.emdataresource.org). Top, Map-Model FSC curves, APOF 1.8 Å models (random light colors) versus reference model (bold cherry red). Map-Model FSC measures overall agreement of the experimental density map with a density map derived from the coordinate model (model map)[12]. Curves are calculated from Fourier coefficients of the two maps and plotted versus frequency (resolution^{-1}). The resolution value corresponding to FSC = 0.5 (black horizontal line) is typically reported. Smaller values indicate better fit. Bottom, scores comparison tool, ADH models. Interactive score distribution sliders reveal at a glance how well submitted models performed relative to each other. Parallel lanes display score distributions for each evaluated metric in a manner conceptually similar to the graphical display for key metrics used in wwPDB validation reports[4,32]. Score distributions are shown for four representative metrics, one from each evaluation track. Model scores are plotted horizontally (semi-transparent diamonds) with color coding to indicate worse (left, orange) and better (right, green) values. Darker, opaque diamonds indicate multiple overlapping scores. Scores for two individual models are also highlighted: the interactive display enables individual models to be identified and compared (red and blue triangles).

small cluster. Ramachandran outliers, widely used to validate protein backbone conformation, were poorly correlated with all other Coordinates-only measures. More than half (33) of submitted models had zero Ramachandran outliers, while only four had zero CaBLAM conformation outliers. Ramachandran statistics are increasingly used as restraints[29,30], which reduces their use as a validation metric. These results support the concept of CaBLAM as an informative score for validating backbone conformation[22].

CaBLAM metrics, while orthogonal to other Coordinates-only measures, were unexpectedly found to perform very similarly to Comparison-to-Reference metrics. The similarity likely arises because the worst modeling errors in this challenge were sequence and backbone conformation mis-assignments. These errors were equally flagged by CaBLAM, which compares models against statistics from high-quality PDB structures, and the Comparison-to-Reference metrics, which compare models against a high-quality reference. To a lesser extent, modeling errors were also flagged by Fit-to-Map metrics (Fig. 4f). Overall, Coordinates-only metrics were poorly correlated with Fit-to-Map metrics (Fig. 4f and Extended Data Fig. 6a).

Protein sidechain accuracy is specifically assessed by Rotamer and GDC-SC, while EMRinger, Q-score, CAD, hydrogen bonds in residue pairs (HBPR > 6), GDC and LDDT metrics include sidechain atoms. For these eight measures, Rotamer was completely orthogonal, Q-score was modestly correlated with the Comparison-to-Reference metrics, and EMRinger, which measures sidechain fit as a function of main chain conformation, was largely independent (Fig. 4f). These results suggest a need for multiple metrics (for example, Q-score, EMRinger, Rotamer) to assess different aspects of sidechain quality.

Evaluating metrics: local scoring. Several residue-level scores were calculated in addition to overall scores. Five Fit-to-Map metrics considered masked density for both map and model around the evaluated residue (CCbox[19], SMOC[18]), density profiles at nonhydrogen atom positions (Q-score[8]), density profiles of nonbranched residue Cγ-atom ring paths (EMRinger[15]) or density values at non-H-atom positions relative to a chosen threshold (Atom Inclusion[14]). In two of these five, residue-level scores were obtained as sliding-window averages over multiple contiguous residues (SMOC, nine residues; EMRinger, 21 residues).

Residue-level correlation analyses similar to those described above (not shown) indicate that local Fit-to-Map scores diverged more than their corresponding global scores. Residue-level scoring was most similar across evaluated metrics for high resolution maps. This observation suggests that the choice of method for scoring residue-level fit becomes less critical at higher resolution, where maps tend to have stronger density/contrast around atom positions.

A case study of a local modeling error (Extended Data Fig. 3) showed that Atom Inclusion[14], CCbox[19] and Q-score[8] produced substantially worse scores within a four-residue α-helical misthread relative to correctly assigned flanking residues. In contrast, the sliding-window-based metrics were largely insensitive (a new

Table 1 | Participating modeling teams

Team ID[a], name	Team members	No. of submitted models	Effort type(s)	Software
10 Yu	X. Yu	4	ab initio+manual	Phenix[21], Buccaneer[37], Chimera[38], Coot[29], Pymol
25 Cdmd	M. Igaev, A. Vaiana, H. Grubmüller	4	optimization automated	CDMD[39]
27 Kumar	D. Kumar	1	ab initio+manual	Phenix, Rosetta[40], Buccaneer, ARP/wARP[41], Coot
28 Ccpem	S. W. Hoh, K. Cowtan, A. P. Joseph, C. Palmer, M. Winn, T. Burnley, M. Olek, P. Bond, E. Dodson	4	ab initio+manual	CCPEM[42], Refmac[13], Buccaneer, Coot, TEMPy[16–18]
35 Phenix	P. Afonine, T. Terwilliger, L.-W. Hung	4	ab initio+manual	Phenix, Coot
38 Fzjuelich	G. Schroeder, L. Schaefer	3	optimization automated	Phenix, Chimera, DireX[43], MDFF[44], CNS, Gromacs
41 Arpwarp	G. Chojnowski	8	ab initio automated, ab initio+manual	Refmac, ARP/wARP, Coot
54 Kihara	D. Kihara, G. Terashi	8	ab initio+manual	Rosetta, Mainmast[45], MDFF, Chimera
60 Deeptracer	L. Wang, D. Si, R. Cao, J. Cheng, S. A. Moritz, J. Pfab, T. Wu, J. Hou	10	ab initio automated, ab initio+manual	Cascaded-CNN[46], Chimera
73 Singharoy	M. Shekhar, G. Terashi, S. Mittal, D. Sarkar, D. Kihara, K. Dill, A. Perez, A. Singharoy	5	ab initio+manual, optimization automated	reMDFF[47], MELD[48], VMD, Chimera, Mainmast
82 Rosetta	F. DiMaio, D. Farrell	8	ab initio automated, ab initio+manual	Rosetta, Chimera
90 Mbaker	M. Baker	2	ab initio+manual	Pathwalker[49], Phenix, Chimera, Coot
91 Chiu	G. Pintilie, W. Chiu	2	optimization+manual	Phenix, Chimera, Coot

[a]Each team was assigned a random two-digit ID for blinded identification.

TEMPy version offers single residue (SMOCd) and adjustable window analysis (SMOCf)[31]). At near-atomic resolution, single residue Fit-to-Map evaluation methods are likely to be more useful.

Residue-level Coordinates-only, Comparison-to-Reference and Comparison-among-Models metrics (not shown) were also evaluated for the same modeling error. The MolProbity server[20,22] flagged the problematic four-residue misthread via CaBLAM, *cis*-Peptide, Clashscore, bond and angle scores, but all Ramachandran scores were either favored or allowed. The Comparison-to-Reference LDDT and LGA local scores and the Davis-QA model consensus score also strongly flagged this error. The example demonstrates the value of combining multiple orthogonal measures to identify geometry issues, and further highlights the value of CaBLAM as an orthogonal measure for backbone conformation.

Group performance. Group performance was examined by modeling category and target by combining *Z*-scores from metrics determined to be meaningful in the analyses described above (Methods and Extended Data Fig. 6). A wide variety of map density features and algorithms were used to produce a model, and most were successful yet allowing a few mistakes, often in different places (Extended Data Figs. 1–4). For practitioners, it might be beneficial to combine models from several ab initio methods for subsequent refinement.

Discussion

This third EMDR Model Challenge has demonstrated that cryo-EM maps with a resolution ≤3 Å and from samples with limited conformational flexibility have excellent information content, and automated methods are able to generate fairly complete models from such maps, needing only small amounts of manual intervention.

Inclusion of maps in a resolution series enabled controlled evaluation of metrics by resolution, with a completely different map providing a useful additional control. These target selections enabled observation of important trends that otherwise could have been missed. In a recent evaluation of predicted models in the CASP13 competition against several roughly 3 Å cryo-EM maps, TEMPy and Phenix Fit-to-Map correlation measures performed very similarly[31]. In this challenge, because the chosen targets covered a wider resolution range and had more variability in background noise, the same measures were found to have distinctive, map feature-sensitive performance profiles.

Most submitted models were overall either equivalent to or better than their reference model. This achievement reflects significant advances in the development of modeling tools relative to the state presented a decade ago in our first model challenge[2]. However, several factors beyond atom positions that become important for accurate modeling at near-atomic resolution were not uniformly addressed; only half included refinement of atomic displacement factors (B factors) and a minority attempted to fit water or bound ligands.

Fit-to-Map measures were found to be sensitive to different physical properties of the map, including experimental map resolution and background noise level, as well as input parameters such as density threshold. Coordinates-only measures were found to be largely orthogonal to each other and also largely orthogonal to Fit-to-Map measures, while Comparison-to-Reference measures were generally well correlated with each other.

The cryo-EM modeling community as represented by the challenge participants have introduced a number of metrics to evaluate models with sound biophysical basis. Based on our careful analyses of these metrics and their relationships, we make four recommen-

Table 2 | Evaluated metrics

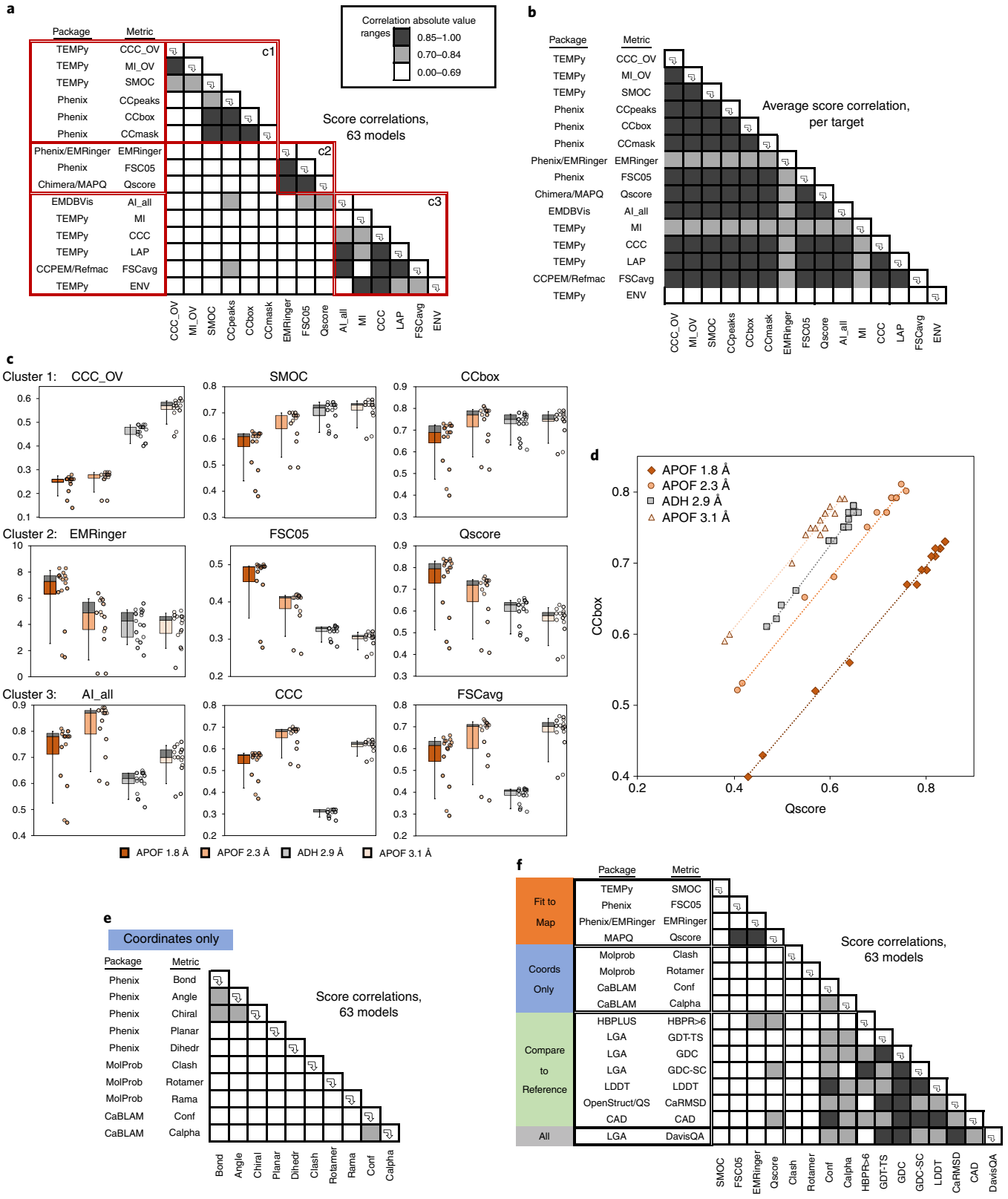
Metric class	Package metric definition
Fit-to-Map	
Correlation Coefficient, all voxels	Phenix **CCbox** full grid map versus model-map density correlation coefficient[19] TEMPy **CCC** full grid map versus model-map density correlation coefficient[17]
Correlation Coefficient, selected voxels	Phenix **CCmask** map versus model-map density, only modeled regions[19] Phenix **CCpeaks** map versus model-map density, only high-density map and model regions[19] TEMPy **CCC_OV** map versus model-map density, overlapping map and model regions[18] TEMPy **SMOC** Segment Manders' Overlap, map versus model-map density, only modeled regions[18]
Correlation Coefficient, other density function	TEMPy **LAP** map versus model-map Laplacian filtered density (partial second derivative)[16] TEMPy **Mutual Information** (MI) map versus model-map Mutual Information entropy-based function[16] TEMPy **MI_OV** map versus model-map Mutual Information, only modeled regions[18]
Correlation Coefficient, atom positions	Chimera/MAPQ **Q-score** map density at each modeled atom versus reference Gaussian density function[8]
FSC	Phenix **FSC05** Resolution (distance) of Map-Model FSC curve read at point FSC = 0.5 (ref. [19]) CCPEM/Refmac **FSCavg** FSC curve area integrated to map resolution limit[13,42]
Atom Inclusion	EMDB/VisualAnalysis **AI all** Atom Inclusion, percentage of atoms inside depositor-provided density threshold[14] TEMPy **ENV** Atom Inclusion in envelope corresponding to sample molecular weight; penalizes unmodeled regions[16]
Sidechain Density	Phenix **EMRinger** evaluates backbone by sampling map density around Cγ-atom ring paths for nonbranched residues[15]
Coordinates-only	
Configuration	Phenix **Bond** r.m.s.d. of bond lengths[21] Phenix **Angle** r.m.s.d. of bond angles[21] Phenix **Chiral** r.m.s.d. of chiral centers[21] Phenix **Planar** r.m.s.d. of planar group planarity[21] Phenix **Dihedral** r.m.s.d. of dihedral angles[21]
Clashes	MolProbity **Clashscore** Number of steric overlaps ≥0.4 Å per 1,000 atoms[20]
Conformation	MolProbity **Rotamer** sidechain conformation outliers[20] MolProbity **Rama** Ramachandran ϕ,ψ main chain conformation outliers[20] MolProbity **CaBLAM** outliers CO and Cα-based virtual dihedrals[22] MolProbity **Calpha** outliers Cα-based virtual dihedrals and Cα virtual bond angle[22]
Versus Reference Model	
Atom Superposition	Local Global Alignment (LGA) **GDT-TS** Global Distance Test Total Score, average percentage of model Cα that superimpose with reference Cα, multiple distance cutoffs[23] LGA **GDC** Global Distance Calculation, average percentage of all model atoms that superimpose with reference, multiple distance cutoffs[23] LGA **GDC-SC** Global Distance Calculation for sidechain atoms only[23] OpenStructure/QS **CaRMSD** r.m.s.d. of Cα atoms[25]
Interatomic Distances	LDDT **LDDT** Local Difference Distance Test, superposition-free comparison of all-atom distance maps between model and reference[24]
Contact Area	CAD **CAD** Contact Area Difference, superposition-free measure of differences in interatom contacts[26] HBPLUS[50] **HBPR > 6**, hydrogen bond precision, nonlocal. fraction of correctly placed hydrogen bonds in residue pairs with >6 separation in sequence
Comparison among models	
Atom Superposition, Multiple	**DAVIS-QA** average of pairwise LGA GDT-TS scores among submitted models[27]

Fig. 4 | Evaluation of metrics. Model metrics (Table 2) were compared with each other to assess how similarly they performed in scoring the challenge models. **a**–**d**, Fit-to-Map metrics analyses. **a**, Pairwise correlations of scores for all models across all map targets ($n=63$). **b**, Average correlation of scores per target (average over four correlation coefficients, one for each map target with T1, $n=16$; T2, $n=15$; T3, $n=15$; T4, $n=17$). Correlation-based metrics are identified by bold labels. In **a**, table order is based on a hierarchical cluster analysis (Methods). Three red-outlined boxes along the table diagonal correspond to identified clusters (no. c1–c3). For ease of comparison, order in **b** is identical to **a**. **c**, Representative score distributions are plotted by map target, ordered by map target resolution (see legend at bottom; T1, $n=16$; T2, $n=15$; T4, $n=17$; T3, $n=15$). Each row represents one of the three clusters defined in (a). Each score distribution is represented in box-and-whisker format (left) along with points for each individual score (right). Lower boxes represent Q1–Q2 (25th–50th percentile, in target color as shown in legend); upper boxes represent Q2–Q3 (25th–75th percentile, dark gray). Boxes do not appear when quartile limits are identical. Whiskers span 10th to 90th percentile. To improve visualization of closely clustered scores, individual scores (y values) are plotted against slightly dithered x values. **d**, Scores for one representative pair of metrics are plotted against each other (CCbox from cluster 1 and Q-score from Cluster 2). Diagonal lines represent linear fits by map target. **e**, Coordinates-only metrics comparison. **f**, Fit-to-Map, Coordinates-only and Comparison-to-Reference metrics comparison. Correlation levels in **a**,**b**,**e**,**f** are indicated by shading (see legend at top). See the Methods for additional details.

dations regarding validation practices for cryo-EM models of proteins determined at near-atomic resolution as studied here between 3.1 and 1.8 Å, a rising trend for cryo-EM (Fig. 1a).

Recommendation 1. For researchers optimizing a model against a single map, nearly any of the evaluated global Fit-to-Map metrics (Table 2) can be used to evaluate progress because they are all largely equivalent in performance. The exception is TEMPy, ENV is more appropriate at lower resolutions (>4 Å).

Recommendation 2. To flag issues with local (per residue) Fit-to-Map, metrics that evaluate single residues are more suitable than those using sliding-window averages over multiple residues (Evaluating metrics: local scoring).



Recommendation 3. The ideal Fit-to-Map metric for archive-wide ranking will be insensitive to map background noise (appropriate masking or alternative data processing can help), will not require input of estimated parameters that affect score value (for example, resolution limit, threshold) and will yield overall better scores for maps with trustworthy higher-resolution features. The three cluster 2 metrics identified in this challenge (Fig. 4a 'c2' and Fig. 4c center row) meet these criteria.

- Map-Model FSC[12,19] is already in common use, and can be compared with the experimental map's independent half-map FSC curve.
- Global EMRinger score[15] can assess nonbranched protein sidechains.
- *Q*-score can be used both globally and locally for validating nonhydrogen atom *x*,*y*,*z* positions[8].

Other Fit-to-Map metrics may be rendered suitable for archive-wide comparisons through conversion of raw scores to *Z*-scores over narrow resolution bins, as is currently done by the PDB for some X-ray-based metrics[4,32].

Recommendation 4. CaBLAM and MolProbity *cis*-peptide detection[22] are useful to detect protein backbone conformation issues. These are particularly valuable tools for cryo-EM, since maps at typical resolutions (2.5–4.0 Å, Fig. 1a) may not resolve backbone carbonyl oxygens (Fig. 2).

In this challenge, more time could be devoted to analysis when compared with previous rounds because infrastructure for model collection, processing and assessment is now established. However, several important issues could not be addressed, including evaluation of overfitting using half-map based methods[13,33–35], effect of map sharpening on Fit-to-Map scores[8,36], validation of ligand fit and metal ion/water identification and validation at atomic resolution including H atoms. EMDR plans to sponsor additional model challenges to continue promoting development and testing of cryo-EM modeling and validation methods.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-01051-w.

Received: 11 June 2020; Accepted: 21 December 2020;
Published online: 4 February 2021

References

1. Mitra, A. K. Visualization of biological macromolecules at near-atomic resolution: cryo-electron microscopy comes of age. *Acta Cryst. F* **75**, 3–11 (2019).
2. Lawson, C. L., Berman, H. M. & Chiu, W. Evolving data standards for cryo-EM structures. *Struct. Dyn.* **7**, 014701 (2020).
3. Henderson, R. et al. Outcome of the first electron microscopy validation task force meeting. *Structure* **20**, 205–214 (2012).
4. Read, R. J. et al. A new generation of crystallographic validation tools for the Protein Data Bank. *Structure* **19**, 1395–1412 (2011).
5. Montelione, G. T. et al. Recommendations of the wwPDB NMR Validation Task Force. *Structure* **21**, 1563–1570 (2013).
6. Lawson, C. L. & Chiu, W. Comparing cryo-EM structures. *J. Struct. Biol.* **204**, 523–526 (2018).
7. wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
8. Pintilie, G. et al. Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nat. Methods* **17**, 328–334 (2020).
9. Herzik, M. A. Jr, Wu, M. & Lander, G. C. High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nat. Commun.* **10**, 1032 (2019).
10. Masuda, T., Goto, F., Yoshihara, T. & Mikami, B. The universal mechanism for iron translocation to the ferroxidase site in ferritin, which is mediated by the well conserved transit site. *Biochem. Biophys. Res. Commun.* **400**, 94–99 (2010).
11. Kryshtafovych, A., Adams, P. D., Lawson, C. L. & Chiu, W. Evaluation system and web infrastructure for the second cryo-EM Model Challenge. *J. Struct. Biol.* **204**, 96–108 (2018).
12. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
13. Brown, A. et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Cryst. D* **71**, 136–153 (2015).
14. Lagerstedt, I. et al. Web-based visualisation and analysis of 3D electron-microscopy data from EMDB and PDB. *J. Struct. Biol.* **184**, 173–181 (2013).
15. Barad, B. A. et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).
16. Vasishtan, D. & Topf, M. Scoring functions for cryoEM density fitting. *J. Struct. Biol.* **174**, 333–343 (2011).
17. Farabella, I. et al. TEMPy: a Python library for assessment of three-dimensional electron microscopy density fits. *J. Appl. Crystallogr.* **48**, 1314–1323 (2015).
18. Joseph, A. P., Lagerstedt, I., Patwardhan, A., Topf, M. & Winn, M. Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. *J. Struct. Biol.* **199**, 12–26 (2017).
19. Afonine, P. V. et al. New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Cryst. D* **74**, 814–840 (2018).
20. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Cryst. D* **66**, 12–21 (2010).
21. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst. D* **75**, 861–877 (2019).
22. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
23. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).
24. 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**, 2722–2728 (2013).
25. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. & Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.* **7**, 10480 (2017).
26. Olechnovic, K., Kulberkyte, E. & Venclovas, C. CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins* **81**, 149–162 (2013).
27. Kryshtafovych, A., Monastyrskyy, B. & Fidelis, K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins* **82**, 7–13 (2014).
28. Prisant, M. G., Williams, C. J., Chen, V. B., Richardson, J. S. & Richardson, D. C. New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink 'waters,' and NGL Viewer to recapture online 3D graphics. *Protein Sci.* **29**, 315–329 (2020).
29. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Cryst. D* **66**, 486–501 (2010).
30. Headd, J. J. et al. Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. *Acta Cryst. D* **68**, 381–390 (2012).
31. Kryshtafovych, A. et al. Cryo-electron microscopy targets in CASP13: overview and evaluation of results. *Proteins* **87**, 1128–1140 (2019).
32. Gore, S. et al. Validation of structures in the Protein Data Bank. *Structure* **25**, 1916–1927 (2017).
33. DiMaio, F., Zhang, J., Chiu, W. & Baker, D. Cryo-EM model validation using independent map reconstructions. *Protein Sci.* **22**, 865–868 (2013).
34. Pintilie, G., Chen, D. H., Haase-Pettingell, C. A., King, J. A. & Chiu, W. Resolution and probabilistic models of components in cryoEM maps of mature P22 bacteriophage. *Biophys. J.* **110**, 827–839 (2016).
35. Hryc, C. F. et al. Accurate model annotation of a near-atomic resolution cryo-EM map. *Proc. Natl Acad. Sci. USA* **114**, 3103–3108 (2017).
36. Terwilliger, T. C., Sobolev, O. V., Afonine, P. V. & Adams, P. D. Automated map sharpening by maximization of detail and connectivity. *Acta Cryst. D* **74**, 545–559 (2018).
37. Hoh, S., Burnley, T. & Cowtan, K. Current approaches for automated model building into cryo-EM maps using Buccaneer with CCP-EM. *Acta Cryst. D* **76**, 531–541 (2020).
38. Pettersen, E. F. et al. UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
39. Igaev, M., Kutzner, C., Bock, L. V., Vaiana, A. C. & Grubmuller, H. Automated cryo-EM structure refinement using correlation-driven molecular dynamics. *eLife* **8**, https://doi.org/10.7554/eLife.43542 (2019).
40. Frenz, B., Walls, A. C., Egelman, E. H., Veesler, D. & DiMaio, F. RosettaES: a sampling strategy enabling automated interpretation of difficult cryo-EM maps. *Nat. Methods* **14**, 797–800 (2017).

