



Ensemble-Based Framework for Analyzing Dynamically Dominated Allostery

Inaugural-Dissertation

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

vorgelegt von
Christopher Pfleger
aus Frankfurt-Höchst

Düsseldorf, November 2014

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

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

Referent: Prof. Dr. Holger Gohlke

Korreferent: Prof. Dr. Martin Lercher

Tag der mündlichen Prüfung:

Eidesstattliche 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.

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

Düsseldorf, im November 2014

To my parents

“Everything that is living can be understood in terms of the jiggling and wiggling of atoms.”

Richard Feynman

TABLE OF CONTENTS

TABLE OF CONTENTS.....	i
LIST OF PUBLICATIONS.....	iii
ABBREVIATIONS	iv
ZUSAMMENFASSUNG	vi
ABSTRACT	viii
1 INTRODUCTION.....	1
2 BACKGROUND.....	4
2.1 Cooperativity in allosteric regulation.....	4
2.2 Classical view of allosteric regulation involves conformational changes.....	8
2.3 Allosteric regulation in the absence of conformational changes.....	10
2.4 Allosteric signaling via pathways	12
2.5 Allosteric targets in drug discovery	13
2.6 Allosteric regulation deduced from experiments	16
2.6.1 X-ray crystallography	16
2.6.2 NMR experiments	17
2.6.3 Fluorescence resonance energy transfer	18
2.7 Allosteric regulation deduced from computational approaches.....	19
2.7.1 Molecular dynamics simulations	20
2.7.2 Elastic network models	22
2.7.3 Statistical coupling analysis.....	23
2.7.4 Network analysis.....	23
2.8 Rigidity theory and analysis.....	24
2.8.1 Introduction to rigidity theory.....	24
2.8.2 Modeling biomacromolecules as constraint networks.....	27
2.8.3 Distance constraint model.....	28
2.8.4 Allostery deduced from rigidity analysis.....	30
3 SCOPE OF THE THESIS.....	32
4 PUBLICATION I - Global and Local Indices for Characterizing Biomolecular Flexibility and Rigidity	34
4.1 Background	34
4.2 Simulating the thermal unfolding.....	35
4.3 Global and local indices for characterizing biomacromolecular stability.....	35
4.4 Conclusion and significance.....	38
5 PUBLICATION II - Efficient and Robust Analysis of Biomacromolecular Flexibility Using Ensembles of Network Topologies Based on Fuzzy Noncovalent Constraints.....	40
5.1 Background	40
5.2 Theory	40
5.3 Validation of fuzzy noncovalent constraints.....	42
5.4 Conclusion and significance.....	44

6	PUBLICATION III - Constraint Network Analysis (CNA): A Python Software Package for Efficiently Linking Biomacromolecular Structure, Flexibility, (Thermo)Stability, and Function	45
6.1	Background	45
6.2	PyFIRST as an Interface	45
6.3	Workflow of the CNA software	46
6.4	Showcase example: Flexibility characteristics of HEWL	47
6.5	Conclusions and significance	47
7	PUBLICATION IV – Allosteric Coupling deduced from Altered Rigidity Percolation in Biomacromolecules.....	49
7.1	Background	49
7.2	Overall strategy	49
7.3	<i>In silico</i> mutational perturbation of eglin c agrees with experimental findings... ..	50
7.4	Probing V- and K-type allosteric mechanisms in PTP1B and LFA-1	52
7.5	Negative cooperativity in LFA-1 deduced from rigidity analyses	53
7.6	Conclusion and significance.....	54
8	PUBLICATION V - Pocket-Space Maps To Identify Novel Binding-Site Conformations in Proteins.....	55
8.1	Background	55
8.2	Workflow of the PocketAnalyzer ^{PCA} approach.....	55
8.3	PocketAnalyzer ^{PCA} applied to aldose reductase and neuraminidase.....	56
8.4	Conclusion and significance.....	58
9	PUBLICATION VI - Hot Spots and Transient Pockets: Predicting the Determinants of Small-Molecule Binding to a Protein-Protein Interface	59
9.1	Background	59
9.2	Overall strategy	59
9.3	Identification of transient pockets	60
9.4	Conclusion and significance.....	61
10	SUMMARY.....	63
11	PERSPECTIVES.....	65
12	ACKNOWLEDGEMENT.....	68
13	PUBLICATIONS	69
13.1	Reprint permissions for publications.....	69
13.2	Publication I	70
13.3	Publication I – Supporting Information	85
13.4	Publication II	93
13.5	Publication II – Supporting Information	104
13.6	Publication III.....	118
13.7	Publication IV	128
13.8	Publication IV – Supporting Information.....	154
13.9	Publication V.....	160
13.10	Publication V – Supporting Information.....	175
13.11	Publication VI	194
13.12	Publication VI – Supporting Information.....	209
14	CURRICULUM VITAE	231
15	REFERENCES.....	233

LIST OF PUBLICATIONS

This thesis is based on the following publications:

Craig, I.R., Pfleger, C., Gohlke, H., Essex, J.W., Spiegel, K.; *Pocket-Space Maps To Identify Novel Binding-Site Conformations in Proteins*. **J. Chem. Inf. Model.** (2011), 51, 2666–2679.

Impact factor reported in 2010: 3.822

§Metz, A., §Pfleger, C., Kopitz, H., Pfeiffer-Marek, S., Baringhaus, K.-H., Gohlke, H.; *Hot Spots and Transient Pockets: Predicting the Determinants of Small-Molecule Binding to a Protein-Protein Interface*. **J. Chem. Inf. Model.** (2012), 52, 120-133.

Impact factor reported in 2011: 4.675

Pfleger, C., Radestock, S., Schmidt, E., Gohlke, H.; *Global and Local Indices for Characterizing Biomolecular Flexibility and Rigidity*. **J. Comput. Chem.** (2013), 34, 220-233.

Impact factor reported in 2012: 3.835

§Pfleger, C., §Rathi, P.C., Klein, D., Radestock, S., Gohlke, H.; *Constraint Network Analysis (CNA): A Python Software Package for Efficiently Linking Biomacromolecular Structure, Flexibility, (Thermo)Stability, and Function*. **J. Chem. Inf. Model.** (2013), 53, 1007-1015.

Impact factor reported in 2012: 4.304

Pfleger, C., Gohlke, H.; *Efficient and Robust Analysis of Biomacromolecular Flexibility Using Ensembles of Network Topologies Based on Fuzzy Noncovalent Constraints*. **Structure** (2013), 21, 1-10.

Impact factor reported in 2012: 5.994

Pfleger, C., Minges, A.R.M., Gohlke, H.; *Allosteric Coupling deduced from Altered Rigidity Percolation in Biomacromolecules*. **Submitted (2014)**.

§ Both authors contributed equally to this work.

ABBREVIATIONS

ALR	aldolase reductase
ATP	adenosine triphosphate
AMP	adenosine monophosphate
cAMP	cyclic adenosine triphosphate
AMPA	α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor
BQM	3-({4-[(1 <i>E</i>)-3-morpholin-4-yl-3-oxoprop-1-en-1-yl]-2,3-bis(trifluoromethyl)phenyl}sulfanyl)aniline
CaM	calmodulin
CAP	catabolite activator protein
CheY	chemotaxis protein
CNA	constraint network analysis
CO ₂	carbon dioxid
CS	citrate synthase
DCM	distance constraint model
DHFR	dihydrofolate reductase
DNA	deoxyribonucleic acid
ENM	elastic network models
ENT	ensembles of network topologies
FIRST	floppy inclusion and rigidity substructure topology
FNC	fuzzy non-covalent constraints
FRET	fluorescence resonance energy transfer
FRODA	framework rigidity optimized dynamic algorithm
GABA	γ -aminobutyric acid
GluR2	glutamate receptor subunit 2
GPCR	G protein–coupled receptor
HCN2	hyperpolarization-activated cyclic nucleotide–regulated ion channel
HEWL	hen egg-white lysozyme
Hsp70	70 kDa heat shock protein
ICAM	intercellular adhesion molecule
IL-2	interleukin-2
IL-2R α	interleukin-2 receptor subunit α
IPMS	α -isopropylmalate synthase
KNF	Koshland, Némethy, and Filmer
KSP	kinesin spindel protein
LFA-1	lymphocyte function-associated antigen 1

MD	molecular dynamics
MM-PBSA	molecular mechanics Poisson-Boltzmann surface area
MPM	mechanical perturbation method
MSA	multiple sequence alignments
MWC	Monod, Wyman, and Changeux
NA	neuraminidase
NMR	nuclear magnetic resonance
NSR	nuclear spin relaxation
O ₂	oxygen
PCA	principal component analysis
PDB	protein data bank
PEP	phosphoenolpyruvate
PFK-1	phosphofructokinase 1
PK	pyruvate kinase
PPI	protein-protein interaction
PPIM	protein-protein interaction modulator
pO ₂	partial pressure of oxygen
PRS	perturbation-response scanning
PTI	pancreatic trypsin inhibitor
PTP	protein-tyrosine phosphatase
PTP1B	protein-tyrosine phosphatase 1B
RDC	residual dipolar coupling
RXR	retinoid X receptor
SCA	statistical coupling analysis
SWIG	simplified wrapper and interface generator
TBDT	TonB-dependent transporters
TCPTP	T-cell protein tyrosine phosphatase
VS	virtual screening

ZUSAMMENFASSUNG

Die Vorhersage von allosterischer Regulation in Biomolekülen ist von besonderem Interesse in der modernen Wirkstoffentwicklung, z. B. besitzen allosterische Effektoren einen Selektivitätsvorteil gegenüber „konventionellen“ kompetitiven Liganden. Allerdings stellen die verschiedenen Mechanismen der allosterischen Regulation eine große Herausforderung für die Identifikation neuer allosterischer Bindestellen dar. Die klassische Sichtweise der allosterischen Regulation beinhaltet das Auftreten einer Konformationsänderung wohingegen in der neuen Sichtweise die Änderung der Proteindynamik hinzugefügt wurde. Insbesondere, der letztgenannte Fall ist aufgrund fehlender Konformationsänderungen herausfordernd und anhand statischer Kristallstrukturen schwer zu identifizieren. Bis zum heutigen Zeitpunkt existiert keine Methode für die routinemäßige Untersuchung solcher Dynamik-gesteuerter Allostery hinsichtlich Vorhersage von Freien Energien für die Kooperativität und Beschreibung der Signalweiterleitung zwischen entfernten Stellen.

In dieser Arbeit habe ich einen Ensemble-basierten Störungsansatz entwickelt, um Freie Energien für die Kooperativität und allosterische Signalweiterleitung in Biomolekülen vorherzusagen. Der Kern dieses Ansatzes ist eine Graphen-basierter Methode für die effiziente Identifikation von flexiblen und rigiden Regionen in Biomolekülen. Für eine sinnvolle Verknüpfung von solchen Stabilitätsanalysen mit biologisch relevanten Informationen, habe ich zunächst Richtlinien für die Anwendung und Interpretation von bekannten und neuen Index-Definitionen ausgearbeitet (**Publikation I**). Ein Nachteil von Graphen-basierten Stabilitätsanalysen ist ihre hohe Empfindlichkeit hinsichtlich der verwendeten Eingabestruktur. Um die Empfindlichkeit zu reduzieren, wurden bisher Strukturenssembles aus rechenintensiven MD Simulationen verwendet. Als eine Alternative habe ich eine effiziente Methode entwickelt, die die thermische Fluktuationen in Biomolekülen simuliert (**Publikation II**). Dies führte zu einer Weiterentwicklung des *Constraint Network Analysis* (CNA) Verfahrens hinsichtlich einer effizienteren Ensemble-basierten Stabilitätsanalyse. Zusätzlich beinhaltet die Weiterentwicklung eine automatisierte Berechnung von globalen und lokalen Indices sowie eine robuste Prozedur für die korrekte Netzwerkdarstellung von Ligand-Molekülen. Diese Arbeiten führten schließlich zu der Entwicklung des benutzerfreundlichen Programmes CNA (**Publikation III**).

Das Programm CNA wurde für die Untersuchung von Stabilitätsänderungen durch Einfügen von *in silico* Störungen in Interaktionsnetzwerken verwendet (**Publikation IV**). Die Validierungsstudie an Eglin c zeigte eine sehr gute Korrelation bezüglich vorhergesagten und

experimentellen freien Energien mutationsbedingter Stabilitätsänderungen sowie eine hohe Übereinstimmung der Pfadvorhersage mit Dynamikänderungen von Resten in einer NMR Studie. Anschließend wurde der Störungsansatz auf Protein Tyrosin Phosphatase 1B (PTP1B) und Leukozyten Funktionsassoziierte Antigen-1 (LFA-1) angewendet. Für beide Systeme wurden anhand von *in silico* Störungen in der allosterischen Bindestelle langreichweitige Stabilitätsänderungen gefunden, die allosterische und orthosterische Stelle verbinden. Die Berechnung der Freien Energie für die Kooperativität in LFA-1 zeigte, dass eine nicht-additive Stabilisierung durch das Binden des allosterischen Effektors und des natürlichen Substrats vorliegt. Diese nicht-additive Stabilisierung entspricht der zugrundeliegenden negativen Kooperativität, die experimentell für LFA-1 gefunden wurde.

Für zukünftige Projekte bietet dieser Störungsansatz eine ausgezeichnete Möglichkeit für die effiziente Untersuchung von Systemen mit unbekanntem allosterischen Mechanismus. Zunächst werden Biomoleküle mittels PocketAnalyzer (**Publikation V**) nach neuen Bindestellen untersucht. Die Eignung von PocketAnalyzer wurde in einer Studie für die Identifikation von transienten Bindetaschen in der Protein-Protein Interaktionsfläche von Interleukin-2 gezeigt (**Publikation VI**). Anschließend werden identifizierte Bindetaschen hinsichtlich ihrer allosterischen Funktion überprüft, z. B. durch gezieltes Einfügen von Störungen im Interaktionsnetzwerk der Bindetaschen. Die Änderung der thermischen Fluktuation kann anschließend mittels einer modifizierten Variante des Ansatzes aus Publikation II untersucht werden. Die Informationen über die chemischen Eigenschaften potentieller allosterischer Bindestaschen kann abschließend in einem Virtual Screening Ansatz verwendet werden, um Leitstrukturen für neuartige allosterische Effektoren zu identifizieren.

ABSTRACT

Understanding allosteric regulation in biomacromolecules is of great interest of current drug design efforts as they provide many opportunities to overcome, e.g. selectivity problems of conventional orthosteric site ligands. However, different mechanisms for allosteric regulation, i.e. conformational change versus changes in dynamics, complicate the identification of novel allosteric sites. In particular, the latter mechanism does not require a conformational change, and hence, is difficult to deduce from static X-ray structures alone. Until now, no computational method is available to analyze this type of dynamically dominated allostery in terms of free energies of cooperativity and how information is transmitted between allosteric and orthosteric site.

In this thesis, I developed a novel ensemble-based perturbation approach for analyzing free energies of cooperativity and allosteric signaling. The core of the approach is a graph-theory based network representation that efficiently identifies flexible and rigid regions in biomacromolecules. To obtain the maximal advantage of such a rigidity analysis, I presented guidelines for the usage of available and new indices and how to link them with biological relevant information (**publication I**). Because, rigidity analyses are very sensitive to the structural input, which has so far been overcome by analyzing ensembles from computationally costly MD simulations, I developed an efficient method for simulating the thermal dynamics of biomacromolecules (**publication II**). Hence, I improved the efficiency of an ensemble-based variant of the *Constraint Network Analysis* (CNA) approach, leading to the CNA software package (**publication III**). This improvement also includes an automated calculation of global and local indices and a robust procedure for the network construction of ligand molecules.

Next, I used the CNA software to analyze changes in stability upon *in silico* perturbations on ensembles of network topologies (**publication IV**). The validation study of mutational perturbations in eglin c showed a good correlation between predicted free energies and free energies for stability changes from experiments. In addition, the perturbation approach almost perfectly reproduces a continuous pathway of dynamically coupled residues as found in NMR experiments. The approach was then applied on protein tyrosine phosphatase 1B (PTP1B) and lymphocyte function-associated antigen 1 (LFA-1). Upon *in silico* perturbation, altered rigidity characteristics revealed long-range effects in both systems. Remarkably, condensed clusters of residues has been identified that form continuous pathways connecting allosteric and orthosteric sites in both systems. Finally, the predicted free energy of cooperativity

between allosteric and orthosteric site of LFA-1 revealed a non-additive stabilization in agreement with the underlying mechanism of negative cooperativity in LFA-1.

The perturbation approach provides an excellent tool for studying systems with yet unknown allosteric mechanism in a routine way. First, biomacromolecules under investigation would be analyzed by the PocketAnalyzer program (**publication V**) to identify novel binding pockets. The performance of PocketAnalyzer was validated by studying transient pockets in the protein-protein interface of interleukin-2 (**publication VI**). Next, detected pockets would be probed for an allosteric function by perturbing the constraint network, i.e. by adding extra constraints at the identified pockets, and analyzing the altered rigidity of the biomacromolecule by a modified variant of the method introduced in publication II. Finally, the molecular features of the potential allosteric pockets could be used for subsequent virtual screening experiments in order to identify leads for novel allosteric effectors.

1 INTRODUCTION

Complex networks of biomacromolecular interactions regulate many cellular processes such as enzymatic catalysis and biosynthesis, metabolism through feedback regulation, molecular recognition, signaling, and energy transduction (1, 2). Understanding these regulatory mechanisms is not only interesting from a scientific point of view but also valuable in rational drug design. There are two ways to affect regulatory mechanisms by ligand binding: first, ‘conventional ligands’ can bind to the active site of proteins (further referred to as orthosteric site) and hence compete with the natural substrate. Second, ‘effector ligands’ can bind to a regulatory site that is distant from the orthosteric site, and is referred to as allosteric site.

Accordingly, allosteric regulation is the coupling between separated sites in biomacromolecules such that an action at one site changes the function at a distant site. This action can affect enzymatic activity by altering the turnover rate (V-type effect) and/or by altering the affinity for its substrate (K-type effect) (3). The function-altering of biomacromolecules is caused by a change of the structure and/or dynamics (4). Allosteric effects can be triggered (I) by binding processes ranging from small molecules (5) to very large systems such as peptides (6) or proteins (7), (II) by changes in the environment such as pH, temperature or ionic strength/concentration (8-10), and (III) due to covalent modifications such as tethering, glycosylation, phosphorylation and ubiquitination (11-14).

Typical examples of allosteric regulation are feedback loops to control metabolic pathways. Here, an effector acts as inhibitor in negative regulation or as activator in positive regulation. In negative regulation, a product of the metabolic pathway functions as an inhibitor for one of the earlier synthesis steps, whereas in positive regulation, binding of an allosteric activator switches on one of the crucial enzymes of the pathway. One example that involves both negative and positive regulation is phosphofructokinase 1 (PFK-1) (15). PFK-1 is of central importance in the mammalian glycolytic pathway by catalyzing the irreversible conversion of fructose 6-phosphate to fructose 1,6-biphosphate. Highly concentrated ATP and phosphoenolpyruvate (PEP), which is a product of a later step on the glycolytic pathway, are inhibitors of PFK-1 that bind to an allosteric site and trap PFK-1 in the inactive state. In contrast, a high level of AMP acts as allosteric activator of PFK-1 by introducing a structural change toward the active state.

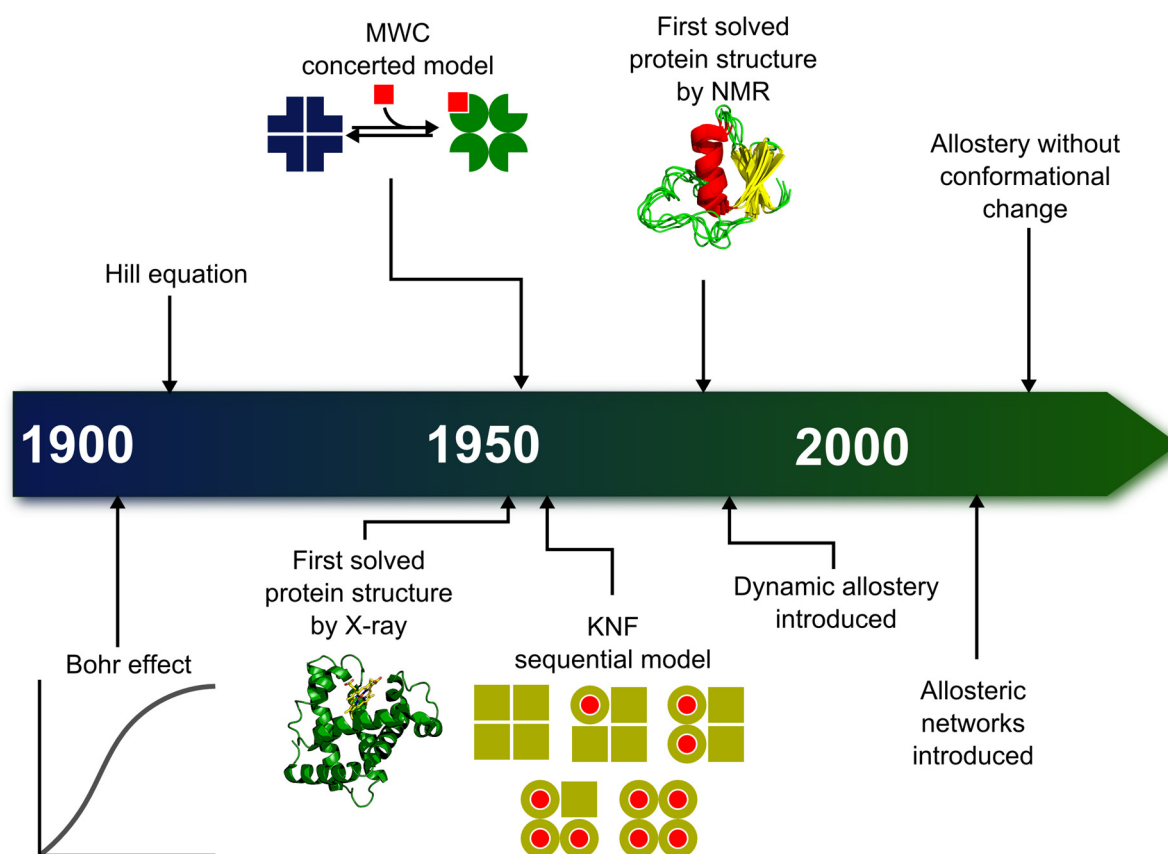


Figure 1: Historical timeline of allosteric regulation. In 1904, Christian Bohr studied the cooperative binding of oxygen to hemoglobin, which was mathematically described by Hill in 1910. After the first 3-dimensional structure was elucidated in 1958, the two main concepts of allosteric regulation emerged, termed the MWC or ‘concerted’ model and the KNF or ‘sequential’ model. Improvements in experiments gave insights into aspects of dynamics in biomacromolecules and, hence, inspired the model of dynamically-driven allostery. Accordingly, allosteric information is transmitted via networks of interactions and in the absence of conformational changes. The figure has been adapted from ref. (16).

The history of our understanding of allosteric regulation in biomacromolecules is illustrated in Figure 1. The idea of cooperative binding events arose long before the first structural information of a biomacromolecule was available. In 1904, Christian Bohr studied the binding of oxygen to hemoglobin under different conditions (17). He analyzed the relation of oxygen saturation as well as CO₂ concentration and pH for the binding of oxygen to hemoglobin, which is called the Bohr effect. The first mathematical description of the cooperative binding of oxygen to hemoglobin was developed by Hill in 1910 (18). In 1958, John C. Kendrew elucidated the first 3-dimensional structure of myoglobin by X-ray crystallography (19). Two years later, the first structural information of hemoglobin followed, by Max Perutz (20). This study generated insights in the structural arrangement of the heme groups in hemoglobin, which revealed the distant positions from each other. At the same time, Jean-Pierre Changeux analyzed the negative feedback mechanism of L-threonine deaminase, which is inhibited by the end-product L-isoleucine in the isoleucine biosynthesis (21). There,

Changeux proposed the existence of two topographically distant sites, which form the active and inhibitory site. Binding of L-isoleucine at the inhibitory site then influences the active site. The term allostery, which originates from the two Greek words *allos* (other) and *stereos* (solid) (22), occurred the first time in 1961. This led, in combination with the structural information about hemoglobin, to the two classical concepts of allosteric regulation proposed by Monod, Wyman, and Changeux (MWC model) (3) and Koshland, Némethy, and Filmer (KNF model) (23) (section 2.2). In both models, it was assumed that allosteric regulation is a general property of multimers and involves a conformational change. Two decades later, this view was extended by a model that suggests allosteric regulation by changes in dynamics as carrier for allosteric signaling (24). This view implies that allosteric regulation can take place in the absence of conformational changes. However, experimental evidence of this model was missing because X-ray crystallography only provides limited information about dynamics. Improved experimental techniques such as nuclear magnetic resonance (NMR) allow the exploration of dynamics in biomacromolecules (25). In 2006, Popovych *et al.* provided the first example of a system where allosteric regulation was mediated exclusively by changes in dynamics (26).

Overall, these studies revealed critical insights into the basis of allosteric regulation. However, predicting if biomacromolecules can be modulated by allosteric effects, and how, is still not possible in a routine way. This provides the motivation for this thesis.

2 BACKGROUND

First, I will review the current status of allosteric regulation, then I will provide an overview about commonly used experimental and computational techniques to study allosteric effects, and finally, I will provide a detailed introduction to rigidity theory, which forms the basis of my thesis.

2.1 Cooperativity in allosteric regulation

Historically, allosteric regulation has been associated with cooperativity between subunits in multimers. Christian Bohr was the first to study the cooperative binding of oxygen to hemoglobin under different conditions (17). If oxygen binds to hemoglobin, it stabilizes the oxy-(R state) hemoglobin and favors the binding of further oxygens. The behavior of saturation with oxygen in hemoglobin reveals a sigmoidal shape and not a hyperbolic shape as observed for the similar myoglobin (Figure 2 A). This effect allows hemoglobin to function as an oxygen transporter because binding and release of oxygen depends on the environment. For instance, binding of oxygen is favored at higher partial pressure of oxygen (pO_2) such as in the lungs, while it releases oxygen in tissues with lower pO_2 . The monomeric myoglobin is lacking this property because oxygen can bind with high affinity even at reduced pO_2 and hence, the oxygen saturation is almost insensitive for small changes in pO_2 .

Beside transportation of oxygen to the tissues, hemoglobin transports carbon dioxide (CO_2) back from the tissue to the lungs. This process is mainly influenced by the pH value and partial pressure of CO_2 in the tissue. For instance, metabolically active tissues have an increased level of CO_2 . The CO_2 dissolves in water and exists in equilibrium with bicarbonate and protons (H^+). However, under physiological conditions this reaction is relatively slow and hence, is catalyzed by carbon anhydrase. The increased level on H^+ causes a pH drop and a protonation of hemoglobin. This protonation shift the conformation to the deoxy-(T state) hemoglobin, which favors the release of further oxygens. While the majority of CO_2 is carried as bicarbonate through the blood serum to the lungs, a small fraction of CO_2 (~10%) reversibly bind to the T state of hemoglobin. Binding of CO_2 at the amino-terminals further stabilizes the T state because the resulting negatively charged carbamates support the salt bridge formations. In the lungs, the oxygen affinity of hemoglobin increases because of the higher pO_2 level. Together with a higher pH, this causes the deprotonation of hemoglobin and the dissociation of bound CO_2 . The dissolved bicarbonate in the blood serum reacts with the released H^+ to carbonic acids, which then dissociates to water and CO_2 .

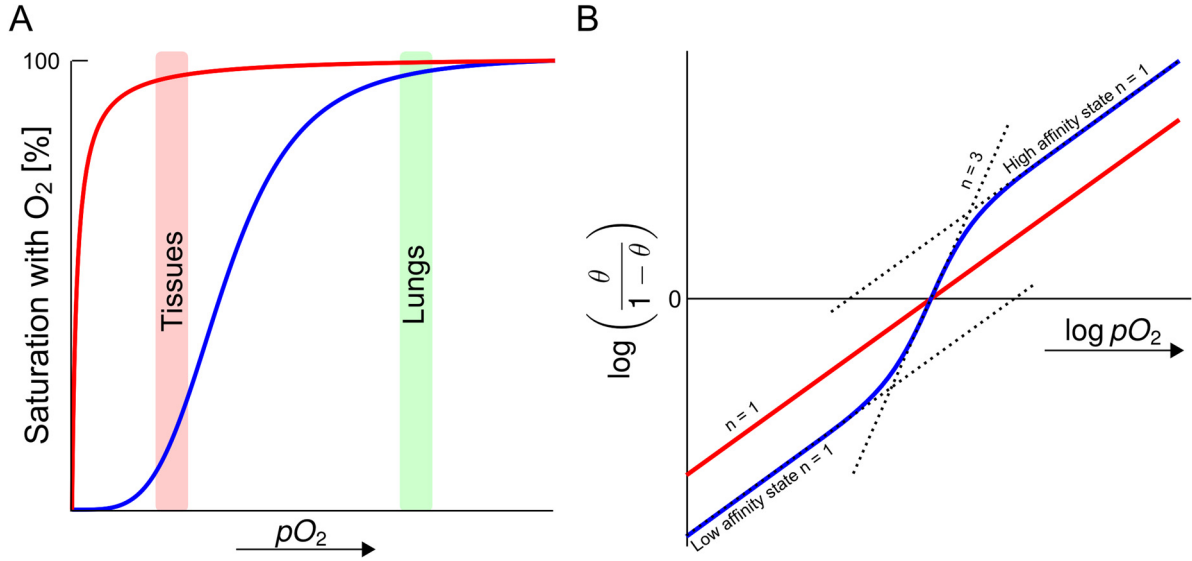


Figure 2: Cooperative binding of oxygen to hemoglobin. (A) Saturation with oxygen in hemoglobin (blue) reveals a sigmoidal curve and not a hyperbolic shape as in myoglobin (red). The sigmoidal curve indicates that the saturation in hemoglobin depends on the partial pressure of oxygen (pO_2) in lungs and tissues. The rapid saturation in myoglobin indicates that oxygen efficiently binds at low pO_2 and hence, is almost insensitive for changes in pO_2 . (B) The graph shows the Hill plot for oxygen binding to hemoglobin (blue) and myoglobin (red). θ is the fraction of saturated binding sites with oxygen. A Hill coefficient $n = 1$ indicates non-cooperativity as in myoglobin. In hemoglobin the maximum cooperativity between the low and high affinity state is ($n = 3$). The figure is adopted from ref. (27).

In 1910, Archibald Hill provided the first formulation to quantify the cooperative binding process of oxygen to hemoglobin (18). In the case of a biomacromolecule P with n binding sites the equilibrium upon ligand L binding is described as



In a hypothetical case where biomacromolecules with infinite cooperativity, i.e. biomacromolecules are either fully saturated or fully unbound, it follows that the dissociation constant K_d is

$$K_d = \frac{[P] \cdot [L]^n}{[PL_n]}. \quad (2.2)$$

The fraction of the saturation of binding sites θ can be written as

$$\theta = \frac{n[PL_n]}{n([P] + [PL_n])}. \quad (2.3)$$

Combining equation (2.2) and (2.3) results in

$$\theta = \frac{n[P][L]^n K_d^{-1}}{n[P](1 + [L]^n K_d^{-1})} . \quad (2.4)$$

Subsequent reordering results in the Hill equation

$$\theta = \frac{[L]^n}{[L]^n + K_d} . \quad (2.5)$$

The Hill coefficient n and dissociation constant K_d can be graphically determined by further reordering and taking the logarithm

$$\log\left(\frac{\theta}{1 - \theta}\right) = n \log[L] - \log K_d . \quad (2.6)$$

The maximum slope of the Hill plot gives the Hill coefficient n (Figure 2 B). If the Hill plot is a straight line and n equals one, then binding occurs in a non-cooperative manner. If $n < 1$, binding of a first molecule decreases the affinity for further binding processes and, hence, is negatively cooperative. Otherwise, if $n > 1$, the affinity for further binding processes is increased and hence, the binding is positively cooperative. However, it is not seen in reality that the Hill coefficient equals the number of actual binding sites, e.g. the Hill coefficient for hemoglobin never exceeds 2.8-3.0.

While the above mentioned cooperativity relates to conformational changes of the quaternary structure, Cooper and Dryden stress the aspect of dynamics as a carrier for allosteric signaling (24). Consequently, this mechanism allows allosteric regulation in the absence of conformational changes. The binding of a ligand causes a *stiffening* of the biomacromolecular structure, which then affects further binding processes. A quantity of allosteric cooperativity is the allosteric free energy

$$\Delta\Delta G = (G_2 - G_1) - (G_1 - G_0) , \quad (2.7)$$

where G_0 is the free energy of the unbound biomacromolecule, G_1 of the single bound and G_2 of the double bound state. Accordingly, a non-zero allosteric free energy indicates cooperativity in the system. For instance, if the allosteric free energy is positive ($\Delta\Delta G > 0$) then binding of the first ligand increases the burden for binding of the second ligand and, hence, binding is negatively cooperative. Otherwise, if the allosteric free energy is negative ($\Delta\Delta G < 0$) then ligand binding reduces the burden for subsequent ligand binding and, hence,

binding is positively cooperative. The allosteric cooperativity in a system where identical ligands bind to distant sites is given by

$$\Delta\Delta G = -kT \cdot \ln\left(\frac{Q_1^2}{Q_0Q_2}\right), \quad (2.8)$$

where Q is the canonical partition function of the system, either of the unbound biomacromolecule (Q_0), in the single bound (Q_1) or in the double bound (Q_2) state. kT is the product of the Boltzmann constant k and temperature T . The canonical partition function Q is the sum of statistical weights for each state in an ensemble. Q is defined as

$$Q = \sum_i e^{-\frac{E_i}{kT}}, \quad (2.9)$$

and combines information about mechanics, i.e. the energies E_i among all possible states i of the system, and thermodynamics through the temperature T .

From this, Cooper and Dryden assume two main sources that contribute to the cooperativity between two sites in biomacromolecules. The first source relates to changes in the vibrational spectrum and the second source to conformational changes, which leads to

$$\Delta\Delta G = -kT \cdot \left\{ \ln\left(\frac{q_1^2}{q_0q_2}\right)_{vib} + \ln\left(\frac{q_1^2}{q_0q_2}\right)_{conf} \right\}, \quad (2.10)$$

where q is either the vibrational (q_{vib}) or the conformational partition function (q_{conf}).

Source 1 – changes in the vibrational spectrum: The spectrum of vibrations of a system is composed of the frequency of atomic motions. Although the individual contributions for the allosteric free energy of each dynamic mode upon shifting the frequency is relatively small (about $-0.01 kT$ per mode as guessed by the authors), the total number of low-frequency modes in biomacromolecules can achieve cooperative free energies of a few kcal mol⁻¹.

Source 2 – changes in the conformational dynamics due to ligand binding: At the native state, biomacromolecules are composed of a multitude of possible conformational states. Upon ligand binding a certain conformation is stabilized over other states in the native state ensemble. This results into a shift of the mean probability distribution of conformational states. Accordingly, the authors assume that if ligand binding at distant sites affects the same atom in the biomacromolecule, then this atom is contributing to the cooperativity. To this end,

the authors approximated atomic motions by Gaussian fluctuations about the mean positions of each atom. Even small conformational changes, i.e. barely observable with current experimental techniques, can result in total allosteric free energies of several kcal mol⁻¹. This stresses the aspect of allosteric communication in the absence of gross conformational changes.

Overall, Cooper and Dryden suggested a mechanism of allosteric regulation in the absence of conformational changes. By separating the interaction free energies into entropy and enthalpy contributions, they concluded that cooperativity effects could be primarily entropic (24). It follows that positively cooperative binding of the first ligand results in the major loss of entropy, and thereby lowers the entropic burden for subsequent binding processes. In contrast, if binding of the first ligand does not quench the entire dynamics of the biomacromolecule, the major loss of entropy can occur during the second binding process; such an effect is called negative cooperativity (28).

2.2 Classical view of allosteric regulation involves conformational changes

In 1965, the MWC model for allosteric regulation was proposed by Monod *et al.* (3). There, the authors assumed that allostery is a property of symmetric multimers that predominantly adopt two conformations (T and R state) (Figure 3). In the absence of a ligand, the ratio of the inactive T and the active R state is determined by an allosteric equilibrium where the R state is of higher energy, and thus less populated (Figure 4 A). The ligand can only bind if the biomacromolecule is temporarily in the R state, and then binding pulls down the minimum of the energy landscape at the R state. Accordingly, the active R state of the biomacromolecule is now higher populated. Consequently, one can explain this mechanism as a conformational selection, in which the ligand preferable binds to one conformation and locks this state. In this case, all subunits change their conformation in a concerted manner to preserve the symmetry of the entire multimer. The MWC model proposes an all-or-nothing view on allostery because the multimer is either in the inactive (OFF) or active (ON) state.

One year after the MWC model, the KNF, or sequential, model was introduced by Koshland *et al.*, which explains the mechanism of allosteric regulation by changing the conformation of one subunit at a time (Figure 3) (23). As in the MWC model, allosteric regulation occurs in multimers by triggering a conformational change. However, the KNF model does not assume a pre-existing ensemble of the T and R state, but rather that the R state is not visited in the absence of the ligand (Figure 4 B). Here, ligand binding only causes a conformational change from T to R state in the subunits where the ligand actually binds. The

binding of a ligand then thermodynamically favors the binding in one of the adjacent subunits until all subunits are in the R state. Upon ligand binding, the energy landscape of the biomacromolecule reveals a minimum that is not present in the T state and relates to the R state, which then becomes the predominant state. Consequently, the KNF model can be understood as a type of induced fit mechanism. The KNF model also covers the case of negative cooperativity where binding of the first molecule hampers the binding of further molecules (29, 30). The MWC model does not capture this alternative in cooperativity.

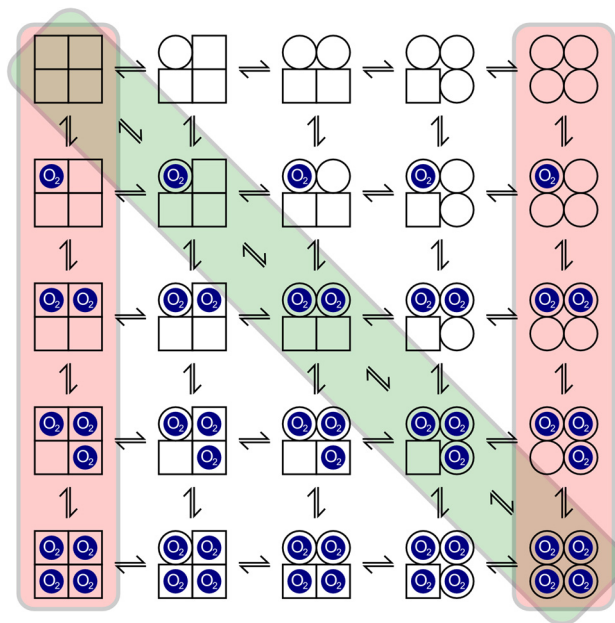


Figure 3: Classical models for allosteric regulation. Upon ligand binding, e.g. oxygen, each subunit can undergo a conformational change from the inactive (\square) to the active (\circ) state. The vertical reddish boxes represent the concerted MWC model and the diagonal greenish box represents the sequential KNF model.

Despite fundamental differences, both models share some common characteristics: (I) allosteric regulation occurs in multimers, (II) they involve a conformational change, (III) each subunit in an multimer can adopt two predominant states, which switch in a concerted (MWC) or sequential (KNF) manner. Both models are supported by experiments (31, 32) but are not mutually exclusive, as shown by a mechanism proposed by Perutz (33, 34).

Note that different mechanisms are likely to happen in different biomacromolecules, which means that the KNF and the MWC model are valuable models to explain allosteric regulation in biomacromolecules. For this thesis, I want to explain dynamically dominated allostery, hence I developed a computational approach (**publication IV**) that is based on aspects of flexibility and rigidity, i.e. static properties that denote the *possibility* of motions (section 2.8). Accordingly, biomacromolecules that are regulated by one of the two “classical views” are not addressed in the present thesis.

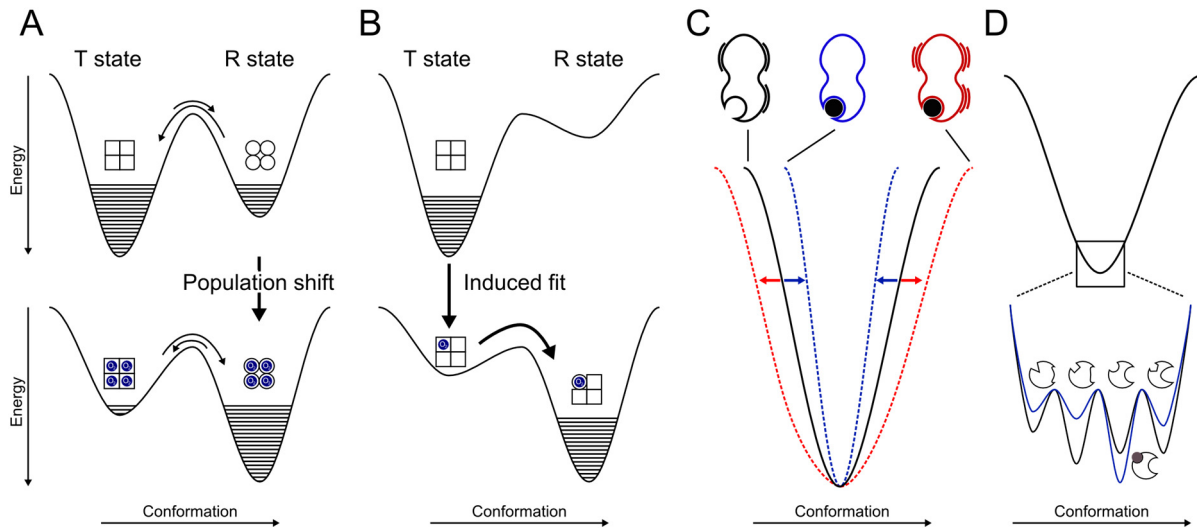


Figure 4: Illustration of the energy landscape in allosteric regulation. The energy landscape of biomacromolecules are composed of two minima (A, B) representing the inactive T and active R state. (A) In the MWC model, both states exist in an equilibrium, with a preference for the T state (depicted by horizontal lines) in the absence of a ligand. Ligand binding shifts the equilibrium to the R state. (B) In the KNF model, an induced fit mechanism shifts the energy landscape from the T to the R state, which is not sampled in the absence of a ligand. The figure exemplarily shows the first binding process in a multimeric system. The energy landscape of allosteric regulation without conformational changes is illustrated in (C) and (D). (C) The energy landscape is either narrowed (blue) or broadened (red) upon ligand binding. This way the vibrational dynamics are either decreased or increased. (D) In the case of changes in the conformational dynamics due to ligand binding, the biomacromolecule at the global minima exists in an equilibrium of states (black). The landscape is then narrowed to a single state upon ligand binding (blue).

2.3 Allosteric regulation in the absence of conformational changes

In the last decade, the “classical view” of allosteric regulation was extended by the “new view” (28, 35, 36). In the “new view”, it is suggested that the native state of a biomacromolecule is represented by a conformational ensemble, due to intrinsic flexibility and internal motions (37). Binding of an allosteric effector shifts the energy landscape of the population and stabilizes a certain state. However, Cui and Karplus questioned the newness of this view. Actually, the shift in the population is the basis of the original MWC model (38) as well as in an expanded MWC/Weber model (39). Nevertheless, three important aspects emerged from this “new view” of allosteric regulation: (I) allosteric regulation is not limited to multimers but is also a property of monomers (13), (II) allosteric regulation can exist in almost all biomacromolecules (35), (III) the absence of a conformational change does not exclude the possibility of allosteric regulation (4).

In particular, the last point emphasizes the question of how allosteric regulation works. In this context, Cooper and Dryden stress the aspects of dynamics rather than conformational changes as a key feature in allosteric regulation (24). The authors proposed an allosteric mechanism in the absence of conformational changes by changing the thermal fluctuation

amplitudes (Figure 4 C) or by “stiffening” a single state, out of multiple possible states around the native state (Figure 4 D). They showed that already small changes of fast-motions (ps-ns timescale) upon ligand binding could cause a change of the free energy by a few kcal mol⁻¹ (section 2.1). Because of the absence of conformational changes and, accordingly, a missing enthalpy term for breaking and re-(forming) of interactions, the entropy term must be the predominant effect of the allosteric regulation; in other words, the allosteric regulation is entropy-driven. Kern and Zuiderweg reviewed the role of dynamics in allosteric regulation already in 2003 (28). This view of allosteric regulation has been confirmed over the last two decades in several studies (4, 16, 38, 40-44). These studies indicate that no structural differences can exist between active and inactive states and, therefore, underpin the existence of allosteric regulation purely by changes in dynamics. Accordingly, allosteric regulation can be also understood as a thermodynamic phenomenon (45) and classified in entropy- (changes in dynamics without or with only subtle conformational changes), entropy and enthalpy- (change in dynamics accompanied by minor conformational changes) or enthalpy- (large conformational changes, e.g. domain motions) driven allostery (Figure 5) (4).

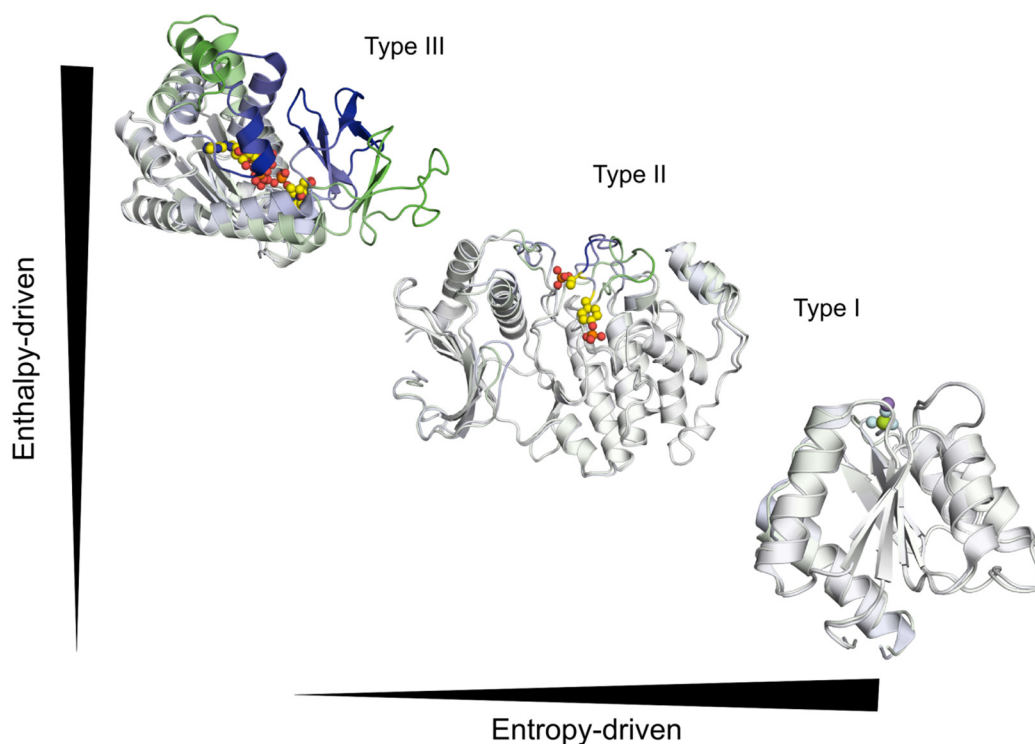


Figure 5: Classification of conformational effects in allosteric regulation. Illustration depicts the superpositioned structures of the open (green) and closed (blue) states for three allosterically regulated systems. Effector molecules are shown as spheres. The type I allosteric regulation happens in the absence of a conformational change and is mainly entropy-driven as shown for the chemotaxis protein CheY. The type II allosteric regulation involves small conformational changes such as loop movements and hence, is entropy and enthalpy-driven as shown for mitogen-activated protein kinase. Finally, in type III allosteric regulation, interactions must break and (re-)form to achieve a large conformational change such as the domain movement in adenylate kinase.

In 2006, Popovych *et al.* provided the first experimental evidence that allosteric signaling through changes in biomacromolecular dynamics exists in the catabolite activator protein (CAP) (26). A second example for long-range alteration of dynamics is the protease inhibitor eglin C (46, 47). Though itself non-allosteric, NMR studies of eglin C reveal dynamically coupled residues that have been identified by mutational perturbation experiments. Here the authors observed a type of dynamic response upon single-point mutations in which residues form a continuous pathway of dynamically coupled residues while the overall conformation does not change (Figure 6).

2.4 Allosteric signaling via pathways

Allosteric signaling between distant sites requires a physically connected pathway of residues to transmit information. However, this raises the question of how allosteric signaling works if no obvious conformational changes are visible between the inactive and the active state. While in the “classical view”, binding of an allosteric effector alters the conformation in a biomacromolecular assembly, in the “new view” the alteration occurs by changes in the dynamics. Williams *et al.* showed that binding of monoclonal antibodies to the hen egg white lysozyme (HEWL) leads to changes in the dynamics in a coordinated manner (48). Here, perturbations at the epitope are propagated to distant regions. Freire expanded this view in a theoretical study in which the transmission of cooperative interactions involves only a subset of residues within the biomacromolecule (49). The idea of a network of physically interconnected and/or thermodynamically linked residues for allosteric signaling is intuitively very appealing, and there are several computational and experimental studies that provide evidence for this concept (5, 40, 50-57). One example for a pathway of dynamically coupled residues is shown in Figure 6. Mutating residue 34 in eglin c to an alanine results in a dynamical change of several residues (indicated by NMR spin relaxation properties) (46, 47). Remarkably, those changes are not dispersed over the entire structure but rather form a continuous pathway that connects the mutation site with sites up to 11 Å apart (46, 47). Beside experimental techniques, several computational approaches have been introduced for analyzing allosteric signaling. One example is a sequence-based statistical method in which thermodynamics are estimated by characterizing the pattern of evolutionary constraints on and between amino acid positions within protein families (58) (section 2.7.3). The idea of allosteric signaling via pathways inspires the network representation of biomacromolecules (section 2.7.4). From this, graph-theoretical methods, e.g. shortest path analysis or analysis of

graph centrality, can be applied on the network representation to identify pathways of allosteric signaling.

Finally, the question arises as to whether allosteric regulations are an exclusively single pathway phenomenon or if they are composed by multiple pathways. The group of Nussinov has stressed the aspect of multiple pathways (59, 60). Because allosteric regulation has also been conceptually understood as a purely thermodynamic phenomenon (45), it does not require the existence of a specific pathway for allosteric signaling, but rather can be composed of multiple pathways. This view is supported by the existence of multiple conformational and dynamic states, which indicate multiple pathways for allosteric signaling in biomacromolecules (61-72).

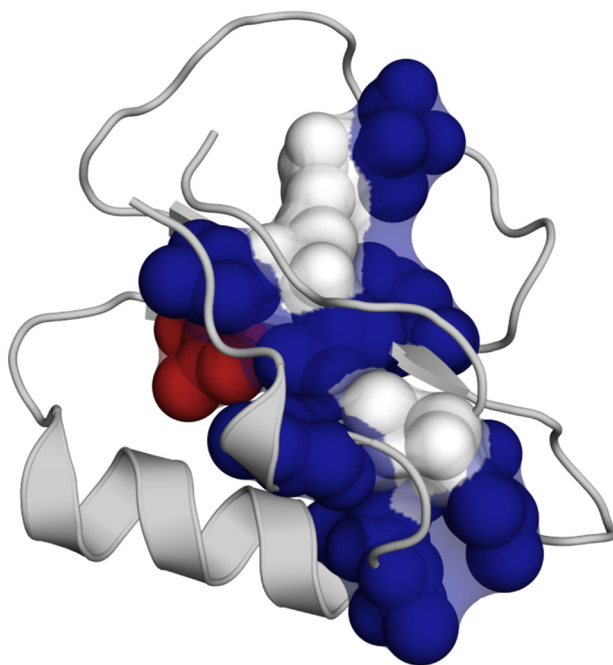


Figure 6: Altered dynamics in eglin c. Continuous pathway identified in NMR experiments of the mutant V34A (red) (46). The residues shown in white do not respond on the V34A mutation but are considered pathway residues as reported in ref. (46).

2.5 Allosteric targets in drug discovery

Targeting the function of pharmaceutically interesting biomacromolecules by allosteric effectors is a promising strategy in drug discovery (Figure 7)(73). Allosteric effectors have several advantages over conventional orthosteric site ligands. Orthosteric site ligands must compete with the natural substrate for binding. Accordingly, causing a pharmacological effect requires a sufficiently high affinity and persistency of the competitive ligand to stay in complex with the biomacromolecule. The existence of well-defined binding pockets, such as enzyme pockets, facilitates rational design of competitive ligands. However, those ligands and

natural substrates are somehow chemically similar because both mimic similar interactions at the orthosteric site. In particular, across protein families such as kinases, orthosteric sites are relatively conserved, and hence, share similar chemical features (74, 75). Unfortunately, this reduces the selectivity of competitive ligands for the actual target and can cause severe side effects by also inhibiting “off-targets”.

It has been shown that these problems can be overcome by targeting allosteric sites. Allosteric sites are under less evolutionary pressure to maintain the function of the biomacromolecule. Ideally, this allows the design of allosteric effectors that have no chemical equivalent in the world of small-molecules. Targeting allosteric sites is particularly useful in the case of highly conserved orthosteric sites across receptor subtypes such as G protein-coupled receptors (GPCR) (76, 77), kinases (78), and phosphatases (79). For example, the protein-tyrosine phosphatase 1B (PTP1B) is an attractive target because it is a negative regulator of the insulin receptor phosphorylation (80). Several orthosteric site ligands have been identified that compete with the phosphorylated substrate (79, 81). Unfortunately, among the protein-tyrosine phosphatase (PTP) superfamily, several members reveal high sequence similarities. T-cell protein tyrosine phosphatase (TCPTP) is the closest one, with an overall (active site) sequence identity of 72% (94%) (82). The cross reactivity was confirmed by PTP1B and TCPTP knock-out experiments of mice where the double knock-out turned out to be lethal. Alternatively, an allosteric site has been identified and targeted by a compound with micromolar affinity for PTP1B and improved selectivity over TCPTP (80).

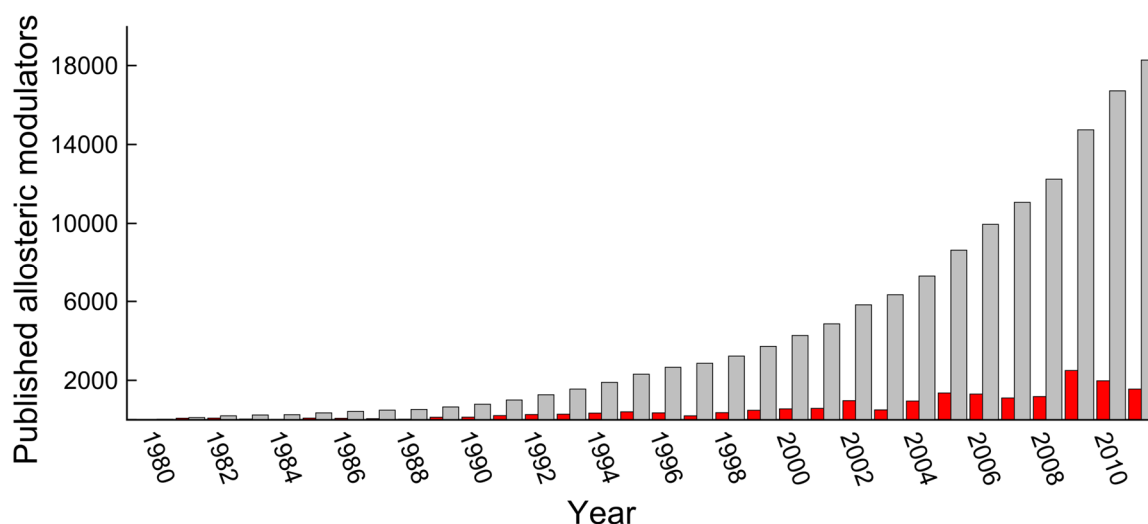


Figure 7: Number of published allosteric modulators. Histogram shows number of yearly (red) and total (gray) number of entries of allosteric modulators in ChEMBL-14. The dataset was taken from ref. (73), and publication years were retrieved from PubMed ids.

Another pharmacological advantage of allosteric effectors is the saturation effect. This is also known as the “ceiling” effect because the allosteric effector has only a pharmacological effect until it occupies all allosteric sites. Hence, overdoses do not affect the allosteric effect any further but increase the duration of the effect. Accordingly, allosteric effectors, although of low affinity, can cause pharmacological effects by administration of higher doses. This has the advantage of providing a long duration of a constant effect in saturated concentration.

A third advantage exclusively relates to allosteric activators due to their ability to address tissue responses selectively. Accordingly, allosteric activators only amplify an effect in the presence of endogenous ligands. The GABA_A receptor is an example where allosteric sites have been suggested to amplify the effect by GABA binding to the receptor (83, 84). One class of allosteric activators for GABA_A receptors are benzodiazepines, which are among the most commonly prescribed psychotropic drugs (83, 85). If GABA neurotransmitters are released, the effect of the allosteric effector is maximal, while the effect is minimal in the absence of GABA.

Finally, protein-protein interactions are generally difficult targets for drug design (86-88). The lack of distinct binding pockets as well as the overall large interface size hampers the rational design of competitive ligands (89, 90). If protein-protein interactions are under allosteric control, targeting allosteric sites with small molecules can affect the ability of protein-protein complex formation. Furthermore, the presence of already distinct pockets for allosteric regulation facilitates the design of potential effectors. The LFA-1 domain is part of β_2 -integrin and binds to intercellular adhesion molecules (ICAM). The LFA-1/ICAM complex formation can be interrupted by either competitive, small molecules (91, 92) or alternatively by several allosteric effectors that bind to a distant allosteric site in LFA-1 (93-95).

Nussinov *et al.* stress the complexity of allosteric regulation in cells and their consequences for drug design (96). Instead of just understanding the intra-biomacromolecular signaling, they point out the importance of allosteric networks, i.e. information flows by a series of interactions through cellular assemblies. These networks are composed of many interconnected pathways of interactions, and changes in these interactions may affect the interconnectivity. As a result, drugs can block a specific pathway. This view is supported by the identification of many different allosteric sites in GPCRs, which allows tuning the pharmacological response (97, 98). Accordingly, with a range of allosteric effectors it should be possible to stabilize receptors in specific biologically active states, which allows the chemical adjustment of receptor activity in more sophisticated ways than with orthosteric ligands. However, understanding the mechanism and complexity of cellular allosteric

regulation first requires a precise knowledge of how each step in this network is controlled in an allosteric manner. One interesting starting point for such an analysis is provided by the AlloSteric Database which contains in total 1,316 allosteric target entries and 22,356 allosteric modulator entries with 17,246 entries reported as allosteric drug-like (99). However, assuming that allosteric regulation is a general phenomenon in almost all biomacromolecules, there is a large number of systems with hitherto undiscovered allosteric mechanisms.

2.6 Allosteric regulation deduced from experiments

Much of the knowledge about allosteric regulation has been deduced from analyzing experimentally determined structures of the inactive and active state. Over the last decades, techniques were continuously improved and expanded to incorporate aspects of dynamics instead of just static structural information. In the following, I will briefly review some of the most common experimental techniques to deduce structural information that can be used to analyze allosteric regulation.