41. Chojnowski, G., Pereira, J. & Lamzin, V. S. Sequence assignment for low-resolution modelling of protein crystal structures. *Acta Cryst. D* **75**, 753–763 (2019).
42. Burnley, T., Palmer, C. M. & Winn, M. Recent developments in the CCP-EM software suite. *Acta Cryst. D* **73**, 469–477 (2017).
43. Wang, Z. & Schröder, G. F. Real-space refinement with DireX: from global fitting to side-chain improvements. *Biopolymers* **97**, 687–697 (2012).
44. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673–683 (2008).
45. Terashi, G. & Kihara, D. De novo main-chain modeling for EM maps using MAINMAST. *Nat. Commun.* **9**, 1618 (2018).
46. Si, D. et al. Deep learning to predict protein backbone structure from high-resolution cryo-EM density maps. *Sci. Rep.* **10**, 4282 (2020).
47. Singharoy, A. et al. Molecular dynamics-based refinement and validation for sub-5 A cryo-electron microscopy maps. *eLife* **5**, https://doi.org/10.7554/eLife.16105 (2016).
48. MacCallum, J. L., Perez, A. & Dill, K. A. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl Acad. Sci. USA* **112**, 6985–6990 (2015).
49. Chen, M. & Baker, M. L. Automation and assessment of de novo modeling with Pathwalking in near atomic resolution cryoEM density maps. *J. Struct. Biol.* **204**, 555–563 (2018).
50. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793 (1994).

[1]Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA. [2]Genome Center, University of California, Davis, CA, USA. [3]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [4]Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA. [5]Department of Biochemistry and Molecular Biology, The University of Texas Health Science Center at Houston, Houston, TX, USA. [6]Department of Integrated Computational Structural Biology, The Scripps Research Institute, La Jolla, CA, USA. [7]York Structural Biology Laboratory, Department of Chemistry, University of York, York, UK. [8]Scientific Computing Department, UKRI Science and Technology Facilities Council, Research Complex at Harwell, Didcot, UK. [9]Department of Computer Science, Pacific Lutheran University, Tacoma, WA, USA. [10]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA. [11]European Molecular Biology Laboratory, c/o DESY, Hamburg, Germany. [12]Laufer Center, Stony Brook University, Stony Brook, NY, USA. [13]Department of Biochemistry and Institute for Protein Design, University of Washington, Seattle, WA, USA. [14]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA. [15]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA. [16]Department of Computer Science, Saint Louis University, St. Louis, MO, USA. [17]Los Alamos National Laboratory, Los Alamos, NM, USA. [18]Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany. [19]Department of Biological Sciences, Purdue University, West Lafayette, IN, USA. [20]Department of Computer Science, Purdue University, West Lafayette, IN, USA. [21]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX, USA. [22]Biodesign Institute, Arizona State University, Tempe, AZ, USA. [23]School of Advanced Sciences and Languages, VIT Bhopal University, Bhopal, India. [24]The European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK. [25]Department of Chemistry, University of Florida, Gainesville, FL, USA. [26]Division of Computing & Software Systems, University of Washington, Bothell, WA, USA. [27]Department of Bioengineering, Stanford University, Stanford, CA, USA. [28]Department of Biochemistry, Duke University, Durham, NC, USA. [29]Structural Biology of Cells and Viruses Laboratory, Francis Crick Institute, London, UK. [30]Institute of Biological Information Processing (IBI-7: Structural Biochemistry) and Jülich Centre for Structural Biology (JuStruct), Forschungszentrum Jülich, Jülich, Germany. [31]Division of CryoEM and Biomaging, SSRL, SLAC National Accelerator Laboratory, Stanford University, Menlo Park, CA, USA. [32]Physics Department, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [33]Center for Development of Therapeutics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [34]New Mexico Consortium, Los Alamos, NM, USA. [35]Department of Biological Structure, University of Washington, Seattle, WA, USA. [36]Biophysics Graduate Program, University of California, San Francisco, CA, USA. [37]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA. [38]SMPS, Janssen Research and Development, Spring House, PA, USA. [39]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ, USA. [40]Department of Biological Sciences and Bridge Institute, University of Southern California, Los Angeles, CA, USA.
✉e-mail: cathy.lawson@rutgers.edu; wahc@stanford.edu

Methods

Challenge process and organization. Informed by previous challenges[2,6,11], the 2019 Model Challenge process was substantially streamlined in this round. In March, a panel of advisors with expertise in cryo-EM methods, modeling and/or model assessment was recruited. The panel worked with EMDR team members to develop the challenge guidelines, identify suitable map targets from EMDB and reference models from the PDB and recommend the metrics to be calculated for each submitted model.

The challenge rules and guidance were as follows: (1) ab initio modeling is encouraged but not required. For optimization studies, any publicly available coordinate set can be used as the starting model. (2) Regardless of the modeling method used, submitted models should be as complete and as accurate as possible (that is, equivalent to publication-ready). (3) For each target, a separate modeling process should be used. (4) Fitting to either the unsharpened/unmasked map or one of the half-maps is strongly encouraged. (5) Submission in mmCIF format is strongly encouraged.

Members of cryo-EM and modeling communities were invited to participate in mid-April 2019 and details were posted on the challenges website (challenges.emdataresource.org). Models were submitted by participant teams between 1 and 28 May 2019. For APOF targets, coordinate models were submitted as single subunits at the position of a provided segmented density consisting of a single subunit. ADH models were submitted as dimers. For each submitted model, metadata describing the full modeling workflow were collected via a Drupal webform, and coordinates were uploaded and converted to PDBx/mmCIF format using PDBextract[51]. Model coordinates were then processed for atom/residue ordering and nomenclature consistency using PDB annotation software (Feng Z., https://sw-tools.rcsb.org/apps/MAXIT) and additionally checked for sequence consistency and correct position relative to the designated target map. Models were then evaluated as described below (Model evaluation system).

In early June, models, workflows and initial calculated scores were made available to all participants for evaluation, blinded to modeler team identity and software used. A 2.5-day workshop was held in mid-June at Stanford/SLAC to review the results, with panel members attending in person. All modeling participants were invited to attend remotely and present overviews of their modeling processes and/or assessment strategies. Recommendations were made for additional evaluations of the submitted models as well as for future challenges. Modeler teams and software were unblinded at the end of the workshop. In September, a virtual follow-up meeting with all participants provided an overview of the final evaluation system after implementation of recommended updates.

Coordinate sources and modeling software. Modeling teams created ab initio models or optimized previously known models available from the PDB. Models optimized against APOF maps used PDB entries 2fha, 5n26 or 3ajo as starting models. Models optimized against ADH used PDB entries 1axe, 2jhf or 6nbb. Ab initio software included ARP/wARP[41], Buccaneer[37], Cascaded-CNN[46], Mainmast[45], Pathwalker[49] and Rosetta[40]. Optimization software included CDMD[39], CNS[52], DireX[43], Phenix[21], REFMAC[13], MELD[48], MDFF[44] and reMDFF[47]. Participants made use of VMD[53], Chimera[38], COOT[29] and PyMol for visual evaluation and/or manual model improvement of map-model fit. See Table 1 for software used by each modeling team. Modeling software versions/websites are listed in the Nature Research Reporting Summary.

Model evaluation system. The evaluation system for 2019 challenge (model-compare.emdataresource.org) was built on the basis of the 2016/2017 Model Challenge system[11], updated with several additional evaluation measures and analysis tools. Submitted models were evaluated for >70 individual metrics in four tracks: Fit-to-Map, Coordinates-only, Comparison-to-Reference and Comparison-among-Models. A detailed description of the updated infrastructure and each calculated metric is provided as a help document on the model evaluation system website. Result data are archived at Zenodo[54]. Analysis software versions/websites are listed in the Nature Research Reporting Summary.

For brevity, a representative subset of metrics from the evaluation website are discussed in this paper. The selected metrics are listed in Table 2 and are further described below. All scores were calculated according to package instructions using default parameters.

Fit-to-Map. The evaluated metrics included several ways to measure the correlation between map and model density as implemented in TEMPy[16–18] v.1.1 (CCC, CCC_OV, SMOC, LAP, MI, MI_OV) and the Phenix[21] v.1.15.2 map_model_cc module[19] (CCbox, CCpeaks, CCmask). These methods compare the experimental map with a model map produced on the same voxel grid, integrated either over the full map or over selected masked regions. The model-derived map is generated to a specified resolution limit by inverting Fourier terms calculated from coordinates, B factors and atomic scattering factors. Some measures compare density-derived functions instead of density (MI, LAP[16]).

The *Q*-score (MAPQ v.1.2 (ref. [8]) plugin for UCSF Chimera[38] v.1.11) uses a real-space correlation approach to assess the resolvability of each model atom in the map. Experimental map density is compared to a Gaussian placed at each atom position, omitting regions that overlap with other atoms. The score is calibrated by the reference Gaussian, which is formulated so that a highest score of 1 would be given to a well-resolved atom in a map at an approximately 1.5 Å resolution. Lower scores (down to −1) are given to atoms as their resolvability and the resolution of the map decreases. The overall *Q*-score is the average value for all model atoms.

Measures based on Map-Model FSC curve, Atom Inclusion and protein sidechain rotamers were also compared. Phenix Map-Model FSC is calculated using a soft mask and is evaluated at FSC = 0.5 (ref. [19]). REFMAC FSCavg[13] (module of CCPEM[42]) integrates the area under the Map-Model FSC curve to a specified resolution limit[13]. EMDB Atom Inclusion determines the percentage of atoms inside the map at a specified density threshold[14]. TEMPy ENV is also threshold-based and penalizes unmodeled regions[16]. EMRinger (module of Phenix) evaluates backbone positioning by measuring the peak positions of unbranched protein C_γ atom positions versus map density in ring paths around C_α–C_β bonds[15].

Coordinates-only. Standard measures assessed local configuration (bonds, bond angles, chirality, planarity, dihedral angles; Phenix model statistics module), protein backbone (MolProbity Ramachandran outliers[20]; Phenix molprobity module) and sidechain conformations, and clashes (MolProbity rotamers outliers and Clashscore[20]; Phenix molprobity module).

New in this challenge round is CaBLAM[22] (part of MolProbity and as Phenix cablam module), which uses two procedures to evaluate protein backbone conformation. In both cases, virtual dihedral pairs are evaluated for each protein residue i using C_α positions $i-2$ to $i+2$. To define CaBLAM outliers, the third virtual dihedral is between the CO groups flanking residue i. To define Calpha-geometry outliers, the third parameter is the C_α virtual angle at i. The residue is then scored according to virtual triplet frequency in a large set of high-quality models from PDB[22].

Comparison-to-Reference and Comparison-among-Models. Assessing the similarity of the model to a reference structure and similarity among submitted models, we used metrics based on atom superposition (LGA GDT-TS, GDC and GDC-SC scores[23] v.04.2019), interatomic distances (LDDT score[24] v.1.2), and contact area differences (CAD[26] v.1646). HBPLUS[50] was used to calculate nonlocal hydrogen bond precision, defined as the fraction of correctly placed hydrogen bonds with more than six separations in sequence (HBPR > 6). DAVIS-QA determines for each model the average of pairwise GDT-TS scores among all other models[27].

Local (per residue) scores. Residue-level visualization tools for comparing the submitted models were also provided for the following metrics: Fit-to-Map, Phenix CCbox, TEMPy SMOC, *Q*-score, EMRinger and EMDB Atom Inclusion; Comparison-to-Reference, LGA and LDDT; and Comparison-among-Models, DAVIS-QA.

Metric score pairwise correlations and distributions. For pairwise comparisons of metrics, Pearson correlation coefficients (P) were calculated for all model scores and targets ($n = 63$). For average per-target pairwise comparisons of metrics, P values were determined for each target and then averaged. Metrics were clustered according to the similarity score $(1 - |P|)$ using a hierarchical algorithm with complete linkage. At the beginning, each metric was placed into a cluster of its own. Clusters were then sequentially combined into larger clusters, with the optimal number of clusters determined by manual inspection. In the Fit-to-Map evaluation track, the procedure was stopped after three divergent score clusters were formed for the all-model correlation data (Fig. 4a), and after two divergent clusters were formed for the average per-target clustering (Fig. 4b).

Controlling for model systematic differences. As initially calculated, some Fit-to-Map scores had unexpected distributions, owing to differences in modeling practices among participating teams. For models submitted with all atom occupancies set to zero, occupancies were reset to one and rescored. In addition, model submissions were split approximately 50/50 for each of the following practices: (1) inclusion of hydrogen atom positions and (2) inclusion of refined B factors. For affected fit-to-map metrics, modified scores were produced excluding hydrogen atoms and/or setting B factors to zero. Both original and modified scores are provided at the web interface. Only modified scores were used in the comparisons described here.

Evaluation of group performance. Rating of group performance was done using the group ranks and model ranks (per target) tools on the challenge evaluation website. These tools permit users, either by group or for a specified target and for all or a subcategory of models (for example, ab initio), to calculate composite *Z*-scores using any combination of evaluated metrics with any desired relative weightings. The *Z*-scores for each metric are calculated from all submitted models for that target ($n = 63$). The metrics (weights) used to generate composite *Z*-scores were as follows.

Coordinates-only. CaBLAM outliers (0.5), Calpha-geometry outliers (0.3) and Clashscore (0.2). CaBLAM outliers and Calpha-geometry outliers had the best correlation with Comparison-to-Reference parameters (Fig. 4f), and Clashscore is an orthogonal measure. Ramachandran and rotamer criteria were excluded since they are often restrained in refinement and are zero for many models.

Fit-to-Map. EMRinger (0.3), Q-score (0.3), Atom Inclusion (0.2) and SMOC (0.2). EMRinger and Q-score were among the most promising model-to-map metrics, and the other two provide distinct measures.

Comparison-to-Reference. LDDT (0.9), GDC_all (0.9) and HBPR >6 (0.2). LDDT is superposition-independent and local, while GDC_all requires superposition; H-bonding is distinct. Metrics in this category are weighted higher, because although the reference models are not perfect, they are a reasonable estimate of the right answer.

Composite *Z*-scores by metric category (Extended Data Fig. 6a) used the Group Ranks tool. For ab initio rankings (Extended Data Fig. 6b), *Z*-scores were averaged across each participant group on a given target, and further averaged across T1 + T2 and across T3 + T4 to yield overall *Z*-scores for high and low resolutions group 54 models were rated separately because they used different methods. Group 73's second model on target T4 was not rated because the metrics are not set up to meaningfully evaluate an ensemble. Other choices of metric weighting schemes were tried, with very little effect on clustering.

Molecular graphics. Molecular graphics images were generated using UCSF Chimera[38] (Fig. 2 and Extended Data Fig. 3) and KiNG[55] (Extended Data Figs. 1, 2 and 4).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The map targets used in the challenge were downloaded from the EMDB, entries EMD-20026 (file emd_20026_additional_1.map.gz), EMD-20027 (file emd_20027_additional_2.map.gz), EMD-20028 (file emd_20028_additional_2.map.gz) and EMD-0406 (file emd_0406.map.gz). Reference models were downloaded from the PDB, entries 3ajo and 6nbb. Submitted models, model metadata, result logs and compiled data are archived at Zenodo at https://doi.org/10.5281/zenodo.4148789, and at https://model-compare.emdataresource.org/data/2019/. Interactive summary tables, graphical views and .csv downloads of compiled results are available at https://model-compare.emdataresource.org/2019/cgi-bin/index.cgi. Source data are provided with this paper.

References

51. Yang, H. et al. Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. *Acta Cryst. D* **60**, 1833–1839 (2004).
52. Brünger, A. T. Version 1.2 of the crystallography and NMR system. *Nat. Protoc.* **2**, 2728–2733 (2007).
53. Hsin, J., Arkhipov, A., Yin, Y., Stone, J. E. & Schulten, K. Using VMD: an introductory tutorial. *Curr. Protoc. Bioinformatics* **24**, https://doi.org/10.1002/0471250953.bi0507s24 (2008).
54. Lawson, C. L. et al. 2019 EMDataresource model metrics challenge dataset. *Zenodo* https://doi.org/10.5281/zenodo.4148789 (2020).
55. Chen, V. B., Davis, I. W. & Richardson, D. C. KING (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program. *Protein Sci.* **18**, 2403–2409 (2009).

Acknowledgements

EMDataResource (C.L.L., A.K., G.P., H.M.B. and W.C.) is supported by the US National Institutes of Health (NIH)/National Institute of General Medical Science, grant no. R01GM079429. The Singharoy team used the supercomputing resources of the Oak Ridge Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science at the Department of Energy under contract no. DE-AC05-00OR22725. The following additional grants are acknowledged for participant support: grant no. NIH/R35GM131883 to J.S.R. and C.W.; grant no. NIH/P01GM063210 to P.D.A., P.V.A., L.-W.H., J.S.R., T.C.T. and C.W.; National Science Foundation grant no. (NSF)/MCB-1942763 (CAREER) and NIH/R01GM095583 to A.S.; grant nos. NIH/R01GM123055, NIH/R01GM133840, NSF/DMS1614777, NSF/CMMI1825941, NSF/MCB1925643, NSF/DBI2003635 and Purdue Institute of Drug Discovery to D. Kihara; grant no. NIH/R01GM123159 to J.S.F.; Max Planck Society German Research Foundation grant no. IG 109/1-1 to M.I.; Max Planck Society German Research Foundation grant no. FOR-1805 to A.C.V.; grant nos. NIH/R37AI36040 and Welch Foundation/Q1279 to D. Kumar (PI: BVV Prasad); grant no. NSF/DBI2030381 to D. Si.; Medical Research Council grant no. MR/N009614/1 to T.B., C.M.P. and M.W.; Wellcome Trust grant no. 208398/Z/17/Z to A.P.J. and M.W.; Biotechnology and Biological Sciences Research Council grant no. BB/P000517/1 to K.C. and Biotechnology and Biological Sciences Research Council grant no. BB/P000975/1 to M.W.

Author contributions

P.D.A., P.V.A., J.S.F., F.D.M., J.S.R., P.B.R., H.M.B., W.C., A.K., C.L.L., G.D.P. and M.F.S. formed the expert panel that selected targets, reference models and assessment metrics, set the challenge rules and attended the face-to-face results review workshop. K.Z. generated the APOF maps for the challenge. M.A.H. provided the published ADH map. C.L.L. designed and implemented the challenge model submission pipeline, and drafted the initial, revised and final manuscripts. Authors as listed in Table 1 built and submitted models and presented modeling strategies at the review workshop. A.K. designed and implemented the evaluation pipeline and website, and calculated scores. A.K., C.L.L., B.M., M.A.H., J.S.R., C.J.W., P.V.A. and J.S.F. analyzed models and model scores. A.P., Z.W., T.C.T., A.P.J., G.D.P., P.V.A. and C.J.W. contributed the software, and provided advice on use and scores interpretation. C.L.L., A.K., G.D.P. and J.S.R. drafted the figures. A.K., H.M.B., G.D.P., W.C., M.F.S., M.A.H. and J.S.R. contributed to manuscript writing. All authors reviewed and approved the final manuscript.

Competing interests

X.Y. is an employee of Janssen Research and Development. All other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-020-01051-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-020-01051-w.

Correspondence and requests for materials should be addressed to C.L.L. or W.C.