2.6.1 X-ray crystallography

X-ray crystallography provides detailed information about the location of atoms in biomacromolecules and is one of most widely used techniques to study allostery in biomacromolecules. The classical view of allostery involves a conformational transition between the inactive and the active state (section 2.2). However, X-ray crystallography provides only static structural information from the two states and, therefore, cannot directly monitor the transition. Accordingly, dynamical aspects have to be provided by other techniques such as NMR or molecular dynamics simulations (see section 2.6.2 and 2.7.1). The α -isopropylmalate synthase (IPMS) is a regulatory protein of the leucine biosynthetic pathway in *Mycobacterium tuberculosis* and a potential target for the design of new anti-tuberculosis drugs. X-ray crystallography reveals a lysine-binding site that is located in the regulatory domain of IPMS (100). Binding of lysine at the regulatory domain affects the enzymatic function in the catalytic domain. Several examples of enzymes that belong to an end-product regulation undergo large conformational changes upon binding of the end-product (101-103). In the case of IPMS, the conformation of the leucine bound and the unbound state do not show significant differences. A link between regulatory and catalytic domains has been suggested by hydrogen-deuterium exchange (H/D exchange) experiments (104). Here the authors identified a network of residues connecting the allosteric site in the regulatory domain with the active site in the catalytic domain based on changes in their dynamics.

The B-factor of structures determined by X-ray provides information about thermal vibrations of each atom. Hence, B-factors are an important indicator of biomacromolecular flexibility and dynamics. Wool *et al.* used information from B-factors to analyze altered function in mammalian pyruvate kinase (PK) by mutational perturbation (105). Upon mutation, the PK shows changes in dynamics not only at the mutation site but also at distant sites, which suggests a mechanism of long-range communication (106). In a different study, B-factors were used to analyze the allosteric inhibition by monostrol in kinesin spindle protein (KSP) (107). Here the authors could show that binding of monostrol at the allosteric site partially rigidifies the system. However, B-factors and X-ray experiments in general have some limitations: (I) biomacromolecules must be crystalized with very high quality, which is often the limiting step in crystallography, (II) the sample of the biomacromolecule is turned out of the natural solvent, (III) crystal packing effects can falsely suggest reduced mobility shown by lower B-factors.

2.6.2 NMR experiments

Solution nuclear magnetic resonance (NMR) spectroscopy allows investigating structural and dynamic changes in atomistic resolution and at room temperature. The large variety of NMR techniques provide meaningful insights in different timescales of biomacromolecular functions that range from ps-ns timescales (thermodynamic quantities including entropy) up to seconds scale (protein folding) (108). This makes NMR experiments a very useful technique to analyze allosteric regulation in the range from purely enthalpy-driven allostery, involving large conformational changes, to purely entropy-driven allostery by monitoring altered dynamics.

Residual dipolar coupling (RDC) analysis monitors molecular motions at the ps-ms timescale, which provides long-range structural information. Thus, RDC allows the study of biomacromolecular motions and dynamics across almost all timescales (108). The RDC approach was applied to analyze rotational motions of the subdomains IIB, IA, and IIA in Hsp70. This motion leads to an opening of a cleft between the subdomains and has been hypothesized to be important for allosteric signaling (109).

In the case of entropy-driven allostery, i.e. allosteric regulation in the absence of conformational change, nuclear spin relaxation (NSR) is appropriate to monitor biomacromolecular dynamics at the ps-ns timescale. At this timescale, biomacromolecules exhibit bond vibration and rapid side-chain/backbone motions. The order parameter S^2 is used to quantify the amplitude of internal motions at the ps–ns timescale. A S^2 value of 1 denotes

no local motion at a certain site, and a value of 0 denotes completely unrestricted or disordered motions. Information from S^2 values can be related to conformational entropy in allosteric signaling. One striking example of the usage of S^2 values is the identification of negative cooperativity of cyclic adenosine monophosphate (cAMP) binding to the catabolite activator protein (CAP) (26, 110). This is also the first experimental evidence of entropy-driven allostery. A second example in which S^2 values have been used is the investigation of the allosteric effect in the PDZ domain, which confirms results in a previous theoretical study about a communication pathway in PDZ (section 2.7.3).

Different NMR experiments are often combined to get complementary information of biomacromolecules. One example is the already mentioned non-allosteric eglin c. Eglin c was analyzed by NSR, RDC, and chemical shifts to test the effect of several mutants (46, 47). The NSR analysis identified altered dynamics for residues that are either dispersed or form a continuous pathway of connected residues. The RDC and chemical shift analysis revealed a conformational rearrangement of the backbone structure in the case of dispersed dynamical effects but an almost unchanged conformation in the case of a continuous pathway of altered dynamics.

Hydrogen–deuterium exchange (H/D exchange) is an appropriate technique to study biomacromolecular motions by monitoring the exchange of hydrogen atoms between the biomacromolecule and the solvent. H/D exchange experiments are typically performed by monitoring the exchange of solvent exposed backbone hydrogen by deuterium. Using H/D exchange combined with mass spectrometry can reveal small perturbations in biomacromolecules' motions caused by an allosteric effect or protein modification (111).

Finally, NMR chemical shifts are appropriate for studying allosteric regulation (112, 113). Chemical shifts monitor the resonance frequencies of the nuclei, and report on the local chemical environment of each nucleus. Because chemical shifts are highly sensitive to small changes in the biomacromolecular environment, e.g. binding of an allosteric effector, they provide an excellent way to analyze allosteric signaling effects (54, 55, 113, 114).

2.6.3 Fluorescence resonance energy transfer

Fluorescence resonance energy transfer (FRET) is a variant of fluorescence spectroscopy. This technique allows real-time monitoring of distances between donor and acceptor fluorophores attached to a biomacromolecule. The relation between the distance of fluorophores and the energy transfer was first described by Förster (115). The measurable distances are typically in the range of 20 to 100 Å. Briefly, a laser beam excites the donor

fluorophore. When it returns to the ground state, the excited donor emits a photon. If donor and acceptor fluorophores are in proximity, the acceptor absorbs the emitted photon. Accordingly, the emitted photon causes a nonradiative energy transfer between donor and acceptor. The efficiency, E , of the energy transfer is determined from measured emission spectra, and the distance R can be finally calculated from the Förster equation eq. (2.11)

$$E = \frac{1}{\left(1 + \left(\frac{R}{R_0}\right)^6\right)}, \quad (2.11)$$

where R_0 is the fluorophore-dependent Förster radius at which the energy efficiency is 50%. This makes the FRET analysis very sensitive to even small changes in the distance of the two fluorophores.

FRET analysis was used to investigate the conformational change in the ligand binding domain of the GluR2 subunit of the α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor (116). The authors showed that the wild-type undergoes a cleft closure upon AMPA binding that correlates to the extent of receptor activation. In a second example, Johnson reviewed single-molecule fluorescence spectroscopy methods to gain insights into conformations and dynamics of the calcium binding protein calmodulin (CaM) and CaM-regulated proteins (117). In a third example, Taraska *et al.* used transition metal ion FRET to propose a model for allosteric movements of the mouse hyperpolarization-activated cyclic nucleotide-regulated ion channel HCN2 (118).

Overall, the problem of FRET analyses is the need for a chemical modification of the biomacromolecule by attaching two fluorophores. Such a modification can cause a shift in a biomacromolecules static and dynamics that blur processes associated with allosteric regulation.

2.7 Allosteric regulation deduced from computational approaches

Over the last years, computer techniques have become an important tool in science. Consequently, the Nobel Prize in Chemistry 2013 was awarded “*for the development of multistate models for complex chemical systems*”. Computational techniques do not only complement experimental data, e.g. by reproducing already known aspects, but also provide new insights into the underlying mechanisms and function of biomacromolecules. In this section, I will briefly review some of the most widely used techniques to analyze aspects of (I) dynamics, i.e. molecular dynamics simulations (section 3.7.1) and elastic network models

(section 3.7.2), (II) evolutionary information, i.e. statistical coupling analysis (section 3.7.3), and (III) network connectivity (section 3.7.4). In addition, results obtained by these methods for analyzing or even detecting allosteric regulation will be provided.

2.7.1 Molecular dynamics simulations

Molecular dynamics (MD) simulations have become one of the most important tools for understanding dynamics in biomacromolecules (119, 120). In MD simulations, the time evolution of a system is computed by solving Newton's equation of motion and results in a trajectory of all particles in the system. MD simulations provide a link between structure and dynamics by enabling the exploration of the conformational energy landscape accessible to biomacromolecules (121). The first MD simulation of a biomacromolecule was reported in 1977 for the bovine pancreatic trypsin inhibitor (PTI) in vacuum with a length of only 9 ps (122). One drawback of MD simulations is the high computational cost, i.e. for a system with 23,558 atoms it still takes several weeks to calculate trajectory lengths in the μ s range using 64 processors (123). This is due to the recalculation of all forces acting on the atoms at each step of the simulation.

In one of the earliest MD studies to analyze allosteric regulation the burden of computational costs was reduced by using targeted MD simulations. Here, the authors performed MD simulations to analyze the transition between the T and R state of hemoglobin. At this time, unrestrained MD simulations were not possible due to the size of hemoglobin (64 kDa) and the large change in the quaternary structure (μ s-ms timescale) (124). The authors overcame this problem by applying a targeted MD simulation that forces an artificially rapid transition from the T to the R state (125). The analysis revealed a natural propensity for a dimer rotation in the T state, which is required in the T to R transition. However, a steric hindrance blocks the complete rotation of the dimer in the absence of the ligand. If the ligand binds, the hindrance is removed and the complete rotation towards the R state becomes possible.

The PDZ domain is a well-studied system with respect to allosteric signaling. The system does not show substantial differences between the bound and the unbound state indicating an entropy-driven allosteric mechanism (43, 44). MD simulations have been used to investigate the mechanism of allosteric signaling. In the first study, the authors performed an anisotropic thermal diffusion method (126). Here, the simulation of the PDZ domain starts at very low temperatures. Simulation at low temperatures aims to reduce the thermal noise, and thus, to improve the signal-to-noise ratio. From increasing the kinetic energy by heating at a single

residue in the binding pocket, the authors observed a pathway by tracking the most significant thermal diffusion. A predicted pathway pointed to the same region of the PDZ domain as suggested from statistical coupling analyses (section 2.7.3) but revealed a more direct connection. In a second example, the authors performed a pump-probe MD simulation (127). The pump-probe simulations measure the transmission of motional energy between residues. Exciting atoms or residues with a set of oscillating forces then triggers motional energy. The impulse percolates through the biomacromolecule and, afterwards, is probed by using Fourier transformation of the atomic motions. Like in the first example, the authors identified a pathway of coupled residues that show a similar pattern as the pathway from statistical coupling analysis. In both studies, the authors performed a perturbation approach by using non-equilibrium MD simulations, i.e. by simulating at very low temperatures or by applying external forces. In a third study on PDZ, an unrestrained MD simulation was carried out (128). Here the authors identified dynamically coupled residues by analyzing the correlation between residue-residue interaction pairs of the PDZ domain. From this, two pathways of allosteric signaling were identified: one novel pathway and a second one that agrees with the pathway predicted by the other two MD methods mentioned above.

A convenient way to gain insights into correlated motions along MD trajectories is the construction of dynamic cross-correlation maps. Cross-correlation maps have been used to study correlated motions in the catabolite activator protein (CAP) of the *apo* structure and in complex with single and double bound cAMP (129). The authors detected entropic effect from dynamic changes upon binding of cAMP on both subunits of CAP, despite cAMP only binding to one of the two subunits.

Although MD simulations provide a comprehensive and realistic description of biomacromolecules, the complexity of motions and dynamics hampers the interpretation due to the wealth of information. Dynamic processes that occur simultaneously at distant sites do not necessarily refer to an allosteric coupling between those sites. Consequently, an accurate separation of the signal from the noise is a prerequisite for analyzing allosteric regulation. One of the first studies that applied such a filtering was done by Sessions *et al.* (130). Here the authors separate the low-frequency motions from the high-frequency ones along MD trajectories by using digital signal processing techniques. As an alternative method, coupling of structural dynamics, e.g. by ligand binding, can be deduced from correlated motions (131-133). Identification of correlated motions requires a robust filtering of the signal from the noise. In particular, this is the case for entropy-driven allostery, i.e. in the absence of a conformational change. McClendon *et al.* introduced a method to quantify correlated motion

using a mutual information metric (131). The mutual information matrix is used to identify pairs of residues with correlated conformations in ensembles of multiple short MD simulations. This method applied on interleukin-2 revealed clusters of coupled residues in known cooperative binding sites. As mentioned above such an analysis requires robust filtering to avoid any artificial correlations.

2.7.2 Elastic network models

Elastic network models (ENM) provide an alternative and successful technique for analyzing biomacromolecular dynamics (134-137). The advantage of ENMs is to analyze conformational transitions in biomacromolecules at low computational costs; this is achieved by a coarse-grained representation of the biomacromolecules (138). In ENM, the atomistic potential, as used in MD simulations, is replaced by Hookean springs based on a harmonic pairwise potential between interacting atoms or residues.

Zheng *et al.* introduced a perturbation approach that uses evolutionary information in sequences to compute probability-based spring constants for each pair of sites making a contact (139). Upon perturbing (a) residue(s), a response at a distant site can be felt that denotes functionally important residues. By applying this perturbation approach on polymerase, the authors identified a network of distal residues that are most important for modulating the open to close transition, suggesting a mechanism of long-range communication. In a second study, the authors identified a sparse network of strongly conserved residues that transmit allosteric signals in DNA polymerase, the myosin motor protein, and the chaperonin of *Escherichia coli* (140). Several other approaches use ENM in combination with a perturbation approach, e.g. by changing the spring constant representing the force between two sites (141-144). Atilgan *et al.* introduced a perturbation variant of ENM. Here, the ENM is combined with concepts from linear response theory to compute the residue fluctuation profile by inserting random forces on all residues sequentially (145). The so-called perturbation-response scanning (PRS) approach has proven to be a powerful approach to capture conformational changes upon binding (146). The same approach was able to identify key residues for allosteric signaling in PDZ. The identified residues overlapped to a high extent, with the pathway predicted by a statistical coupling analysis (144).

However, despite overall success, ENM have some limitations. For example, due to the extreme simplifications by a network of C α atoms, the ENM neglects inter-residual interactions via side-chains.

2.7.3 Statistical coupling analysis

In 1999, Lockless and Ranganathan introduced the statistical coupling analysis (SCA) technique which exposes functional information buried in evolutionary records (58). The SCA approach strengthens the idea of continuous allosteric signaling via pathways of connected residues (section 2.4) and inspired several experimental and computational studies. Remarkably, the SCA does not require any structural information. Rather, the SCA deduces functional information in a protein family, such as coevolving residues, from multiple sequence alignments (MSA). Applying this approach on the PDZ protein family revealed a sparse network of residues representing a so-called evolutionary conserved pathway of energetic connectivity. To this end, the MSA was perturbed by extracting sequences that have a functionally important residue. That way, the authors identified residues that are statistically coupled with the perturbation site. The complete analysis of PDZ showed that most residues in this family do not do any coupling, but that a small set of residues exists which do couple. Remarkably, the results disclose a long-range coupling from sites in the core to residues on the opposite site of the PDZ domain, which was later supported by NMR measurements (44). The SCA method was successfully applied in several studies on G protein-coupled receptors (GPCRs) (147), chymotrypsin (147), hemoglobins (147), guanine nucleotide-binding proteins (148), retinoid X receptor (RXRs) heterodimeric receptors (149), TonB-dependent transporters (TBDTs) (150), and dihydrofolate reductase (DHFR) (151).

Although all these studies demonstrate the success of SCA, other studies cast some doubts on how effectively heuristic approaches like SCA can denote physically meaningful co-variation in residues (152, 153). Furthermore, to obtain meaningful results, it is required to have reliable and comprehensive multiple sequence alignments, which is not always possible.

2.7.4 Network analysis

The idea of allosteric signaling via intra-molecular pathways inspired approaches that are using a network representation of biomacromolecules. Here, 3D structures of biomacromolecules are embedded in a graph representation (154-165). The nodes then represent atoms or residues, and the edges represent the type of interaction between two nodes. The underlying idea is that allosteric signaling occurs from one site to another along a continuous pathway of edges. The high connectivity of these network representations requires smart and accurate techniques for filtering potential pathway residues from non-pathway residues. A common method is to search for shortest paths within the network connecting the allosteric and orthosteric site (55, 166-175). The identification of residues that frequently

occur on a shortest path emphasizes a high impact on allosteric signaling. In an alternative concept, the hierarchical structure of biomacromolecules is analyzed. In other words, biomacromolecules are composed of modules (subgraphs) with higher intra but low inter-connectivity, and residues between these modules tend to be involved in allosteric signaling (171, 174, 176, 177).

Following another network concept, biomacromolecules are modeled as constraint networks, where vertices represent atoms and edges represent covalent bonds and angles (178). To accurately model biomacromolecular flexibility, non-covalent interactions must also be included in this network (179-182). Flexible and rigid regions are then determined from the number and spatial distribution of bond-rotational degrees of freedom (183, 184). The *pebble game* algorithm (183-185), implemented in the FIRST (Floppy Inclusion and Rigidity Substructure Topology) software (179), efficiently assigns each bond as either being part of a flexible or a rigid region. A rigid region results from a collection of interlocked bonds that have no relative motion. If the rigid region has redundant constraints, it is overconstrained. Otherwise, it is isostatically rigid. In a flexible region the dihedral rotation of one bond is not locked in by other bonds. The theory underlying this approach is rigorous (186) and the parameterization for modeling the constraint network has frequently and successfully been applied in various analyses of biomolecules (187-190). In addition, comparisons between the constraint network and dynamical approaches have been performed in the past. These include the analysis of changes in the flexibility of proteins upon complex formation by constraint network analysis and MD simulations (182) as well as a large-scale comparison of protein essential dynamics from MD simulations and coarse-grained normal mode analyses (191), which use information from constraint network analysis as input (190). Because the concept of constraint networks forms the core of the perturbation approach developed in the course of this thesis, a detailed explanation of the underlying theory is given in the following section 2.8.

2.8 Rigidity theory and analysis

The sections 2.8.1 and 2.8.2 were adopted from ref. (192), which I co-authored.

2.8.1 Introduction to rigidity theory

The first mathematical formulations of structural rigidity date back to the last century. In 1864, James Clerk Maxwell investigated the conditions under which mechanical structures made of joints and connecting struts would be stable or instable (193). To this end, Maxwell used the method of constraint counting to deduce the mechanical stability without doing any

detailed, local calculation. The counting is a mean-field method to calculate the number of floppy modes F in a d -dimensional network of N sites. The term “floppy modes” denotes the (independent) internal degrees of freedom of each site in the network that can move without violating any of the constraints. Because the translational and rotational motions in the Euclidian space itself requires degrees of freedom, those global motions, d translations and $d(d - 1)/2$ rotations, must be neglected. For a network with N sites lacking any constraints, all degrees of freedom correspond to floppy modes whose number is $dN - d(d + 1)/2$, where the latter term denotes the global degrees of freedom. If a constraint is added to the network, and if this constraint is independent of all other constraints, then it removes one floppy mode. Accordingly, if all constraints in a network are independent, as assumed by Maxwell, the total number of internal floppy modes F_{Maxw} in a network with N_c constraints can be calculated by eq. (2.12)

$$F_{Maxw} = dN - N_c - d(d + 1)/2 . \quad (2.12)$$

In general, not all constraints are independent, which would lead to an underestimation of F . Non-independent constraints are between sites that are already mutually rigid, and thus, do not further decrease the number of floppy modes. Accordingly, non-independent constraints N_R are redundant and lead to eq. (2.13).

$$F = dN - N_c + N_R - d(d + 1)/2 . \quad (2.13)$$

Furthermore, redundant constraints introduce stress in the network and, hence, such regions are called *over-constrained* or *stressed*. In contrast, regions with fewer constraints than internal degrees of freedom are called *under-constrained*. Finally, regions with as many independent constraints as internal degrees of freedom are called *isostatically rigid* ($F = 0$).

In 1970, a theorem by Laman (194) had a major impact in that it allowed the local determination of the degrees of freedom in 2-dimensional networks, even in the presence of redundant constraints, by applying constraint counting to all sub-graphs within the network. As such, a generic 2-dimensional network is *minimally rigid* if and only if the number of constraints is $2N - 3$, and every non-empty subgraph induced by $N_s \geq 2$ sites spans at most $2N_s - 3$ constraints. Based on Lamans theorem, Hendrickson suggested an algorithm that exactly counts the number of floppy modes in 2-dimensional networks and, hence, is appropriate to decompose a network into rigid regions and flexible links in between (178). Further developments on this algorithm led to the efficient *2D pebble game* algorithm

implemented by Thorpe and Jacobs (183). However, this implementation of the *pebble game* algorithm can fail if applied on 3-dimensional networks such as the double banana network (185). None of the sub-graphs in this network has more than $3N_s - 6$ constraints connecting N_s site; this leads to the wrong conclusion that this network is rigid although it has one rotational hinge. However, this constraint counting can be extended to a certain subtype of 3-dimensional networks with a molecule-like character, the so-called *bond-bending* networks or molecular frameworks (195, 196). In these networks, bond angles (distances between second-nearest neighbor sites) are constrained in addition to bond lengths (distances between first-nearest neighbor sites), which makes them particularly applicable to biomacromolecules. For both the 2-dimensional and 3-dimensional *bond-bending* networks, combinatorial algorithms, called *pebble games*, were devised for the determination of network flexibility and rigidity according to eq. 2 (183-185). These algorithms have been implemented in ProFlex and in early versions of the FIRST software package. As an example, the construction of a *bond-bending* network of a molecule is depicted in Figure 8 A. In this network, fixed bond lengths and angles are modeled as distance constraints between nearest and next-nearest neighbor atoms.

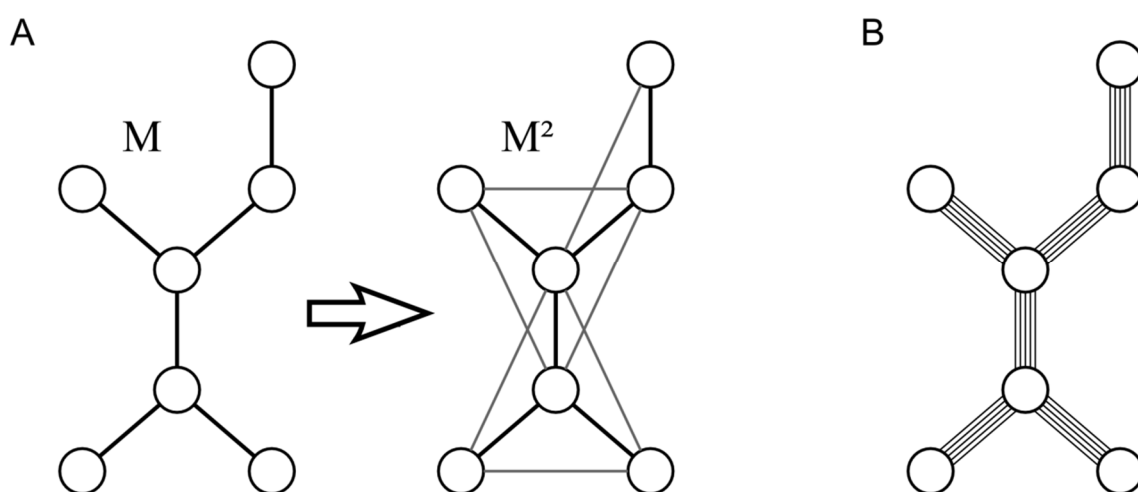


Figure 8: Network representation of molecules. (A) Representation of a molecule M as *bond-bending* M^2 network and (B) the same molecule as *body-and-bar* network.

As an alternative to *bond-bending* networks, *body-and-bar* representations lead to simpler algorithms of the *pebble game* and are used in a recent implementation of FIRST (Figure 8 B). In a *body-and-bar* representation every atom is considered as a rigid body having six degrees of freedom (198). Two rigid bodies are then connected by a set of bars representing the type of interaction, and every bar removes one degree of freedom (179, 180, 199). The number of floppy modes is computed by eq. (2.14)

$$F = 6N - N_{ibar} - 6 , \quad (2.14)$$

where N_{ibar} is the total number of independent bars in the network. A single covalent bond is modelled by five bars, leaving one degree of freedom, which represents the dihedral rotation. Non-rotatable bonds such as double bonds or peptide bonds lock all six degrees of freedom of the two atoms, and thus, the potential dihedral rotation. Apart from algorithmic advantages, the *body-and-bar* representation also has the methodological advantage that constraint strength can be modeled semi-quantitatively: strong bonds are modeled with more bars, whereas weaker bonds get fewer bars (198).

2.8.2 Modeling biomacromolecules as constraint networks

Biomacromolecules can be effectively represented either as *bond-bending* or *body-and-bar* networks. Because all results presented in this thesis are based on the latter representation of biomacromolecules, the modelling of different types of interactions will be now explained with this representation. Network representations of biomacromolecules differ from those of other materials, e.g. glasses, as they are composed of covalent and weaker non-covalent interactions. Although weaker, the importance of non-covalent interactions for the stability and 3D structure of biomacromolecules arises from their large number (Figure 9 A). Accordingly, modelling non-covalent interactions as constraints must be accurate. In the FIRST implementation used in this thesis, non-covalent bond constraints are associated to hydrogen bonds, salt bridges, and hydrophobic contacts. As described in the previous section, atoms are modeled as rigid bodies while covalent and non-covalent bonds are modeled as sets of bars (Figure 9 B). A covalent bond has five bars allowing for the dihedral rotation of this bond. Peptide and double bonds have six bars to lock their rotation. Non-covalent interactions such as hydrogen bonds (including salt bridges) and hydrophobic interactions are modeled as five bars and two bars, respectively (198). Weaker interactions such as van der Waals interactions are not modeled as constraints. If a hydrogen bond is included in the network representation depends on its geometry and energy. To this end, FIRST calculates the hydrogen bond (including salt-bridges) energies using an empirical potential (page 114) (200). The empirical potential for hydrogen bonds can be converted into a temperature scale T as proposed by Radestock and Gohlke (201).

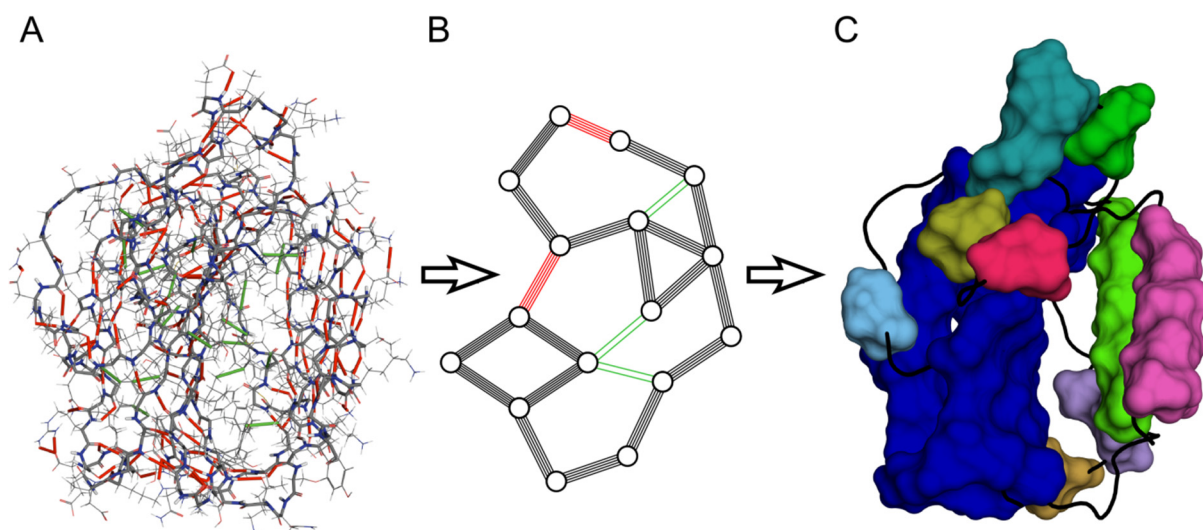


Figure 9: Rigidity analyses on constraint networks. A PDB structure of lymphocyte function-associated antigen 1 with added hydrogen atoms is used as input (A) from which a *body-and-bar* network (B) is modeled. Covalent bonds are depicted in black, hydrogen bonds in red, and hydrophobic tethers in green. Each bond is identified either as a part of rigid region or as a flexible joint. The analysis finally results in a rigid cluster decomposition (C) where each rigid clusters is depicted as a uniformly colored body.

Due to their less specific character, the quantification of the strength of hydrophobic interactions is challenging. Currently, hydrophobic interactions are simply included if two atoms are within a specific distance criterion. The decision for modelling hydrophobic interactions via two bars was done based on comparison with experimental data, i.e. where this parameterization best reproduced results of structurally weak parts in biomacromolecules as derived from experiments (202, 203). In a further study, the temperature dependence of hydrophobic tethers has been modelled by increasing the number of bars for hydrophobic tethers at higher temperatures (204). This follows the idea that hydrophobic interactions become stronger with increasing temperature (205).

Once the *body-and-bar* representation of the biomacromolecule is set up, the flexible and rigid regions are identified as described in section 2.8.1 (Figure 9C).