Peer review information Allison Doerr was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

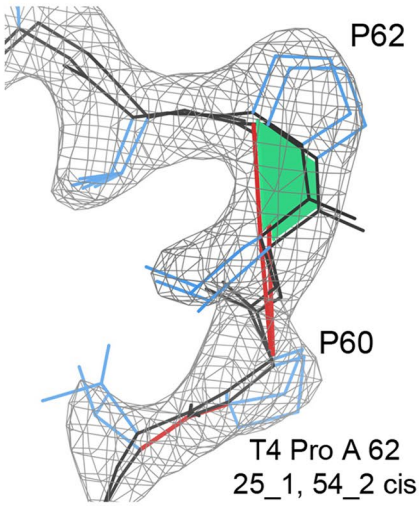
Reprints and permissions information is available at www.nature.com/reprints.

a

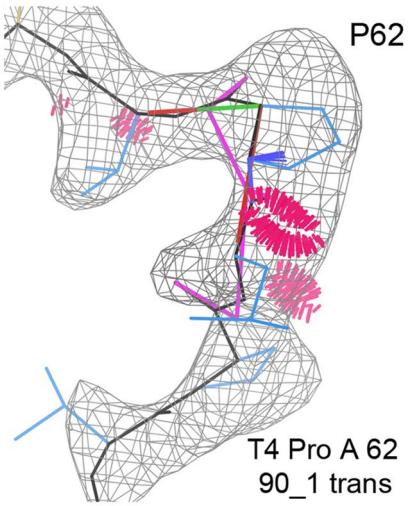
Model: cisP, twistP, cis-nonP, twist-nonP not *ab initio*

Gp 10_1	T1: 1, 0, 0, 0	T2: 1, 0, 0, 0	T3: 0, 0, 0, 0	T4: 2, 0, 0, 0
Gp 25_1	T1: 1, 0, 0, 0	T2: 1, 0, 0, 0	T3: 1, 0, 0, 0	T4: 2, 0, 1, 0 b
Gp 27_1				T4: 2, 0, 0, 0
Gp 28_1	T1: 1, 0, 0, 0	T2: 1, 0, 0, 0	T3: 1, 0, 0, 0	T4: 2, 0, 0, 0
Gp 35_1	T1: 1, 0, 0, 0	T2: 1, 0, 0, 0	T3: 1, 0, 0, 0	T4: 2, 0, 0, 0
Gp 38_1	T1: 0, 1, 0, 0	T2: 0, 1, 0, 0	T3: 0, 1, 0, 0	
Gp 41_1	T1: 0, 0, 0, 0	T2: 0, 0, 0, 0	T3: 0, 0, 0, 0	T4: 0, 0, 0, 3 d
Gp 41_2	T1: 0, 1, 0, 0	T2: 0, 1, 0, 0	T3: 0, 0, 0, 0	T4: 0, 1, 0, 1
Gp 54_1	T1: 0, 0, 5, 0	T2: 1, 0, 3, 0	T3: 0, 0, 4, 0	T4: 1, 0, 23, 0
Gp 54_2	T1: 0, 0, 0, 0	T2: 0, 0, 0, 0	T3: 0, 0, 0, 0	T4: 3, 0, 15, 3 b
Gp 60_1	T1: 0, 0, 0, 0	T2: 0, 0, 0, 0	T3: 0, 0, 0, 2	T4: 2 ,0, 0, 4
Gp 60_2	T1: 0, 0, 0, 0	T2: 0, 0, 0, 0	T3: 0, 0, 0, 2	T4: 2, 0, 0, 2
Gp 60_3	T1: 0, 0, 0, 0	T2: 0, 0, 0, 0		
Gp 73_1	T1: 1, 0, 0, 0	T2: 0, 0, 0, 2	T3: 0, 0, 0, 1	T4: 2, 0, 0, 1
Gp 73_2				T4: 20, 0, 0, 11 ensemble
Gp 82_1	T1: 0, 0, 0, 1	T2: 1, 0, 0, 0	T3: 1, 0, 1, 3	T4: 2, 0, 0 ,14
Gp 82_2	T1: 0, 0, 0, 1	T2: 1, 0, 0, 0	T3: 1, 0, 0, 4	T4: 2, 0, 0, 14
Gp 90_1			T3: 1, 0, 0, 0	T4: 0, 0, 0, 0 c
Gp 91_1	T1: 1, 0, 0, 0			T4: 2, 0, 0, 0

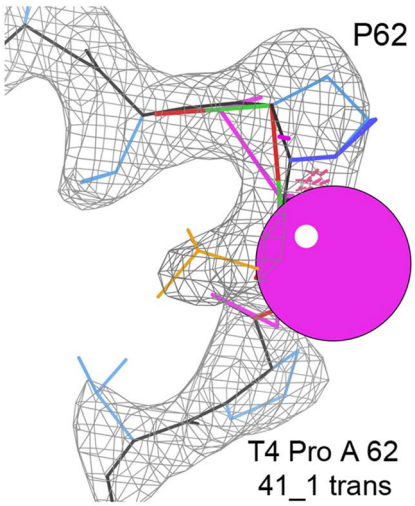
b



c

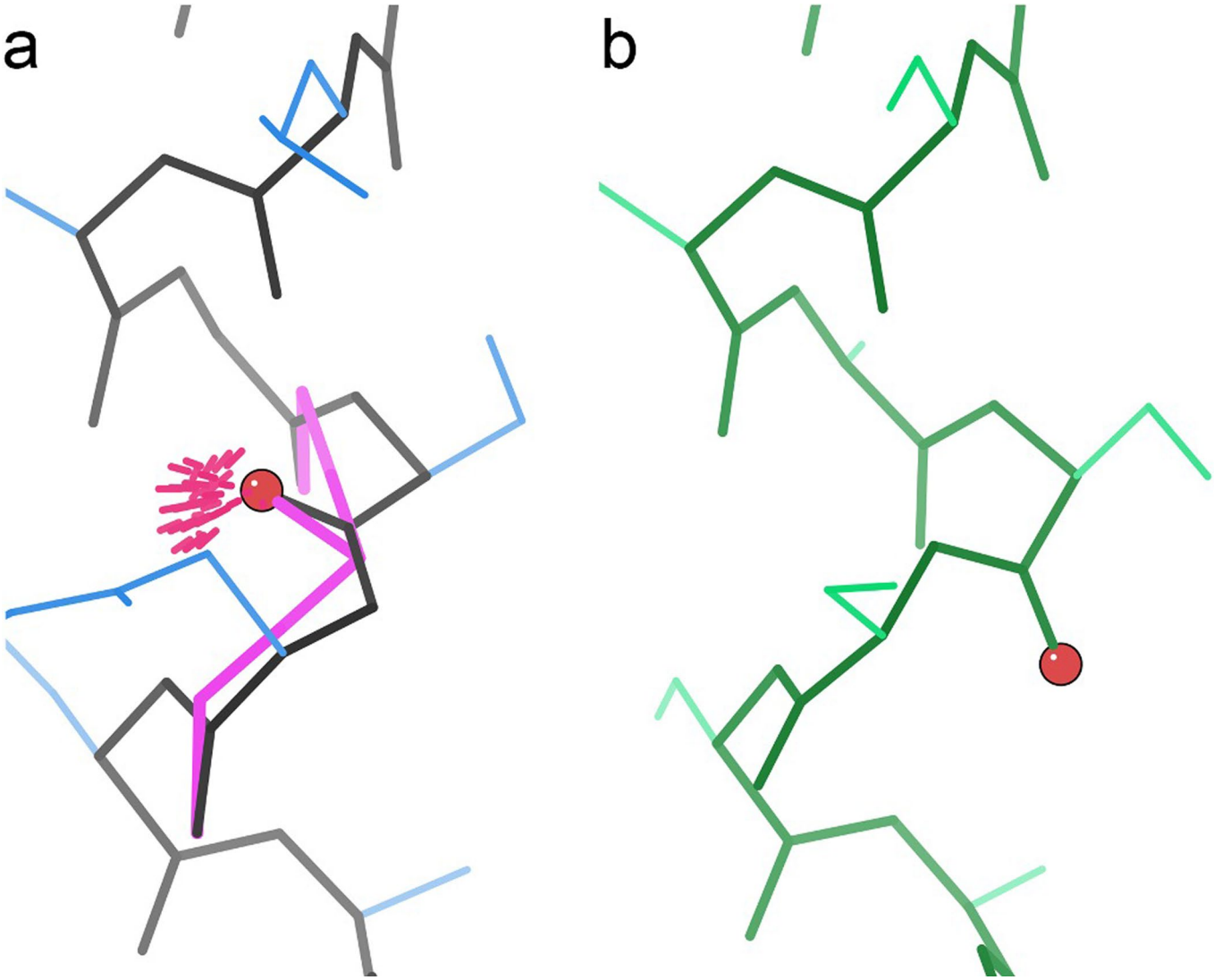


d



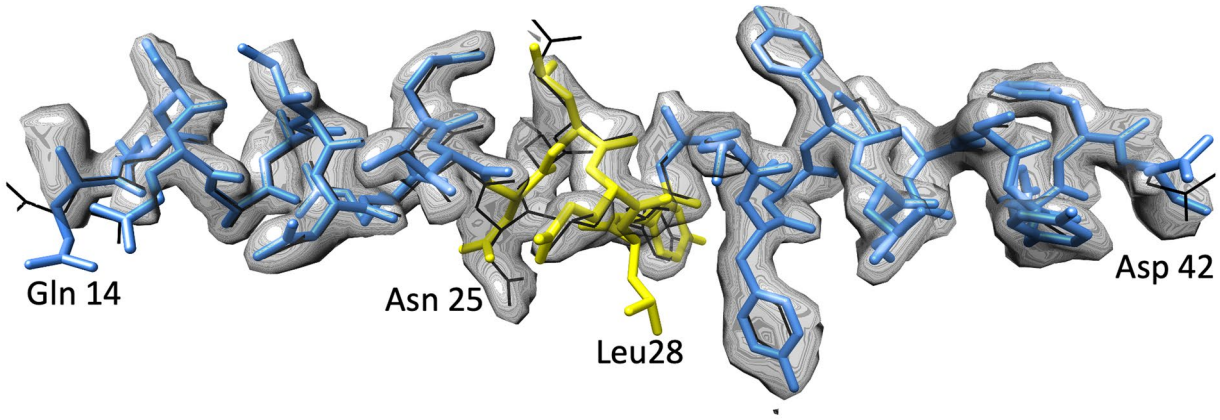
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Evaluation of peptide bond geometry. All 63 Challenge models were evaluated using MolProbity. APOF and ADH each have one *cis* peptide bond per subunit before a proline residue. (**a**) Counts of peptide bonds with each of the following conformational properties: cisP: *cis* peptide before proline, twistP: non-planar peptide (>30°) before proline, cis-nonP: *cis* peptide before non-proline, twist-nonP: non-planar peptide bond before non-proline. Incorrect cis-nonP usually occurred where the model was misfit (see Extended Data Figs. 2 and 3), while incorrect *cis* or *trans* Pro usually produced poor geometry. Values inconsistent with reference models are highlighted. Statistically, 1 in 20 proline residues are genuinely *cis*; only 1 in 3000 non-proline residues are genuinely *cis*, and strongly non-planar peptide bonds (>30°) are almost never genuine[28]. Models are identified by the submitting group (Gp #, group id as defined in Table 1), model number (some groups submitted multiple models), and Target (T1-T3: APOF, T4: ADH). Optimized models are shaded blue. Only two groups (28, 31) had all peptides correct for all 4 targets. Models illustrated in panels **b**-**d** are indicated by labeled boxes. (**b**) Correct *cis* peptide geometry for Pro A62 in two ADH models. (**c**) Incorrect trans peptide geometry, with huge clashes up to 1.25 Å overlap (clusters of hot pink spikes), 2 CaBLAM outliers (magenta CO dihedral lines), and poor density fit. (**d**) Incorrect trans peptide geometry, with huge 1.9 Å C_β deviation at Leu 61 (magenta ball) because of incorrect hand of Cα, and 2 CaBLAM outliers. Molecular graphics were generated using KiNG.

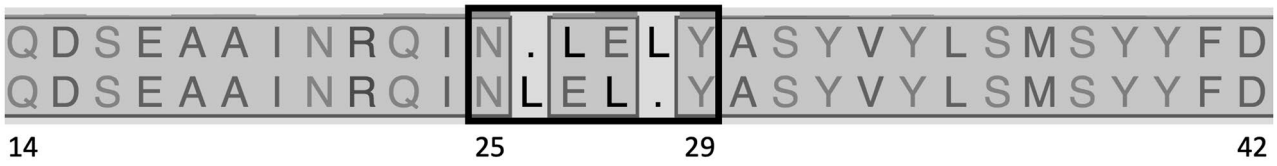


Extended Data Fig. 2 | Classic CaBLAM outlier with no Ramachandran outlier. a, Mis-modeled peptide (identified by red ball at carbonyl oxgen position) is flagged by two successive CaBLAM outliers (magenta dihedrals), a bad clash (hot-pink spikes), and a bond-angle outlier (not shown), but no Ramachandran outlier. **b**, Correctly modeled peptide, involving a near-180° flip of the central peptide to achieve regular α-helical conformation. Ser 38 of T1/APOF model 60_1 is shown in (a); model 35_1 shown in (b). This example illustrates the most easily correctable situations: (1) for a CaBLAM outlier inside helix or β-sheet, regularize the secondary structure; (2) for two successive CaBLAM outliers, try flipping the central peptide. Molecular graphics were generated using KiNG. Note that sidechains are truncated by graphics clipping.

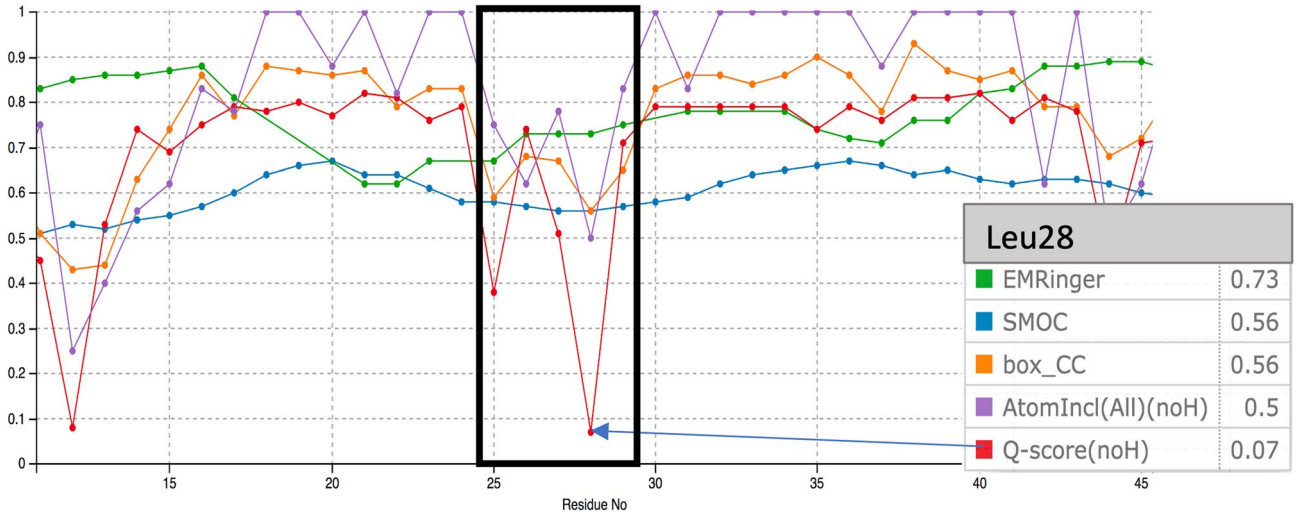
a



b



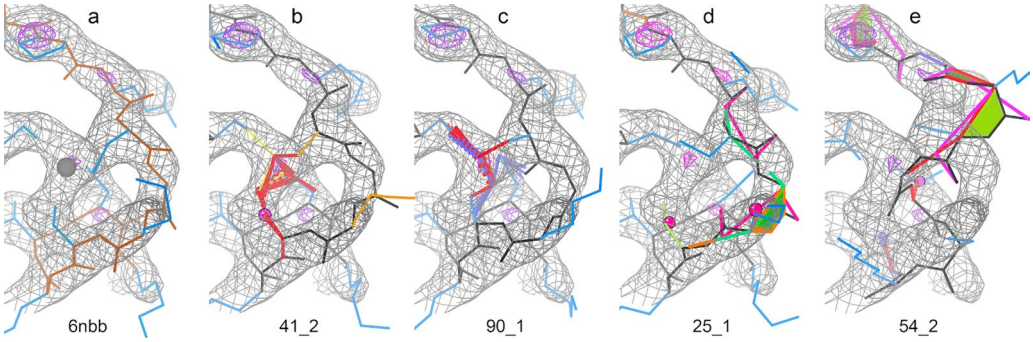
c



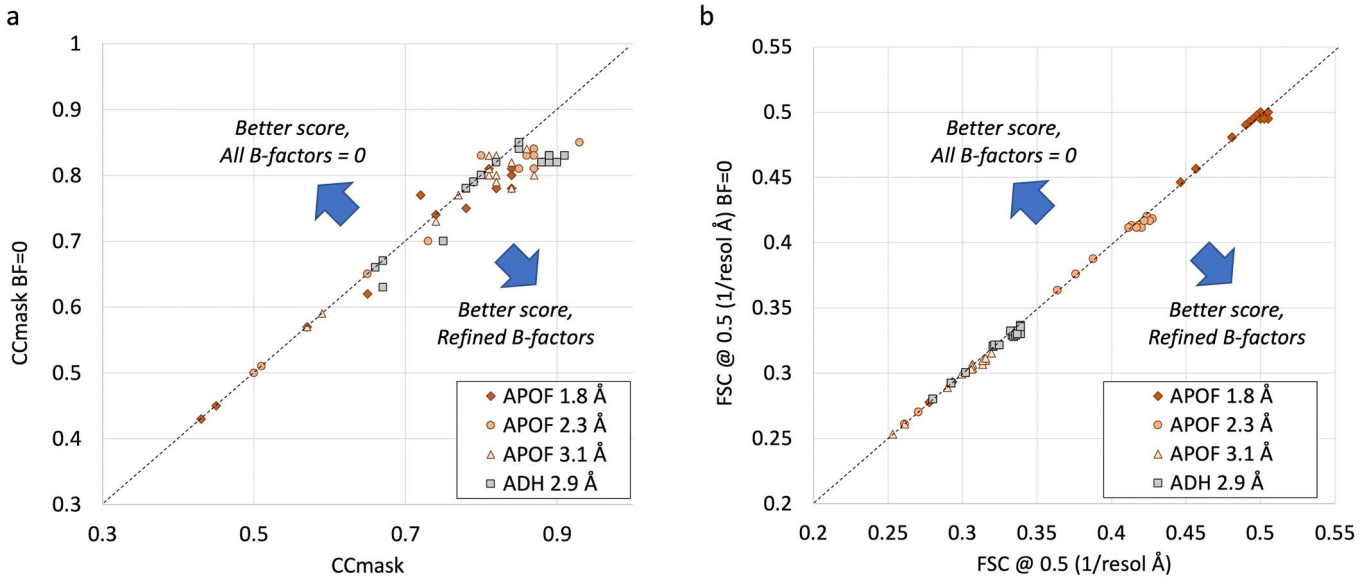
d

Metric	Ile 24	Asn25	Leu26	Glu27	Leu28	Tyr29	Ala 30
Bond Lengths	-	-	Outlier(s)	-	Outlier(s)	Outlier(s)	-
Bond Angles	-	Outlier(s)	Outlier(s)	-	Outlier(s)	Outlier(s)	-
Clashes	-	-	-	-	Clash	-	-
Ramachandran	Favored	Favored	Favored	Allowed	Favored	Favored	Favored
CaBLAM	Favored	Disfavored	Disfavored	Favored	Disfavored	Disfavored	Favored
Cis Peptides	-	-	-	-	-	CisNonPro	-

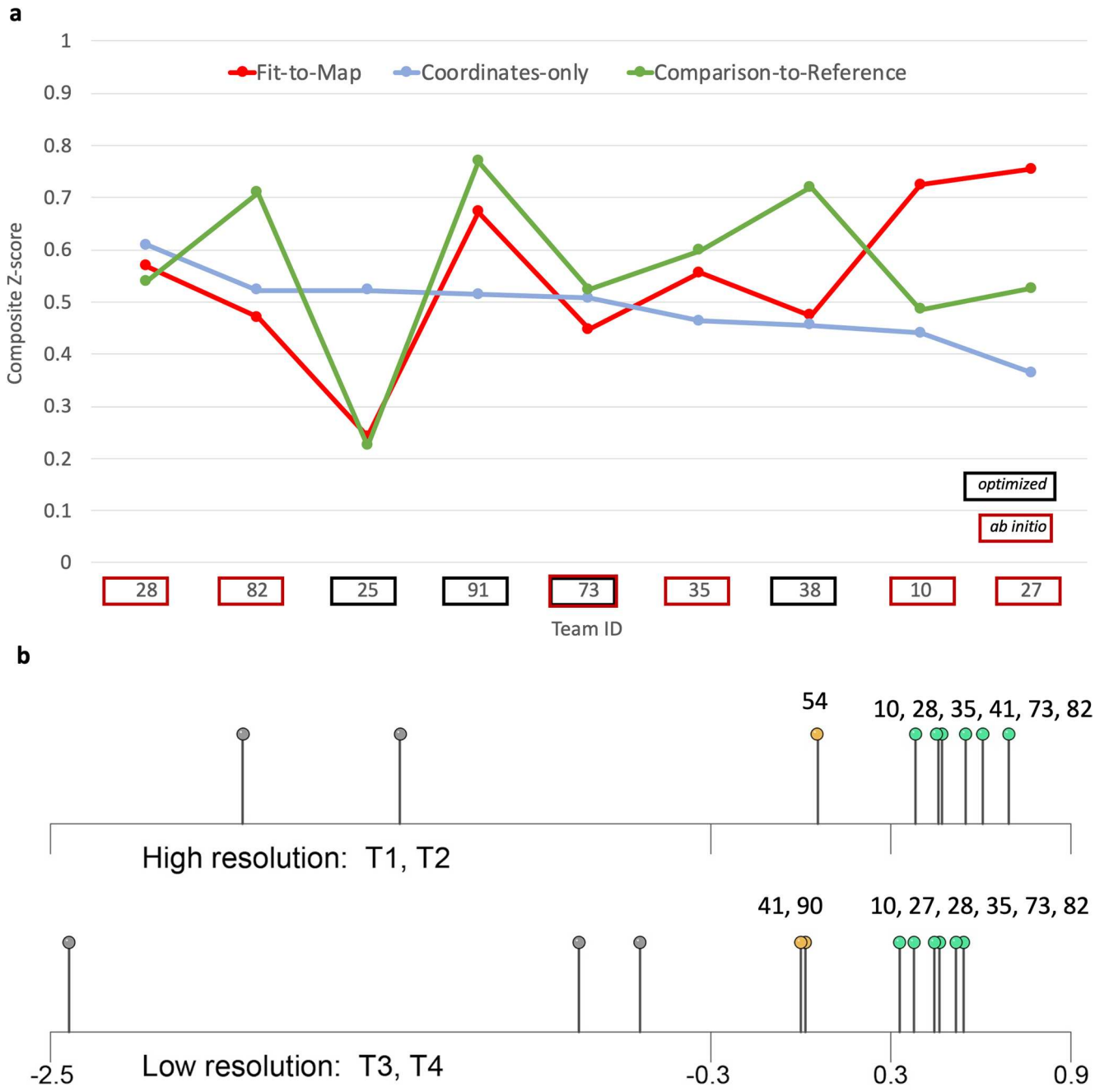
Extended Data Fig. 3 | Evaluation of a short sequence misalignment within a helix. Local Fit-to-Map and Coordinates-only scores are compared for a 3-residue sequence misalignment inside an α-helix in an *ab initio* model submitted to the Challenge (APOF 2.3 Å 54_1). **a**, Model residues 14–42 vs target map (blue: correctly placed residues, yellow: mis-threaded residues 25–29, black: APOF reference model, 3ajo). **b**, Structure-based sequence alignment of the *ab initio* model (top) vs. reference model (bottom). **c**, Local Fit-to-Map scores (screenshot from Challenge model evaluation website Fit-to-Map Local Accuracy tool). Curves are shown for Phenix Box_CC (orange), EMDB Atom Inclusion (purple), Q-score (red) EMRinger (green), and SMOC (blue). The score values for model residue Leu 28 are shown in the box at right. **d**, Residue scores were calculated using the Molprobity server. The mis-threaded region is boxed in (b-d). Panels (a) and (b) were generated using UCSF Chimera.



Extended Data Fig. 4 | Modeling errors around omitted Zinc ligand in ADH. Target 4 (ADH) density map with examples of modeling errors caused by omission of Zinc ligand. **a**, Reference structure with Zinc metal ion (gray ball) coordinated by 4 Cysteine residues (blue sidechains). **b**-**e**, Submitted models missing Zinc (labels indicate the group_model ids). All have geometry and/or conformational violations as flagged by MolProbity CaBLAM (magenta pseudobonds), cis-nonPro (green parallelograms), Ramachandran (green pseudobonds), Cbeta (magenta spheres), and angle (blue and red fans). Model (**b**) has backbone conformation very close to correct, while (**b**) and (**c**) both have flags indicating bad geometry of incorrect disulfide bonds. Models (**c**) and (**d**) have backbone distortions, and (**e**) is mistraced through the Zn density. Molecular graphics were generated using KiNG.