2.8.3 Distance constraint model

Jacobs introduced the distance constraint model (DCM), which is similar in spirit to the approach introduced in publication II (chapter 5). Here, Jacobs combines concepts from rigidity analysis with free energy decompositions (187). Accordingly, DCM is a mechanical model to capture the essential elements that govern the thermodynamic properties of biomacromolecules. Thermal fluctuations in constraint networks are modeled by fluctuating constraints at finite temperature. Covalent interactions are quenched distance constraints because they never break under physiological conditions and thus do not contribute to thermal

fluctuations. Non-covalent interactions frequently break and (re-)form, which is responsible for the soft matter-like character of biomacromolecules (206). In the DCM model, only hydrogen bonds and salt-bridges are considered as non-covalent constraints while hydrophobic contacts are neglected. The non-covalent constraints fluctuate by identifying native crosslinks and assigning a discrete variable to denote whether the constraint is present or not. Each fluctuating constraint in DCM is then assigned an enthalpy and entropy contribution. The sequence of how fluctuating constraints are placed is based on the entropy assignments from strongest to weakest along the network construction. A molecular framework \mathcal{F} is defined by a set of constraints with a particular topological arrangement. The free energy of a certain molecular framework $G(\mathcal{F})$ can be then computed by eq. (2.15)

$$G(\mathcal{F}) = H(\mathcal{F}) - TS(\mathcal{F}) . \quad (2.15)$$

Because the total enthalpy $H(\mathcal{F})$ is an additive property, it is the sum over all constraints. To account for the non-additive property of entropy, DCM considers the non-local cooperativity of interactions explicitly by using rigidity analyses to identify independent constraints (section 2.8.1). Consequently, the total entropy $S(\mathcal{F})$ is the linear sum of entropy components over a set of independent constraints.

The advantage of DCM is the ability to calculate mechanical network properties in a thermodynamically meaningful way, e.g. pairwise residue-to-residue couplings intrinsic to the native state ensemble (207-211). In an extensive study, the DCM was applied on three bacterial chemotaxis protein Y (CheY) proteins to explore the allosteric response across protein families (212). For this, a mechanical perturbation method (MPM) was introduced to simulate the binding of ligands by adding extra constraints to a certain site in the constraint network. The authors concluded that perturbed residues with large changes in stability characteristics are likely for allosteric signaling. From this, several hot spots for allosteric signaling have been identified that demonstrate the complex nature of allostery and the conservation of allostery across short evolutionary distances. In a second study, the same approach was applied to identify long-range effects in lysozyme (213).

However, one downside of DCM is the requirement of a protein specific parameterization of the enthalpy and entropy parameters. The accurate parameterization requires *a priori* knowledge of experimentally determined heat capacity curves C_p . In case of CheY C_p curves were not available. Hence, the authors used approximated C_p curves where the peak matches experimental T_m values.

2.8.4 Allostery deduced from rigidity analysis

Biomacromolecules are generally marginally stable (214), i.e. their network topologies are close to the rigidity percolation threshold at which a few constraints more or less can result in the network being largely rigid or already floppy. At this point, the network has the maximum information content and, hence, is most sensitive to changes. Accordingly, adding constraints rigidifies the network and induces stabilization in a former sensitive state. This was, for example, demonstrated by rigidity analyses of the complex formation of Ras/Raf (182). Remarkably, the stabilization is not locally restricted but also stabilizes regions that do not make any direct interaction with the protein-protein interface. This finding manifests, for the first time, the long-range aspect of rigidity percolation. Such long-range effects are also expected to be important in allosteric signaling (Figure 10).

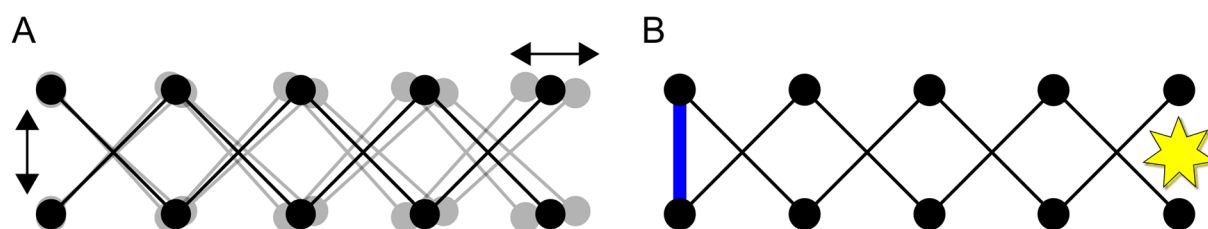


Figure 10: Rigidity percolation through a network of interactions. (A) Pantograph has one degree of freedom, which allows the movement of the network. (B) Ligand binding (blue bar) removes the degree of freedom and causes a rigidification of the network, which can be felt at a distant site (yellow star).

Using rigidity analyses, mechanical coupling for signal transmission has been identified in the ribosomal tunnel region (215). Signals are transmitted through structurally stable regions that are responsible for an induced conformational change in a domino-like manner. One aspect of this study has to be considered, in that the rigidity analysis was applied on single structures and that such an analysis is very sensitive with respect to the input structural information (182, 216, 217). Accordingly, the introduced perturbation approach in this thesis uses robust ensemble-based rigidity analysis to overcome the sensitivity problem of network-based approaches. The benefit of using ensemble-based rigidity analysis has been demonstrated in several studies by others and myself (182, 204, 217). Nevertheless, two independent experimental studies could later confirm this type of mechanism for signal transmission (218, 219). This demonstrates the capability of rigidity analyses to depict functionally important features in biomacromolecules even from single structures. In a different study, results from rigidity analyses demonstrated that the inactive T state is less rigid than the active R state in pairs of 16 crystal structures that exert an allosteric mechanism (220). However, such a comparison can be misleading because the analysis considers conformational changes between the T and the R state, which can be very subtle.

Consequently, a tightly packed structure results in a more rigid state upon ligand binding. The definition of the perturbation approach introduced in this thesis does not consider conformational changes of the biomacromolecular conformation, and solely perturbations in the constraint network are guiding changes in biomacromolecular stability.

3 SCOPE OF THE THESIS

In this work, I developed an ensemble-based perturbation approach for analyzing dynamically dominated allostery from altered biomacromolecular rigidity. Previous work made clear that analyzing dynamically dominated allostery is challenging, and has not been possible in a routine way (section 2.6 - 2.7). Being able that way to identify pathways for allosteric signaling and to compute free energies of cooperativity provides an exciting opportunity to consider entropic contributions to allosteric regulation, which have been difficult to deduce from a structure alone (40).

Rigidity analyses provide an excellent way to analyze the influence of ligand binding on biomacromolecular dynamics and/or stability (section 2.8). For deriving maximal advantage from these analysis, measures of biomacromolecular flexibility and rigidity are required (I) that are sensitive for monitoring changes in stability even due to small modifications of the constraint network and (II) that enable linking mechanical stability with biomacromolecular dynamics. In the literature, several global and local indices are available to link results from rigidity analysis with biologically relevant characteristics of a structure. Because of overlapping or vague index definitions however, I set out to present concise definitions of these indices, analyze the relation between them, review the scope and limitations of the different indices, and compare their informative value. In addition, I set out to develop new indices that extend the application domain of rigidity analysis. These results are reported in **publication I**.

A limitation of rigidity analysis is that single-structure analyses can be misleading, because even subtle conformational changes can have a pronounced effect on the results. This sensitivity problem can be overcome by rigidity analyses on structural ensembles (182, 204, 216). However, the generation of a structural ensemble, e.g. by MD simulations, compromises the efficiency of rigidity analyses. Hence, I set out to develop an approach based on ensembles of network topologies derived from a single input structure. The underlying algorithm is reported in **publication II**.

The Constraint Network Analysis (CNA) approach, introduced by S. Radestock and H. Gohlke (201, 221), goes beyond the mere identification of flexible and rigid regions in biomacromolecules. CNA infers biologically relevant characteristics from rigidity analyses, which are essential for understanding the relationship between biomacromolecular structure, (thermo-)stability, and function. The core of CNA is the efficient graph-theory based FIRST software (section 2.8.1 - 2.8.2). A limitation of the

early CNA implementation is the huge input/output (I/O) overhead, which increases the calculation time. Further limitations of CNA are that an automated ensemble-based analysis was not available, and that ligand molecules are not treated in a proper way. Hence, I set out to avoid the I/O overhead, which is a prerequisite for the development of an efficient and ensemble-based CNA, and to develop a robust treatment of ligand molecules during network construction. This leads to the automated and user-friendly CNA software package, which was implemented in collaboration with Prakash Chandra Rathi (**publication III**).

Concepts of rigidity analysis were already applied on studying dynamically dominated allostery (section 2.8.4). However, these studies have shortcomings in that (I) less robust single-structure analyses were performed (215), (II) active and inactive states of structures were directly compared, which implies that results can be biased by conformational effects (220), and (III) *a priori* knowledge of experimental data was required to incorporate with allosteric free energies (212, 213). I set out to develop a robust ensemble-based perturbation approach for analyzing dynamically dominated allostery. To simulate the change in flexibility and rigidity by allosteric effectors, *in silico* perturbations were applied on ensembles of network topologies extracted from MD trajectories. So far, the alternative method for generating ensembles of network topologies (**publication II**) is not parameterized for protein-ligand interactions, and hence, was not applied to probe allostery. The aims of the perturbation approach were (I) to probe the allosteric cooperativity between distant sites and (II) to predict putative residues that mediate the allosteric signaling. The theory underlying this approach and its application on eglin c, LFA-1, and PTP1B are reported in **publication IV**.

So far, dynamically dominated allostery has been analyzed *retrospectively*, e.g. for systems which are known to be allosteric regulated. Therefore, I set out to develop a computational pipeline that allows the study of systems with yet unknown allosteric mechanism. In course of this pipeline, I developed the grid-based PocketAnalyzer program in order to identify novel binding pockets in biomacromolecules as reported in **publication V**, the performance of which was validated in **publication VI**. As a perspective, I show preliminary tests with respect to how “dummy” ligands can be modeled based on information from PocketAnalyzer and how “dummy” ligands can be used to probe dynamically dominated allostery in a prospective way (chapter 11).

4 PUBLICATION I - Global and Local Indices for Characterizing Biomolecular Flexibility and Rigidity

Pfleger, C., Radestock, S., Schmidt, E., Gohlke, H.

J. Comput. Chem. (2013), 34, 220-233

Original publication, see pages 70 - 92; contribution: 40%

4.1 Background

The precise knowledge about biomacromolecular flexibility and rigidity and how flexibility/rigidity changes, e.g. upon ligand binding, is of great interest in rational protein engineering and for structure-based ligand design. Rigidity analyses as described in section 2.8, efficiently characterize biomacromolecular flexibility and rigidity. For deriving maximal advantage from such analyses, their results need to be linked to biologically relevant characteristics of a structure. With respect to analyze dynamically dominated allostery, measures of biomacromolecular flexibility and rigidity are required (I) that are sensitive for monitoring changes in stability even by small modifications of the constraint network and (II) that enable linking mechanical stability with biomacromolecular dynamics. Several global and local indices were reported in the literature to depict these characteristics. However, sometimes overlapping or vague index definitions were provided. In this publication, I present concise definitions of these indices, analyze their interrelation, scope, limitations, and compare their informative value. The most informative indices are available in the CNA approach (221, 222). In addition, I introduce three indices for the first time that extend the application domain of CNA. As a showcase, I probed the structural stability of the calcium binding protein α -lactalbumin, in the “ground” state and after *in silico* perturbation of the system by removing the calcium ion from the constraint network. From these findings, guidelines were provided for future studies suggesting which of these indices could best be used for analyzing, understanding, and quantifying structural features that are important for biomacromolecular stability and function.

4.2 Simulating the thermal unfolding

The CNA approach provides a method to simulate the thermal unfolding of biomacromolecules by gradually removing non-covalent constraints from initial network representations (180, 201, 202, 221, 224). For a given network state $\sigma = f(T)$ at temperature T , hydrogen bonds (including salt bridges) with an energy $E_{HB} > E_{cut,\sigma}$ are removed from the constraint network (200). Thus, stronger hydrogen bonds will break at higher temperatures than weaker ones. To convert the original, geometry-based hydrogen bond energy scale E_{HB} (200) into a temperature scale T , S. Radestock and H. Gohlke proposed a linear fit by comparing computed phase transition temperatures with experimental melting temperatures (201). The number of hydrophobic contacts is either kept constant or increases with increasing temperature during the thermal unfolding (205). Finally, rigidity analyses are performed on each constraint network state σ resulting in an unfolding trajectory. The thermal unfolding trajectories are then analyzed by CNA, which calculates several global and local indices.

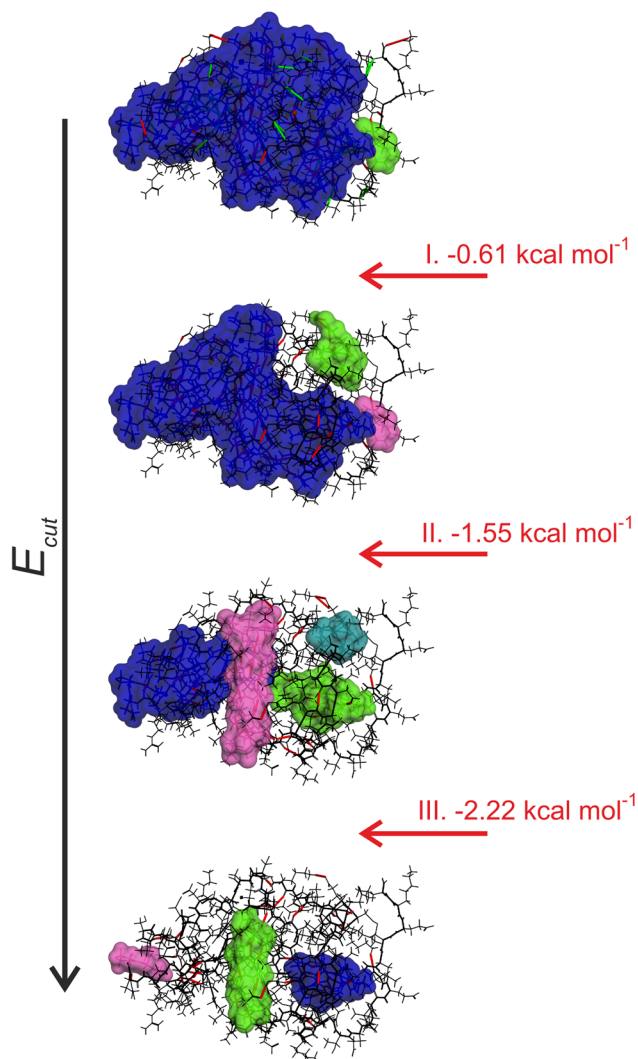


Figure 11: Rigid cluster decomposition along the thermal unfolding trajectory of α -lactalbumin. Rigid clusters are depicted as uniformly colored bodies. The roman numbers relate to the three major steps of rigidity loss. Figure adapted from ref. (223)

4.3 Global and local indices for characterizing biomacromolecular stability

Global flexibility indices monitor the degree of flexibility and rigidity within constraint networks at the macroscopic level. To describe the global percolation behavior of a network, the *microstructure* of the network can be analyzed, i.e. degrees of freedom or

properties of the clusters generated during a thermal unfolding simulation (section 4.2). This led to several indices such as the floppy mode density (Φ), mean rigid cluster size (S), or the rigidity order parameter (P_∞). The cluster configuration entropy (H) is another global index, introduced by Andraud *et al.* as a morphological descriptor for heterogeneous materials (225). H has been adapted from

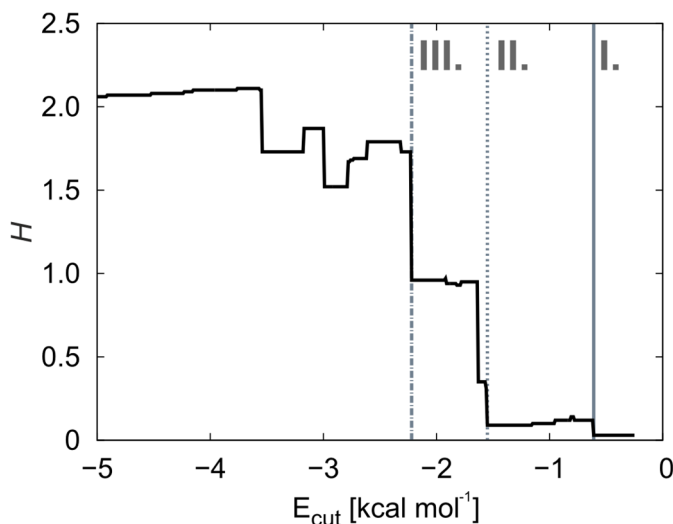


Figure 12: Cluster configuration entropy H for the thermal unfolding of α -lactalbumin. H is plotted as a function of the hydrogen bonding energy cutoff E_{cut} . The roman letters indicate the major steps of rigidity loss as shown in Figure 11. Figure adapted from ref. (223).

Shannon’s information theory, and thus, is a measure of the degree of disorder. By monitoring H during the thermal unfolding simulation of α -lactalbumin, three transitions have been identified (Figure 12). The transitions reflect changes in the network when the largest rigid cluster (I) starts to decay, (II) stops dominating the network, and (III) finally collapses. H was successfully applied by others and myself to analyze unfolding transitions in biomacromolecules that are related to thermostability (201, 204, 217, 221, 224). However, it turned out that H , as well as all other global indices, are not sensitive enough to detect changes between “ground” and perturbed states of α -lactalbumin (unpublished results).

Local indices characterize the network flexibility and rigidity down to the bond level. Accordingly, indices are derived for each covalent bond in the network by monitoring the cutoff energy E_{cut} for hydrogen bonds during a thermal unfolding simulation. The percolation index (p_i) is a local analogue to the rigidity order parameter P_∞ and can be best applied to monitor the percolation behavior through biomacromolecules locally. To this end, p_i monitors when a bond segregates from the giant percolating cluster during the thermal unfolding simulation. The giant percolating cluster is defined as the largest rigid cluster in the network state at the highest E_{cut} with all constraints in place. During the thermal unfolding simulation, the melting of the giant percolating cluster is monitored, and the largest rigid subcluster of the previous giant percolating cluster becomes the new giant percolating cluster of the present network state σ . The comparison of “ground” and perturbed state of α -lactalbumin, reveal a lower structural stability for a region, which is

about 15 Å apart from the ion binding site (Figure 13 A). The rigidity index (r_i) is a generalization of the percolation index p_i . This index monitors when a bond segregates from any rigid cluster. Again, perturbing the network topology demonstrates the sensitivity of this index to detect long-range aspects of altered stability (Figure 13 B). Furthermore, the results demonstrate that the information derived from r_i and the percolation index p_i is complementary. While p_i indicates the region of the biomacromolecule that becomes movable as a rigid body, r_i depicts the hinge regions that encompass the rigid body.

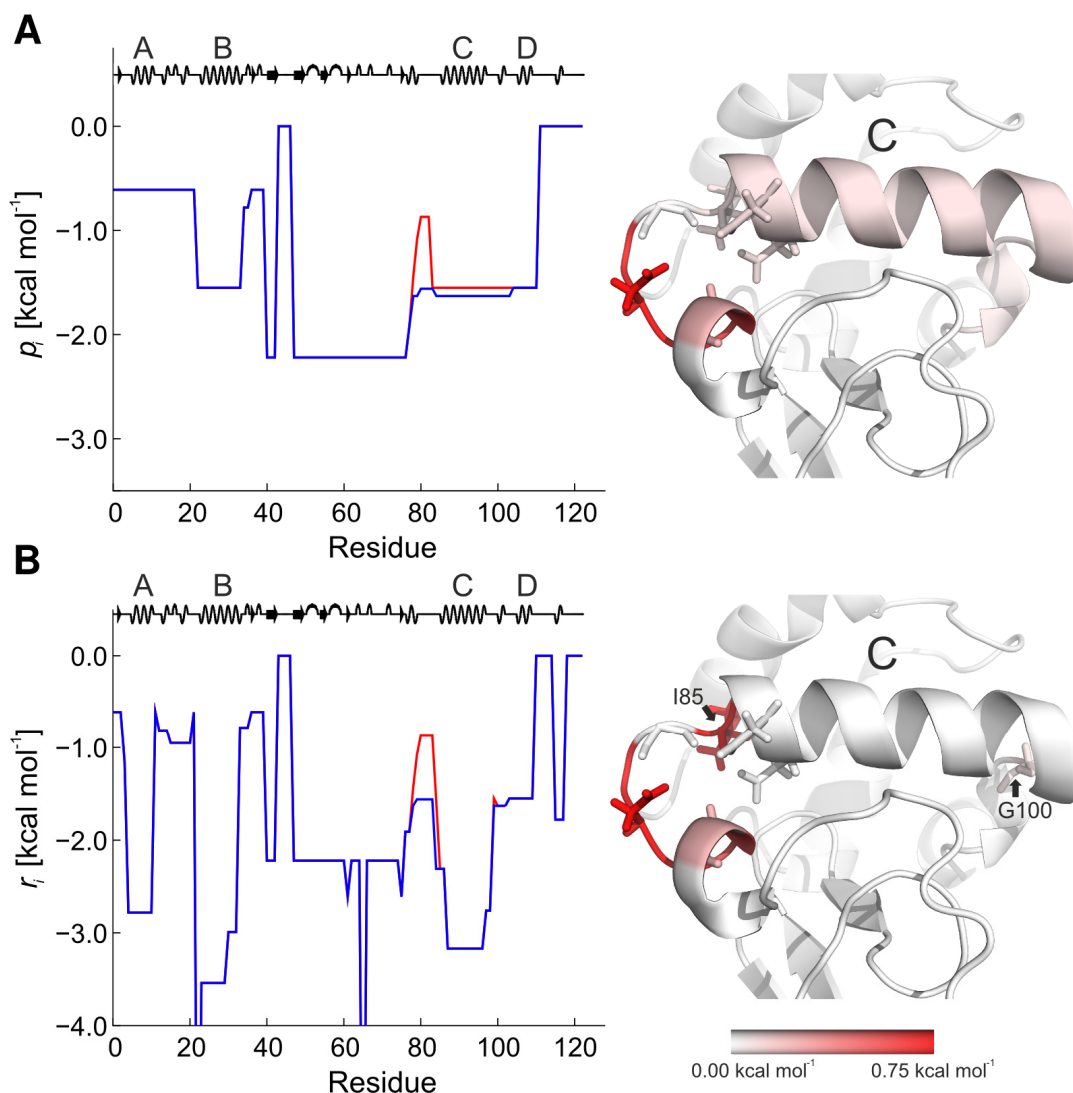


Figure 13: Change in the biomacromolecular stability by removing the calcium ion from α -lactalbumin. (A) Percolation index p_i and (B) rigidity index r_i show the ion bound state in blue and the ion unbound state in red. The lower p_i (r_i) the longer is a residue part of the giant percolating cluster (any rigid cluster). On the right, changes in the respective indices on going from an ion bound to an unbound state are mapped in a color-coded fashion on the structure of α -lactalbumin. Figure adapted from ref. (223).

As a third local index, stability maps (rc_{ij}) are 2-dimensional generalizations of the rigidity index. To derive a stability map, “rigid contacts” between two residues are identified. A rigid contact exists if two residues belong to the same rigid cluster. During a thermal unfolding simulation, stability maps are then constructed by monitoring E_{cut} at which a

rigid contact between two residues is lost. That way, a contact's stability relates to the microscopic stability in the network and, taken together, the microscopic stabilities of all residue-residue contacts result in a stability map. In Figure 14, the stability map of α -lactalbumin for the constraint networks with and without the calcium ion is shown. The map reveals that losses of rigid contacts do not only occur between isolated pairs of residues but also in a cooperative manner. That is, parts of the biomacromolecule break away from the rigid cluster as a whole.

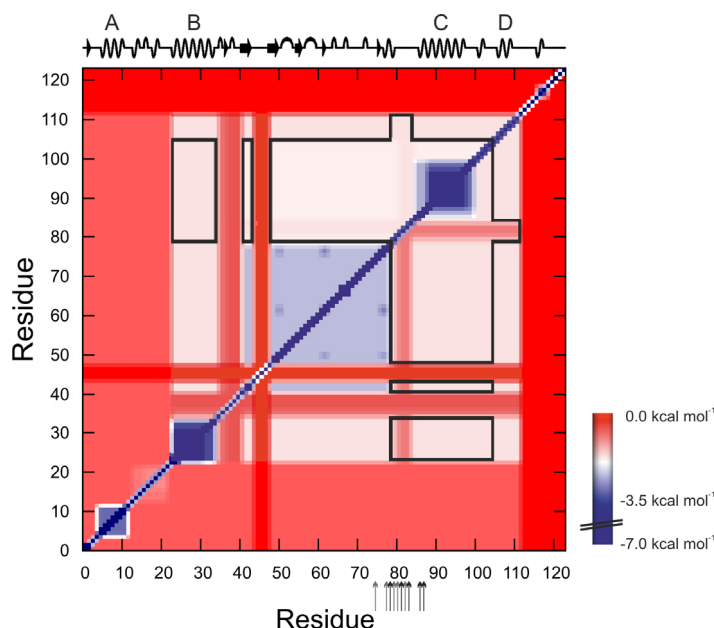


Figure 14: Stability map for α -lactalbumin. The upper half shows stability information of the calcium-bound state and the lower half of the ion unbound state. Red colors indicate pairs of residues where no/weak rigid contact exists. In contrast, blue colors indicate strong rigid contacts. Arrows highlight residues next to the calcium binding site. The black frames indicate regions that are affected by removing the calcium ion. Figure adapted from ref. (223).

4.4 Conclusion and significance

- For the first time, concise definitions of global and local indices were provided to describe stability characteristics in biomacromolecules.
- I introduced and applied three index definitions for the first time, which significantly extended the application domain of CNA.
- The showcase analyses of α -lactalbumin demonstrated the scope, limitations, and informative value of each index. A good agreement is found between the rigidity index r_i and protein regions that have high protection factors according to H/D exchange experiments.
- Based on the showcase analysis we provided guidelines for future studies suggesting which of these indices would be best for analyzing, understanding,

and quantifying structural features that are important for protein stability and function.

- We made suggestions for proper index notations in future studies, to prevent misinterpretation and to facilitate the comparison of results obtained from flexibility and rigidity analyses

The comparison of the “ground” and perturbed state of α -lactalbumin showed insightful results about which indices are best able to detect long-range stability changes that are related to allosteric regulation.

5 PUBLICATION II - Efficient and Robust Analysis of Biomacromolecular Flexibility Using Ensembles of Network Topologies Based on Fuzzy Noncovalent Constraints

Pfleger, C., Gohlke, H.

Structure (2013), 21, 1725–1734

Original publication, see pages 93 - 117; contribution: 70%

5.1 Background

Biomacromolecules require a balance of flexibility and rigidity to achieve their diverse functional roles. To this end, biomacromolecules have a soft matter-like character, i.e. noncovalent interactions flicker at room temperature (206). Previous studies demonstrated that rigidity analyses (section 2.8) are very sensitive with respect to the structural information used as input (182, 216). Accordingly, the difference of a few constraints due to the soft matter-like character of biomacromolecules can result in a network being either rigid or flexible. This sensitivity problem can be overcome by analyzing an ensemble of network topologies rather than a single-structure network. To do so, MD-generated conformations were used so far (182, 204). This way, however, a computational costly simulation compromises the efficiency of the rigidity analysis. As an efficient alternative, I present a novel approach (ENT^{FNC}) for performing rigidity analyses on ensembles of network topologies (ENT) generated from a single input structure (217). The ENT is based on definitions for fuzzy noncovalent constraints (FNC) and considers thermal fluctuations without actually sampling conformations. The definitions for fuzzy noncovalent constraints are derived from persistency data from molecular dynamics (MD) simulations.

5.2 Theory

The FNC model consists of two parts related to the modeling of hydrogen bonds (including salt bridges) and hydrophobic tethers in biomacromolecules. Parameters of the FNC model are derived from MD trajectories of ten different HEWL structures.

Part I - flickering of hydrogen bonds: To this end, I determined the persistence characteristics of hydrogen bond interactions along MD trajectories. The persistency of these interactions depends on the type, and thus, hydrogen bonds and salt bridges are treated separately. I further distinguished between hydrogen bonds in different secondary structure

elements. From this, I derived the probability with which a hydrogen bond (salt bridge) will be present across the ensemble of network topologies. The hydrogen bond energy E_{HB} determines the order with which hydrogen bonds are removed during a thermal unfolding simulation (section 4.2). In network topologies derived from a single structure, thermal motions of atoms are neglected that may influence E_{HB} . To account for this effect, Gaussian white noise is added to the initial E_{HB} computed for a hydrogen bond from the single input structure. The amount of Gaussian white noise depends on the type of hydrogen bond, which was parameterized from analyzing hydrogen bond energies along MD trajectories.