Extended Data Fig. 5 | Fit-to-Map Scores with and without refined B-factors (ADP). Two representative metrics are shown: **a**, CCmask correlation, **b**, FSC05 resolution^{-1}. Each plotted point indicates the calculated score for atom positions with B-factors included (horizontal axis) versus the calculated score for atom positions alone (vertical axis). Plot symbols identify map targets. Of 63 models total, 33 included refined B-factors. Differing scores +/- B-factors contribute off-diagonal points (black dotted lines are reference diagonals).



Extended Data Fig. 6 | Group performance evaluations. a, Group composite Z-scores plotted by metric category. The nine teams with highest Coordinate-only composite Z-score rankings are shown, sorted left to right. The plot illustrates that by group/method, Coordinate-only scores are poorly corelated with Fit-to-Map and Comparison-to-Reference scores. In contrast, a modest correlation is observed between Fit-to-Map and Comparison-to-Reference scores. **b**, Averaged model composite Z-scores plotted for ab initio modeling groups at higher resolution (T1 at 1.8 Å, T2 at 2.3 Å) and lower resolution (T3 at 3.1 Å, T4 at 2.9 Å). In each case 6 groups produced very good models ($Z \geq 0.3$; green pins), though not the same set. Runner-up clusters ($-0.3 \leq Z < 0.3$) are shown with gold pins. Individual scores and order shift with alternate choices of evaluation metrics and weights, but the clusters at each resolution level are stable. Composite Z-scores were calculated as described in Methods.

nature research

Corresponding author(s): Catherine L. Lawson, Wah Chiu

Last updated by author(s): Nov 12, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Software for Models prepared by participating groups:

Ab initio
ARP/wARP v.8.0 arpwarp.embl-hamburg.de (groups 27, 41)
CCPEM v.1.2.0/Buccaneer-v.1.16.8* www.ccpem.ac.uk www.ccp4.ac.uk (group 27)
CCPEM v.1.3.0/Buccaneer-v.1.16.8* www.ccpem.ac.uk www.ccp4.ac.uk (groups 10, 28)
Cascaded-CNN v.1.0 github.com/DrDongSi/Ca-Backbone-Prediction (group 60)
Mainmast v.1.0 kiharalab.org/mainmast (groups 54, 73)
Pathwalker v.2.0 blake.bcm.edu/emanwiki/EMAN2/Programs/e2pathwalker (group 90)
Rosetta 3.9 rosettacommons.org (groups 27, 54, 82**)

optimization
CDMD v.gromacs-5.0.7-densfit www.mpibpc.mpg.de/grubmueller/densityfitting (group 25)
CNS v.1.3 cns-online.org/v1.3 (group 38)
DireX v.0.7.1 simtk.org/home/direx (group 38)
Gromacs v.2018.6 gromacs.org (group 38)
Phenix/real_space_refine v.1.15 phenix-online.org (groups 10, 27, 35, 38, 91)
CCPEM v.1.3.0/Refmac v.5.7* www.ccpem.ac.uk www.ccp4.ac.uk (groups 28, 41)
MELD 0.2.3 github.com/maccallumlab/meld (group 73)
MDFF v.0.4 www.ks.uiuc.edu/Research/vmd/plugins/mdff (groups 38, 54, 73
reMDFF v.0.4 github.com/jvant/ReMDFF_Singharoy_Group (group 73)

Visual evaluation/manual model improvement:
VMD v.1.9.3 www.ks.uiuc.edu/Research/vmd (groups 54, 73, 82)
UCSF Chimera v.1.11-v.1.14 www.cgl.ucsf.edu/chimera (groups 10, 38, 60, 73, 90)
PyMol v.2.2.0-v.2.3.0 github.com/schrodinger/pymol-open-source (groups 10, 27)

CCPEM/COOT v.1.3.0 www.ccpem.ac.uk (group 28)
COOT v.0.9-pre www2.mrc-lmb.cam.ac.uk/Personal/pemsley/coot (groups 10, 27, 28, 41, 90, 91)

Model coordinate submission metadata were collected using a Drupal webform.
Model coordinates were collected using pdb-extract.wwpdb.org and processed using MAXIT swtools.rcsb.org/apps/MAXIT

*The CCPEM package requires installation of the CCP4 package (www.ccp4.ac.uk) in order to run Buccaneer and Refmac.
**Full modeling scripts (group 82): https://faculty.washington.edu/dimaio/files/rosetta_em_challenge_2019.tgz

See also Table I

Data analysis

Fit-to-Map
TEMPy v.1.1 tempy.ismb.lon.ac.uk (CCC, CCC_OV, SMOC, LAP, MI, MI_OV, ENV)
Phenix/map_model_cc v.1.15 phenix-online.org (CCbox, CCpeaks, CCmask, FSC05)
Phenix/em_ringer v.1.15 phenix-online.org (EMRinger)
CCPEM v.1.4.1/ Refmac v.5.7* www.ccpem.ac.uk www.ccp4.ac.uk (FSCavg)
EMDB CryoEM Validation Analysis (va) v.0.0.dev8 pypi.org/project/va/0.0.0.dev8 (AI_all)

Coordinates-only
Phenix/molprobity v.1.15 phenix-online.org (CaBLAM, Clashscore, Rotamer, Rama, Calpha)
Phenix/model_statistics v. 1.15 phenix-online.org (Bond, Angle, Chiral, Planar, Dihedral)
MAPQ v.1.2 github.com/gregdp/mapq (Qscore)
KiNG 2.23 kinemage.biochem.duke.edu/software (issue visualization)

Comparison-to-Reference
LGA v.04.2019 proteinmodel.org/AS2TS/LGA/lga.html (GDT-TS, GDC, GDC-SC, DAVIS-QA)
OpenStructure/LDDT v.2.1 www.openstructure.org/download (LDDT)
CAD v.1646 bitbucket.org/kliment/voronota/src/master (CAD)
HBPLUS v.3.06 www.ebi.ac.uk/thornton-srv/software/HBPLUS (HBPR>6)

*The CCPEM package requires installation of the CCP4 package (www.ccp4.ac.uk) in order to run Refmac.

See also Online Methods and Table II.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The map targets used in the Challenge were downloaded from EM Data Bank, entries:
EMD-20026 (file: emd_20026_additional_1.map.gz),
EMD-20027 (file: emd_20027_additional_2.map.gz),
EMD-20028, (file: emd_20028_additional_2.map.gz), and
EMD-0406. (file: emd_0406.map.gz)
Reference models were downloaded from Protein Data Bank, entries 3ajo and 6nbb.

Submitted models, model metadata, result logs, and compiled data are archived at Zenodo: https://doi.org/10.5281/zenodo.4148789, and at https://model-compare.emdataresource.org/data/2019/. Interactive summary tables, graphical views, and csv downloads of compiled results are available at https://model-compare.emdataresource.org/2019/cgi-bin/index.cgi.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Sample size was determined by the number of model coordinate submissions. Sample size was sufficient to meet the goal of qualitatively comparing model quality across current methods in use, and assessing usefulness of different model metrics. The Challenge was not designed to quantitatively and exhaustively explore all variables.

Data exclusions	All 63 submitted models were evaluated, with the exception that model hydrogen atom positions and refined B-factors were excluded from the reported Fit-to-Map analyses.
Replication	Participating groups were asked to complete the same four modeling tasks, yielding 15-17 models per task. Each model was created independently, so there are no exact replicates.
Randomization	Not applicable--No attempt was made to randomize the data.
Blinding	Initial evaluations of the submitted coordinates were blinded to the identity of the participating groups and software used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

B. Embedded Publication II

ARTICLES
https://doi.org/10.1038/s41594-020-0442-4

nature structural & molecular biology

Cryo-EM structure of islet amyloid polypeptide fibrils reveals similarities with amyloid-β fibrils

Christine Röder[1,2,3,6], Tatsiana Kupreichyk[1,2,3,6], Lothar Gremer[1,2,3], Luisa U. Schäfer[1,2], Karunakar R. Pothula[1,2], Raimond B. G. Ravelli[4], Dieter Willbold[1,2,3], Wolfgang Hoyer[1,2,3] ✉ and Gunnar F. Schröder[1,2,5] ✉

Amyloid deposits consisting of fibrillar islet amyloid polypeptide (IAPP) in pancreatic islets are associated with beta-cell loss and have been implicated in type 2 diabetes (T2D). Here, we applied cryo-EM to reconstruct densities of three dominant IAPP fibril polymorphs, formed in vitro from synthetic human IAPP. An atomic model of the main polymorph, built from a density map of 4.2-Å resolution, reveals two S-shaped, intertwined protofilaments. The segment 21-NNFGAIL-27, essential for IAPP amyloidogenicity, forms the protofilament interface together with Tyr37 and the amidated C terminus. The S-fold resembles polymorphs of Alzheimer's disease (AD)-associated amyloid-β (Aβ) fibrils, which might account for the epidemiological link between T2D and AD and reports on IAPP–Aβ cross-seeding in vivo. The results structurally link the early-onset T2D IAPP genetic polymorphism (encoding Ser20Gly) with the AD Arctic mutation (Glu22Gly) of Aβ and support the design of inhibitors and imaging probes for IAPP fibrils.

Pancreatic islet amyloid deposits are a hallmark of T2D. Islet amyloid, first reported almost 120 years ago as islet hyaline[1], is found in >90% of individuals with T2D[2,3]. The main constituents of islet amyloid are fibrillar aggregates of the 37-residue polypeptide hormone IAPP, also called amylin. IAPP is detected in many organs, including the brain, but is mainly localized in the beta-cells of pancreatic islets, where it is co-synthesized and co-secreted with insulin[3,4]. IAPP is involved in glucose homeostasis and metabolism, with putative functions as a regulator of insulin and glucagon secretion, satiety and gastric emptying[3,5]. Formation of toxic IAPP amyloid aggregates has been associated with dysfunction and death of beta-cells, placing T2D in the group of protein misfolding disorders[2,3,5–8]. However, the nature of the toxic IAPP species and the mechanisms of beta-cell death are not well determined[9]. Potential toxic effects of IAPP amyloid include induction of apoptosis[10], chronic inflammation[11], defects in autophagy[12,13], endoplasmic reticulum stress[14,15] and membrane disruption[16]. Apart from its association with T2D, IAPP amyloid might also play a role in type 1 diabetes[10,17].

IAPP interacts with amyloidogenic proteins that trigger other protein misfolding disorders[18–20]. Of particular interest is its relation to the Aβ peptide, the main component of senile plaques found in the brain tissue of patients with AD. IAPP and Aβ are infamous not only for their strong aggregation propensity and the insolubility of their aggregates[3], but also for their primary sequence similarity[21]. IAPP and Aβ colocalize in Aβ deposits in the brain tissue of patients with AD[19]. Mutual cross-seeding of IAPP and Aβ aggregation observed in transgenic mice further supports a role of the IAPP–Aβ interaction in pathogenesis[19,20].

Structural information on IAPP amyloid is fundamental for improving understanding of the mechanism of amyloid formation, for defining toxic IAPP species and for elucidating IAPP–Aβ cross-seeding[5,7]. Furthermore, high-resolution IAPP fibril structures can inform the design of fibril growth inhibitors and support the development of soluble, nontoxic IAPP analogs for co-formulation with insulin and leptin for treatment of type 1 diabetes and obesity, respectively[5]. Current structural models of IAPP fibrils at physiological pH based on, for example, solid-state NMR (ssNMR) of full-length IAPP and X-ray crystallography of IAPP fragments consistently place the majority of the 37 amino acid residues into the fibril core, while the N terminus is located at the periphery[22–27]. Conversely, the available models also exhibit substantial differences, which could be either a consequence of the limited, distinct restraints obtained by the different techniques applied or a reflection of IAPP fibril polymorphism[5,28]. Here, we have applied cryo-EM to determine the structure of IAPP amyloid fibrils grown at physiologically relevant pH. We provide a structural analysis of three dominant polymorphs, including an atomic model of the main polymorph comprising residues 13–37 in a density map of 4.2-Å resolution.

Results

Polymorphism of IAPP fibrils. For this work, amyloid fibrils were prepared from synthetic human IAPP including the amidated C terminus. Islet amyloid in T2D is typically extracellular, but IAPP aggregation is supposedly initiated intracellularly, possibly in the secretory granules at a pH of 5.0–6.0 (refs. [3,29]); therefore, IAPP fibrils were prepared at a pH of 6.0. Long, well-ordered fibrils were obtained, as shown by atomic force microscopy (AFM) imaging (Extended Data Figs. 1 and 2). We could differentiate at least five different polymorphs in the AFM images and in subsequently performed cryo-EM experiments. Of these five polymorphs, three were

[1]Institute of Biological Information Processing (IBI-7: Structural Biochemistry), Forschungszentrum Jülich, Jülich, Germany. [2]Jülich Centre for Structural Biology (JuStruct), Forschungszentrum Jülich, Jülich, Germany. [3]Institut für Physikalische Biologie, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [4]The Multimodal Molecular Imaging Institute, Maastricht University, Maastricht, the Netherlands. [5]Physics Department, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [6]These authors contributed equally: Christine Röder, Tatsiana Kupreichyk. ✉e-mail: wolfgang.hoyer@uni-duesseldorf.de; gu.schroeder@fz-juelich.de

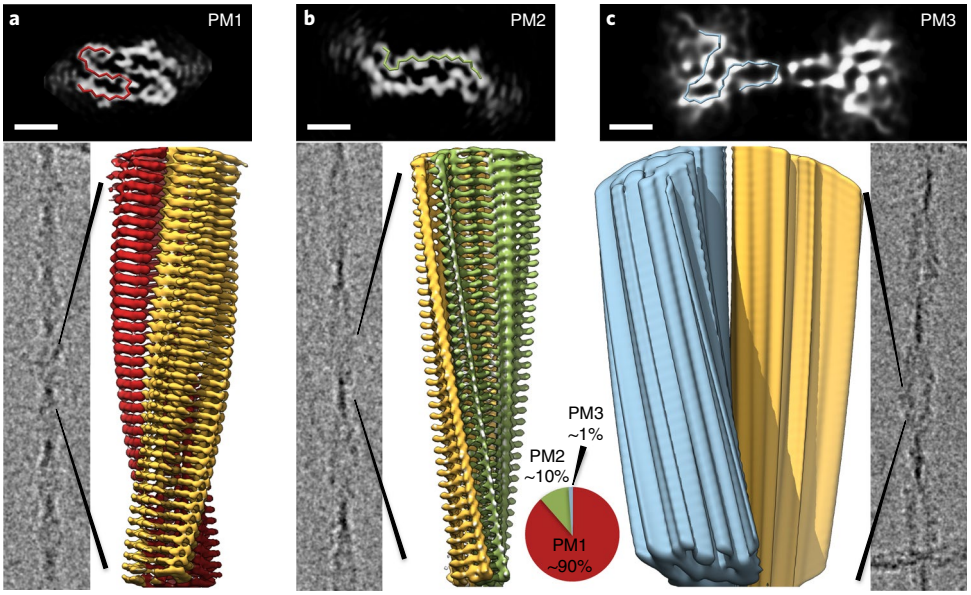


Fig. 1 | Comparison of reconstructed IAPP polymorphs. a–c, PM1 (**a**), PM2 (**b**) and PM3 (**c**). For each polymorph, three panels are shown: a slice of a 3D reconstruction superimposed on the respective Cα chain for one monomer (black; scale bars, 2 nm); a micrograph displaying the respective polymorph (gray); and 3D density (red/yellow, PM1 (**a**); green/yellow, PM2 (**b**); blue/yellow, PM3 (**c**)). The pie chart visualizes the fraction of each polymorph in the dataset.

present in sufficient amounts for further analysis (Fig. 1, Table 1, and Extended Data Figs. 1 and 2). The main polymorph, polymorph 1 (PM1), makes up ~90% of all fibrils, while polymorph 2 (PM2) and polymorph 3 (PM3) represent up to ~10% and ~1%, respectively, of the total number of fibrils in the dataset.

PM1 has a right-handed helical symmetry with a pitch of 48 nm and a width of 2.5–4.5 nm (Figs. 1a and 2a). Three-dimensional (3D) reconstruction of 1,161 individual fibril images using a helical pseudo-2_1 symmetry led to 4.2-Å resolution, which was sufficient to unambiguously build an atomic model with helical parameters of 2.35 Å (helical rise) and 178.23° (helical twist). The fibril consists of two stacks of S-shaped IAPP monomers winding around each other. Further details of the structure and molecular characteristics of PM1 are described later.

PM2 also consists of two protofilaments and exhibits pseudo-2_1 symmetry (Fig. 1b). With a maximum and minimum width of 52 Å and 17 Å, respectively, PM2 shows a more pronounced twist in the projection images (Fig. 1b) and is remarkably flatter than PM1 (Fig. 1a). The helical pitch is 94 nm, and AFM experiments suggest a left-handed twist. In contrast to the S-shaped PM1, the density map indicates an extended conformation of two IAPP monomers in PM2. The protofilament interface consists of a continuous sequence region of at least 18 amino acids. The density map with approximately 4.2-Å resolution would in principle allow for model building of 21 amino acid residues, but the sequence assignment is ambiguous. Therefore, we modeled all 17 possible sequence assignments in both forward and backward backbone trace directions, leading to $17 \times 2 = 34$ different models. All 34 models were refined in DireX[30] using cross-validation in the resolution range of 3.0–4.0 Å for calculation of the C_{free} value[31]. Results were ranked by C_{free} value (Extended Data Fig. 3). According to this criterion, the most probable model for PM2, which also exhibits the highest C_{work} value, shares important features with the PM1 model, as discussed below.

Compared to the other polymorphs, PM3 was not well represented in the micrographs. The overall features of PM3, namely the broad width (110 Å) and pronounced twist (159-nm pitch), lead to a dumbbell shape (Fig. 1c). From the 4,591 particles extracted, we could reconstruct a density with 8.1-Å resolution. Because the resolution was rather low, we were not able to build an atomic model but only hypothesize a possible Cα backbone trace (Fig. 1c). Nonetheless, the density also clearly indicates two symmetric protofilaments and reveals that the 10-Å-wide protofilament interface of PM3, presumably consisting of three residues, is very small compared to those of PM1 and PM2 (Fig. 1).

Fibril architectures of PM1 and PM2. In PM1, each monomer exhibits an overall S-fold that comprises residues Ala13–Tyr37 (Fig. 2). Up to residue 12, the N-terminal part including the disulfide bond between Cys2 and Cys7 is largely disordered and, therefore, does not reveal clear density (Fig. 1a). The side view of PM1 shows the typical cross-β pattern of amyloid fibril structures with a spacing of 4.7 Å between the layers (Fig. 2c). The cross-β layers are well resolved in the density, as shown in Fig. 2d. On the secondary structure level, we observed three β-sheets: residues 14–20, 26–32 and 35–37. Figure 3b shows the comparison of our model with former secondary structure predictions based on sequence analysis[22], NMR[23,26,27], electron paramagnetic resonance (EPR)[25] and X-ray crystallography experiments[24].

The cross-section of the PM1 fibril displays two monomeric S-folds related by approximate 2_1 symmetry (Fig. 2b). The double-S shape is stabilized by both hydrophobic and polar interactions. The central part of the protofilament interface contains a hydrophobic cluster comprising residues Phe23, Gly24, Ala25 and Leu27 as well as Phe23′, Gly24′, Ala25′ and Leu27′ (Fig. 2b and Extended Data Fig. 4). Additionally, the backbone of Phe23 and Ala25 forms hydrogen bonds at the center of the fibril, thereby connecting one subunit with two neighboring subunits above and below in the opposing protofilament (Fig. 3c). More precisely, there is a hydrogen bond between the carbonyl group of Phe23 of chain i and the amide group of Ala25 of chain $i+1$ and another hydrogen bond between the amide group of Ala25 of chain i and the carbonyl group of Phe23 of chain $i-1$. The backbone around Gly24 does not maintain the cross-β hydrogen-bonding pattern along the fibril. The aforementioned interactions are formed by residues located in the sequence motif (N)NFGAIL, shown earlier to be important for fibrillization of IAPP[5,32–34]. This motif is located in the central part

Table 1 | Cryo-EM data collection, refinement and validation statistics

	PM1 (EMD-10669, PDB 6Y1A)	PM2 (EMD-10670)	PM3 (EMD-10671)
Data collection and processing			
Magnification	110,000	110,000	110,000
Voltage (kV)	200	200	200
Dose rate (e^{-} Å^{-2} s^{-1})	0.9	0.9	0.9
Exposure time (s)	46	46	46
Movie frames (no.)	1,800	1,800	1,800
Defocus range (μm)	−1.0 to −2.2	−1.0 to −2.2	−1.0 to −2.2
Pixel size (Å)	0.935	0.935	0.935
Symmetry imposed	helical, pseudo 2_1	helical, pseudo 2_1	helical, pseudo 2_1
Helical rise (Å)	2.351	2.352	2.323
Helical twist (°)	178.23	179.10	179.47
Helical pitch (Å)	479.5	940	1590
Final fibril images (no.)	1,161	1,480	99
Final particle images (no.)	37,120	24,011	4,591
Map resolution (Å)	4.2	4.2	8.1
FSC threshold	0.143	0.143	0.143
Refinement			
Initial density model used	Noise-filled cylinder	Noise-filled cylinder	Noise-filled cylinder
Model composition			
Non-hydrogen atoms	2,975		
Protein residues	416		
R.m.s. deviations			
Bond lengths (Å)	0.0039		
Bond angles (°)	0.60		
Validation			
MolProbity score	1.99		
Clashscore	15.2		
Poor rotamers (%)	0		
Ramachandran plot			
Favored (%)	95.7		
Allowed (%)	4.3		
Disallowed (%)	0		

of the structure, in the turn between the first two β-sheets (Figs. 2b and 3a,b). Within this turn, the kink around Phe23 and Asn21 is stabilized by hydrogen bonds between Asn22 and Ser19, as well as between Asn22 and Gly24 (Figs. 2b and 3c). Additionally, Ile26 might support this turn by hydrophobic interactions with Val17. In the second turn, between β-sheets 2 and 3, Asn31 together with Ser29, Asn35 and Tyr37 creates a hydrophilic cluster at the C terminus of IAPP with possible interactions between Asn31 and Ser29, as well as Asn31 and Asn35. In addition, Tyr37 might interact with both Asn35 and Ser29 (Fig. 2b). Moreover, the amidated C terminus itself forms a polar ladder (Fig. 3d). This ladder is further connected to Asn21′ of the opposite protofilament with slightly longer and, therefore, weaker hydrogen bonds. The overall cross-β arrangement is further stabilized by Asn14, Asn21 and Asn31, which form polar ladders alongside the fibril axis. Asn22 does not form a polar ladder, but instead its Nδ2 atom forms a hydrogen bond with the carbonyl group of Gly24 within the same monomer (Fig. 3c). It should be noted that the detailed analysis of the hydrogen-bonding network is derived from the atomic model, which is an interpretation of the experimental density map.