CNA

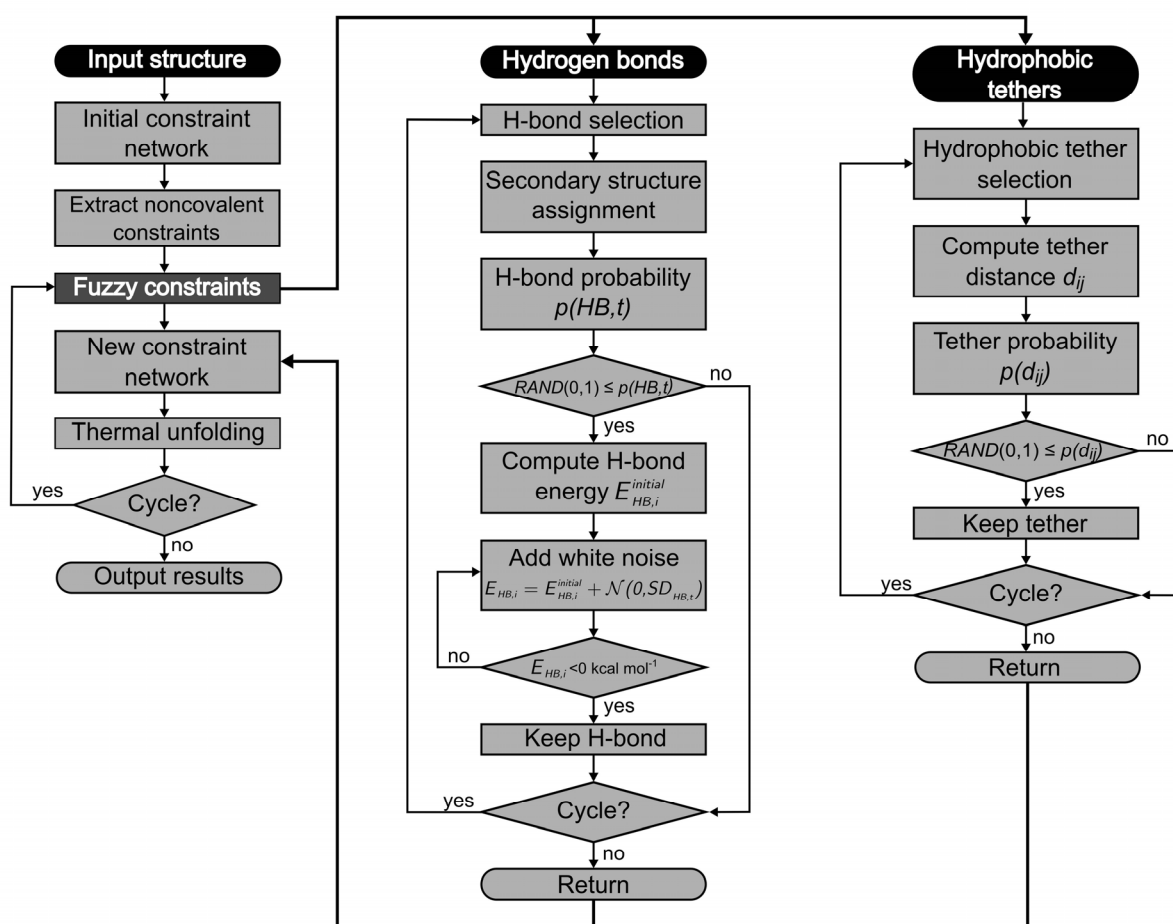


Figure 15: Workflow of the ENT^{FNC} approach. RAND(0,1) draws a random number with equal probability from the range [0,1]. Figure adapted from ref. (217). Figure taken from ref. (217).

Part II – flickering of hydrophobic tethers: Hydrophobic tethers are modelled to be less specific than polar ones. To this end, a fuzzy constraint representation has been developed in which tethers between closer atoms are included with a higher probability in a network topology than those between atoms further apart. The probability of tethers to be present in a network topology is sampled from a Gaussian distribution. Gaussian distributions have

previously been applied for modeling the strength of pairwise interactions between hydrophobic atoms (226-228).

With this FNC model, the ENT^{FNC} is generated from a single input structure and analyzed in five steps (Figure 15): (1) an initial network topology is generated; (2) information about noncovalent constraints are extracted and the secondary structure, a hydrogen bond or salt bridge is involved in, is assigned; (3) the number and distribution of noncovalent constraints is modified according to the definitions of FNC; (4) a new network topology is built; (5) global and local stability characteristics are calculated (**publication I**). Steps 3–5 are repeated until a user-specified number of networks is generated. Finally, global and local stability characteristics are averaged over the generated ensemble.

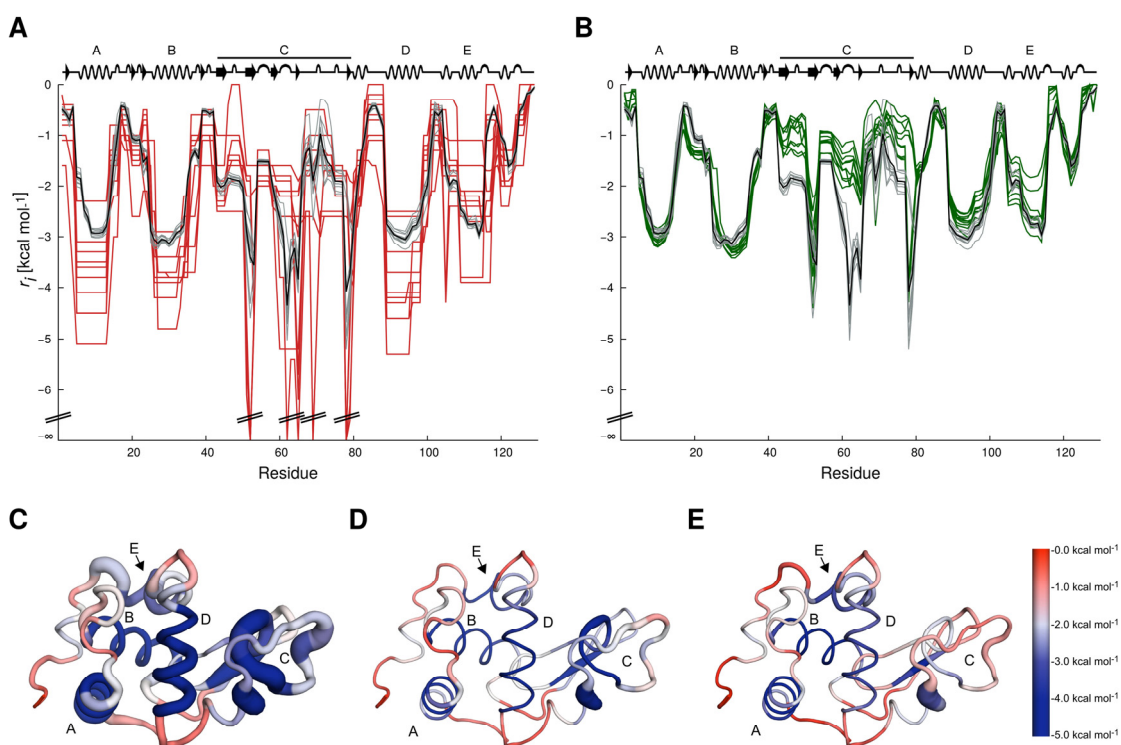


Figure 16: (A) Rigidity index r_i for the single-structure analyses of the ten HEWL structures (red), r_i curves for the ten ENT^{MD} analyses of HEWL (gray), and the average over all ENT^{MD} analyses (black) (B) Rigidity index r_i for the ENT^{FNC} analyses of the ten HEWL structures (green). (C–E) The mean r_i values and standard deviations from the single-structure analyses (C), ENT^{MD} analyses (D), and ENT^{FNC} analyses (E) are mapped onto a HEWL structure. The colors show the r_i values and the diameter of the putty plot the standard deviation of the r_i s at each residue position. The diameter is scaled with respect to the maximum standard deviation of all three analyses. Figure adapted from ref. (217). Figure adapted from ref. (217).

5.3 Validation of fuzzy noncovalent constraints

The approach has been validated by comparing the results (ENT^{FNC}) with those obtained for conformational ensembles generated by MD simulations (ENT^{MD}) of 300 ns length starting from HEWL structures.

Main results for the definitions of fuzzy noncovalent interactions: The average number of polar interactions in network topologies generated with FNC differs by at most 5% from those generated from MD ensembles. The definition of fuzzy hydrophobic tethers leads to a related difference of at most 14%. The higher difference in the latter case may reflect that hydrophobic interactions are less specific than polar ones (216).

Main results of the comparison of the rigidity analyses on ENT^{FNC} compared with those on ENT^{MD} : Averaging over an ENT^{FNC} leads to robust, almost starting structure-independent rigidity analyses, as also found for the ENT^{MD} approach (217). Furthermore, local flexibility and rigidity characteristics determined for the ENT^{FNC} agree almost perfectly (88% of the HEWL residues show r_i values that differ by $<1.0 \text{ kcal mol}^{-1}$) with those from the ENT^{MD} approach in terms of the magnitudes of the calculated rigidity indices r_i (Figure 16). These findings indirectly confirm the appropriateness of the FNC definitions. Furthermore, they suggest that the ENT^{FNC} approach is viable for overcoming the problem of the sensitivity of rigidity analyses with respect to the input structure.

In a case study, the ENT^{FNC} approach was used for computing relative thermostability values of citrate synthases (CS) and lipase A structures. In both cases, the results are encouraging for two reasons: (I) both protein systems were not used in the course of parameterizing the FNC; this demonstrates the transferability of the ENT^{FNC} approach; (II) both protein systems strongly differ in terms of the extent of the sequence similarity among the members. This indicates that, while the ENT^{FNC} approach is largely insensitive with respect to small conformational changes of input structures, it remains sensitive enough to pick up effects on the thermostability due to small sequential variations.

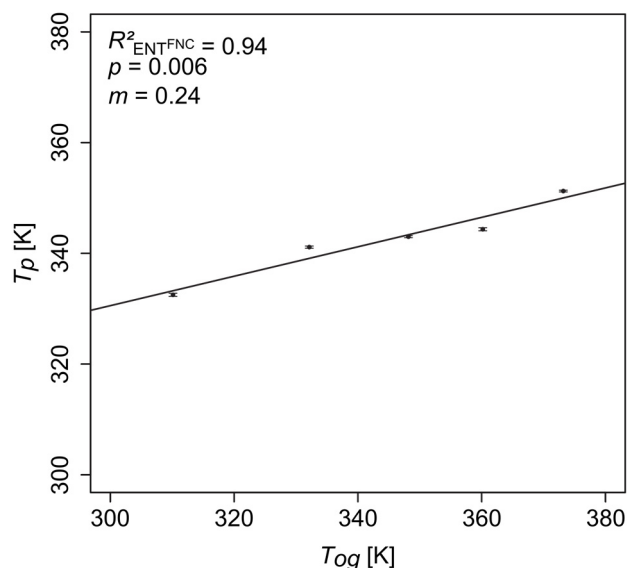


Figure 17: Correlation between predicted T_p from ENT^{FNC} and optimal growth temperatures T_o of citrate synthase structures. Figure adapted from ref. (217).

5.4 Conclusion and significance

- In this study, I introduced the novel ENT^{FNC} approach, which significantly improves the robustness of rigidity analyses by considering thermal fluctuations of biomacromolecules. Intriguingly, the results from rigidity analyses become almost starting structure-independent.
- The ENT^{FNC} approach demonstrates a high predictive power in which a good agreement between local flexibility and rigidity characteristics from ENT^{FNC} and MD simulations-generated ensembles is found. Regarding global characteristics, convincing results were obtained when relative thermostabilities of citrate synthase and lipase A proteins were computed. These results are encouraging because both protein systems were not used in the course of parameterizing the FNC and even for sequentially highly similar lipase A proteins the ENT^{FNC} remains sensitive enough to pick up effects on the thermostability.
- The approach does not require a protein-specific parameterization as required in the related DCM approach (see section 2.8.3).
- The low computational demand makes it especially valuable for the analysis of large data sets, e.g. for data-driven protein engineering. For instance, the ENT^{FNC} approach speeds up the calculation by a factor of ~ 4000 (~ 100) for HEWL (CS) comparing to the ENT^{MD} approach.

So far, the definitions for FNC do not contain a parameterization for protein-ligand interactions and hence, the approach was not applied to probe dynamically dominated allostery.

6 PUBLICATION III - Constraint Network Analysis (CNA): A Python Software Package for Efficiently Linking Biomacromolecular Structure, Flexibility, (Thermo)Stability, and Function

§Pfleger, C., §Rathi, P.C., Klein, D., Radestock, S., Gohlke, H.

J. Chem. Inf. Model. (2013), 53, 1007-1015

Original publication, see pages 118 - 127; contribution: 35%

§ Both authors contributed equally to this work.

6.1 Background

Biomacromolecular flexibility and its opposite, rigidity, are crucial for understanding the relationship between biomacromolecular structure, (thermo-)stability, and function. The CNA approach was first introduced by S. Radestock and H. Gohlke (201, 221) and aims on deriving the maximal information from rigidity analysis (chapter 2.8). To this end, CNA, carries out thermal unfolding simulations of biomacromolecules for linking information about mechanical stability with biological relevant characteristics (section 4.2). CNA functions as a front- and back-end to the FIRST software and allows one (I) to set up a variety of constraint networks, (II) to process results from rigidity analysis, and (III) to calculate various global and local indices for characterizing biomacromolecular stability. However, in this version, FIRST functions as an external program, and CNA post-processes the results obtained from FIRST. This leads to an I/O bottleneck and thus, increase the calculation time. Consequently, applying CNA in a routine way on large-scale datasets was not possible. Hence, we developed the Python-based CNA software package (229), that provides efficient and robust results from rigidity analyses and automatic detection of phase transition points (section 4.3) and unfolding nuclei (“weak spots”) of the structure.

6.2 PyFIRST as an Interface

In the initial implementation of CNA, FIRST functions as an external program, and the resulting I/O overhead drastically increases the overall calculation time. In the new implementation of the CNA software presented in this publication, the I/O problem was solved by developing the *pyFIRST* interface module. This interface module provides the full functionality of the C++-based FIRST program within the Python environment of the CNA

software. As a result, the *pyFIRST* interface module improves the calculation time from minutes to seconds for systems of several hundred residues. The interface module was implemented using the SWIG (Simplified Wrapper and Interface Generator) software tool (230). The SWIG tool automatically generates wrapper code for C/C++ programs, which then acts as an interface for other high-level programming languages such as Python.

6.3 Workflow of the CNA software

The CNA software can be performed on a single 3-dimensional structure of a biomacromolecule. Unfortunately, this can lead to varying results of the rigidity analysis, as we (217, 231), and others (216), observed. To overcome this problem, CNA can work on an ensemble of network topologies. Ensembles can be generated by extracting conformations from MD trajectories (204), experimental sources or, as an alternative, by fluctuating non-covalent constraints derived from a single input structure (**publication II**, chapter 5). For the accurate network construction of ligand molecules, we implemented an automated method, which determines the correct bond order and then merge the covalent constraint information for the ligand with the network information of the biomacromolecule. Next, the thermal unfolding simulation is carried out by sequentially removing non-covalent constraints from the network (see section 4.2) For each step of the thermal unfolding simulation a rigidity analysis is performed and post-processed to calculate different global and local indices (**publication I**). Finally, results from such analysis are written to output files. In case of ensemble-based analyses, all results are averaged over the entire ensemble. The overall workflow of CNA is shown in Figure 18.

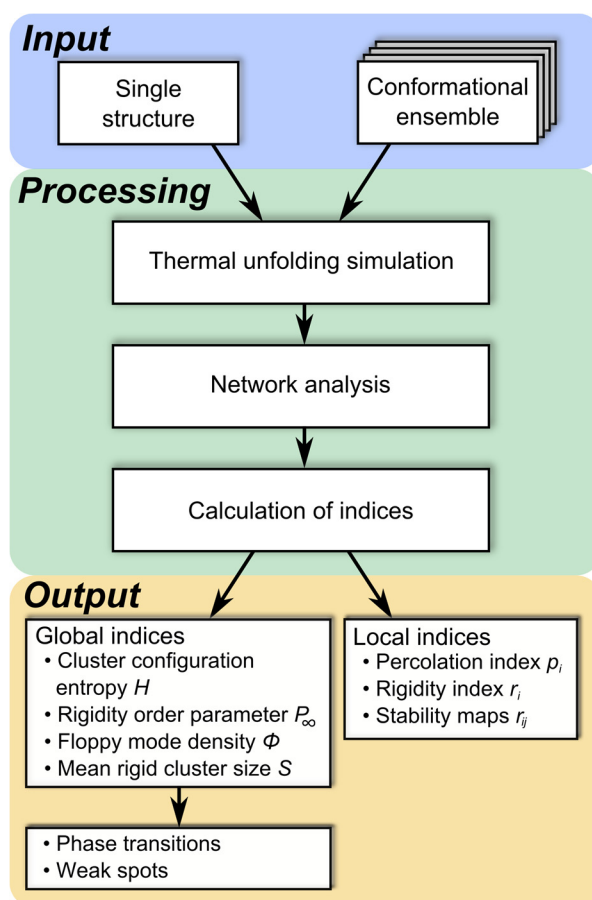


Figure 18: Schematic workflow of the CNA software. Figure taken from ref. (229).

6.4 Showcase example: Flexibility characteristics of HEWL

The value of applying CNA to ensembles of conformations generated by thermal fluctuations was validated by using the HEWL protein as a test case. The validation was based upon two analysis types: (I) the analysis of a constraint network derived from a single input structure and (II) the analysis of an ensemble of constraint networks derived from a MD trajectory. Several global and local indices were compared with experimental results, of which only three will be briefly mentioned: (I) the cluster configuration entropy H is a global index and monitors the loss of network stability during the thermal unfolding simulation (see section 4.3). At the transition point, unfolding nuclei were inferred to be weak spots of the structure. Encouragingly, the identified unfolding nuclei are located in a region that plays a crucial role in stabilizing the tertiary structure of HEWL (232); (II) the local rigidity index r_i characterizes the stability of a biomacromolecule on the bond level (see section 4.3). The identified stable regions in HEWL are in good agreement with those identified by H/D experiments (233), while flexible regions are in agreement with those suggested to be involved in a hinge bending movement during the catalytic cycle (234); (III) stability maps rc_{ij} are 2-dimensional itemizations of the rigidity index and report when a “rigid contact” between two residues of the network is lost during the thermal unfolding simulation (see section 4.3). Remarkably, weaker contacts are identified for residues in disordered regions as indicated by NMR experiments (235).

6.5 Conclusions and significance

- In this study, we presented the user-friendly Python-based CNA software package that automatically sets up a variety of network representations, process the results from rigidity analysis, and calculates several global and local indices.
- For the first time, CNA integrates robust and efficient ensemble-based analysis and automatic index calculation. In addition, we extended the application domain of rigidity analyses in that phase transition points (“melting points”) and unfolding nuclei (“structural weak spots”) are determined automatically.
- The CNA software is highly efficient such that a single-structure analysis requires only few seconds. This has been achieved by linking CNA and FIRST via the *pyFIRST* interface module, minimizing the I/O overhead.
- We implemented a robust handling of small-molecules during the network construction. This is important when it comes to estimate the influence of ligands

on biomacromolecular stability, e.g. for probing signal transmission across biomacromolecules for analyzing dynamically dominated allostery.

- The showcase example on HEWL demonstrates that results from CNA are in good agreement with those from experiments. Hence, CNA provides an important platform for further projects

For this thesis, CNA was applied to analyze changes in stability upon *in silico* perturbations in constraint networks. Monitoring altered stability aims at depicting long-range aspects of rigidity percolation, which are expected to be important in allosteric signaling.

7 PUBLICATION IV – Allosteric Coupling deduced from Altered Rigidity Percolation in Biomacromolecules

Pfleger, C., Minges, A.R.M., Gohlke, H.

Submitted manuscript (2014)

Original manuscript, see pages 128 - 159; contribution: 70%

7.1 Background

Targeting the function of pharmaceutically interesting biomacromolecules by allosteric effectors is a promising strategy in drug discovery (section 2.5). Over the last two decades, the classical view of allosteric regulation (section 2.2) was extended by considering changes in dynamics as carrier for allosteric signaling (section 2.3). However, understanding dynamically dominated allostery from a single structure alone can be challenging due to the absence of conformational changes. Rigidity analyses provide an excellent way to analyze the influence of ligand binding on biomacromolecular dynamics and/or stability (section 2.8). This concept was already applied on studying dynamically dominated allostery (section 2.8.4). However, these studies have shortcomings in that (I) less robust single-structural analyses were performed (215), (II) direct comparison of active and inactive state of structures was done, which can lead to biased results by conformational effects (220), and (III) *a priori* knowledge of experimental data was required to incorporate rigidity analyses with allosteric free energies (212, 213). In this publication, I present a robust ensemble-based framework for analyzing dynamically dominated allostery. To simulate the change in flexibility and rigidity by allosteric effectors, *in silico* perturbations were applied on ensembles of network topologies extracted from MD trajectories. This way, the conformation of the biomacromolecule is unchanged and solely the perturbations in the constraint network are guiding changes in biomacromolecular stability. The approach allowed us (I) to compute free energies of allosteric cooperativity and (II) to predict putative residues that mediate the allosteric signaling. The application of the perturbation approach is demonstrated on eglin c, LFA-1, and PTP1B.

7.2 Overall strategy

The CNA software (publication III) was used to perform rigidity analysis on eglin c, PTP1B, and LFA-1, which exhibit a dynamic response upon mutating residues and/or are known to be allosteric regulated with subtle or absent conformational changes. The results

were post-processed to compute free energies of cooperativity and to identify pathways of residues for allosteric signaling. Changes in residues' flexibility and rigidity were studied by perturbing the constraint network at functionally important sites, in general the known allosteric site. To overcome the sensitivity problem of rigidity analysis (182, 216, 217), CNA was applied to conformational ensembles extracted from MD trajectories starting from the wild-type structure of eglin c and the effector bound structures of PTP1B and LFA-1. The constraint networks were then perturbed *in silico* by mutations (eglin c) or by removing the allosteric effectors (PTP1B and LFA-1) from the network. In this way, changes in biomacromolecular stability arise solely from the perturbation in the network topology and conformational changes do not bias results. From these mechanical stability characteristics, we computed free energies of altered rigidity. To this end, stability maps (section 4.3) were used, which displays the hierarchy of stability in biomacromolecules. The mechanical energy E_{CNA} of a system is defined as the sum over all rigid contacts in the stability map. The difference between $E_{CNA,perturbed}$ and $E_{CNA,ground}$ of the system then gives the mechanical energy ΔE_{CNA} , which is the altered rigidity by the presence of an allosteric effector. From this energy, we calculated a mechanical perturbation free energy ΔG_{CNA} , which is used to deduce the cooperative free energy $\Delta\Delta G$. The underlying positive or negative regulation in biomacromolecules is derived following the formulation of Cooper and Dryden (section 2.1). In order to identify pathways for allosteric signaling, we performed a per-residue decomposition of the free energy. This led to the free energy $\Delta G_{i,CNA}$ for each residue i , which was used to predict putative residues that mediate the allosteric signaling between distant sites.

7.3 *In silico* mutational perturbation of eglin c agrees with experimental findings

The validation of the perturbation approach is based on the serine protease inhibitor eglin c. To this end, predicted free energies upon *in silico* mutational perturbations were compared with free energies of unfolding from chemical denaturation experiments (46). This results in a fair correlation ($R^2 = 0.80$) between predicted and experimentally determined free energies of unfolding (Figure 19 A). Overall, the results demonstrate the ability of our approach to predict the coupling from altered rigidity.

NMR studies of eglin c reveal continuous pathways of dynamically coupled residues upon mutations in the absence of a conformational change (46, 47). The per-residue $\Delta G_{i,CNA}$ was derived from the comparison of stability maps in the ground (wild-type) and perturbed

(mutant) state. Stability maps provide insights into residues that are flexibly and rigidly correlated across the structure (publication II, section 4.3). Remarkably, residues with the strongest altered rigidity (> 0.2 kcal mol⁻¹) form a continuous pathway with the most far-reaching effect up to 15.9 Å apart from the mutation site. Furthermore, the predicted pathway derived from altered rigidity agrees with findings from NMR experiments (46, 47). Phrasing this as a binary classification problem, I could discriminate between pathway and non-pathway residues with an accuracy of ~85%.

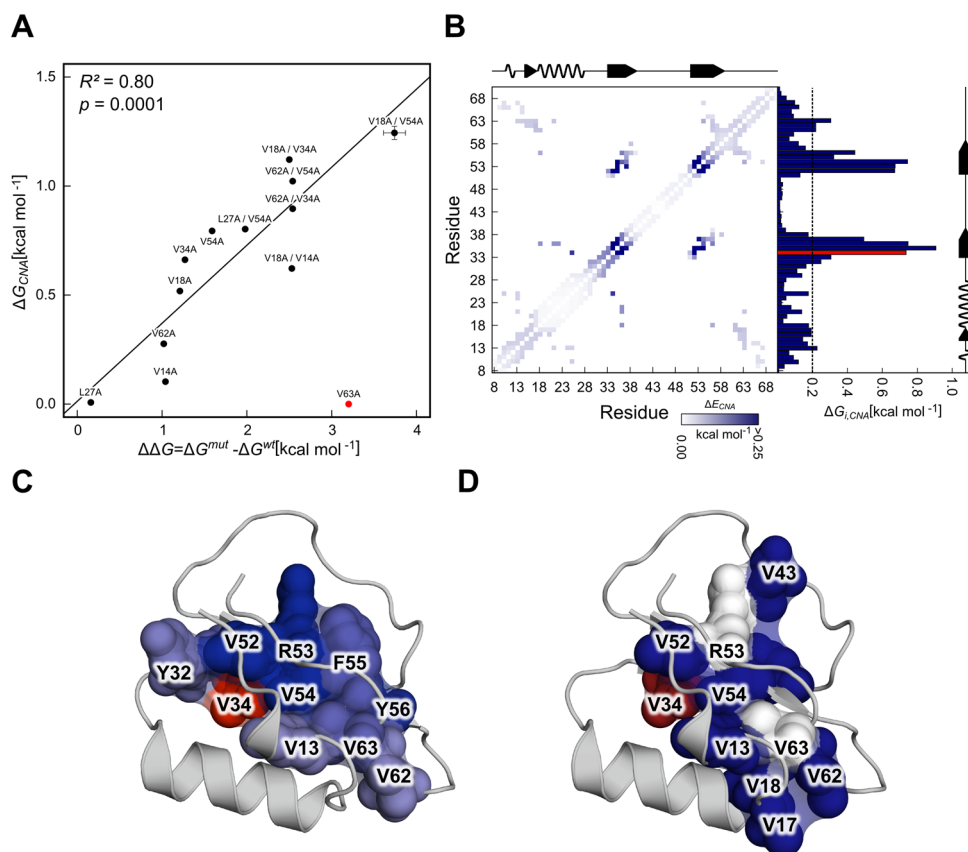


Figure 19: *In silico* mutational perturbation in eglin c. (A) Correlation between predicted ΔG_{CNA} 's and $\Delta\Delta G$'s from chemical denaturation experiments. The SEM is exemplary shown for V18A/V54A. The mutant V63A (red) is considered as outlier because this mutant has no effect upon *in silico* perturbation with subtle changes in dynamics as observed in NMR experiments (46). (B) Differences between stability maps of the ground (wild-type) and perturbed (V34A) state of eglin c. The attached histogram shows the per-residue $\Delta G_{i,CNA}$. Residues above a threshold of 0.2 kcal mol⁻¹ (dashed line in B) are mapped on the structure of eglin c (C) and are in good agreement with dynamically coupled residues as derived from NMR experiments (D) (46). The site of mutation (V34) is shown in red (C, D), blue colors show predicted $\Delta G_{i,CNA}$ values in (C) and white residues in (D) are those which are not showing any altered dynamics but are included in the network as suggested by the authors (46).

7.4 Probing V- and K-type allosteric mechanisms in PTP1B and LFA-1

Encouraged by the results obtained on eglin c, the perturbation approach was applied on PTP1B and LFA-1. The human PTP1B is part of the signaling transduction cascade leading to the phosphorylation of the insulin receptor. The LFA-1 domain is part of β_2 -integrin and binds to intercellular adhesion molecules (ICAM). While PTP1B possesses a V-type allosteric mechanism, LFA-1 possesses a K-type mechanism. In order to identify pathways for allosteric signaling, I calculated the per-residue $\Delta G_{i,CNA}$ from stability maps of the effector bound and perturbed constraint networks. From this, I identified multiple residues (“broad pathway”) which have pronounced altered rigidity characteristics (>0.2 kcal mol $^{-1}$) (Figure 20 A, B). Interestingly, in both systems the altered rigidity characteristics connect the perturbation site, i.e. allosteric site, with the orthosteric site. To determine the most likely pathways for allosteric signaling, I applied methods from the field of graph analysis to probe the allosteric communication in PTP1B and LFA-1. To this end, differences between stability

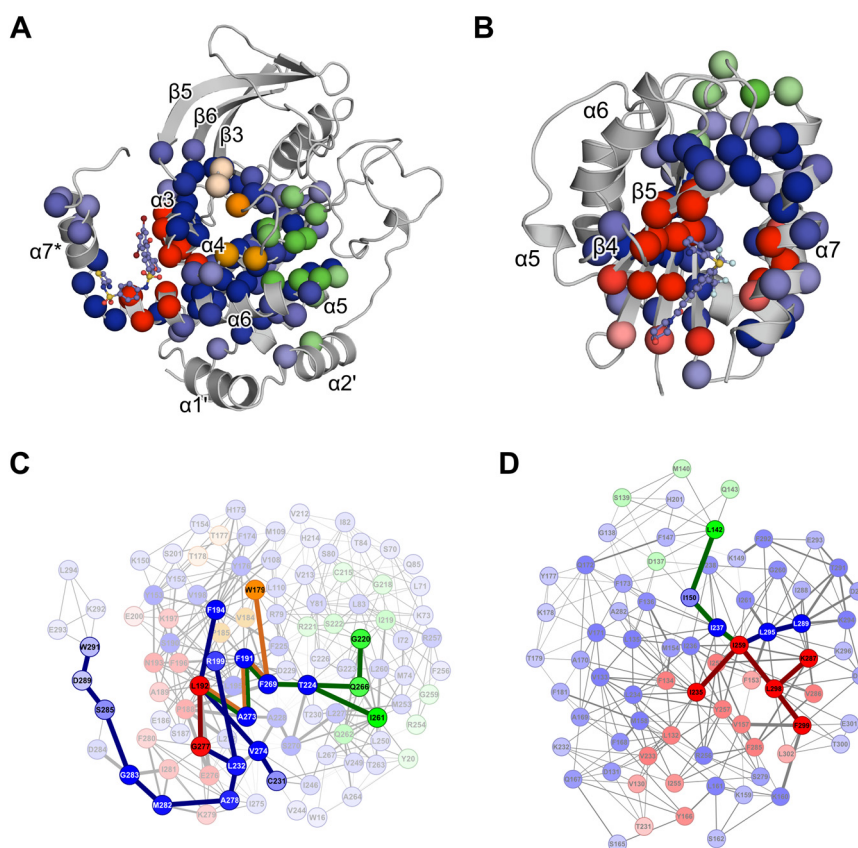


Figure 20: Probing the allosteric mechanism in PTP1B and LFA-1. Residues with strong altered rigidity are mapped on the structures of PTP1B (A) and LFA-1 (B). The color shows if a residue is part of the allosteric (red), orthosteric (green), WPD loop (orange, in PTP1B), or any other site (blue). Darker colors indicate stronger altered rigidity. The same information as in (A) and (B) is represented as stress-minimized graphs for PTP1B (C) and LFA-1 (D). The graph embedding is minimized based on the pairwise C_α distance. Highlighted residues show the highest allosteric communication through the graph as derived from the measure of the betweenness centrality.

maps in the ground and perturbed state were represented by a stress-minimized graph. In such a graph, nodes represent the residues and edges represent the *change* in the correlation of stability between pairs of residues. The graph embedding is minimized based on the pairwise C α distance. Next, centrality indices were applied to characterize the allosteric communication within the graph and to depict pathways for allosteric signaling in PTP1B and LFA-1. I found condensed clusters (“small pathway”) for allosteric signaling in both systems. Intriguingly, these clusters point to the orthosteric site or functionally important residues for allosteric regulation. These findings are striking because even if the underlying graph-theory based approach provides solely *static* information, i.e. information on flexibility and rigidity, the perturbation approach introduced here correctly identified relevant residues for known conformational changes.