IAPP contains an unusually large number of the polar residues asparagine, serine and threonine[5]. We found that these residues form polar streaks within the fibril core of PM1 (Extended Data Fig. 4). The polar streaks are characterized by extensive networks of hydrogen bonds, as discussed earlier. The segregation of polar and apolar residues into distinct clusters within the fibril core likely contributes to the high stability of IAPP amyloid. In IAPP, this segregation is facilitated by the preorganization of amino acid residues in polar and apolar clusters within the primary structure, in the fashion of a block copolymer with polar blocks 19-SSNN-22 and 28-SSTN-31 and apolar block 23-FGAIL-27.

All-atom molecular dynamic (MD) simulations were performed to evaluate the overall stability of the model. In two independent 250-ns simulations, the model remained stable (Extended Data Fig. 5) with an all-atom r.m.s. deviation (r.m.s.d.) of a single subunit from the deposited model of ~2 Å and an r.m.s. fluctuation (r.m.s.f.) of residues 16–37 of 0.8 Å. The N-terminal part including Phe15 was already substantially more mobile (Extended Data Fig. 5a,b,e). Notably, we observed ladder formation for Asn22 in the MD simulation, which was not supported by the density map.

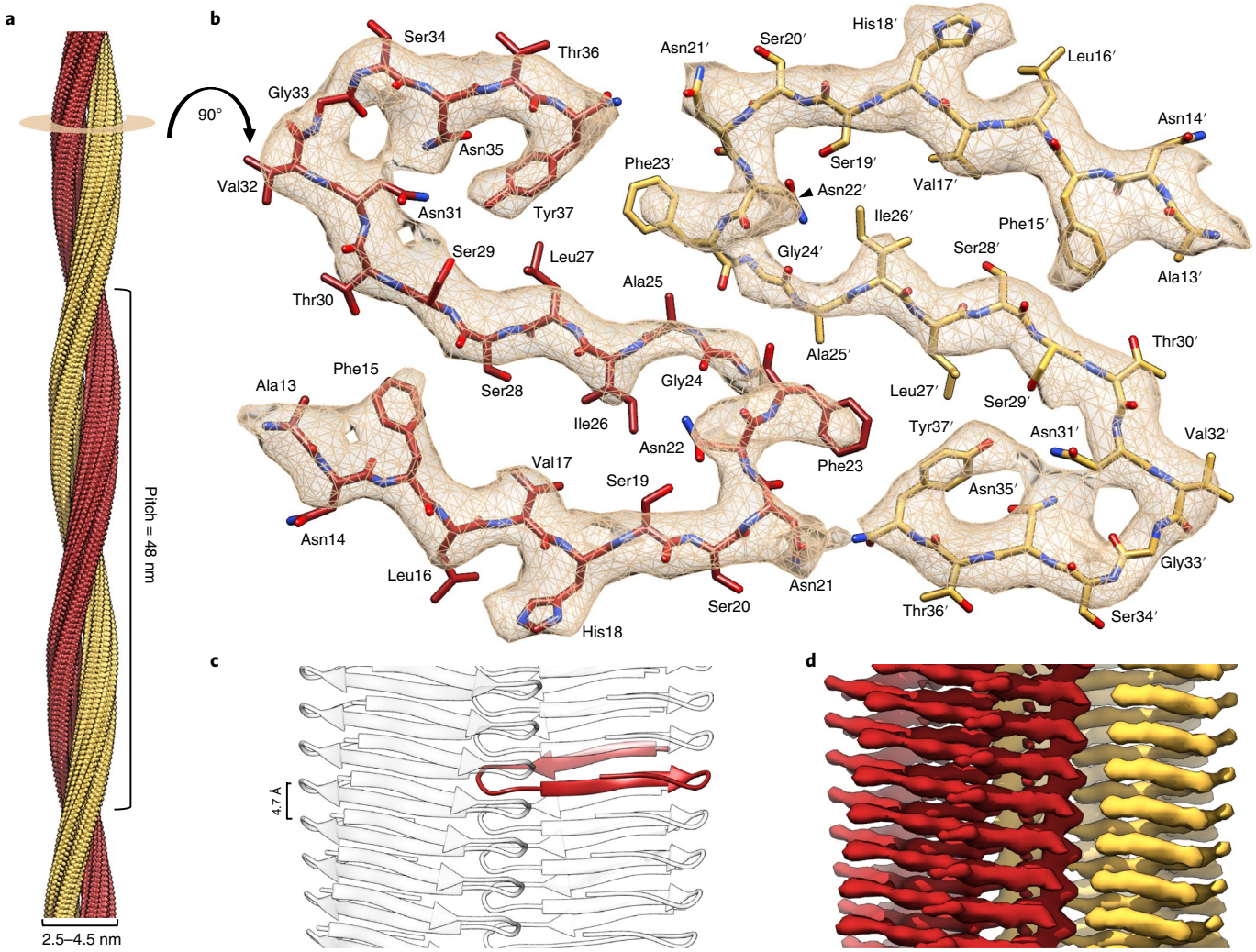


Fig. 2 | Architecture of the main polymorph, PM1. a, PM1 exhibits a helical pitch of 48 nm and a minimum and maximum width of 2.5 and 4.5 nm, respectively. The fibril consists of two protofilaments (red and yellow). **b**, Cross-sectional view: two symmetry-related monomers, with an atomic model of residues 13–37 built into the 4.2-Å-resolution density (contour level of 1.5σ). **c**, Side view: one monomer is highlighted (red) to show its integration in the fibrillar structure. Cross-β layers are separated by 4.7 Å. **d**, Side view: reconstructed density corresponding to **c**.

For earlier structures of amyloid fibrils, we discussed the need for a minimal fibril unit, which is the smallest fibril structure fragment in which the capping subunits at both ends would have established the same full contact interface with other constituting monomers as the capping subunits of an extended fibril[35,36]. Here, the minimal fibril unit consists of only three monomers, which is the smallest possible unit. One subunit is in contact exclusively with its neighboring monomers above and below and with its opposing monomers through protofilament interface contacts (Fig. 2c). Indeed, we did not observe any interlocking of different cross-β layers, which was postulated to have a stabilizing effect on other amyloid fibrils[35,36].

The IAPP folds in PM1 and PM2 are clearly distinct, yet the most probable model for PM2 shares important features with the PM1 model (Extended Data Fig. 3). First, the NFGAIL motif forms the center of the fibril interface. Second, the N terminus is rather flexible and thus not resolved in the density map. The first visible residue in the density of the PM2 model is Phe15. In contrast to PM1, not only the N terminus but also the two C-terminal residues Thr36 and Tyr37 are not clearly resolved and are potentially mobile. In between the two protofilaments is a relatively large cavity lined by hydrophobic residues Phe23, Ala25 and Ile26. It is not clear whether this gap is water filled.

Similar S-folds in IAPP and Aβ fibrils. Colocalization of IAPP and Aβ has been observed in patients with T2D and AD[19]. The epidemiological link between diabetes and dementia might be explained by cross-seeding of IAPP and Aβ aggregation[19,20,37,38]. Different sites on amyloid fibrils are relevant for cross-seeding: cross-elongation (that is, the elongation of a fibril with a heterologous protein) occurs at the fibril end, while cross-nucleation (that is, the fibril-catalyzed formation of a heterologous fibril nucleus) may occur both at the fibril end and along the fibril surface. Like IAPP, Aβ forms different fibril polymorphs, according to ssNMR and cryo-EM studies[35,39–43]. Comparing IAPP PM1 to multiple $A\beta_{1-42}$ polymorphs containing S-shaped folds[35,40,44], we found that the backbones superimpose (Fig. 4b,c). The structural similarity of the backbones is highest when superimposing the models in an antiparallel arrangement (Fig. 4c). The similarity between IAPP and $A\beta_{1-42}$ fibril folds regarding topology and size might promote cross-seeding at the fibril end, which could further be supported by the sequence similarity of IAPP and Aβ[21]. The sequence similarity is highest around the Gly-Ala-Ile segment at positions 24–26 of IAPP and positions 29–31 of Aβ. In both IAPP and Aβ, this segment is located in the solvent-excluded center of the S-fold (Fig. 4d). A further segment that can be superimposed in parallel arrangement is the N-terminal strand of the

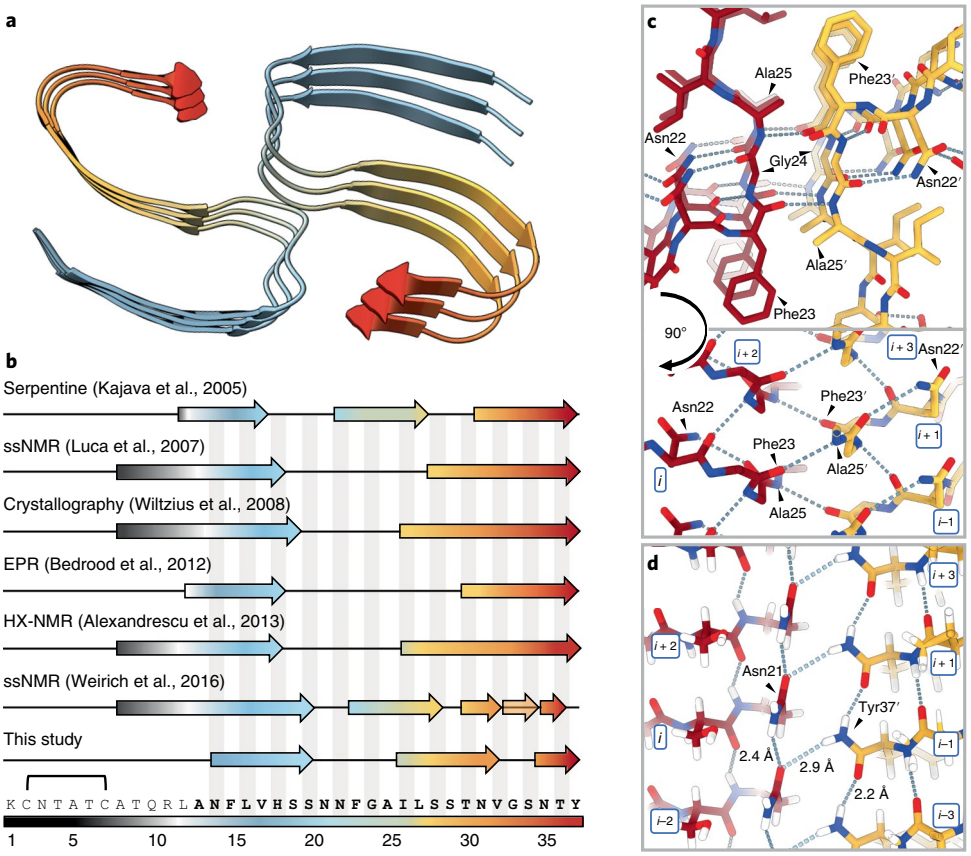


Fig. 3 | Secondary structure and hydrogen bonding in PM1. a, Secondary structure of the IAPP model. Tilted cross-section of three fibril layers with representation of the three β-sheets. **b**, Comparison of our PM1 structure to former models derived from sequence-based prediction[22], EPR[25], ssNMR[27], X-ray crystallography[24] and hydrogen-exchange NMR (HX-NMR) studies[26]. Arrows indicate β-sheets (one potential β-sheet is shown as transparent). The disulfide bridge between residues Cys2 and Cys7 is indicated. Fibril formation was performed at pH 7.4 (refs. [23,25–27]) or 6.5 (ref. [24]), while our IAPP fibrils were formed at pH 6.0. **c**, Top view of the fibril core showing the two NFGAIL motifs in the opposing protofilaments (red and yellow), as well as the hydrogen-bonding network (dashed lines). Bottom, side view of the fibril core illustrating the interlocking of the protofilaments by hydrogen bonds. **d**, Side view of the fibril showing the hydrogen-bonding interaction of Asn21 with the amidated C terminus of the opposing protofilament (yellow), as well as polar ladders of Asn21 and Tyr37′ along the fibril axis.

S-fold in IAPP PM1 and in the LS-shaped $A\beta_{1-42}$ polymorph, corresponding to 14-NFLVHSSNN-22 of IAPP and 16-KLVFFAEDV-24 of Aβ (Fig. 4d).

A serine-to-glycine substitution at position 20 (Ser20Gly), the only known IAPP genetic polymorphism in humans, is associated with early onset of T2D[45,46]. The Ser20Gly substitution enhances aggregation and toxicity of IAPP and leads to increased beta-cell apoptosis[47–50]. Substitution of serine with glycine has been suggested to promote turn formation at residue 20, favoring the amyloid fibril conformation[51,52]. In line with this notion, Ser20 is located at the edge of the turn comprising residues 20–25 in PM1. Interestingly, when comparing the S-fold of IAPP with the LS-fold of Aβ[35] (Fig. 4c,d), the Ser20Gly substitution in IAPP and the Arctic mutation (encoding a Glu22Gly substitution) of Aβ[53], which causes early-onset AD, are located at corresponding positions (Fig. 4d). This suggests that these two replacements with glycine might have analogous conformational consequences.

Discussion

The IAPP fibril samples investigated here displayed fibril polymorphism. While all three main polymorphs consist of two (pseudo) symmetric, helically intertwined protofilaments, they exhibit substantial differences in the protein fold (Fig. 1). PM1 consists of a compact S-shaped fold, PM2 features an extended IAPP conformation and the PM3 cross-section shows two compact motifs connected by an extended bridge. Marked differences are also observed between the protofilament interfaces: in PM1, the interface consists of one of the turns and the C-terminal end of the S-fold; in PM2, the entire extended IAPP segment that constitutes the fibril core is involved in the protofilament interface; and in PM3, a very narrow interface of probably only three residues is observed. Despite these differences, certain IAPP sequence segments might contribute similarly to distinct fibril polymorphs—in both PM1 and the most probable PM2 model, residues 22-NFGAIL-27 form the central fibril core.

In an early report[28] of IAPP fibril polymorphism, the most common polymorph consisted of two protofilaments coiled around each other with a helical pitch of 50 nm, while another polymorph showed a helical pitch of 100 nm. These values are in good agreement with PM1 (48 nm) and PM2 (94 nm). Despite these similarities, when comparing the cryo-EM results with previous structural data, it must be considered that variations may arise from differences between both the applied techniques and the polymorphs present in the samples. In line with previous studies[22–27], we found that the IAPP N terminus including the disulfide bond between Cys2 and Cys7 is not part of the fibril core, neither in PM1 (Fig. 3b) nor in the most probable model of PM2 (Extended Data Fig. 3). While well-defined density starts from residue 13 in the cryo-EM data, some studies reported the fibril core to begin around residue 8 (Fig. 3b). However, HX-NMR data indicated that residues 8–14

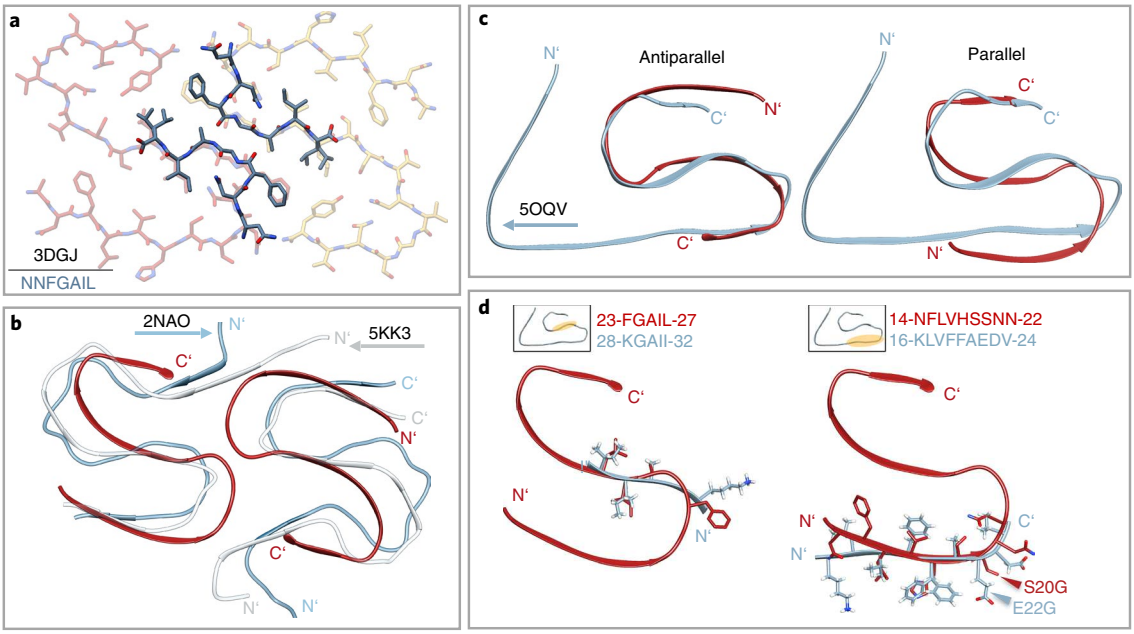


Fig. 4 | Structural comparison of PM1 with fibrillar IAPP peptide and Aβ fibril models. a, Overlay of our model with a crystal structure (dark; 3DGJ) of the NNFGAIL peptide from Wiltzius et al.[24]. **b**, Overlay of IAPP (red) with NMR structures by Wälti et al.[40] (light blue; 2NAO) and Colvin et al.[44] (gray; 5KK3). **c**, Antiparallel (left) and parallel (right) overlay of one IAPP PM1 monomer (red) with the atomic model of $A\beta_{1-42}$ (light blue; 5OQV) from Gremer et al.[35]. **d**, Detailed parallel overlay of sequence segments in IAPP (red) and $A\beta_{1-42}$ (light blue; 5OQV) from Gremer et al.[35]. Small boxes indicate where the respective sequence motif is located in the $A\beta_{1-42}$ model. Left, the FGAIL motif of IAPP shows high sequence identity to KGAII in $A\beta_{1-42}$. Right, the NFLVHSSNN motif of IAPP corresponds to the KLVFFAEDV motif of $A\beta_{(1-42)}$ with high structural similarity. Disease-related substitutions in IAPP (Ser20Gly) and $A\beta_{1-42}$ (Glu22Gly) are located at corresponding positions.

were less protected than those in the central fibril core[26]. In agreement with previous data, residues 13–37 are largely in β-sheet conformation in PM1, although variation exists with respect to the precise location of β-strands (Fig. 3b). A common feature of the PM1 cryo-EM structure and previous models is a turn in segment 20-SNNFG-24 (refs. [23–27]). A second turn is formed in PM1 in segment 32-VGS-34 and was also supported by ssNMR and HX-NMR studies[26,27]. Both turns establish an S-shaped fold of IAPP in PM1. Consequently, the tyrosyl ring of the C-terminal Tyr37 packs against Phe23′ in the adjacent protofilament, which is in line with distance restraints for IAPP fibrils obtained by fluorescence resonance energy transfer[54]. In addition, these energy-transfer experiments proposed a maximum distance of 11 Å between Tyr37 and a second phenylalanine, coinciding with the Tyr37–Phe15 distance in the PM1 model[54]. The C-terminal amide stabilizes the S-shaped fibril structure by forming a polar ladder and a hydrogen bond with Asn21 in the adjacent protofilament (Fig. 3d), in line with enhanced amyloid formation upon C-terminal amidation of IAPP[55,56].

The sequence region at positions 20–29 is particularly important for the amyloidogenicity of IAPP[5,32,33]. This can be rationalized with the PM1 fibril structure. First, residues 22-NFGAILSS-29 constitute the solvent-excluded central core of PM1 fibrils (Fig. 2b). Second, residues 21-NNFGAIL-27 form, together with Tyr37 and the amidated C terminus, the protofilament interface. In previous structural models of IAPP fibrils, the region encompassing residues 20–29 was associated with formation of a partially ordered loop rather than a β-structure, which was surprising considering the sensitivity of IAPP amyloid formation to mutations mapping to this region[7]. The PM1 fibril structure shows that residues 20–25 indeed form a turn, albeit one that is an integral part of the fibril core, featuring an extensive hydrogen-bonding network (Fig. 3c,d). Residues 26–29, on the other hand, are part of the central β-sheet of IAPP PM1 fibrils. Remarkably, the structure of the 21-NNFGAIL-27 segment in PM1 is highly similar in atomic detail to a crystal structure of the NNFGAIL peptide[24] (Fig. 4a). This applies both to the fold of the individual polypeptide molecules and to the peptide–protofilament interface, which displays extensive main chain–main chain interactions between the 23-FGA-25 segments. The similarity of the NNFGAIL structure between the peptide crystal and the PM1 fibril indicates that the 21-NNFGAIL-27 segment drives IAPP amyloid formation.

In contrast to the human protein, IAPP proteins from several other species were found to be non-amyloidogenic[5]. The non-amyloidogenic rat and mouse IAPP contain six amino acids that are different from the human sequence[33]. Five of these are located in the sequence region encompassing residues 23–29, which is part of the central core of PM1 fibrils, as discussed above. The differing amino acids include three prolines in rat and mouse IAPP at positions 25, 28 and 29. As proline disrupts secondary structures, these proline residues are incompatible with the PM1 structure, consistent with the low amyloidogenicity of rat and mouse IAPP. The insights gained from the rat and mouse IAPP sequences were exploited in the design of pramlintide, a non-amyloidogenic IAPP analog carrying proline substitutions at positions 25, 28 and 29 (ref. [57]). Pramlintide is co-administered with insulin in type 1 diabetes to improve glucose level regulation. Similarly, the combination of a non-amyloidogenic IAPP analog and leptin could be a promising treatment option for obesity[58]. However, these drugs would benefit from increased solubility[59]. The structural data on IAPP fibrils presented here may aid in the design of non-amyloidogenic, soluble IAPP analogs by suggesting potential sites for chemical modifications of IAPP that counteract fibril formation.

This study presents the 4.2-Å-resolution structure of an IAPP fibril polymorph consisting of two S-shaped protofilaments but also highlights the polymorphism of IAPP fibrils. The dominant S-shaped PM1 can rationalize many of the characteristics of IAPP fibrils described by various groups, suggesting that PM1 is a common polymorph or that it at least represents general features of

prevalent IAPP polymorphs. The study provides detailed insight into the link between the IAPP amino acid sequence and fibril structure; furthermore, it reveals similarities between IAPP and Aβ fibril structures, which are particularly striking in consideration of the link between diabetes and AD. The structural information gained may serve as a basis to define the mechanisms of amyloid formation and toxicity of IAPP. Moreover, the PM1 fibril may be used as a target structure to design imaging probes for IAPP fibrils and inhibitors of IAPP fibril growth.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41594-020-0442-4.

Received: 12 February 2020; Accepted: 30 April 2020;
Published online: 15 June 2020

References

1. Opie, E. L. On the relation of chronic interstitial pancreatitis to the islands of Langerhans and to diabetes mellitus. *J. Exp. Med.* **5**, 397–428 (1901).
2. Jurgens, C. A. et al. β-cell loss and β-cell apoptosis in human type 2 diabetes are related to islet amyloid deposition. *Am. J. Pathol.* **178**, 2632–2640 (2011).
3. Westermark, P., Andersson, A. & Westermark, G. T. Islet amyloid polypeptide, islet amyloid, and diabetes mellitus. *Physiol. Rev.* **91**, 795–826 (2011).
4. Wimalawansa, S. J. Amylin, calcitonin gene-related peptide, calcitonin, and adrenomedullin: a peptide superfamily. *Crit. Rev. Neurobiol.* **11**, 167–239 (1997).
5. Akter, R. et al. Islet amyloid polypeptide: structure, function, and pathophysiology. *J. Diabetes Res.* **2016**, 2798269 (2016).
6. Mukherjee, A., Morales-Scheihing, D., Butler, P. C. & Soto, C. Type 2 diabetes as a protein misfolding disease. *Trends Mol. Med.* **21**, 439–449 (2015).
7. Cao, P., Abedini, A. & Raleigh, D. P. Aggregation of islet amyloid polypeptide: from physical chemistry to cell biology. *Curr. Opin. Struct. Biol.* **23**, 82–89 (2013).
8. Halban, P. A. et al. β-cell failure in type 2 diabetes: postulated mechanisms and prospects for prevention and treatment. *J. Clin. Endocrinol. Metab.* **99**, 1983–1992 (2014).
9. Zraika, S. et al. Toxic oligomers and islet beta cell death: guilty by association or convicted by circumstantial evidence? *Diabetologia* **53**, 1046–1056 (2010).
10. Zhang, S. et al. The pathogenic mechanism of diabetes varies with the degree of overexpression and oligomerization of human amylin in the pancreatic islet β cells. *FASEB J.* **28**, 5083–5096 (2014).
11. Masters, S. L. et al. Activation of the NLRP3 inflammasome by islet amyloid polypeptide provides a mechanism for enhanced IL-1β in type 2 diabetes. *Nat. Immunol.* **11**, 897–904 (2010).
12. Rivera, J. F. et al. Human IAPP disrupts the autophagy/lysosomal pathway in pancreatic β-cells: protective role of p62-positive cytoplasmic inclusions. *Cell Death Differ.* **18**, 415–426 (2011).
13. Gupta, D. & Leahy, J. L. Islet amyloid and type 2 diabetes: overproduction or inadequate clearance and detoxification? *J. Clin. Invest.* **124**, 3292–3294 (2014).
14. Casas, S. et al. Impairment of the ubiquitin–proteasome pathway is a downstream endoplasmic reticulum stress response induced by extracellular human islet amyloid polypeptide and contributes to pancreatic β-cell apoptosis. *Diabetes* **56**, 2284–2294 (2007).
15. Hull, R. L. et al. Amyloid formation in human IAPP transgenic mouse islets and pancreas, and human pancreas, is not associated with endoplasmic reticulum stress. *Diabetologia* **52**, 1102–1111 (2009).
16. Janson, J., Ashley, R. H., Harrison, D., McIntyre, S. & Butler, P. C. The mechanism of islet amyloid polypeptide toxicity is membrane disruption by intermediate-sized toxic amyloid particles. *Diabetes* **48**, 491–498 (1999).
17. Paulsson, J. F. et al. High plasma levels of islet amyloid polypeptide in young with new-onset of type 1 diabetes mellitus. *PLoS ONE* **9**, e93053 (2014).
18. Martinez-Valbuena, I. et al. Interaction of amyloidogenic proteins in pancreatic β cells from subjects with synucleinopathies. *Acta Neuropathol.* **135**, 877–886 (2018).
19. Oskarsson, M. E. et al. In vivo seeding and cross-seeding of localized amyloidosis: a molecular link between type 2 diabetes and Alzheimer's disease. *Am. J. Pathol.* **185**, 834–846 (2015).
20. Moreno-Gonzalez, I. et al. Molecular interaction between type 2 diabetes and Alzheimer's disease through cross-seeding of protein misfolding. *Mol. Psychiatry* **9**, 1327–1334 (2017).
21. O'Nuallain, B., Williams, A. D., Westermark, P. & Wetzel, R. Seeding specificity in amyloid growth induced by heterologous fibrils. *J. Biol. Chem.* **279**, 17490–17499 (2004).
22. Kajava, A. V., Aebi, U. & Steven, A. C. The parallel superpleated β-structure as a model for amyloid fibrils of human amylin. *J. Mol. Biol.* **348**, 247–252 (2005).
23. Luca, S., Yau, W. M., Leapman, R. & Tycko, R. Peptide conformation and supramolecular organization in amylin fibrils: constraints from solid-state NMR. *Biochemistry* **46**, 13505–13522 (2007).
24. Wiltzius, J. J. W. et al. Atomic structure of the cross-β spine of islet amyloid polypeptide (amylin). *Protein Sci.* **17**, 1467–1474 (2008).
25. Bedrood, S. et al. Fibril structure of human islet amyloid polypeptide. *J. Biol. Chem.* **287**, 5235–5241 (2012).
26. Alexandrescu, A. T. Amide proton solvent protection in amylin fibrils probed by quenched hydrogen-exchange NMR. *PLoS ONE* **8**, e56467 (2013).
27. Weirich, F. et al. Structural characterization of fibrils from recombinant human islet amyloid polypeptide by solid-state NMR: the central FGAILS segment is part of the β-sheet core. *PLoS ONE* **11**, e0161243 (2016).
28. Goldsbury, C. S. et al. Polymorphic fibrillar assembly of human amylin. *J. Struct. Biol.* **119**, 17–27 (1997).
29. Hutton, J. C. The internal pH and membrane potential of the insulin-secretory granule. *Biochem. J.* **204**, 171–178 (1982).
30. Wang, Z. & Schröder, G. F. Real-space refinement with DireX: from global fitting to side-chain improvements. *Biopolymers* **97**, 687–697 (2012).
31. Falkner, B. & Schröder, G. F. Cross-validation in cryo-EM-based structural modeling. *Proc. Natl Acad. Sci. USA* **110**, 8930–8935 (2013).
32. Westermark, P., Engstrom, U., Johnson, K. H., Westermark, G. T. & Betsholtz, C. Islet amyloid polypeptide: pinpointing amino acid residues linked to amyloid fibril formation. *Proc. Natl Acad. Sci. USA* **87**, 5036–5040 (1990).
33. Betsholtz, C. et al. Sequence divergence in a specific region of islet amyloid polypeptide (IAPP) explains differences in islet amyloid formation between species. *FEBS Lett.* **251**, 261–264 (1989).
34. Tenidis, K. et al. Identification of a penta- and hexapeptide of islet amyloid polypeptide (IAPP) with amyloidogenic and cytotoxic properties. *J. Mol. Biol.* **295**, 1055–1071 (2000).
35. Gremer, L. et al. Fibril structure of amyloid-β(1–42) by cryo–electron microscopy. *Science* **358**, 116–119 (2017).
36. Röder, C. et al. Atomic structure of PI3-kinase SH3 amyloid fibrils by cryo-electron microscopy. *Nat. Commun.* **10**, 3754 (2019).
37. Janson, J. et al. Increased risk of type 2 diabetes in Alzheimer's disease. *Diabetes* **53**, 474–481 (2004).
38. Yang, Y. & Song, W. Molecular links between Alzheimer's disease and diabetes mellitus. *Neuroscience* **250**, 140–150 (2013).
39. Colvin, M. T. et al. Atomic resolution structure of monomorphic $A\beta_{42}$ amyloid fibrils. *J. Am. Chem. Soc.* **138**, 9663–9674 (2016).
40. Wälti, M. A. et al. Atomic-resolution structure of a disease-relevant $A\beta_{1-42}$ amyloid fibril. *Proc. Natl Acad. Sci. USA* **113**, 4976–4984 (2016).
41. Tycko, R. Molecular structure of aggregated amyloid-β: insights from solid state nuclear magnetic resonance. *Cold Spring Harb. Perspect. Med.* **6**, a024083 (2016).
42. Xiao, Y. et al. $A\beta_{1-42}$ fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease. *Nat. Struct. Mol. Biol.* **22**, 499–505 (2015).
43. Kollmer, M. et al. Cryo-EM structure and polymorphism of Aβ amyloid fibrils purified from Alzheimer's brain tissue. *Nat. Commun.* **10**, 4760 (2019).
44. Colvin, M. T. et al. Atomic resolution structure of monomorphic $A\beta_{42}$ amyloid fibrils. *J. Am. Chem. Soc.* **138**, 9663–9674 (2016).
45. Sakagashira, S. et al. Missense mutation of amylin gene (S20G) in Japanese NIDDM patients. *Diabetes* **45**, 1279–1281 (1996).
46. Seino, S. S20G mutation of the amylin gene is associated with type II diabetes in Japanese. Study Group of Comprehensive Analysis of Genetic Factors in Diabetes Mellitus. *Diabetologia* **44**, 906–909 (2001).
47. Meier, D. T. et al. The S20G substitution in hIAPP is more amyloidogenic and cytotoxic than wild-type hIAPP in mouse islets. *Diabetologia* **59**, 2166–2171 (2016).
48. Cao, P. et al. Sensitivity of amyloid formation by human islet amyloid polypeptide to mutations at residue 20. *J. Mol. Biol.* **421**, 282–295 (2012).
49. Sakagashira, S. et al. S20G mutant amylin exhibits increased in vitro amyloidogenicity and increased intracellular cytotoxicity compared to wild-type amylin. *Am. J. Pathol.* **157**, 2101–2109 (2000).
50. Ma, Z. et al. Enhanced in vitro production of amyloid-like fibrils from mutant (S20G) islet amyloid polypeptide. *Amyloid* **8**, 242–249 (2001).
51. Xu, W., Jiang, P. & Mu, Y. Conformation preorganization: effects of S20G mutation on the structure of human islet amyloid polypeptide segment. *J. Phys. Chem. B* **113**, 7308–7314 (2009).
52. Mirecka, E. A. et al. β-hairpin of islet amyloid polypeptide bound to an aggregation inhibitor. *Sci. Rep.* **6**, 33474 (2016).
53. Nilsberth, C. et al. The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Aβ protofibril formation. *Nat. Neurosci.* **4**, 887–893 (2001).

54. Padrick, S. B. & Miranker, A. D. Islet amyloid polypeptide: identification of long-range contacts and local order on the fibrillogenesis pathway. *J. Mol. Biol.* **308**, 783–794 (2001).
55. Chen, M. S. et al. Characterizing the assembly behaviors of human amylin: a perspective derived from C-terminal variants. *Chem. Commun.* **49**, 1799–1801 (2013).
56. Yonemoto, I. T., Kroon, G. J. A., Dyson, H. J., Balch, W. E. & Kelly, J. W. Amylin proprotein processing generates progressively more amyloidogenic peptides that initially sample the helical state. *Biochemistry* **47**, 9900–9910 (2008).
57. Kruger, D. F. & Gloster, M. A. Pramlintide for the treatment of insulin-requiring diabetes mellitus: rationale and review of clinical data. *Drugs* **64**, 1419–1432 (2004).
58. Roth, J. D. et al. Leptin responsiveness restored by amylin agonism in diet-induced obesity: evidence from nonclinical and clinical studies. *Proc. Natl Acad. Sci. USA* **105**, 7257–7262 (2008).
59. Wang, H., Abedini, A., Ruzsicska, B. & Raleigh, D. P. Rationally designed, nontoxic, nonamyloidogenic analogues of human islet amyloid polypeptide with improved solubility. *Biochemistry* **53**, 5876–5884 (2014).

Methods

Sample preparation. Human IAPP (H-KCNTATCATQRLANFLVHSSNNFGA ILSSTNVGSNTY-NH_2; molecular mass 3903.4 Da) with an amidated C terminus and a disulfide bond between Cys2 and Cys7 was custom synthesized (Pepscan, Lelystad). Identity and purity (93.1%) were confirmed by reverse-phase HPLC (RP-HPLC) and mass spectroscopy. RP-HPLC of a reduced sample confirmed that the disulfide bond between Cys2 and Cys7 was fully established in the non-reduced sample. To ensure monomeric starting material, the peptide was dissolved at 2 mg ml^{-1} in 1,1,1,3,3,3-hexafluoro-2-propanol at room temperature for 1 h and lyophilized. Afterward, 1 mg peptide powder was dissolved in 0.5 ml aqueous 6 M guanidine hydrochloride solution, and size-exclusion chromatography was performed on a Superdex 75 Increase 10/300 column (GE Healthcare) equilibrated with 10 mM 2-(*N*-morpholino)ethanesulfonic acid (MES)/NaOH buffer at a pH of 6.0 using an ÄKTA Purifier system (GE Healthcare). The monomeric peak fraction was collected, aliquotted, flash frozen in liquid nitrogen and stored at −80 °C for further use. The purity of the IAPP monomer fraction was 93.8% according to RP-HPLC. IAPP fibrils were prepared from the stock solution by diluting to a final concentration of 100 µM peptide with 10 mM MES/NaOH buffer (pH 6.0, 6 mM NaN_3). Fibrillation occurred by incubation within 7 d at room temperature under quiescent conditions in 1.5-ml Protein LoBind tubes (Eppendorf). As a control, we also prepared fibrils from an IAPP monomer sample of increased purity (96.9% after size-exclusion chromatography) due to an additional preparative RP-HPLC purification step preceding monomerization. All three dominant polymorphs were recovered in this sample, indicating that increasing peptide purity does not affect aggregation kinetics or thermodynamics in a way that would result in monomorphic fibrillation.

Atomic force microscopy. IAPP fibrils in 10 mM MES/NaOH buffer (pH 6.0, 6 mM NaN_3) were diluted to a peptide concentration of 10 µM monomer equivalent. Afterward, 5 µl of the fibril solution was applied to freshly cleaved muscovite mica and incubated under a humid atmosphere for 10 min. After three washing steps with 100 µl ddH_2O, the samples were dried with a stream of N_2 gas. Imaging was performed in intermittent contact mode (AC mode) in a Nano Wizard 3 atomic force microscope (JPK, Berlin) using a silicon cantilever and tip (OMCL-AC160TS-R3, Olympus) with a typical tip radius of 9 ± 2 nm, a force constant of 26 N m^{-1} and a resonance frequency of approximately 300 kHz. The images were processed using JPK data processing software (version spm-5.0.84). For the height profiles presented, a polynomial fit was subtracted from each scan line, first independently and then using limited data range.

Cryo-electron microscopy image acquisition. Cryo-EM sample preparation was performed on glow-discharged holey carbon films (Quantifoil R 1.2/1.3, 300 mesh). A 2.5-µl sample containing 100 µM IAPP in 10 mM MES/NaOH buffer (pH 6.0, 6 mM NaN_3) was applied to the carbon grid and incubated for 1 min. Subsequently, the sample was blotted for 5 s (blotting force 5) before cryo-plunging using a Vitrobot (FEI). With 110,000-fold nominal magnification, 1,330 micrographs were recorded on a Tecnai Arctica electron microscope operating at 200 kV with a field emission gun using a Falcon III (FEI) direct electron detector in electron counting mode directed by EPU data collection software (version 1.5). Each movie was composed of 50 fractions, and each fraction contained 36 frames, resulting in a total of 1,800 frames recorded per micrograph. The sample was exposed to an integrated flux of 0.9 e^- Aring^{-2} s^{-1} for 46.33 s. Applied defocus values ranged from −1 to −2.2 µm. The pixel size was calibrated to 0.935 Å as described previously[36]. Details of data acquisition are summarized in Table 1.

Cryo-electron microscopy image processing and helical reconstruction. For all polymorphs, MotionCor2 (ref. [60]) was used for movie correction, and contrast transfer function parameters were fitted with CTFFIND4 (ref. [61]). Fibrils were manually picked, and segments were extracted with an inter-box distance of 10% of the box sizes. Box sizes were chosen as 220 Å, 200 Å and 220 Å for PM1, PM2 and PM3, respectively. Further image processing, including 3D reconstructions, was performed with RELION 3.0.5 (refs. [62,63]).

For all polymorphs (PM1, PM2 and PM3), we used a noise-filled cylinder as an initial density model. Initial rounds of density refinement used the relion_refine command without the auto_refine option ($K=1$) and a T value of 20. Final refinements were conducted with a T value of 200. Gold-standard refinements were performed as described previously[35] by selecting entire fibrils and splitting the dataset accordingly into an even and an odd set. Fourier shell correlation curves were computed between two half maps. According to the 0.143 criterion, the obtained resolutions were 4.2 Å (PM1), 4.2 Å (PM2) and 8.1 Å (PM3) (Extended Data Figs. 6–8). Image processing and reconstruction details for all polymorphs are presented in Table 1.

Model building and refinement of PM1. For PM1, a single-chain atomic model was built with Coot[64,65] by placing a polyalanine model de novo into the density. The density was clearly resolved and unambiguously defined the backbone trace. After manual optimization of the protein backbone, side chains were added and rotamers were manually refined with respect to Ramachandran outliers and potential clashes. Five copies of the final single-chain model were placed into the EM density map. The final model, containing six symmetry-related monomers of IAPP PM1, was used for real-space refinement in PHENIX[66] with manually assigned β-sheets. Subsequently, the model was refined by multiple rounds of optimization in Coot, PHENIX and MDFF[67,68]. MDFF was performed using an explicit solvent. The structure was embedded in a box of water, and ions were added to the system (concentration, 1.5 M). Secondary structure, *cis*-peptide and chirality restraints were applied. The scaling factor of the map potential was set to $g=0.3$, and a time period of 10 ns was simulated. The final model of PM1 was obtained by averaging the coordinates of the MDFF trajectory and a final energy minimization with the non-crystallographic symmetry restraints and position restraints using CNS[69,70], including hydrogen atoms. *B* factors were assigned based on r.m.s.f. values calculated from the MDFF trajectory. After model evaluation using MolProbity[71], molecular graphics and further analyses were performed using Chimera[72] and ChimeraX[73]. The final statistics of the refinement are shown in Table 1.

Model building and refinement of PM2. Because of the difficulties in assigning residues to the density of PM2, two polyalanine backbones, each containing 21 residues, were built in both forward and backward trace directions in Coot[64,65]. A total of 17 possible assignments of segments from the IAPP sequence to the 21 residues were visible in the density. Accordingly, we performed 17 side chain assignments for each backbone using Scwrl4 (ref. [74]). The resulting 34 models were energy minimized with CNS[69] and refined into the density map using DireX[30]. The C_{free} value[31] is the real-space map correlation coefficient computed from the density map filtered with a bandpass of 3.0- to 4.0-Å resolution and served as a criterion to rank the models (Extended Data Fig. 3). The model that scored best according to this ranking was further refined using MDFF[67,68] with the same settings as those for PM1. Refinement was finalized by averaging the coordinates of the MDFF trajectory.

Molecular dynamics simulation. MD simulations were performed to test the stability of the PM1 model. The starting structure for the simulation was built using CHARMM-GUI solution builder[75,76] by inserting the cryo-EM structure of PM1 into a cubic water box containing 38,907 water molecules and further adding 10 chloride ions to neutralize the system. We carried out two independent all-atom simulations using GROMACS[77] (version 2019.3) and CHARMM36 force fields for protein[78], water[79] and ions[80]. The systems were first minimized using the steepest descent algorithm in 5,000 steps to remove bad contacts, followed by 500 ps (time step, 1 fs) of equilibration in an ensemble with constant volume and temperature. Later, two production runs of 250 ns were conducted under conditions of constant pressure and temperature, with a time step of 2 fs, by applying LINCS constraints to the bonds containing hydrogen atoms[81]. The temperature of the systems was maintained at 300 K using a Nosé-Hoover thermostat[82,83], and the pressure was maintained at 1 bar with a Parrinello–Rahman barostat[84]. Short-range electrostatic and van der Waals interactions were computed up to a cutoff of 12 Å using potential-shift and force-switch methods, respectively. Long-range electrostatic interactions beyond the 12 Å cutoff were computed using the particle-mesh Ewald algorithm[85].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The structure of IAPP PM1 has been deposited in the Protein Data Bank under accession code PDB 6Y1A. The cryo-EM density maps have been deposited in the Electron Microscopy Data Bank under accession codes EMD-10669 (PM1), EMD-10670 (PM2) and EMD-10671 (PM3).

References

60. Zheng, S. Q., Palovcak, E., Armache, J.-P., Cheng, Y. & Agard, D. A. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
61. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
62. He, S. & Scheres, S. H. W. Helical reconstruction in RELION. *J. Struct. Biol.* **198**, 163–176 (2017).
63. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
64. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
65. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
66. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
67. Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B. & Schulten, K. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **49**, 174–180 (2009).

68. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673–683 (2008).
69. Brunger, A. T. et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921 (1998).
70. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nat. Protoc.* **2**, 2728–2733 (2007).
71. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
72. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
73. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
74. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **77**, 778–795 (2009).
75. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).
76. Lee, J. et al. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **12**, 405–413 (2016).
77. Abraham, M. J. et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
78. Best, R. B. et al. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.* **8**, 3237–3256 (2012).
79. MacKerell, A. D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
80. Beglov, D. & Roux, B. Finite representation of an infinite bulk system: solvent boundary potential for computer simulations. *J. Chem. Phys.* **100**, 9050–9063 (1994).
81. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
82. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81**, 511–519 (1984).
83. Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A Gen. Phys.* **31**, 1695–1697 (1985).
84. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
85. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an *N*·log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

Acknowledgements

We thank P.J. Peters and C. López-Iglesias for advice and helpful discussions, H. Duimel for help with sample preparation and the M4I Division of Nanoscopy of Maastricht University for microscope access and support. The authors gratefully acknowledge the computing time granted by the Jülich Aachen Research Alliance High-Performance Computing (JARA-HPC) Vergabegremium and VSR commission on the supercomputer JURECA at Forschungszentrum Jülich. We acknowledge support from a European Research Council (ERC) Consolidator grant (no. 726368; W.H.), the Alzheimer Forschung Initiative e.V. and Alzheimer Nederland (project no. 19082CB; R.B.G.R. and G.F.S.), the Russian Science Foundation (RSF; project no. 20-64-46027; L.G. and D.W.) and the Helmholtz Association Initiative and Networking Fund (project no. ZT-I-0003; K.R.P. and G.F.S.).