7.5 Negative cooperativity in LFA-1 deduced from rigidity analyses

The allosteric cooperativity in LFA-1 has been analyzed following the formulation of Cooper and Dryden (section 2.1). In negative cooperativity, binding of the first molecule does not quench all dynamics and hence, the major loss of entropy must occur upon binding of the second molecule (24). An ensemble of conformations was extracted from a MD trajectory which used a modelled complex of LFA-1, BQM (allosteric effector), and ICAM-1 (natural substrate; including a magnesium ion) as a starting structure. From this, perturbed ensembles were generated for *apo* LFA-1, LFA-1/BQM, and LFA-1/ICAM-1 (Figure 21). Remarkably, mechanical perturbation free energies for LFA-1 upon binding of BQM or ICAM-1 shows a non-addictive character. This results in a cooperative free energy $\Delta\Delta G$ of 3.2 kcal mol⁻¹. In accordance with the formulation of entropy-driven cooperativity by Cooper and Dryden (section 2.1), the non-zero positive free energy of cooperativity indicates a rigidity coupling between the allosteric and orthosteric site

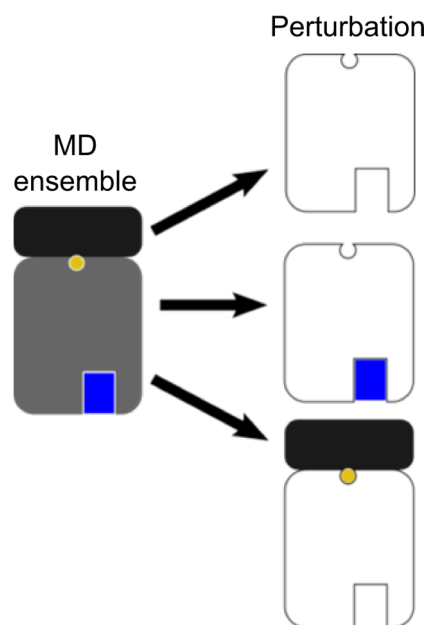


Figure 21: Probing the allosteric cooperativity in LFA-1. A structural ensemble was extracted from a MD trajectory starting from the modelled complex of LFA-1 (gray), allosteric effector BQM (blue), ICAM-1 (black), and magnesium ion (yellow). From this, perturbed ensembles of *apo* LFA-1, LFA-1/BQM, and LFA-1/ICAM-1 (including the magnesium ion) were generated.

in LFA-1, correctly predicted as a negative cooperativity in allosteric regulation.

7.6 Conclusion and significance

- In this publication, I presented an ensemble-based perturbation approach to analyze dynamically dominated allostery starting from a single-structure as input.
- For the first time, free energies of cooperativity and allosteric signaling were deduced from rigidity analysis.
- The perturbation approach benefits from two aspects: (I) robust results by averaging over *ensembles of conformations* instead of using single input structures (182, 204, 217). This is noteworthy, because the approach is still sensitive enough to depict the right allosteric effects even due to small modifications on network topologies; (II) a direct assessment of biomacromolecular flexibility and rigidity is provided, i.e. the results require no filtering of spurious correlations as necessary when analyzing correlated motions within MD trajectories.
- In all three tested systems, Eglin c, PTP1B, and LFA-1, the results demonstrated that long-range effects of these characteristics are present in biomacromolecules and, even more importantly, that the introduced perturbation approach is sensitive to effects related to allosteric regulation. The results pointed to residues that are most important for allosteric signaling by forming continuous pathways of altered stability in all three test systems.

Even though the ensemble-based perturbation approach was applied retrospectively, i.e. for systems with known allosteric mechanism, the approach provides a good starting point for the analysis of systems with yet unknown allosteric mechanism. To this end, I will outline a computational methodology (see chapter 11) that combines a pocket detection protocol (publication V and VI) and subsequent analysis of the allosteric response by using the perturbation approach introduced in this publication.

8 PUBLICATION V - Pocket-Space Maps to Identify Novel Binding-Site Conformations in Proteins

Craig, I.R., Pfleger, C., Gohlke, H., Essex, J.W., Spiegel, K.

J. Chem. Inf. Model. (2011), 51, 2666–2679

Original publication, see pages 160 - 193; contribution: 30%

8.1 Background

The precise knowledge about the diversity and novelty of binding pockets is of great importance in structure-based ligand design and enables to overcome key issues in drug discovery, such as potency, selectivity, toxicity, and pharmacokinetics. However, the examination of diverse pockets shapes in larger data sets is not possible in a routine way. In this publication (236), we reported an automated approach to select diverse binding pockets from structural ensembles by using the so called PocketAnalyzer^{PCA} approach. The core of this approach is the grid-based pocket detection program ‘PocketAnalyzer’, which identifies binding pockets in biomacromolecules. Identified binding pockets are post-processed in the PocketAnalyzer^{PCA} approach by applying principle component analysis (PCA) and clustering approaches. Since the approach works directly on pocket shape descriptors it overcomes the general problems of diverse pocket selections on proxy coordinates. The utility of the PocketAnalyzer^{PCA} approach was demonstrated by diverse sets of pocket shapes for aldolase reductase (ALR) and viral neuraminidase (236).

8.2 Workflow of the PocketAnalyzer^{PCA} approach

The overall workflow of the PocketAnalyzer^{PCA} approach consists of three steps (Figure 22): (I) the grid-based PocketAnalyzer program is applied on structural ensembles. The underlying algorithm of the PocketAnalyzer is a variant of the LIGSITE algorithm introduced by Hendlich *et al.* (237). Initially, a cubic grid is defined across the entire biomacromolecule and each grid point must meet a certain number of criteria to become part of a binding pocket. First, grid points must be sufficiently well enclosed within the biomacromolecule. Second, grid points must be surrounded by a certain number of other well-buried grid points. Finally, grid points that meet the first two criteria are clustered together and each identified cluster that is above the minimal cluster size is then classified as a binding pocket. In course of this study, the PocketAnalyzer program was applied on structural ensembles generated by MD

simulations but is applicable to any source of atomistic structural information. Next, a row vector is used to encode the inclusion (“1”) or exclusion (“0”) of grid points, representing the identified pockets of each structure in the ensemble. Since, the grid definition is the same for each structure, the row vectors are merged in a pocket shape matrix. Each column then describes, by the inclusion/exclusion of grid points, the diversity of pocket shapes. (II) The pocket shape matrix is subjected to PCA to reduce the dimensionality of the matrix. From PCA, the principal component (PC) eigenvectors gives the dominant deformation modes of the pocket, and the PC projections (or “score”) characterize the distribution of pocket shapes along the PCs. (III) Finally, the set of structures is reduced to a small subset by clustering according to their score along the PCs. The representative structure of each cluster then corresponds to a diverse selection of pocket shapes.

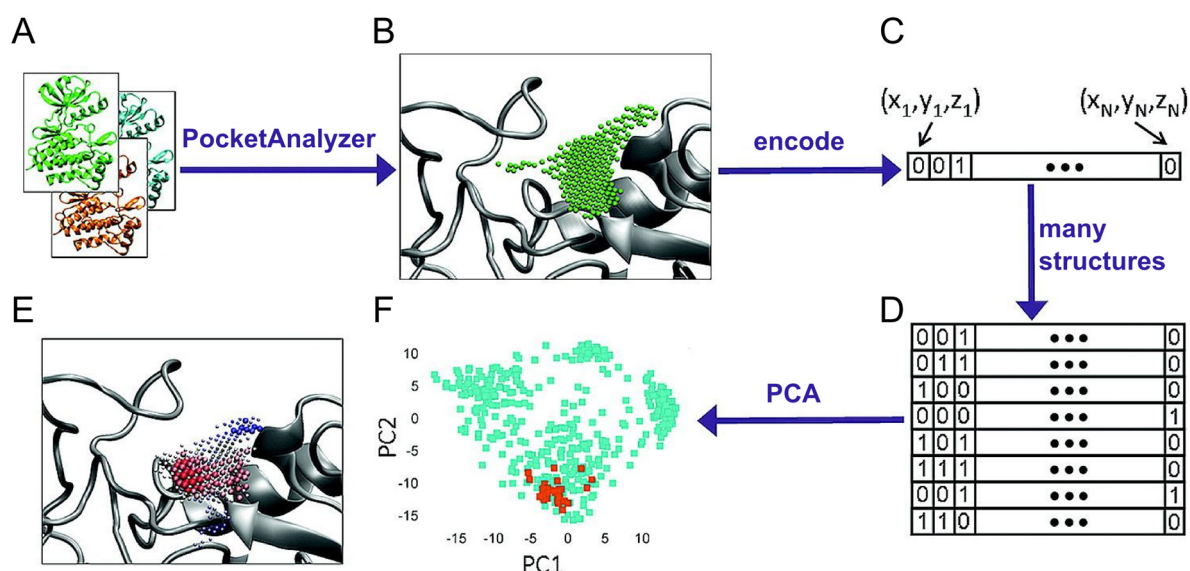


Figure 22: Workflow of the PocketAnalyzer^{PCA} approach. The PocketAnalyzer algorithm is applied to each structure in a structural ensemble (A). For each structure, the pocket shape (B) is encoded as a row vector for inclusion (“1”) or exclusion (“0”) of grid points. (C). A pocket shape matrix is created from merging the row vectors from each protein structure (D). Projecting the row vectors onto principal components derived from the pocket shape matrix generates a map of the pocket conformational space (E). The principal components describe the dominant changes in pocket shape within the set of protein conformations (F). Figure adapted from ref. (236).

8.3 PocketAnalyzer^{PCA} applied to aldose reductase and neuraminidase

The PocketAnalyzer approach was applied to aldose reductase (ALR) and neuraminidase. Both systems exhibit moderate but significant active site flexibility and have been extensively studied, resulting in a well-characterized set of binding pocket conformations. Accordingly, the systems are useful as benchmark systems for diverse pocket selection by PocketAnalyzer^{PCA}.

Mains results for diverse pocket selection on ALR: The aim was to compare pocket shapes derived from independent MD trajectories with those from available ALR crystal structures. From this, we correctly identified three of four crystallographically known pocket shapes. In addition, distinct pocket shapes were identified, which have not been observed experimentally. One striking example is the opening of a channel connecting the active site of ALR with a second cavity (Figure 23). None of the crystallographically available inhibitor-bound ALR structures shows a binding mode that addresses this second cavity. Therefore, we predicted the binding modes of known ALR inhibitors for which no experimental data is available. To this end, those inhibitors were aligned, which exhibit structural similarity to ligands with available crystallographic binding modes. Remarkably, the predicted binding modes of two ALR inhibitors require the opening of a subpocket, which is very similar to our predicted pocket shape.

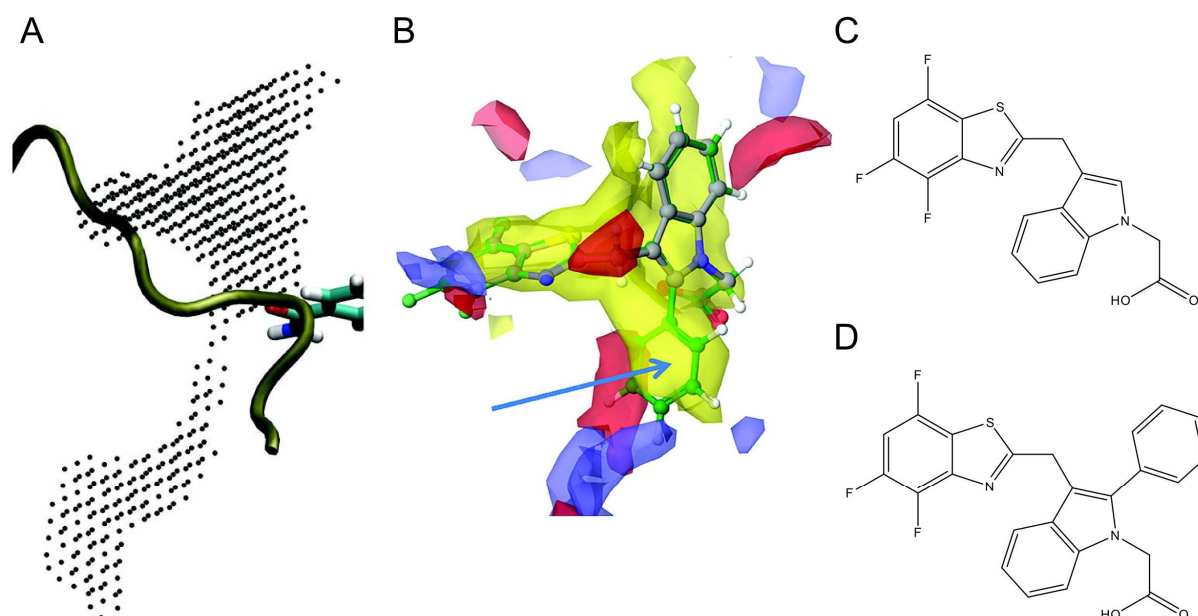


Figure 23: Diverse pocket selection on ALR. (A) Pocket shape as identified by PocketAnalyzer shows the open channel connecting a second cavity with the active site of ALR. (B) Drugability of the same region as in (A) derived from SiteMap analysis (238) showing the hydrophobic (yellow), hydrogen bond donor (blue), and hydrogen bond acceptor (red) fields. Additionally, (B) shows the crystallographic binding mode of lidorestat (C) (gray) and the predicted binding mode of a derivative (D) (green). Figure adapted from ref. (236).

Mains results for diverse pocket selection on neuraminidase: Neuraminidase is composed of three cavities referred to (I) the sialic acid (SA)-binding site, (II) the 150-cavity, and (III) the 430-cavity. The results from PocketAnalyzer^{PCA} show that the most variable region involves the 430-cavity and the 150-cavity, which is in agreement with findings in previous studies (239). Next, we focused on the opening of distinct subpockets in the SA cavity of neuraminidase. From this, we identified three novel pocket conformations, which exhibit the opening of distinct subpockets. As for ALR, known neuraminidase inhibitors were aligned on

crystallographically derived binding-modes of structurally similar ligands. Several predicted binding modes of inhibitors occupy the first subpocket, while none of the predicted binding modes matches the second subpocket. However, some support for the existence of this latter pocket conformation is provided by an earlier computational study (240). To our best knowledge, the third pocket conformation has not been previously observed in either experimental or computational study. None of the predicted binding modes of known neuraminidase inhibitors occupy this subpocket, which may provide a good starting point for the design of selective neuraminidase inhibitors considering this subpocket.

8.4 Conclusion and significance

- A computational methodology, called PocketAnalyzer^{PCA}, has been introduced which automatically describe the diversity of pocket shapes in structural ensembles.
- From a methodological point of view, the PocketAnalyzerPCA approach provides a novel and complementary perspective on protein dynamics that may prove particularly relevant for ligand binding and drug design
- The PocketAnalyzer approach is technically straightforward and allows simultaneous analysis of mutants, isoforms, and homologous proteins. In addition, the approach is applicable to any source of structural information, e.g. experimental data or computer-generated ensembles.
- By directly working on pocket shape descriptors, the approach can quickly highlights conserved and variable regions in the pocket, which is not possible by working on proxy coordinates.
- In a benchmark test, we demonstrated the performance of the PocketAnalyzer^{PCA} approach for the correct identification of known pocket shape conformations in ensembles generated by MD simulations.
- Most striking is the identification of a number of distinct pocket shapes that have not been observed experimentally and therefore represent novel computationally derived binding-site conformations.

As another important application is the identification of novel pockets, which is the first step in probing allosteric regulation in biomacromolecules with yet unknown mechanisms. Thus, its application will be a key element when combined with the CNA software package by predicting which sites of the biomacromolecules should be perturbed; i.e. to identify the location of the potential allosteric pockets.

9 PUBLICATION VI - Hot Spots and Transient Pockets: Predicting the Determinants of Small-Molecule Binding to a Protein-Protein Interface

§Metz, A., §Pfleger, C., Kopitz, H., Pfeiffer-Marek, S., Baringhaus, K.-H., Gohlke, H.

J. Chem. Inf. Model. (2012), 52, 120-133

Original publication, see pages 194 - 230; contribution: 35%

§ Both authors contributed equally to this work.

9.1 Background

The identification of allosteric binding sites is of great interest of current drug design efforts (section 2.5). However, the identification of allosteric binding pockets is challenging due to the occasionally transient character of allosteric pockets, which are possibly not present in determined X-ray structures (241, 242). In **publication V** (chapter 8), I introduced the pocket detection algorithm PocketAnalyzer, which is part of the PocketAnalyzer^{PCA} approach, and provides an excellent way to identify novel binding pockets along structural ensembles. In order to validate PocketAnalyzer, we applied the pocket detection protocol to protein-protein interactions (PPI), which are known as challenging drug-targets. Unlike enzymes, protein-protein interfaces often lack a distinct binding pocket and have very large interface sizes (89, 90). In this publication (243), we presented a computational strategy that considers aspects of energy and plasticity to identify the determinants of small molecule binding, hot spots as well transient pockets. We used interleukin-2 (IL-2) as a model system because crystallographically derived bindings modes of small molecules show the opening of a transient pocket in the protein-protein interface, which is not present in the *apo* IL-2 structure.

9.2 Overall strategy

The overall strategy consisted of four steps: (I) generation of structural ensembles by means of MD and constrained geometrical FRODA simulations, (IIa) detection of hot spots via MM-PBSA calculations on the ensemble generated by the MD simulation, (IIb) detection of transient pockets in conformational ensembles generated by MD and FRODA, (III) molecular docking of protein-protein interaction modulators (PPIMs) guided by the identified transient pockets and detected hot spot residues, (IV) ranking of the PPIMs via effective binding energies as predicted by the MM-PBSA approach.

Alexander Metz processed the energetic aspects in this study, namely, carrying out the MD simulations, identifying *hot spot* residues, and ranking the known interleukin-2 (IL-2) ligands. Additionally, he set up a decoy dataset to perform a retrospective virtual screening (VS) experiment. For this, he used the MM-PBSA 1-trajectory approach, to identify hot spot residues in the IL-2/IL-2R α complex and in five known small-molecule bound IL-2 structures. The analysis resulted in the identification of three out of five known *hot spots* (243) in IL-2/IL-2R α that are equally important for small-molecule binding.

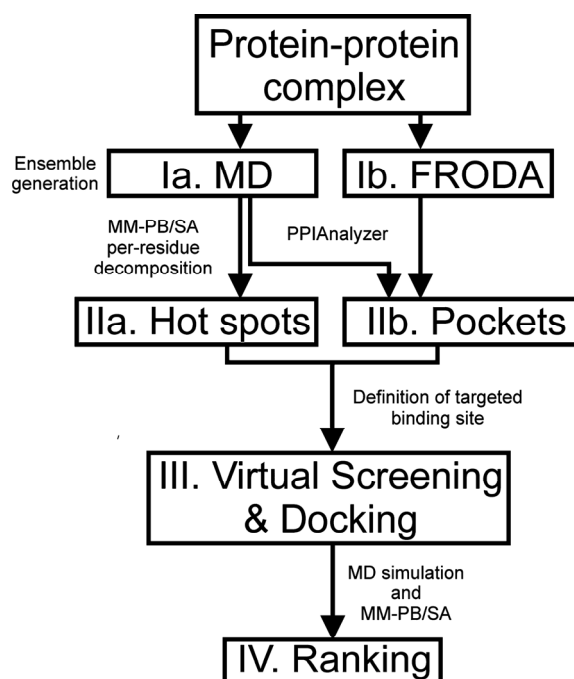


Figure 24: Methodology that uses hot spots and transient pocket prediction to guide docking and ranking. Figure taken from ref. (243).

My contribution to this study was running constrained geometrical FRODA simulations. This includes performing rigidity analyses via FIRST and using the resulting rigid cluster decomposition as an input for the FRODA simulation. Furthermore, I implemented the PPIAnalyzer method to investigate structural features of protein-protein interfaces. The PPIAnalyzer comprises three steps: (I) analyzing structural changes along ensembles generated by either MD or FRODA simulations, (II) testing the stereochemical quality of the generated structures and clustering them with respect to a so called ‘interface similarity’, (III) identifying potential pockets in the remaining snapshots with the help of the PocketAnalyzer program (chapter 8).

9.3 Identification of transient pockets

We used the PPIAnalyzer program to investigate the opening of transient binding pockets in the rather flat protein-protein interface of *apo* IL-2 (243). As mentioned, the sampling was performed (I) via state-of-the-art MD simulations and (II) via constrained geometrical FRODA simulations (189). FRODA relies upon the rigidity analysis performed by FIRST. To this end, FIRST decomposes the biomacromolecule into rigid and flexible regions (section 2.8) and correctly identifies those residues as rigid that are not involved in the pocket opening and those as flexible where the pocket opens. FRODA generates new conformations by a

random atom displacement, i.e. breaking all constraints and subsequent iterative fitting until all constraints are recovered. Notably, several snapshots from FRODA simulation come close to the small-molecule bound conformation of IL-2 ($< 1\text{\AA}$) which was not the case along the MD trajectory of a length of 10 ns.

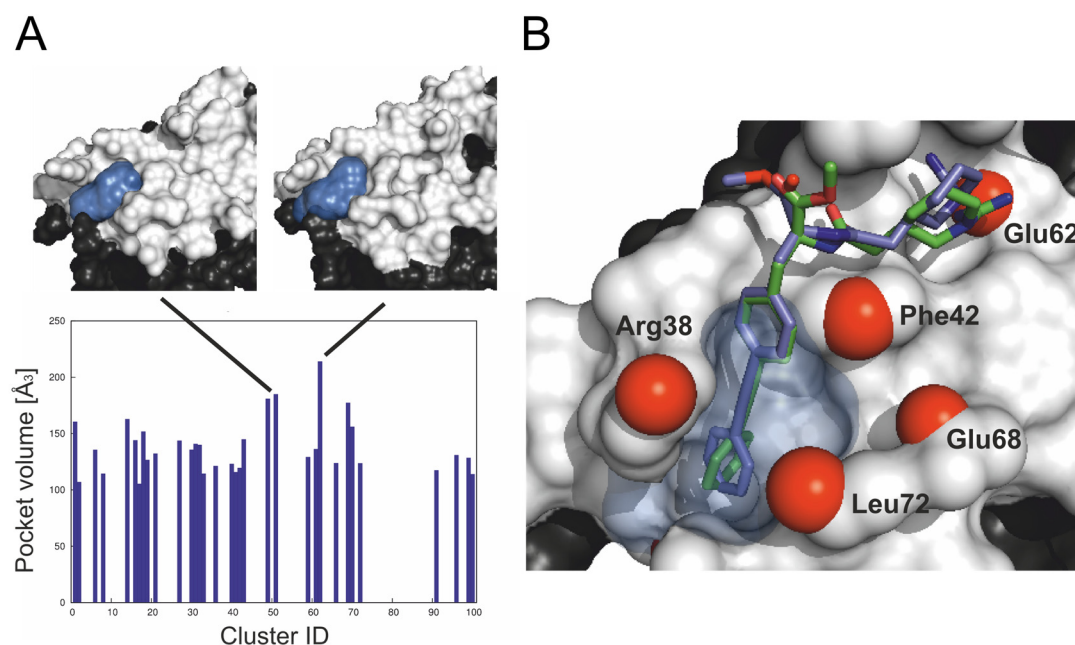


Figure 25: Transient interface pockets in IL-2. (A) Detection of interface pockets in the cluster representatives of IL-2 structures generated by FRODA simulation. The box plots depict pocket volumes computed by PocketAnalyzer. In addition, the two largest pockets found in IL-2 structures are shown. Predicted binding pose of known IL-2 ligand (blue sticks) docked into a FRODA snapshot containing an identified transient pocket. The root-mean-square deviation between the predicted and crystallographic binding pose (green sticks) is 1.28 Å. Figure adapted from ref. (243).

After clustering and identifying transient pockets in the interface region, we chose the 10 conformations with the largest pocket volume from both simulations, respectively. In a subsequent docking experiment, known IL-2 ligands were docked into the representative conformations and compared with the ligand pose present in the crystal structure. Remarkably, at least one conformation from FRODA simulation allowed successful docking in four out of five IL-2 ligands whereas docking into conformations from MD simulations failed in all five cases.

9.4 Conclusion and significance

- For the first time, we presented a computationally methodology that considers aspects of energetics and plasticity that facilitate the identification of small molecules to target PPI.
- Although retrospective in manner, at no step was *a priori* knowledge used from experiments to guide our analysis. The introduced computationally methodology

only requires structural information about a single protein-protein complex to develop PPIMs.

- Remarkably, our results showed that a computational, much cheaper, constrained geometrical simulation method FRODA outperforms state-of-the-art MD simulations in sampling transient pockets in the IL-2 interface. This finding also demonstrates that rigidity analyses by FIRST properly discriminate between residues that must be flexible for pocket opening while non-pocket residues are rather rigid.
- Docking of known IL-2 inhibitors to selected conformations from FRODA simulation results in predicted binding poses, which are in good agreement with crystallographically derived binding-modes. This result is intriguing because identified transient binding pockets and hot spots prediction solely guide the docking.

The significance here to my thesis is (I) that I demonstrated the usage of a coarse-grained geometric simulation method for sampling the opening of transient pockets in a rather flat protein-protein interface and (II) that I was able to identify transient binding pockets in a difficult target by using the PocketAnalyzer program.

10 SUMMARY

In this thesis, I developed an ensemble-based perturbation approach for gaining a deeper structure-based understanding of the relationship between changes in dynamics and allosteric behavior in biomacromolecules. To this end, the perturbation approach uses results from rigidity analysis to analyze dynamically dominated allostery. Contrary to other studies that use rigidity analysis, the introduced ensemble-based perturbation approach does not need (I) a robust filtering, as required in analyzing correlated dynamics (131) by the direct assessment of stability characteristics, nor (II) *a priori* knowledge of any experimental data for a protein-specific parameterization (212, 213). Furthermore, robust ensemble-based analyses instead of single-structure analyses (215) ensure consistent results that are unbiased by conformational effects from direct comparison of the active and inactive states (220).

In order to characterize dynamically dominated allostery from rigidity analyses, I presented indices of biomacromolecular flexibility and rigidity (I) that are sensitive for monitoring changes in stability even due to small modifications of the constraint network and (II) that enable linking mechanical stability with biomacromolecular dynamics (**publication I**). The comparison of the “ground” and perturbed state of α -lactalbumin allowed me to select indices, which are best for depicting long-range changes that are related to allosteric signaling. Remarkably, the local index definitions are able to detect long-range aspects of altered rigidity in α -lactalbumin, which are in good agreement with those from experimental studies.

In order to overcome the sensitivity problem of rigidity analyses, I developed the ENT^{FNC} method, which efficiently generate ensembles of network topologies by using definitions for fuzzy noncovalent constraints (**publication II**). The ENT^{FNC} approach demonstrates its high predictive power in analyzing local and global stability characteristics, which are in remarkable agreement with those from MD simulation-based ensembles as well as experimental studies. So far, this method for generating ensembles of network topologies is not parameterized for protein-ligand interactions and was not applied to probe dynamically dominated allostery.

The findings from **publication I** and **II**, contributed to the significant improvement of the *Constraint Network Analysis* (CNA) approach, leading to the user-friendly Python-based CNA software package (**publication III**). For the first time, CNA provides efficient ensemble-based analyses, automated calculation of global and local indices, and a robust treatment of ligand molecules to analyze biomacromolecular stability. The efficiency of the

ensemble-based CNA was achieved by linking CNA and the graph-theory based FIRST software via the *pyFIRST* interface module, which drastically reduces the calculation time from minutes to seconds for single structures of several hundred residues.

The CNA software is the core of the perturbation approach and was used to analyze changes in stability upon *in silico* perturbations of constraint networks (**publication IV**). Because network approaches are generally sensitive with respect to the structural information used as input, an ensemble-based variant of CNA was used (229). The perturbation approach was validated on eglin c, which is an ideal test case as experimental data is available (I) from chemical denaturation experiments providing free energies of single- and double-mutants and (II) from mutational perturbation experiments that revealed a continuous pathway of dynamically coupled residues (46, 47). Encouragingly, a good correlation ($R^2 = 0.80$) was obtained between predicted free energies and free energies of stability changes upon mutational perturbation in experiments. Furthermore, the pathway, as determined by the experimental data, could be reproduced in $\sim 85\%$ of the involved residues. Next, the approach was applied on PTP1B and LFA-1. *In silico* perturbations upon removing the allosteric effector from the network topology revealed long-range characteristics of altered rigidity in both systems. Although multiple residues (“broad pathway”) reveal altered rigidity characteristics, a condensed cluster (“small pathway”) has been identified to be important for allosteric signaling. Remarkably, in both systems, pathways connecting allosteric and orthosteric sites have at least two residues in between. Finally, the predicted free energy of cooperativity of LFA-1 is $3.2 \text{ kcal mol}^{-1}$, which corresponds to a non-additive stabilization in agreement with the underlying mechanism of negative cooperativity in LFA-1.

So far, the analysis was done retrospectively, e.g. for systems which are known to be allosteric regulated. In the perspectives (chapter 11), I outline a computational methodology for the detection of novel allosteric sites and design of “dummy” ligands, which are then used to probe the allosteric response by this perturbation approach. The preliminary results suggest that this methodology can become an efficient tool in drug discovery: It can analyze dynamically dominated allostery in a routine way and predict the mechanism of allosteric regulation.

11 PERSPECTIVES

In the course of this thesis, I developed a computational pipeline that allows for the identification of potential allosteric sites in biomacromolecules in an efficient and user-friendly way. Until now, the generation of the structural response was triggered by the presence of a bound ligand, which would not be possible in the case of yet unknown allosteric sites. First, to identify allosteric sites, biomacromolecules are sampled for potential pockets using the PocketAnalyzer program (publication V, chapter 8).

To date, the introduced perturbation approach was only tested on conformational ensembles, which were generated by MD simulations. This way, the approach benefits from an accurate all-atom simulation technique that samples a thermodynamic ensemble and provides consistent results by averaging results from rigidity analyses over an ensemble of conformations. However, MD simulations are still computationally expensive, and thus, inappropriate for large-scale studies. Another drawback of MD simulations is that they may fail to sample the opening of pockets in the case of biomacromolecules that follow an “induced fit” mechanism. The reason for this is the occasionally hydrophobic character of binding pockets, which are hard to detect in a classical MD water environment. Simulation techniques that do not consider solvation effects can overcome this problem. Accordingly, the constrained-geometric based FRODA method, but not an MD simulation, was able to open a transient pocket which was absent in the *apo* structure used as input (publication V, chapter 9).

If a potential pocket has been identified by PocketAnalyzer, it could be used to test the allosteric response. A graphical summary of this strategy is illustrated in Figure 26 A-C and will be described briefly in this paragraph. A grid map of an identified pocket is used for modelling a „dummy” ligand, which covers the entire binding site (Figure 26 A). Therefore, a probe atom moves along the grid and tests if atoms of the biomacromolecule are within a specific interaction distance (Figure 26 B). The used atom types can be varied but should include at least polar or hydrophobic atom types. Finally, grid points are removed that are not making any interaction with the biomacromolecule and do not decompose the grid. The remaining grid represents the „dummy” ligand (Figure 26 C), which is then merged with the network representation of the biomacromolecule by adding appropriate constraints. Thus, contrary to the methodology presented in publication III (chapter 5), perturbations of constraint network are done by adding extra constraints rather than removing them. Finally, the allosteric response will be tested by the ENT^{FNC} approach, which considers thermal

fluctuations without sampling conformations (chapter 5). Hence, the ENT^{FNC} approach efficiently analyzes altered thermal fluctuations caused by perturbing the constraint network of a single input structure. For this, the definitions for fuzzy noncovalent constraints will have to be extended by a parameterization for protein-ligand interactions.

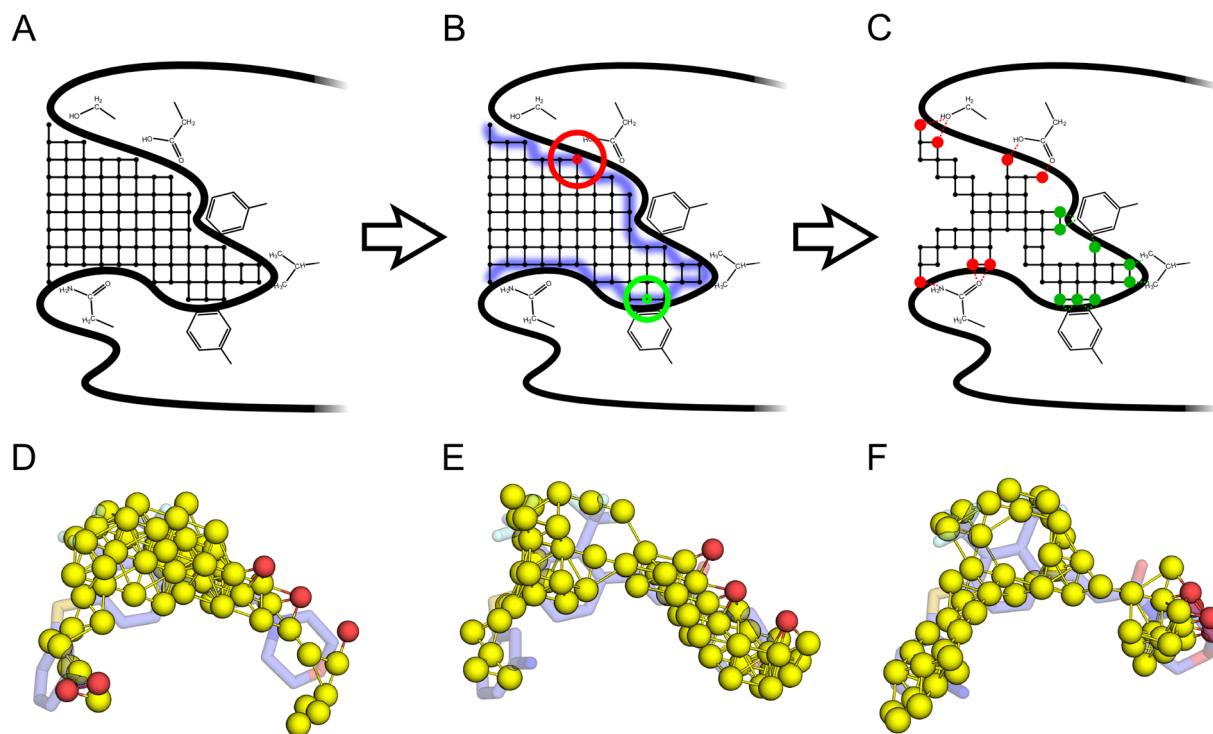


Figure 26: Illustration of modeling a „dummy“-ligand. (A) Initially, a grid map that represents a potential pocket is identified by the PocketAnalyzer program. (B) A probe atom moves along the grid and test whether atoms in the biomacromolecule are within polar (red) or unpolar (green) interaction distance. (C) A type of „dummy“ ligand is obtained by removing atoms from the grid which are not making any interactions with the biomacromolecule and do not decompose the grid. (D-E) Three exemplary „dummy“ ligands derived from a preliminary study. The spheres indicate either carbon (yellow) or oxygen (red) atoms. The actual pose of the allosteric effector BQM is shown as sticks.

In a preliminary study to test the general feasibility of this approach, PocketAnalyzer was applied to conformations extracted from a MD trajectory of LFA-1 in complex with the allosteric effector BQM. The usage of the effector bound LFA-1 state should guarantee the identification of distinct pockets at the allosteric site. In the subsequent perturbation approach, constraints of calculated „dummy“ ligands were merged with the network of each single conformation of the trajectory. Remarkably, calculated „dummy“ ligands quite well reflect the actual ligand poses and molecular features as shown in Figure 26 D-F. Although it turned out that perturbations by „dummy“ ligands over-stabilize the LFA-1 structure, the results from the perturbation approach are already encouraging as they agree with those presented in publication IV (chapter 5) (results not shown).

At this point, the analysis is still preliminary but provides an interesting starting point for a strategy to identify drug-like molecules. Knowledge-based potential fields such as in

DrugScore (244, 245) can be used to assign different atom types to the nodes in the grid object. Accordingly, each grid object that belongs to a novel binding pocket and is coupled with a functionally important site of the biomacromolecule can be used, e.g. for screening based on pharmacophoric features. Overall, the developed perturbation approach and the here outlined methodology provides an important starting point for modern drug-design by (I) detection of novel allosteric sites and (II) identification of potential lead compounds for subsequent optimization.

12 ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisor Prof. Dr. Holger Gohlke for enabling me the doctorate in his working group and for his guidance throughout the work. Furthermore, I am deeply grateful for numerous encouraging and fruitful discussions and his never ending enthusiasm, which constantly encouraged me during the time of my PhD.

I am very grateful to Dr. Sebastian Radestock and Dr. Prakash Chandra Rathi for the joint work on the CNA software package. In this context, I would also like to thank Doris L. Klein for her constructive contributions to the development of CNA and significantly improved the program. The cooperative work always led to very fruitful discussions about rigidity analysis. I am also grateful to Teresa Jimenez Vaquero for providing the initial version of the grid-based pocket detection algorithm, which finally became part of the PocketAnalyzer software.

Special thanks to Dr. Alexander Metz, Dr. Prakash Chandra Rathi, Doris L. Klein, Dr. Ian R. Craig, and Hannes Kopitz for the joint research and shared (co-)authorships in several publications. For the permissions to reprint the publications in this thesis, I thank to the American Chemical Society, John Wiley & Sons, Inc., and Elsevier Ltd. A special thank goes to Daniel Mulnaes for the accurate proof reading of manuscripts and always fruitful discussions that led to a significant improvement of my work.

I am very grateful to Christian Hanke, Dr. Simone Fulle, Daniel Mulnaes, and Daniel A. Cashman for critically reading the manuscript, suggestions and discussions related to this thesis.

Finally, I would like to thank all those people that accompanied my way at the Goethe University, Frankfurt, Christian-Albrechts University, Kiel, and Heinrich-Heine University, Düsseldorf. Most notably, I would like to thank my parents for supporting me in every circumstance and in all my decisions.

13 PUBLICATIONS

13.1 Reprint permissions for publications

Publication I (page 70 to 92):

Reprinted from "*Global and local indices for characterizing biomolecular flexibility and rigidity*", Christopher Pfleger, Sebastian Radestock, Elena Schmidt, Holger Gohlke, Journal of Computational Chemistry, (2013), 34, 220-233, DOI: 10.1002/jcc.23122, Copyright (2012), with permission from Wiley Periodicals, Inc.

Publication II (page 93 to 117):

Reprinted from "*Efficient and Robust Analysis of Biomacromolecular Flexibility Using Ensembles of Network Topologies Based on Fuzzy Noncovalent Constraints*", Christopher Pfleger, Holger Gohlke, Structure (2013), 21, 1725–1734, DOI: 10.1016/j.str.2013.07.012, Copyright (2013), with permission from Elsevier Ltd.

Publication III (page 118 to 127):

Reprinted from "*Constraint Network Analysis (CNA): A Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo)stability, and function*", Christopher Pfleger, Prakash Chandra Rathi, Doris L. Klein, Sebastian Radestock, Holger Gohlke, Journal of Chemical Information and Modeling, (2013), 53, 1007-1015, DOI: 10.1021/ci400044m, Copyright (2013), with permission from the American Chemical Society.

Publication IV (page 128 to 159):

"*Allosteric Coupling deduced from Altered Rigidity Percolation in Biomacromolecules*", Christopher Pfleger, Alexander R. M. Minges, Holger Gohlke (submitted manuscript)

Publication V (page 160 to 193):

Reprinted from "*Pocket-Space Maps To Identify Novel Binding-Site Conformations in Proteins*". Ian R. Craig, Christopher Pfleger, Holger Gohlke, Jonathan W. Essex, Katrin Spiegel, Journal of Chemical Information and Modeling, (2011), 51, 2666-2679, DOI: 10.1021/ci200168b, Copyright (2011), with permission from American Chemical Society.

Publication VI (page 194 to 230):

Reprinted from "*Hot Spots and Transient Pockets: Predicting the Determinants of Small-Molecule Binding to a Protein-Protein Interface*", Alexander Metz, A., Christopher Pfleger, Hannes Kopitz, Stefania Pfeiffer-Marek, Karl-Heinz Baringhaus, Holger Gohlke, Journal of Chemical Information and Modeling, (2012), 52, 120-133, DOI: 10.1021/ci200322s, Copyright (2011) with permission from American Chemical Society.

13.2 Publication I

Global and Local Indices for Characterizing Biomolecular Flexibility and Rigidity

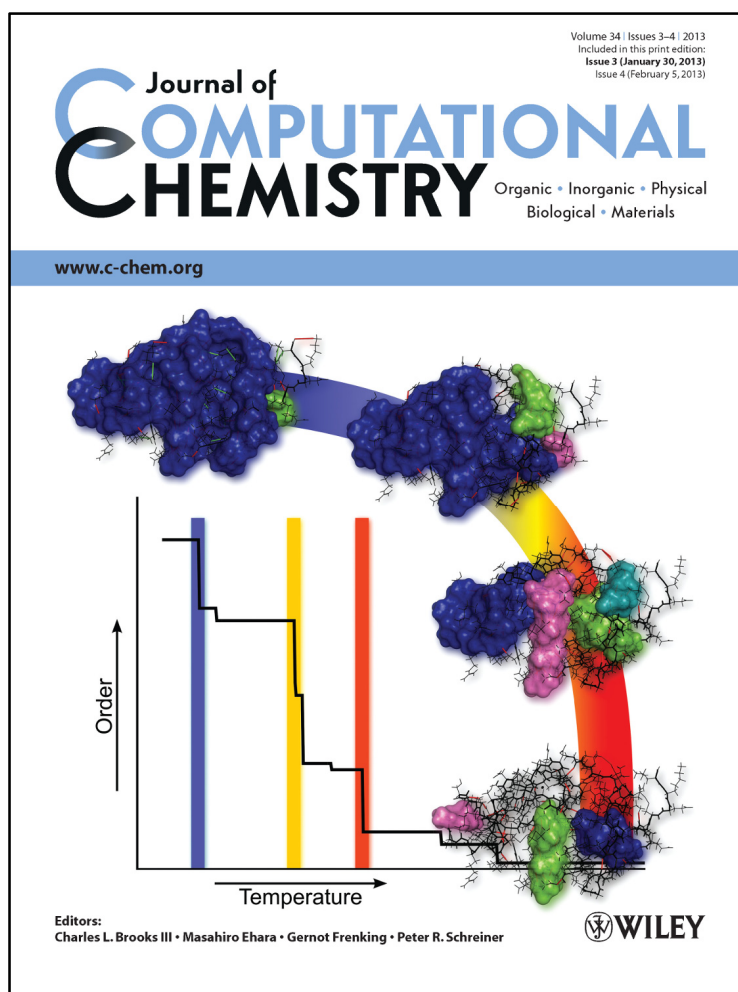
Pfleger, C., Radestock, S., Schmidt, E., Gohlke, H.

J. Comput. Chem. (2013), 34, 220–233

Author contribution to the publications:

My contribution to this publication was to identify overlapping or identical index definitions from literature research, to investigate the scope and limitations of each index, and to develop new index definitions that extend the applicability domain of flexibility and rigidity analysis. From these findings, I provided precise guidelines for the usage of the indices in a showcase example on α -lactalbumin. This results in a contribution of **40%** to this publication.

Cover article:



Global and Local Indices for Characterizing Biomolecular Flexibility and Rigidity

Christopher Pfleger,^[a] Sebastian Radestock,^[a] Elena Schmidt,^[b] and Holger Gohlke*^[a]

Understanding flexibility and rigidity characteristics of biomolecules is a prerequisite for understanding biomolecular structural stability and function. Computational methods have been implemented that directly characterize biomolecular flexibility and rigidity by constraint network analysis. For deriving maximal advantage from these analyses, their results need to be linked to biologically relevant characteristics of a structure. Such links are provided by global and local measures ("indices") of biomolecular flexibility and rigidity. To date, more than 14 indices are available with sometimes overlapping or only vague definitions. We present concise definitions of these indices, analyze the relation between, and the scope and limitations of them, and compare their informative value. For this, we probe the structural stability of the calcium binding protein α -lactalbumin as a showcase, both

in the "ground state" and after perturbing the system by changing the network topology. In addition, we introduce three indices for the first time that extend the application domain of flexibility and rigidity analyses. The results allow us to provide guidelines for future studies suggesting which of these indices could best be used for analyzing, understanding, and quantifying structural features that are important for biomolecular stability and function. Finally, we make suggestions for proper index notations in future studies to prevent the misinterpretation and to facilitate the comparison of results obtained from flexibility and rigidity analyses. © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23122

Introduction

Flexibility and its opposite, rigidity, are important characteristics of protein stability and function.^[1–3] As such, the mechanical heterogeneity of proteins is a prerequisite for proper enzyme function. Consequently, the distribution of flexible and rigid regions is highly conserved within homologous proteins.^[3–5] Likewise, orthologs from meso- and thermophilic organisms are in states of corresponding flexibility at their respective working temperatures.^[6] Being able to identify flexible and rigid regions as well as changes in the flexibility/rigidity on changes in a protein's environment, for example, due to binding of a ligand, temperature or solvent change, is essential for understanding protein stability and function. From an application point of view, such information provides a means for optimizing protein stability and function by rational protein engineering^[6–9] and is valuable for structure-based ligand design.^[10–12]

Flexibility and rigidity are static properties that denote the possibility of motions but do not give any information about directions and magnitudes of actual motions. The flexibility of a biological macromolecule is typically characterized by measuring motions of the structure using techniques such as hydrogen/deuterium (H/D) exchange, neutron scattering, and relaxation measurements by nuclear magnetic resonance (NMR) spectroscopy.^[3] However, the experimental characterization and quantification of biomolecular flexibility remains challenging, especially with respect to the diverse types and time-scales of motions.^[2,3] Moreover, it remains difficult to determine biomolecular flexibility that is related to function because such flexibility can be limited to small but crucial

parts of the structure.^[13] Thus, computational methods provide a valuable complement for analyzing flexibility and rigidity of biomolecules.

On the one hand, methods are applied that analyze conformational ensembles, either determined by crystallography or NMR spectroscopy or generated by molecular dynamics (MD) simulations.^[14] The outcome of these methods obviously depends on the number and diversity of states in the ensemble. On the other hand, methods have been devised that identify flexible and rigid regions from a single input structure, for example, by determining the spatial variation in the local packing density,^[15] or by representing the structure as a connectivity network of interacting residues or atoms.^[16–21]

Following another network concept, protein structures are modeled as constraint networks (molecular frameworks), where vertices represent atoms and edges represent covalent bonds and angles.^[22] For accurately modeling biomolecular flexibility, noncovalent interactions must also be included in this network.^[23–26] Flexible and rigid regions are then determined from the number and spatial distribution of bond-rotational degrees of freedom.^[27,28] The Pebble Game algorithm,^[27–29] implemented in the Floppy Inclusion and Rigidity

[a] C. Pfleger, S. Radestock, H. Gohlke

Department of Mathematics and Natural Sciences, Institute for Pharmaceutical and Medicinal Chemistry, Heinrich-Heine-University, Düsseldorf, Germany
E-mail: gohlke@uni-duesseldorf.de

[b] E. Schmidt

Department of Biological Sciences, Goethe-University, Frankfurt am Main, Germany

© 2012 Wiley Periodicals, Inc.

Substructure Topology (FIRST) software,^[23] efficiently assigns each bond as either being part of a flexible or a rigid region. A rigid region results from a collection of interlocked bonds that have no relative motion. If the rigid region has redundant constraints, it is overconstrained. Otherwise, it is isostatically rigid. In a flexible region, the dihedral rotation of one bond is not locked in by other bonds. The theory underlying this approach is rigorous,^[30] and the parameterization for modeling the constraint network has frequently and successfully been applied in various analyses of biomolecules.^[31–34] In addition, comparisons between the constraint network and dynamical approaches have been performed in the past. These include the analysis of changes in the flexibility of proteins on complex formation by constraint network analysis (CNA) and MD simulations^[26] as well as a large-scale comparison of protein essential dynamics from MD simulations and coarse-grained normal mode analyses,^[35] which use information from CNA as input.^[34]

On diluting constraints in a network, a phase transition occurs at which the network loses its ability to carry stress.^[36] In particular, diluting constraints that model noncovalent interactions in a protein allows simulating the thermal unfolding of the biomolecule.^[6,7,24,37–39]

Linking results from CNA to biologically relevant characteristics of a structure is key for deriving maximal advantage from information on biomolecular flexibility. In this context, biologically relevant characteristics are, for example, the (macroscopic) phase transition point where a biomolecule switches from a structurally stable (largely rigid) state to an unfolded (largely flexible) state or the (microscopic) localization and distribution of structurally weak parts. As links, global and local measures ("indices") of biomolecular flexibility and rigidity are applied. To date, more than 14 indices have been introduced, which are summarized in Supporting Information, Table S1. A subset of four global and four local indices that are particularly useful will be described in more detail below and will be related to case studies where the indices have been successfully applied for explaining and interpreting experimental findings. Sometimes, only vague index definitions have been provided from the original authors, leading to similar or identical indices being used under different names. In other cases, the scope of an index is limited to only a subdomain of flexibility and rigidity analysis. Finally, a comparison of indices with respect to their informative value is elusive except for a comparative study of metrics used within the distance constraint model (DCM).^[9]

Thus, this study aims at (i) providing concise definitions of the indices, (ii) analyzing the relation between, and the scope and limitations of, indices, and (iii) comparing their informative value. We also introduce three new indices that allow extending the applicability domain of flexibility and rigidity analysis to understanding and improving thermostability, analyzing flexibility changes on complex formation and mutations, and investigating how flexibility information is transmitted between sites in a protein. The majority of these indices can be computed by the CNA package^[40] developed in our laboratory, which functions as a front- and backend to the FIRST

software. As a test system, we analyze the structural stability of the calcium binding protein α -lactalbumin both in the "ground state" and after perturbing the system by changing the network topology. Furthermore, we concentrate on monitoring changes in a network along an unfolding trajectory rather than investigating a single static network state. With these showcase analyses, we intend to provide guidelines for future studies suggesting which of these indices could best be used for analyzing, understanding, and quantifying biologically relevant characteristics that are important for protein stability and function.

Materials and Methods

Structure preparation

The structure of α -lactalbumin determined by X-ray crystallography to 1.7 Å resolution^[41] was taken from the Protein Data Bank (PDB code 1HML).^[42] The quality of the structure was checked using the PDBREPORT database.^[43] Hydrogens were added by the REDUCE program,^[44] and, where necessary, Asn, Gln, or His side chains were flipped. The secondary structure information of α -lactalbumin is summarized in Table 1.

Table 1. Domain and secondary structure information for human α -lactalbumin.

Domain or secondary structure name	Residues
α -domain	K1—T38, D83—L123
β -domain	Q39—D82
Helix A	K5—L11
Helix B	L23—S34
Helix C	T86—K98
Helix D	A106—L110
Strand S1	I41—E43
Strand S2	T48—Y50
Strand S3	I55—S56

Constraint network construction

To construct the constraint network, only the protein molecule was used, whereas water and buffer ions were removed. We decided to not include water molecules in the network based on previous findings^[26,45] that showed only a negligible difference in the flexibility characteristics of proteins when structural waters were considered. In fact, when we performed thermal unfolding simulations of α -lactalbumin with and without crystal water (in the former case, only those water molecules with a distance < 3.5 Å to a protein atom were considered), the computed indices did not change (data not shown). However, we note that in certain cases water can have a pronounced effect on a protein's flexibility.^[23] The molecular network of the covalent and noncovalent bond constraints present within the protein was constructed using the FIRST software (version 6.2).^[23] In addition, metal ions were retained when they were part of the structure. Here, a calcium ion is coordinated to the two backbone carbonyl oxygens of K79 and D84, and to the three carboxylate side chains of D82, D87, and D88. Additionally, a zinc ion is coordinated to the carboxyl group of E49. Interactions between ions and protein atoms were treated as covalent bonds and inserted manually.

Hydrogen bonds, salt bridges, and hydrophobic contacts were considered as noncovalent bond constraints as described previously.^[40] Hydrogen bond energies E_{HB} were calculated using an empirical potential.^[46] Hydrophobic interactions between carbon or sulfur atoms were taken into account if the distance between these atoms was less than the sum of their van der Waals radii (C: 1.7 Å, S: 1.8 Å) plus 0.25 Å.^[24] That way, the influence of solvent on the protein stability is considered implicitly.

Then, flexible and rigid regions within the constraint network are identified by FIRST. The algorithm and the underlying theory have been described elsewhere.^[27–29]

Constraint network analysis

CNA has been developed in our laboratory with the aim to analyze the structural features that are important for a biomolecule's stability. CNA functions as a front- and backend to FIRST by (i) setting up a variety of constraint network representations for rigidity analysis, (ii) processing the results obtained from FIRST, and (iii) computing indices for characterizing molecular stability both globally and locally. With respect to (i), CNA provides a method to simulate the thermal unfolding of a protein structure by removing noncovalent constraints from the network in the order of increasing strength.^[6,7] In this study, only hydrogen bonds (including salt bridges) were removed. For each threshold value E_{cut} , a new network state $\sigma = f(T)$ was generated where only hydrogen bonds with an energy $E_{HB} \leq E_{cut}$ were considered. This follows the idea that weaker hydrogen bonds will break at lower temperatures than stronger ones. Each network σ was then decomposed into flexible and rigid regions, producing a thermal unfolding trajectory. The number of hydrophobic contacts was kept constant during the simulation. This is motivated by the fact that hydrophobic interactions actually remain constant or even become stronger when the temperature increases.^[47,48] We performed two thermal unfolding simulations of α -lactalbumin, one in which the calcium ion was included and another one in which the ion was removed from the constraint network. The zinc ion was present in both simulations. We note that we applied CNA to only a single structure in the present study but that it can be applied to an ensemble of conformations, too, then yielding averaged index values.^[39]

Results

Unfolding simulation of α -lactalbumin

α -Lactalbumin is a metallo protein that regulates the lactose biosynthesis by modulating the specificity of galactosyltransferase. The protein contains two distinct ion binding sites: one site is located at the connection of α - and β -domain (K79–D88) and binds a calcium ion; the second site is located at the active site of α -lactalbumin and binds a zinc ion. Several studies have demonstrated that the binding of calcium affects the function of α -lactalbumin,^[49–52] whereas the binding of zinc has only a negligible effect to the structure.^[52] It has been suggested that binding of calcium induces a conformational transition and enhances the thermal stability of α -lactalbumin.^[49] Accordingly, the removal of calcium causes a transition from a well-ordered, rigid to a less-ordered, floppy binding site.^[52]

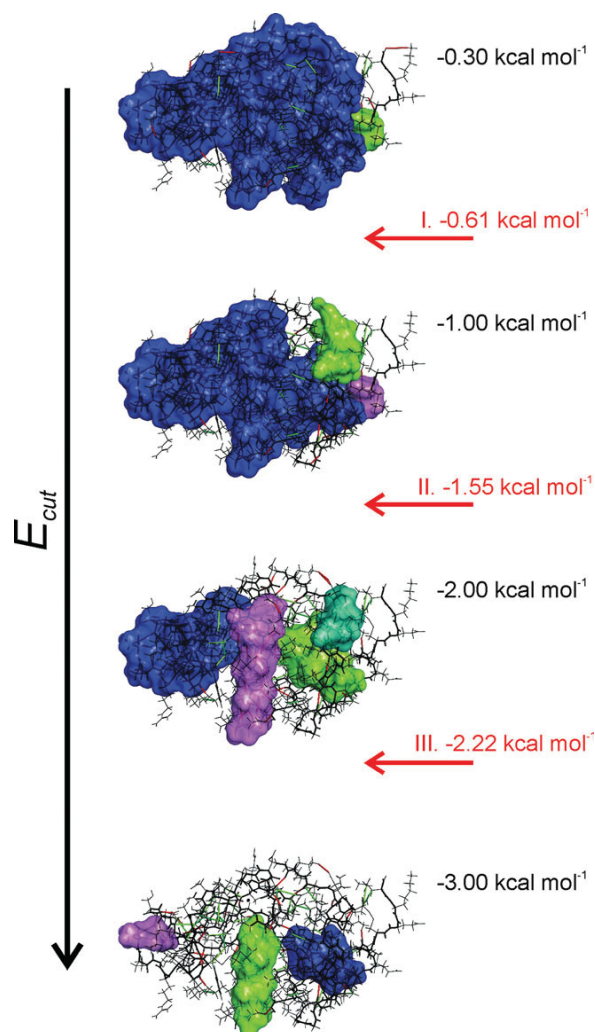


Figure 1. Rigid cluster decomposition along the thermal unfolding trajectory of α -lactalbumin. Rigid clusters are depicted as uniformly colored bodies. The blue body at -0.3 to -3.0 kcal mol⁻¹ represents the giant cluster. Hydrogen bonds and hydrophobic contacts are shown as red and green rods, respectively. I, II, and III indicate the three steps of the rigidity percolation. Figures have been made using PyMOL.^[53] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

To relate information about α -lactalbumin's mechanical flexibility and rigidity to thermal stability, a thermal unfolding of the structure was simulated. Simulating the thermal unfolding of α -lactalbumin with CNA requires less than 2 min on a standard workstation computer. Snapshots from the unfolding trajectory of α -lactalbumin are depicted in Figure 1. They show the loss of rigidity in terms of the decay of rigid clusters with increasing temperature. Arrows point to states where the network undergoes a transition as indicated by a sudden drop in rigidity. The first transition relates to the beginning of the collapse of the largest rigid cluster, which is located in the α -domain of α -lactalbumin. Still, the cluster continues to dominate the network after this transition. At the second transition point, the cluster stops dominating the network and breaks

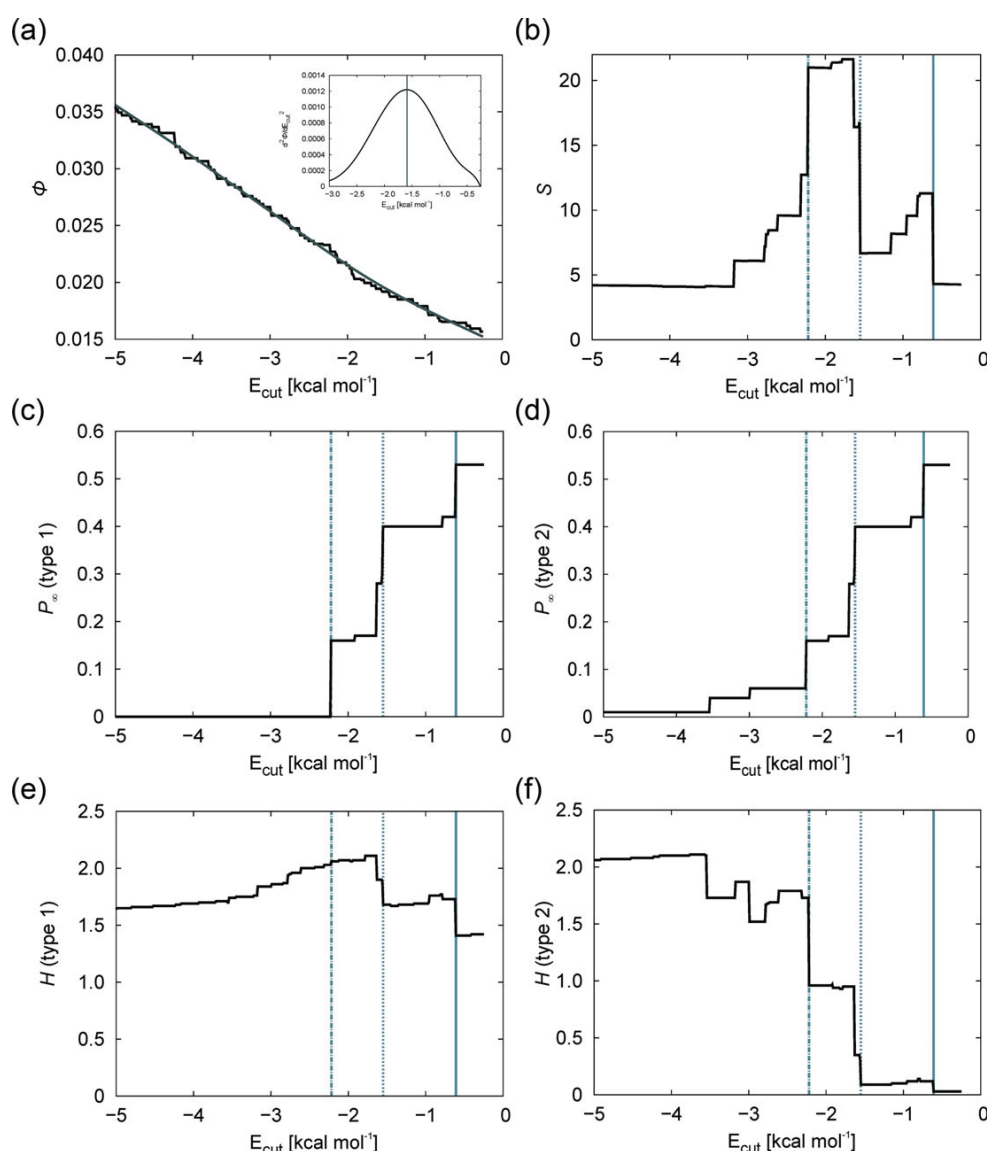


Figure 2. Global indices for the thermal unfolding of α -lactalbumin as a function of the hydrogen bonding energy cutoff E_{cut} : a) floppy mode density (the inset depicts the second derivative), b) mean cluster size, rigidity order parameter, c) type 1, and d) type 2, cluster configuration entropy, e) type 1, and f) type 2. The vertical lines correspond to the three transitions as depicted in Figure 1. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

into smaller rigid components. Thus, rigidity ceases to percolate through the structure, that is, the structure is not able to transmit stress anymore after this transition. Finally, during the last transition, the rigid cluster capturing the β -domain of α -lactalbumin collapses, and nearly the whole system becomes flexible. We will use these transition points for the evaluation and comparison of computed global and local indices.