Author contributions

L.G., W.H., T.K. and G.F.S. conceived the study. T.K. and L.G. performed and analyzed fibril preparation and AFM experiments. R.G.B.R. performed cryo-EM experiments and the initial data analysis. C.R., T.K. and G.F.S. performed image processing and initial reconstruction. C.R. and G.F.S. performed reconstruction, model building and refinement. L.U.S., K.R.P. and G.F.S. performed molecular dynamics simulations and structure fitting. C.R., T.K., G.F.S., W.H., L.G., K.R.P. and L.U.S. wrote the manuscript. D.W. and all other authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

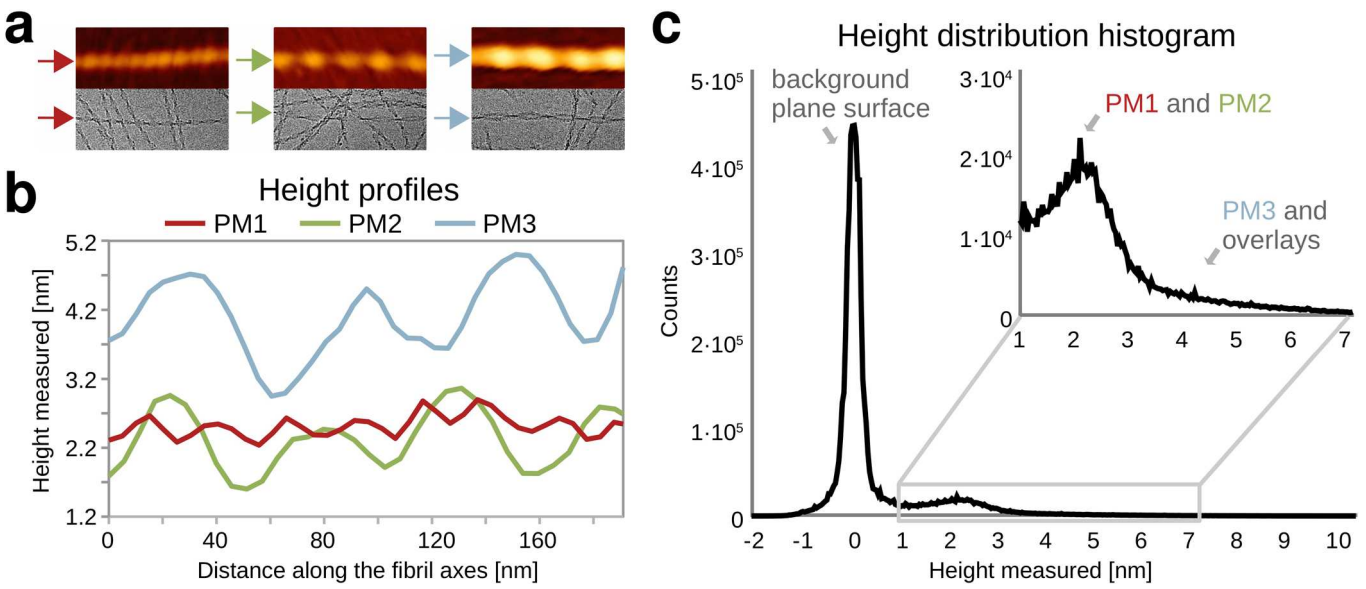
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41594-020-0442-4.

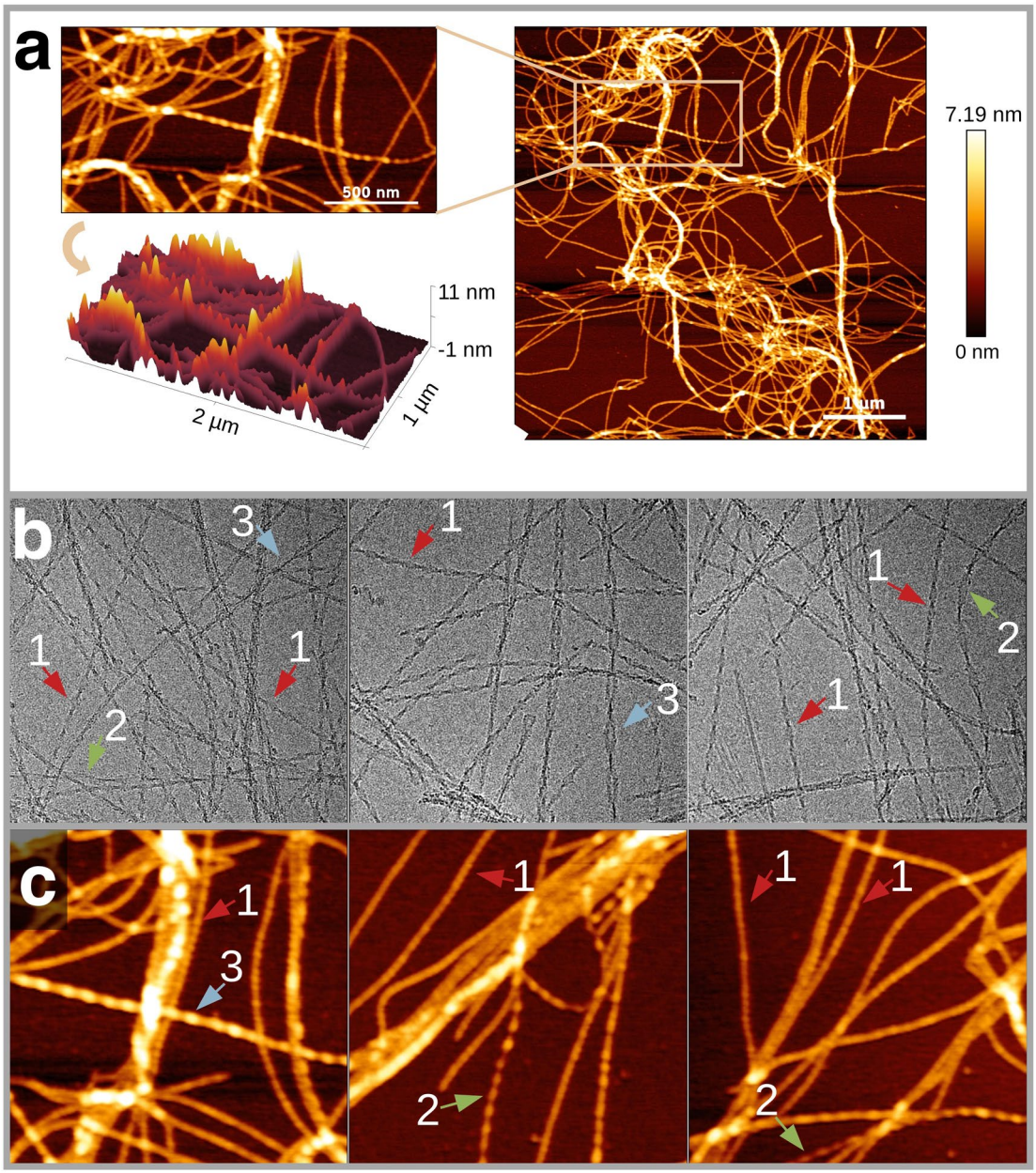
Supplementary information is available for this paper at https://doi.org/10.1038/s41594-020-0442-4.

Correspondence and requests for materials should be addressed to W.H. or G.F.S.

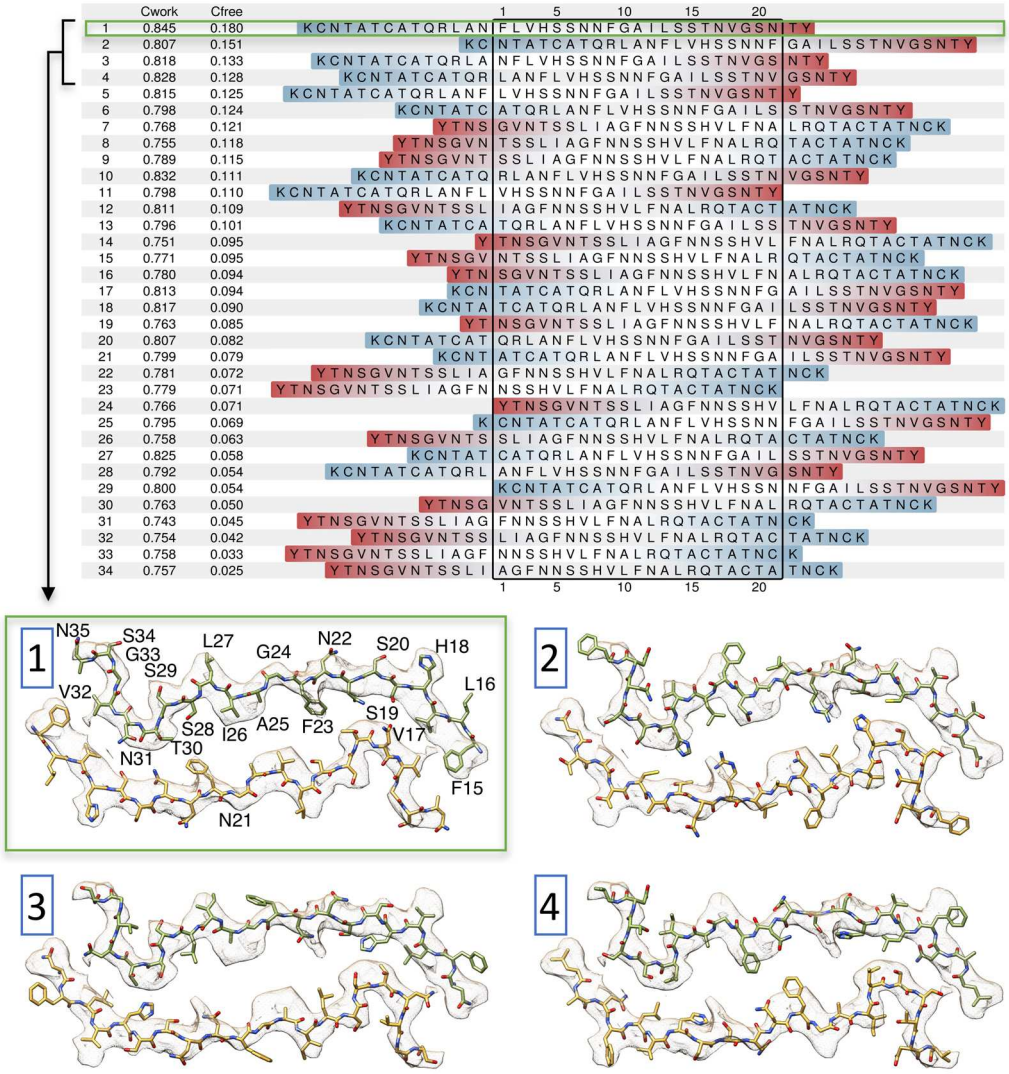
Peer review information Peer reviewer reports are available. Inês Chen was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.



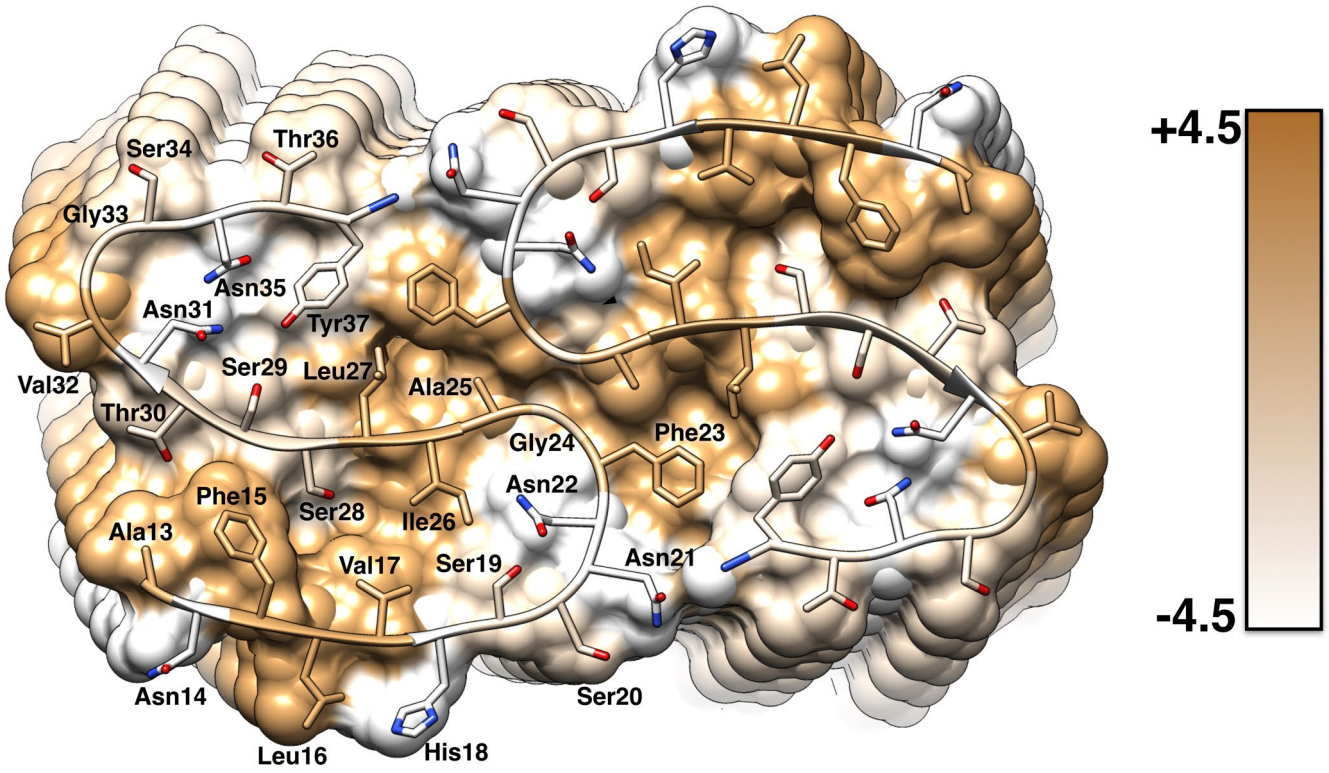
Extended Data Fig. 1 | Comparison of described IAPP polymorphs. a, Single fibril cut-outs of polymorphs PM1, PM2 and PM3 from AFM images (top row) and cryo-EM micrographs (bottom row); single box size is 100 × 250 nm. **b**, Height profiles of individual fibrils extracted from AFM images. **c,** Height distribution histogram, showing the highest number of counts for the plane background surface around 0 nm and a distinct peak around 2.2 nm. The peak around 2.2 nm includes both PM1 and PM2 which are non-distinguishable in sense of height distribution. Moreover, a pronounced shoulder on the right corresponds to the presence of lower amounts of PM3 as well as the overlaps of single PM1/PM2 fibrils. For the height distribution analysis, histograms from six height images of 5 × 5 µm size and a resolution of 1024 × 1024 pixels were obtained, binned and presented in one graph. An example of the image used can be seen in Supplementary Figure 2.



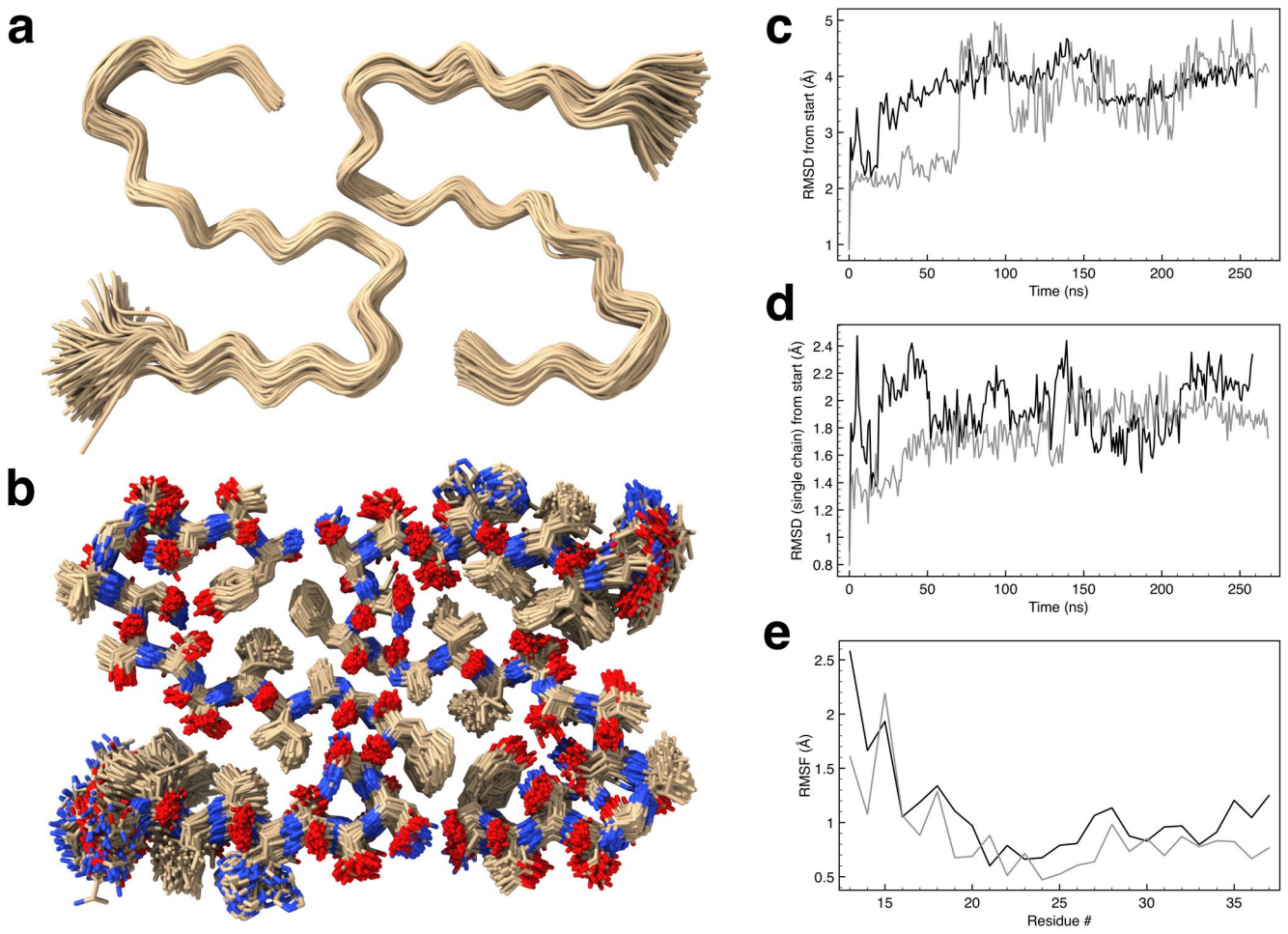
Extended Data Fig. 2 | Overview of IAPP polymorphs. a, Typical height profile AFM image used for polymorph distribution analysis. **b**, Cryo-EM micrographs showing 370 × 370 nm areas. **c**, AFM overview images showing 1 × 1 µm areas. Arrows indicate the presence of PM1 (red), PM2 (green) and PM3 (blue).



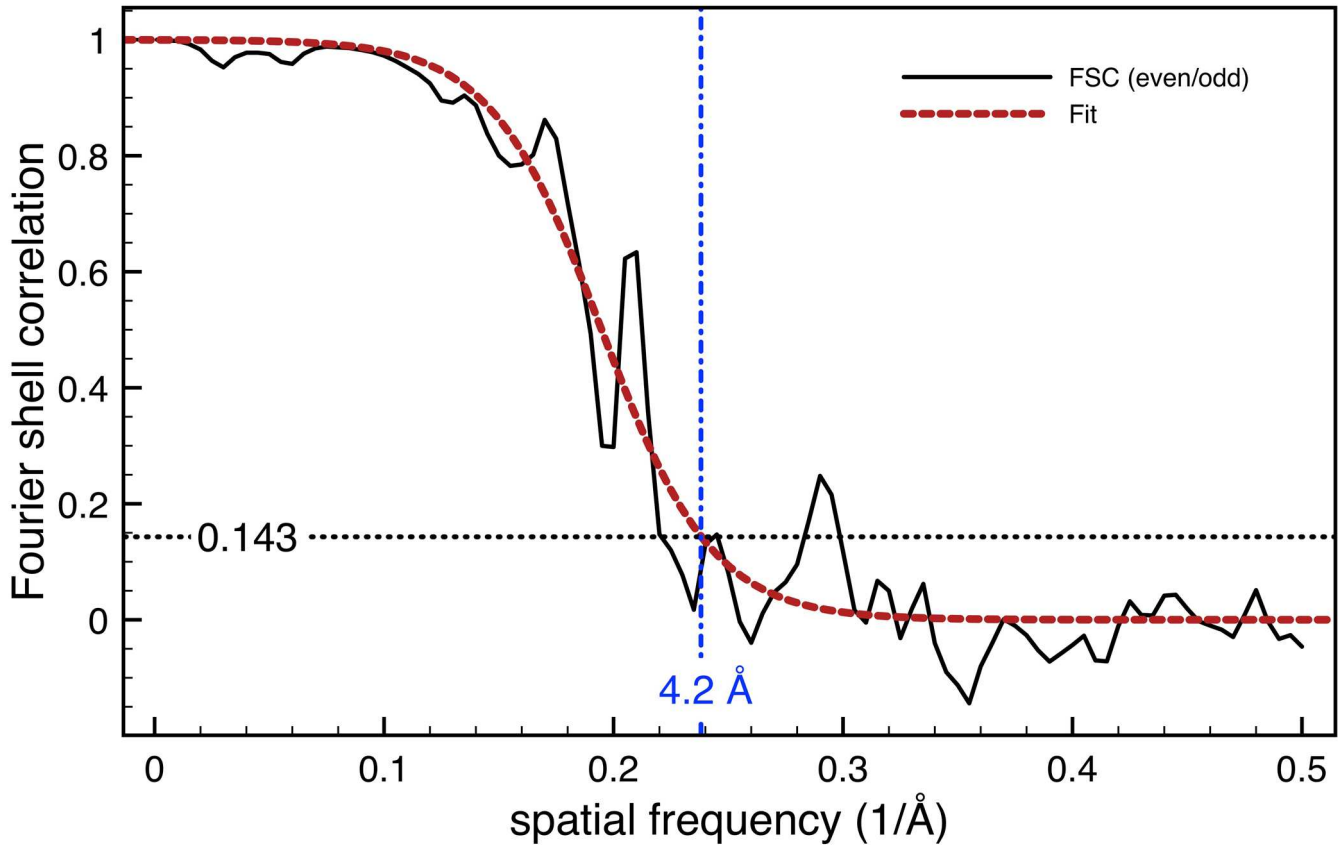
Extended Data Fig. 3 | DireX analysis of polymorph 2 (PM2). The table contains the C_{work} and C_{free} values from DireX fitting of 21-residue-long sequence snippets (black box) of IAPP in both possible Cα-chain directions into a density layer of PM2 together with the respective amino acid sequence. The results are ranked according to their C_{free} values. Highlighted (green box) is the most favorable sequence fit. Atomic models of the four most favorable sequence snippets are shown at the bottom. Note that some models, for example model 2, can be excluded since they are incompatible with the disulfide bond between residues Cys2 and Cys7.



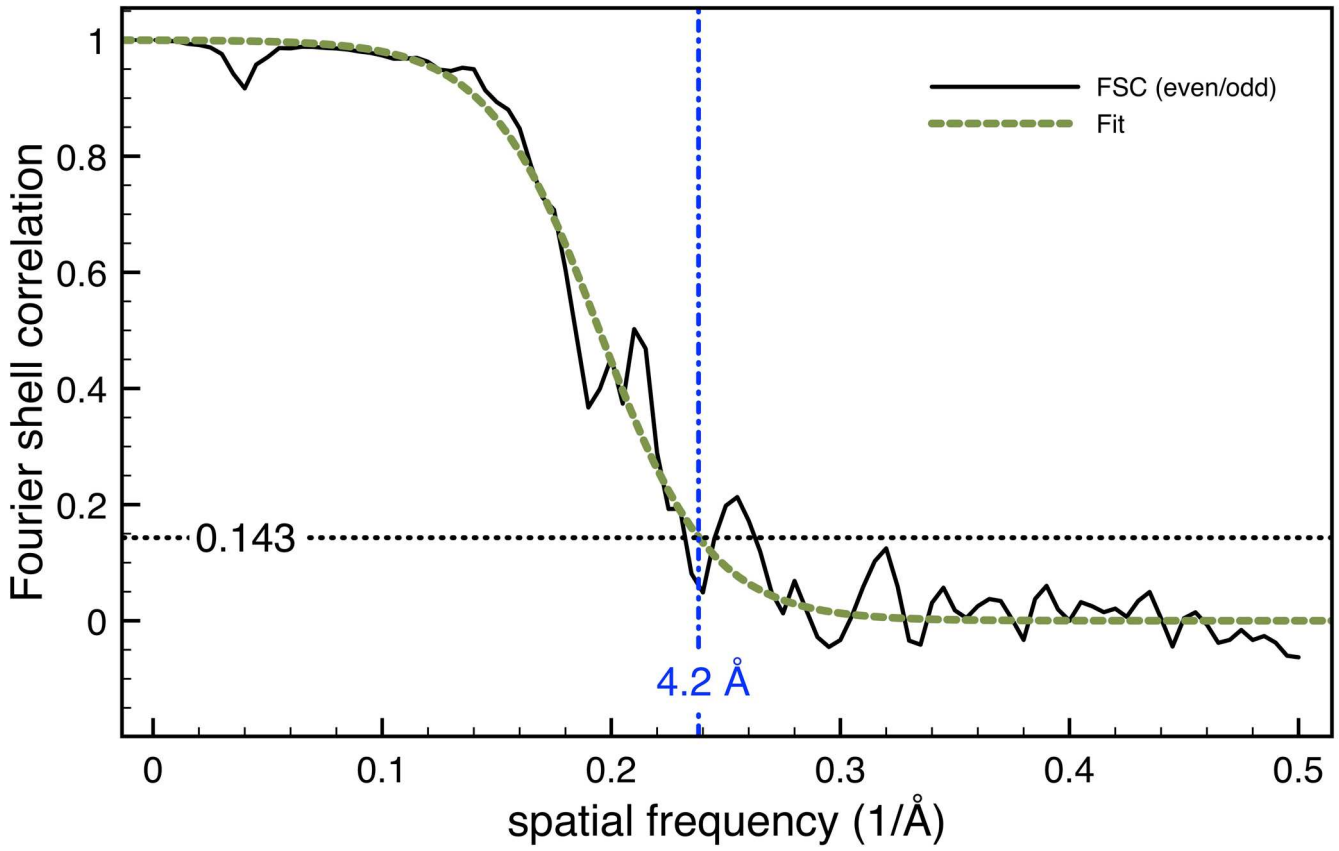
Extended Data Fig. 4 | Hydrophobicity plot of the fibril displayed as top view. Hydrophobicity levels of the IAPP polymorph 1 (PM1) fibril are colored according to Kyte-Doolittle in the hydrophobicity score range −4.5 (white) to 4.5 (gold). One hydrophobic cluster spans the entire diagonal of the fibril cross-section. This hydrophobic streak is surrounded by highly ordered polar clusters.



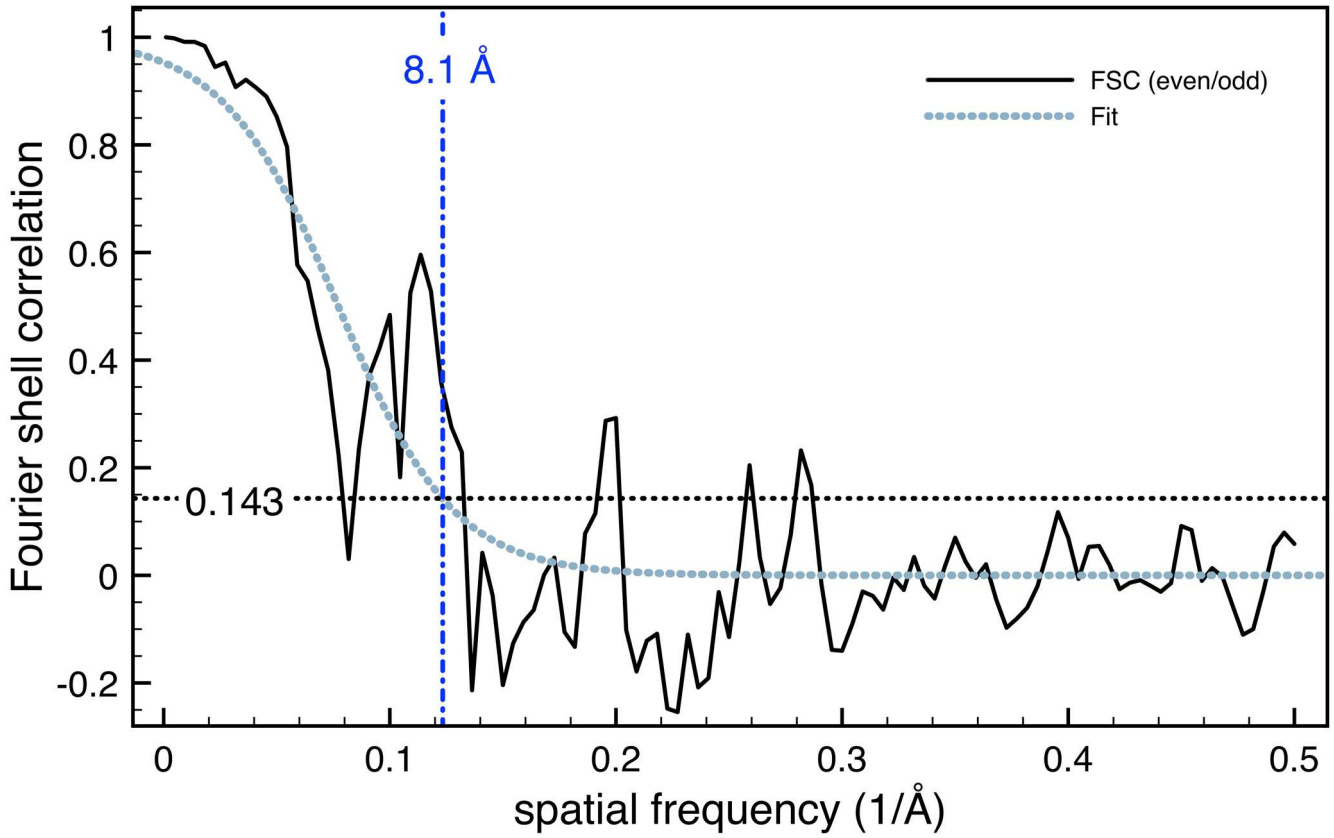
Extended Data Fig. 5 | Results of molecular dynamics simulations of IAPP polymorph 1 (PM1). Superimposed snapshots from a 250 ns simulation displaying only the backbone (**a**) or all atoms (except for solvent and hydrogen) (**b**). **c,** Showing the RMSD from the deposited structure of PM1 (PDB ID 6Y1A) for two 250 ns simulations (black and grey lines, respectively). **d,** Showing the RMSD of a single chain from the deposited structure during the two 250 ns simulations. **e,** Showing the atomic root mean square fluctuations (RMSF) for each residue calculated over each 250 ns simulation.



Extended Data Fig. 6 | FSC Analysis of polymorph 1 (PM1). FSC curves from the even/odd test (solid black) from the gold-standard refinement yields a resolution of 4.2 Å (using the 0.143 criterion). The even/odd FSC curve is fitted (red) with the model function 1/(1+exp((x-A)/B)) (with A = 0.1947 and B = 0.026) to obtain a more robust resolution estimate.



Extended Data Fig. 7 | FSC analysis of polymorph 2 (PM2). FSC curves from the even/odd test (solid black) from the gold-standard refinement yields a resolution of 4.2 Å (using the 0.143 criterion). The even/odd FSC curve is fitted (green) with the model function 1/(1+exp((x-A)/B))) (with A = 0.194789 and B = 0.02427) to obtain a more robust resolution estimate.



Extended Data Fig. 8 | FSC analysis of Polymorph 3 (PM3). FSC curves from the even/odd test (solid black) from the gold-standard refinement yields a resolution of 8.1 Å (using the 0.143 criterion). The even/odd FSC curve is fitted (light blue) with the model function 1/(1+exp((x-A)/B)) (with A = 0.0772 and B = 0.0256) to obtain a more robust resolution estimate.

nature research

Corresponding author(s): Gunnar Schroeder
Wolfgang Hoyer

Last updated by author(s): Apr 26, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	EPU 1.3;
Data analysis	CTFFIND4; MotionCor2; RELION 3.0.5; Phenix 1.11; Coot 0.8.9.1; DireX 0.7.1; Chimera 1.13;

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Atomic model of PM is available at the PDB ID 6Y1A; EM density maps for polymorphs PM1,PM2,and PM3 are available at EMDB under accession codes EMD-10669, EMD-10670,EMD-10671, respectively.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	See Supplementary Tables 2 and 3 for number of fibrils and number of fibril segments that were used in structure determination. All cryo-EM fibril images used for structure determination were obtained from one single fibril sample. To confirm that polymorphs were reproducible in terms of structure (helical pitch, diameter) and distribution, we perfomed electron microscopy and atomic force microscopy on independent IAPP fibril samples.
Data exclusions	No data were excluded.
Replication	All attempts to confirm the reproducibility fibril polymorph structure (helical pitch, diameter) and distribution were successful.
Randomization	not relevant.
Blinding	not relevant.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

C. Supplementary Material for Topology Tracing

C.1. STAR Methods

C.1.1. Step 1 Trace Initialisation

Map preparation The density map is segmented around the single-chain protein using CHIMERA. The segmented density map is normalized and filtered to 5 Å using EMAN [115]. For bead placement, an additional skeletonised version of the filtered map is generated using CHIMERA.

Bead placement We use the DIREX tool DXBEADGEN to place beads into the skeletonised density map. The number of beads are chosen equal to the number of residues in the sequence. Afterwards, DIREX is used to refine the beads into the filtered map applying repelling forces between the beads to avoid too close beads and to ensure a more homogeneous distribution of beads in the map.

Tracing The beads are connected by solving the TSP problem using the Lin-Kernighan algorithm. The LKH program requires a cost matrix, which describes the cost for each connection. To prepare the cost matrix we use a customised PATHWALKER program, which assigns a lower cost to connections through high density regions and vice versa. We generate ten temporary traces using the *–noise 0.1* option in PATHWALKER and build a histogram of connections by counting how often a certain connection appeared in the temporary traces. Based on this histogram a second cost matrix is generated, by assigning a connection cost of (*number of traces*)*100/(*connection count*). Using this histogram-based cost matrix, LKH generates a consensus trace. Finally, the trace is refined using DIREX. During the refinement distance restraints of 3.8 Å and 6 Å are applied between 1-2 and 1-3 bead pairs, respectively to impose a realistic C_α-trace geometry. The resulting trace is referred to as the conventional trace. The beads are indexed according to their order in this conventional trace.

C.1.2. Step 2 Weights Estimation

The output of step 1 is a trace that typically contains several correct subtraces but a few connections may be wrong, so that the global topology of the protein chain might be disrupted. The goal of step 2 is to obtain an initial estimate of where the correct subtraces are and in which order they need to be connected. For this, we compare the distance matrix of the conventional trace from step 1 with the distance matrix defined by predicted inter-residue distances. This comparison will lead to an estimate of the bead-to-sequence assignment.

Prediction of inter-residue distances We use trRosetta to predict the inter-residue distances. No templates were used for the prediction. The output of trRosetta is given in the following format: The distance range between 2 Å and 20 Å is subdivided into 36 bins of 0.5 Å width. For each residue pair a probability distribution is given where the value in each bin corresponds to the predicted probability that the distance between the two residues is in the corresponding distance range of that bin. We convert this distribution into a real-valued distance by only regarding the bin with the highest probability, which yields the predicted distance matrix, $\mathbf{d}_{\text{res}}$.

Calculation of inter-bead distances We calculate the distance matrix, $\mathbf{D}_{\text{bead}}$, for the conventional trace and adapt the values of these distances to match the values of the predicted distances, which are given in bins as described above. This is achieved by ranking the entries of both distance matrices by size and replacing the ranked values of the bead distance matrix with the corresponding ranked values from the predicted distance matrix.

Subdividing distance maps into matrix representations of subtraces The conventional trace is subdivided into subtraces; odd subtrace lengths in the range of $n = 3$ beads to $n = 0.9 \cdot N_{res}$ of the full trace are considered. For all subtraces of length n, we compute a $n \times n$ distance matrix $\hat{\mathbf{D}}_{\text{bead}}$ and a $n \times (N_{\text{bead}} - n)$ distance profile $\hat{\mathbf{P}}_{\text{bead}}$. The distance profile stores for each bead in the subtrace the distribution of distances to all beads that are not part of the subtrace. Analogously in sequence space, we extract the $n \times n$ distance matrix, $\hat{\mathbf{d}}_{\text{res}}$, and the corresponding $n \times (N_{\text{res}} - n)$ distance profile, $\hat{\mathbf{p}}_{\text{res}}$, from the predicted distances for all subsets of n adjacent residues. Similarly, the distance profile stores for each residue in the subset the distribution of predicted distances to all residues that are not part of the subset.

Comparing distance patterns Having matrix representations for subtraces of beads and subsets of residues at hand, we compare both by calculating a similarity

score between them:

$$S_{\text{sim}}^{\text{forward}} = 0.5\sqrt{\sum(\hat{\mathbf{D}}_{\text{bead}} - \hat{\mathbf{d}}_{\text{res}})^2} + 0.5\sqrt{\sum(\hat{\mathbf{P}}_{\text{bead}} - \hat{\mathbf{p}}_{\text{res}})^2}$$

$$S_{\text{sim}}^{\text{backward}} = 0.5\sqrt{\sum(\hat{\mathbf{D}}_{\text{bead}}^{rev} - \hat{\mathbf{d}}_{\text{res}})^2} + 0.5\sqrt{\sum(\hat{\mathbf{P}}_{\text{bead}}^{rev} - \hat{\mathbf{p}}_{\text{res}})^2}$$

$$S_{\text{sim}} = min(S_{\text{sim}}^{\text{forward}}, S_{\text{sim}}^{\text{backward}})$$

Since the correct direction of a subtrace is not known, we need to consider both directions for the comparison. For this, the similarity score is computed for the forward ($S_{\text{sim}}^{\text{forward}}$) and backward ($S_{\text{sim}}^{\text{backward}}$) direction. The distance matrix and distance profile for the backward direction is denoted d^{rev} and p^{rev}, respectively. The smaller of the two values $S_{\text{sim}}^{\text{forward}}$ and $S_{\text{sim}}^{\text{backward}}$ is taken as the similarity score. For each subtrace associated with its central bead b_c, the best match to a subset of adjacent residues associated with the central residue r_c is identified by finding the minimal similarity score S_{sim}. Note that for each central bead, there are multiple subtraces of different lengths, as described above. We define the bead-to-sequence assignment matrix $\mathbf{w}$ and set all entries to zero at the beginning. While testing all possible assignments of a subtrace to all subsets, the corresponding weight w_{b_c,r_c} for the best match is increased by 1. An entry w_{ki} of the assignment matrix then describes the likelihood that bead k is assigned to amino acid sequence position i.

C.1.3. Step 3 Weights Optimisation

Given the assignment matrix, $\mathbf{w}$, from step 2, the goal of step 3 is to optimise the weights such that beads are assigned to residues in a way that the difference between the predicted distance matrix and the distance matrix of the resulting trace is minimised.

Score Minimization The entrys of the $N_{\text{bead}} \times N_{\text{res}}$ assignment matrix $\mathbf{w}$ is first normalized within each row and the again normalized within each column. The following scoring function is then minimised with respect to the weights, w_{ki}, via a gradient descent:

$$S = -\sum_{ijkl} \frac{w_{ki} w_{lj}}{1 + (d_{ij} - D_{kl})^2}$$

Here w_{ki} denotes entries of the assignment matrix while D_{kl} and d_{ij} describe distances between beads k and l and predicted distances between residue i and j, respectively.

Postprocessing of the trace After optimisation, beads are assigned to residues according to their maximal weight. The resulting trace is further processed by removing outliers. A bead is considered an outlier if its distance to one of its neighbours within the trace is larger than 10 Å. A bead is also considered an outlier if it is not part of a subtrace. A subtrace is defined as a group of beads that were assigned to adjacent residues and that have an index difference not larger than 10; this means a subtrace appears as a (fuzzy) diagonal in the assignment matrix.

Runtime As the scoring function is based on an entry-wise comparison of two distance matrices, the optimisation scales with $(N_{\text{res}})^4$. For proteins with more than about 300 residues, the minimum is not found within a reasonable time using the current implementation. Runtime, using ten nodes with 24 Intel Xeon 2.1GHz cores each, for the smallest with 85 residues and largest test case with 229 residues was 9 minutes and 4 hours and 43 minutes, respectively.

C.2. Trials, Errors and Perspectives

It was a long journey to develop the presented method for backbone topology tracing guided by predicted inter-residue distances. Many ideas and approaches were tested during the development, but did not make it into the final version due to insufficient performance and unsatisfactory results. Some of them are sketched in the following chapter.

C.2.1. Guiding the Assignment

During the development of the method several approaches were tested which could guide the assignment of beads to residues. The idea was, that integrating more prior information into the assignment could facilitate the score minimisation and improve runtime as well as accuracy. However, none of the presented ideas showed the desired effect and are therefore not included in the manuscript.

Confidence of predicted distances TrRosetta predicts not a single distance for a pair of residues, but a probability distribution describing the probability that the distance between the residue pair falls in one of 36 different bins. We reduce those distributions to their maximum value, i.e. the bin with the highest probability. However, this simplification might be misleading for residue pairs associated with a rather flat probability distribution. For those distances the reduced single value prediction is not as reliable as for pairs with a distinct maximum in the predicted probability distribution. During the score minimisation distances

with lower reliability should have less impact on the assignment than distances with higher reliability. Therefore, we weighted each distance prediction with its associated probability.

Information about Secondary Structure Secondary structure elements are somehow cornerstones of protein topology. Identifying α helices and β sheets in the sequence as well as in the map can offer helpful restraints to guide the assignment. We performed secondary structure prediction on the sequence level with PsiPred [116] and used Haruspex [117] for secondary structure identification in the density map. In the context of assigning beads to residues, information derived from secondary structure predictions can then be included as the rational that beads in a map region identified as possible helical region or β-sheet region are more likely to correspond to a residue with high probability to be in a helix or in a sheet. Therefore, we weighted combinations of beads and residues with matching secondary structure predictions higher than other combinations.

Direct Neighbour Relations One key idea behind the presented method is, that the problem of finding the correct path or trace of the main-chain through the density map, the tracing problem, has the same solution as the problem of assigning beads to residues. Solving one, immediately solves the second. One can even interpret the tracing problem as assignment problem where each bead is assigned to its direct neighbour that precedes it in the trace. In that sense, relations between direct neighbours may also play a key role in the assignment of beads to residues.

There are several aspects to consider:

If two beads are assigned to neighbouring residues, there should be density along that connection. While connections traversing high-density regions are preferred in the generation of the conventional trace, they are not in the assignment. However, we tested to do that, by associating combinations where a pair of beads, which are in spatial proximity to each other and connected by high-density regions, were assigned to a pair of neighbouring residues with a higher weight in the assignment score.

Further, if a bead is assigned to a certain residue with high probability, the likelihood that beads in close spatial proximity should be assigned to neighbouring residues increases. We implemented this situation by smoothing weights of spatial neighbouring beads once the probability of the assignment of one bead was higher than a threshold.

The distance between two neighbouring C_α atoms is well defined and should be approximately 3.8 Å, residue i and $i + 2$ should have a distance of $\approx$ 5.6 Å. In the method described in the manuscript, we adapt the predicted distance matrix of residues and the calculated distance matrix of beads by transferring values

(given in bins) of the predicted matrix to the bead matrix. To consider the well-defined distances between direct neighbours in the assignment, we tested also to act the other way round, and to transfer values from the bead distances (in Å) to the predicted inter-residue distances, but used the well-defined distance values of 3.8 Å and 5.6 Å for neighbouring residues.

We also tried to include the idea of tracing as assigning neighbours on a more fundamental level and implemented a three-dimensional assignment, where each bead is assigned to a residue and a neighbouring bead with the help of Google's OR-tools.

Number of Beads In the presented method we place as many beads in the density as there are residues in the sequence. However, some of the beads may be located unfavourable such that the corresponding C_α positions may not be represented with sufficient accuracy. To increase the probability to place a bead at each C_α position, we increased the number of beads up to $5n_{seq}$ where n_{seq} is the number of residues in the sequence. The assignment of beads to residues is then not a one-to-one matching, but there are many beads that are not assigned to any residue. The assignment weights then no longer add up to 1 in bead direction (but still in the residue direction). The size of the distance matrix of beads increases and with it the run-time. Moreover, the comparison of subtraces of beads with geometric patterns found in the predicted distances is not straightforward anymore, such that the estimation of start weights is difficult. Due to these conditions, the assignment did not converge in a reasonable time.

Iterative Weight Adaption Another idea was to perform the LKH tracing and the estimation of the assignment weights in an iterative manner. So, we built the conventional trace using LKH, performed the assignment from beads to residues, transformed the assignment weights to connection weights, and fed the connection weights into the LKH solver to perform the tracing again.

C.2.2. Optimising Runtime

A short runtime is always desirable. Particularly for larger proteins runtime can become a problem, such that no satisfactory solution can be found in a reasonable time. We tested several methods to improve runtime, to firstly make the usage of our program more comfortable and secondly to enable the processing of larger proteins.

ADAM optimiser We implemented a gradient descent method to optimise our score and find the best assignment of beads to residues. A more efficient gradient

based minimisation method, which is also broadly used in the context of machine learning is the ADAM optimiser [118]. Adam is a stochastic gradient descent method which applies adaptive learning rates. In the hope to improve runtime we implemented the Adam algorithm instead of the conventional gradient descent. However, we did not yield satisfactory results.

Sparse Matrices During the score minimisation we iterate over the distances between beads and the predicted inter-residue distances. Therefore, the runtime of the minimisation scales with n^2, where n is the number of beads and residues. However, not all distances offer the same amount of topology information. Many entries of the distance maps have the value (bin index) 37 indicating that two residues or beads are not in contact. Because of their number they are not very specific for a bead or residue and therefore less valuable to identify a bead and residue pair with matching distance patterns than smaller distances. Hence, we tested to regard only distances < 37 for the assignment. However, the win in runtime was less than expected and the quality of results suffered.

C.2.3. Applying Topology Tracing

Besides more technical aspects as described before, also ideas about further possible applications of the method come to mind during its development.

Identification of Chains in Multi Chain Complexes Physiological functions of proteins are often not performed by a single chain protein alone, but by protein complexes consisting of multiple protein chains. A fist step of structural interpretation of cryo-EM maps of protein complexes is often the identification which region of the density map corresponds to which protein chain. A rough idea was, that our method could help to do that. For a complex of m chains one would have m predicted distance matrices for inter-residue distances. Placing as many beads as there are residues in all chains in total in the density map and calculating all bead distances would give a large distance matrix for the beads. The idea is then, to compare distance patterns found in sub-groups of beads with distance patterns in the predicted distances and to identify which beads best resemble which protein chain. The question is, how to define those sub-groups of beads. Without prior map segmentation there is no hint which beads would correspond to the same chain.

On-the-fly Modification of Traces in Chimera Our method for topology tracing can not only be used to build a trace but also to assess traces. For that purpose, a .bild file is generated that visualises possible wrong connections. In a next step, we

would like to provide a CHIMERA plug-in which allows to modify the trace while reviewing it. The idea is to open the trace and the .bild file in CHIMERA, review the possible wrong connections and decide via clicking if one wants to accept the change or keep the original geometry. The new trace can then be saved separately.