Global flexibility indices

Floppy mode density. Global flexibility indices monitor the degree of flexibility and rigidity within the constraint network at a macroscopic level. One such global index is the number of inter-

nal independent degrees of freedom (floppy modes, F) that are associated with dihedral rotations in a network.^[24,36] By monitoring F during a thermal unfolding simulation, a phase transition can be detected that relates to the transition from an amorphous, rigid to a polymeric, glassy state in random networks^[36] as well as from a structurally stable (rigid) to an unfolded (flexible) state in proteins.^[24] Usually, F is normalized by the number of (overall) internal degrees of freedom associated with the N atoms, resulting in a floppy mode density [Φ , eq. (1)]

$$\Phi = \frac{F}{6N - 6} \quad (1)$$

Figure 2a shows the change in Φ during the thermal unfolding simulation of α -lactalbumin as a function of E_{cut} . E_{cut} relates to

the mean coordination number $\langle r \rangle$ in the constraint network in that $\langle r \rangle$ is a monotonic function of E_{cut} . Previously, $\Phi = f(\langle r \rangle)$ has been used in such analyses,^[24] predominantly in the case of network glasses.^[54–56] However, for applications to biological systems, $\Phi = f(E_{\text{cut}})$ seems to be a more intuitive choice because E_{cut} can be related to the temperature of the protein system.^[6] In fact, the shape of $d^2\Phi/dE_{\text{cut}}^2$ (inset in Fig. 2a) resembles the curve of the specific heat for a phase transition of second order,^[24,57,58] and its maximum at -1.59 kcal mol⁻¹ corresponds to the second transition in Figure 1. The $d^2\Phi/dE_{\text{cut}}^2$ curve also exhibits a weak shoulder around -0.5 to -1.0 kcal mol⁻¹, which is related to the earlier transition.

Mean rigid cluster size. Originating from percolation theory, moments of the size distribution of rigid clusters (i.e., the microstructure of the network) can be used to analyze macroscopic properties of constraint networks.^[59] In this context, the change of the mean rigid cluster size S can be monitored during the thermal unfolding simulation, with the size of the largest rigid cluster always being excluded from the calculation (Supporting Information, Table S1). This leads to S being zero as long as one rigid cluster dominates the whole network or if all rigid clusters have vanished. Figure 2b shows the changes of S for α -lactalbumin as a function of E_{cut} . Here, S is not zero at high E_{cut} because even then the largest cluster does not cover the whole network. By removing hydrogen bond constraints from the network, the rigid cluster starts to decay, which leads to a steep increase of S at -0.61 kcal mol⁻¹. At this point, the system is still dominated by one large rigid cluster (see also Fig. 1). In the following, the mean cluster size only monitors the collapse of smaller rigid components that have segregated from the largest cluster. The second transition then highlights the (loss of) rigidity percolation at -1.55 kcal mol⁻¹, which matches the second transition in Figure 1. At this point, the largest rigid cluster collapses as reflected by the steepest increase of the curve in Figure 2b. After a plateau phase, S starts to decrease due to the fact that the network is less and less dominated by rigid components.

Rigidity order parameter. As another index originating from percolation theory and derived from the microstructure of a constraint network, the rigidity order parameter P_{∞} has been used to identify transition points during thermal unfolding.^[6,7,54,59–61] Here, the fraction of the network belonging to the giant percolating cluster (type 1) or the actual largest rigid cluster (type 2) is chosen as an order parameter (Supporting Information, Table S1). That is, compared to S , the extremum of the cluster size distribution is considered. The giant percolating cluster is defined as the largest rigid cluster present at high E_{cut} values with all constraints in place (i.e., at low temperatures). As long as the network is in the rigid phase, it is dominated by one rigid cluster and, hence, P_{∞} is close to one. In the floppy phase, with a vanishing largest rigid cluster, P_{∞} is zero.

Figures 2c and 2d show P_{∞} for type 1 and type 2 as a function of E_{cut} . The giant percolating cluster corresponds to the actual largest rigid cluster at -0.30 , -1.00 , and -2.00 kcal mol⁻¹ (Fig. 1). At -3.00 kcal mol⁻¹, this identity ceases to exist because now the actual largest cluster is located in the α -domain, whereas

the giant percolating cluster remains in the β -domain of α -lactalbumin. Accordingly, both P_{∞} are identical until $E_{\text{cut}} = -2.22$ kcal mol⁻¹, where $P_{\infty, \text{type 1}}$ drops to zero. In contrast, $P_{\infty, \text{type 2}}$ continues monitoring the decay of the actual largest rigid cluster until it drops close to zero at -3.55 kcal mol⁻¹. Note the step-wise decrease of the P_{∞} curves that reflects a process of multiple smaller transitions during the thermal unfolding. Both rigidity order parameters show distinct transitions at -0.61 , -1.55 , and -2.22 kcal mol⁻¹, which match the findings from Figure 1.

Entropy associated with the rigid cluster size distribution. The cluster configuration entropy H , introduced by Andraud et al.^[62] as a morphological descriptor for heterogeneous materials, is a particularly useful index for characterizing macroscopic properties of a constraint network in terms of its microstructure. H has been adapted from Shannon's information theory [eq. (2)] and, thus, is a measure of the degree of disorder in the realization of a given state.

$$H = - \sum_s w_s \ln w_s \quad (2)$$

It is defined as a function of the probability that an atom is part of a cluster of size s (s -cluster) [eq. (3)]

$$w_s = \frac{s^k n_s}{\sum_s s^k n_s} \quad (3)$$

with n_s being the cluster number normalized by the total number of atoms (N) [eq. (4)]

$$n_s = \frac{\text{Number of clusters of sizes}}{N} \quad (4)$$

Previously, w_s was calculated for $k = 1$ in the context of a statistical analysis of the complexity in Monte-Carlo sampled networks,^[63] which corresponds to the original definition by Andraud et al. ($H_{\text{type 1}}$). A modified version ($H_{\text{type 2}}$) has been introduced by Radestock and Gohlke^[7] using $k = 2$ (s^2 -cluster).

As long as the largest rigid cluster dominates the system, H is zero because there is only one realization of the system possible. For the same reason, H is zero if all atoms can move independently. In between, H is nonzero because of multiple possible realizations of the system associated with a heterogeneous cluster size distribution. Figures 2e and 2f show $H_{\text{type 1}}$ and $H_{\text{type 2}}$ as a function of E_{cut} . In the case of $H_{\text{type 1}}$, two transitions at -0.61 and -1.55 kcal mol⁻¹ can be identified. These transitions reflect changes in the network when the largest rigid cluster starts to decay or stops dominating the network (see also Fig. 1). In addition, a third transition can be identified at -2.22 kcal mol⁻¹ in the case of $H_{\text{type 2}}$, which corresponds to the collapse of the largest rigid cluster, as can be two later transitions at -2.99 and -3.54 kcal mol⁻¹ indicative of the final loss of the remaining rigid components.

Local indices

Flexibility index. Local indices characterize the network flexibility and rigidity down to the bond level. The flexibility index g_i

implemented in the FIRST program is a local analog of the floppy mode density Φ and quantifies the degree of flexibility or rigidity of a subcomponent of the network.^[23,64] The current FIRST implementation models biomolecules as body-bar networks. Here, atoms are treated as bodies with six degrees of freedom, and covalent and noncovalent interactions are modeled by a different number of bars that connect bonded atoms. Each bond i is part of a network subcomponent in one of four states: (i) a dangling end, (ii) a flexible collective mode, (iii) an isostatically rigid region, or (iv) an overconstrained region. For bonds in a dangling end, $g_i = 1$. For bonds that are part of a flexible collective mode j , $0 < g_i \leq 1$ [eq. (5)]:

$$g_i = \frac{F_j}{6E_j - B_j} \quad (5)$$

F_j is the number of independently rotatable bonds in j , E_j is the number of edges that represent rotatable bonds, and B_j is the total number of bars from flexible bonds. Thus, g_i relates the number of independently rotatable bonds to the number of potentially rotatable bonds in this case. In an isostatically rigid region where the number of internal degrees of freedom equals the number of constraints, $g_i = 0$. Finally, in an overconstrained region k , $-1.0 < g_i < 0$ [eq. (6)]:

$$g_i = -\frac{C_k - (6V_k - 6)}{\frac{6(V_k)(V_k - 1)}{2} - (6V_k - 6)} \quad (6)$$

C_k is the total number of constraints, and V_k indicates the number of atoms in that region. Thus, g_i relates the number of redundant constraints to the maximal number of redundant constraints in this case.

Note that with respect to a previous definition of the flexibility index f_i ,^[23] $g_i = f_i$ for collective modes, but $g_i \neq f_i$ for overconstrained regions. The latter is because f_i relates the number of redundant constraints to the actual number of all constraints (Supporting Information, Table S1).

The index g_i is calculated from a single network state and does not require a thermal unfolding simulation. For visualizing g_i results, it is often useful to derive an atom-based flexibility index as an average over g_i values of bonds an atom is involved in. For example, a flexibility index for C_α atoms can be calculated by averaging over the two backbone bonds ($N-C_\alpha$ and $C_\alpha-C'$). Figure 3 shows g_i of C_α atoms for two different network states of α -lactalbumin before ($E_{\text{cut}} = -1.46$ kcal mol⁻¹) and after ($E_{\text{cut}} = -1.66$ kcal mol⁻¹) the phase transition where the largest rigid cluster stops dominating the network (see also Fig. 1). The g_i for the region K79–D84 increases from 0.00 to 0.12, reflecting that this region now forms a flexible hinge region between two newly formed clusters that originated from the largest rigid cluster. Likewise, g_i for the region G100–L110 increases, which indicates that the largest rigid cluster decays resulting in a larger hinge region.

As a downside of the index definition for overconstrained regions, g_i values close to zero are obtained for these regions due to the denominator being dominated by the maximal number of redundant constraints in the region. This hampers

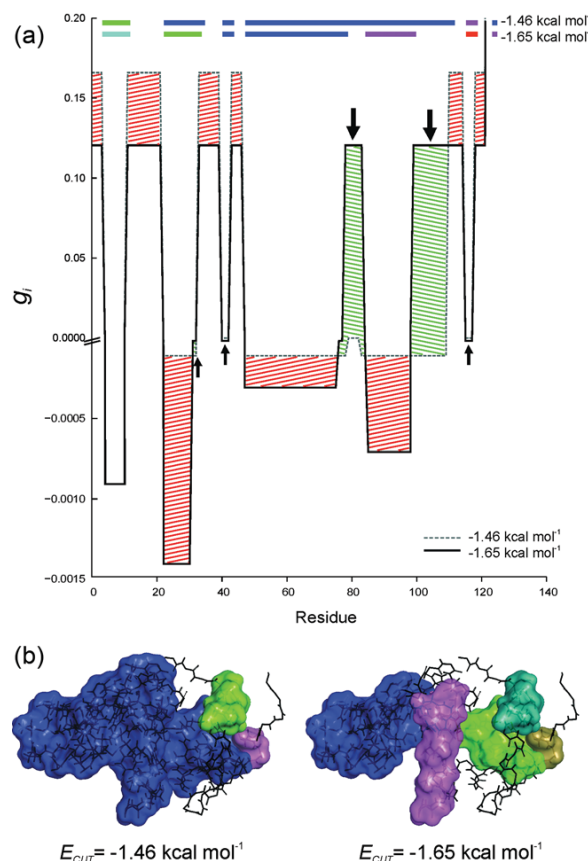


Figure 3. a) Flexibility index g_i determined for C_α atoms of α -lactalbumin before (black line) and after (gray dashed line) a phase transition. Note the different scales for g_i values larger and smaller than zero. Rigid clusters are highlighted by solid blocks at the top. Regions marked by green slanted lines and arrows are less stable after the phase transition. For regions marked by red slanted lines, the index counter intuitively shows an increase in the stability after the phase transition. b) Rigid cluster decomposition of α -lactalbumin at the respective E_{cut} values. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

a differentiation of regions with varying degrees of rigidity, at least if the same scale is used for g_i values in flexible and overconstrained regions. A more fundamental drawback of the index arises from the fact that g_i values for the majority of all regions of the network after the phase transition, that is, in a structurally less stable state, are lower than g_i values for the network before the transition. Counter intuitively, the index indicates that the network has become less flexible within collective modes and more rigid in overconstrained regions, although constraints have been removed from the network.

Percolation index. Here, we introduce for the first time the percolation index p_i , which is a local analog to the rigidity order parameter P_∞ in that it monitors the percolation behavior of a biomolecule on a microscopic level. As such, it allows identifying the hierarchical organization of the giant percolating cluster during a thermal unfolding simulation. The index value p_i is derived for each covalent bond i between two atoms $A_{i\{1,2\}}$ by the E_{cut} value during a thermal unfolding

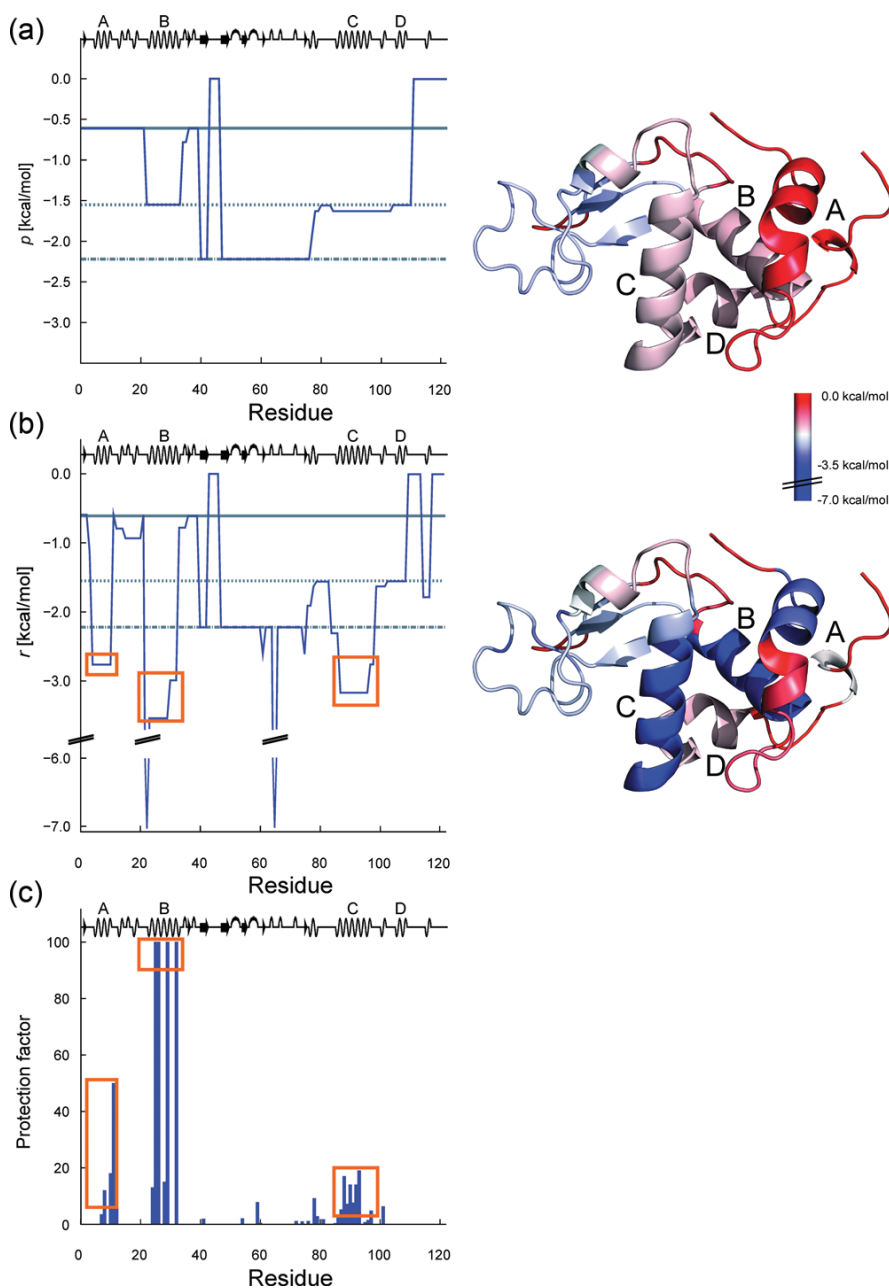


Figure 4. a) Percolation index p_i for α -lactalbumin. The lower p_i the longer is a residue part of the giant percolating cluster during the thermal unfolding simulation. The horizontal lines correspond to the three transitions as depicted in Figure 1. The secondary structure profile of α -lactalbumin is given at the top. b) Rigidity index r_i for α -lactalbumin. The lower r_i the longer is a residue part of a rigid cluster during the thermal unfolding simulation. The secondary structure profile of α -lactalbumin is given at the top. On the right, the respective index is mapped in a color-coded fashion on the α -lactalbumin structure. c) Protection factors from H/D exchange experiments in the molten globule state of α -lactalbumin. The data were taken from Schulman et al.^[65] The orange rectangles in b) and c) mark those protein regions that agree in terms of large structural stability and high protection factors.

simulation at which, the bond segregates from the giant percolating cluster c_{gpc} [eq. (7)]:

$$p_i = \min\{E_{\text{cut}} | A_{i1} \wedge A_{i2} \in c_{\text{gpc}}(E_{\text{cut}})\} \quad (7)$$

For a C_α atom-based representation, the lower of the two p_i values of the two backbone bonds is taken. $p_i = 0$ then indicates that an atom has never been part of the giant percolat-

ing cluster, that is, the atom has always been in a flexible region of the biomolecule. In contrast, the lower p_i the longer is a residue part of the giant percolating cluster during the thermal unfolding simulation. The lowest p_i thus highlights the most stable subcomponent in the network.

Figure 4a shows the C_α atom-based p_i for α -lactalbumin and, hence, on a residue-level how the giant percolating cluster

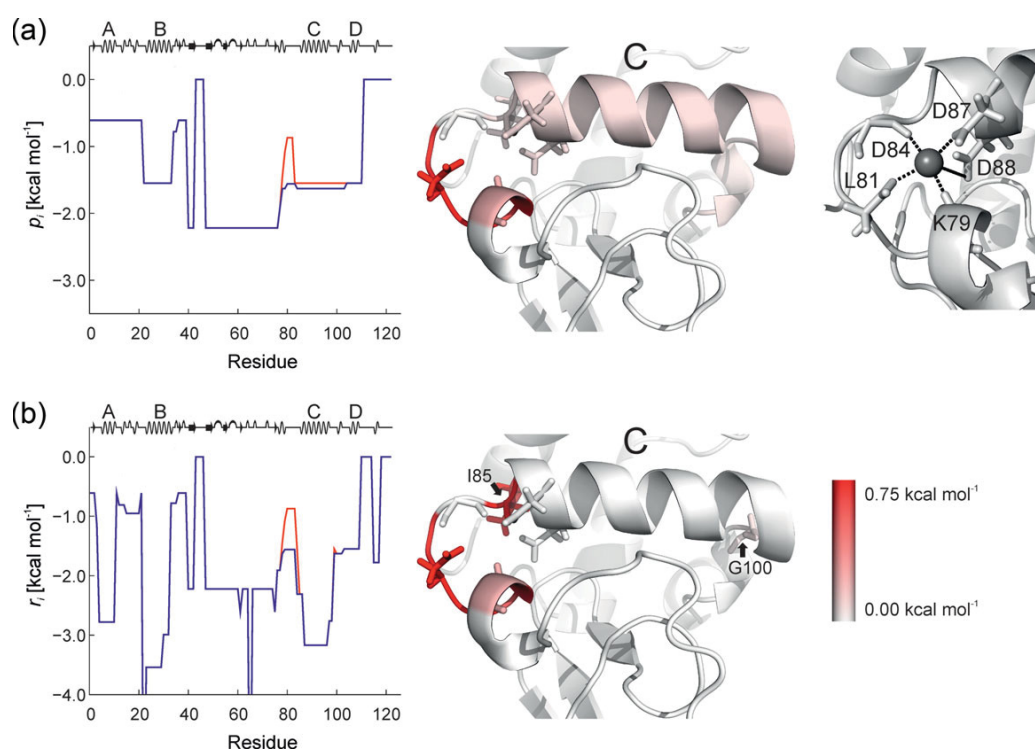


Figure 5. Change in the bimolecular stability by removing the calcium ion from α -lactalbumin as monitored by the a) percolation index p_i and b) rigidity index r_i . The indices show the ion bound state in blue and the ion unbound state in red. On the right, the changes in the respective indices on going from an ion bound to an ion unbound state are mapped in a color-coded fashion on the structure of α -lactalbumin. The calcium ion interacts with K79, L81, D84, D87, and D88 of α -lactalbumin. The secondary structure profile of α -lactalbumin is given at the top.

decays. The decay occurs as a multistep process in which residues collectively segregate from the rigid core. Three main steps can be identified, which correspond to the three transitions in Figure 1: first, the N-terminus including helix A and residues D37–A40 segregate at $p_i = -0.62$ kcal mol $^{-1}$; second, helix B and D segregate at $p_i = -1.56$ kcal mol $^{-1}$; finally, the most stable region in the β -domain collapses at $p_i = -2.22$ kcal mol $^{-1}$.

Additionally, we also analyzed the effect of perturbing the constraint network topology by removing the calcium ion (Fig. 5a). The binding region of the ion (K79–D88) is mostly affected by the removal in that the region now segregates earlier during a thermal unfolding simulation, equivalent to a lower structural stability. In addition, a lower structural stability is found for the region I89–W104. As W104 is about 15 Å away from the ion binding site, this demonstrates the long-range aspect of rigidity percolation in such networks.

Rigidity index. As a generalization of the percolation index p_i , we introduce the rigidity index r_i here for the first time. The index is defined for each covalent bond i between two atoms $A_{i\{1,2\}}$ as the E_{cut} value during a thermal unfolding simulation at which the bond changes from rigid to flexible. Phrased differently, this index monitors when a bond segregates from any rigid cluster c of the set of rigid clusters $C^{\text{E}_{\text{cut}}}$ [eq. (8)]

$$r_i = \min\{E_{\text{cut}} | \exists c \in C^{\text{E}_{\text{cut}}} : A_{i1} \wedge A_{i2} \in c\} \quad (8)$$

For a C_α atom-based representation, the average of the two r_i values of the two backbone bonds is taken. Accordingly, $r_i = 0$

indicates that an atom has always been in a flexible region of the biomolecule. In contrast, the lower r_i the longer is a residue part of a rigid cluster during the thermal unfolding simulation.

Figure 4b shows the C_α atom-based r_i for α -lactalbumin. As expected, secondary structure elements are identified as regions of highest structural stabilities. The most buried helix B ($r_i = -3.00$ to -3.55 kcal mol $^{-1}$) is also the most stable component followed by helices C ($r_i = -3.18$ kcal mol $^{-1}$) and A ($r_i = -2.79$ kcal mol $^{-1}$). These findings are in very good agreement with protein regions that have high protection factors according to H/D exchange experiments in the molten globule state of α -lactalbumin (Fig. 4c). The two spikes in Figure 4b at residues L23 and V66 reveal that both residues are captured in small clusters that remain rigid until the end of the thermal unfolding simulation.

The effect of perturbing the network topology by removal of the calcium ion is shown in Figure 5b. This affects the first residues K79–D84 of the ion binding site in a similar manner as detected by the percolation index p_i (Fig. 5a). In contrast, r_i reveals a stronger effect on residue I85, whereas no change of r_i is observed for residues T86–W104. The sole exception in the latter region is residue G100, which becomes less stable according to r_i . These results demonstrate that the information derived from r_i and the percolation index p_i (see above) is complementary: although the stability within helix C remains unaffected by the ion removal as revealed by r_i (Fig. 5b), the

percolation index p_i shows that the helix as a whole now segregates earlier from the giant percolating cluster (Fig. 5a). The latter index thus shows at which places in the structure the change of the global percolation behavior due to the ion removal manifests. Furthermore, r_i , but not p_i , exposes I85 and G100 as potential hinges for the movement of helix C because these residues become locked in only when the helix is fixed at its top by the ion to other parts of the protein. In summary, these results can be interpreted in that ion removal makes helix C movable as a rigid body as it is then encompassed by two hinges. These findings are in very good agreement with those from an experimental study on bovine α -lactalbumin^[66] where helix C and the adjacent helical element (C77–K80) change their relative orientation and where the opposite face of the calcium binding site is perturbed, as implicated from comparing the apo and calcium-bound structures.

Stability maps. Stability maps rc_{ij} have been introduced by Radestock and Gohlke.^[6] A stability map is a two-dimensional itemization of the rigidity index r_i and is derived by identifying “rigid contacts” between two residues $R_{i,j}$, which are represented by their C_α atoms. A rigid contact exists if two residues belong to the same rigid cluster c of the set of rigid clusters $C^{E_{\text{cut}}}$. During a thermal unfolding simulation, stability maps are then constructed in that, for each residue pair, E_{cut} is identified at which a rigid contact between two residues is lost [eq. (9)].

$$rc_{ij} = \min\{E_{\text{cut}} | \exists c \in C^{E_{\text{cut}}} : R_i \wedge R_j \in c\} \quad (9)$$

That way, a contact's stability relates to the microscopic stability in the network and, taken together, the microscopic stabilities of all residue–residue contacts result in a stability map. Thus, stability maps denote the distribution of flexibility and rigidity within the system, they identify regions that are flexibly or rigidly correlated across the structure, and they provide information on how these properties change with increasing E_{cut} . In Figure 6, the stability map for α -lactalbumin is shown. The upper and lower triangles of the map have been derived for the constraint network with and without the calcium ion, respectively. Again, residues belonging to the calcium binding site are mostly affected by the ion removal. Furthermore, the maps intriguingly reveal that losses of rigid contacts do not only occur between isolated pairs of residues but also in a co-operative manner. That is, parts of the protein break away from the rigid cluster as a whole, as can be seen for helices A, B, and C.

Discussion

We provided concise definitions of indices for deriving biologically relevant characteristics of a biomolecule from rigidity analysis. The majority of these indices are computed by monitoring the changes in a network along an unfolding trajectory of the showcase example α -lactalbumin. The trajectory was computed by consecutively removing hydrogen bond constraints using the CNA package. Here, only a single input structure was used as a starting point for building up the constraint network. Different conformations of a biomolecule can

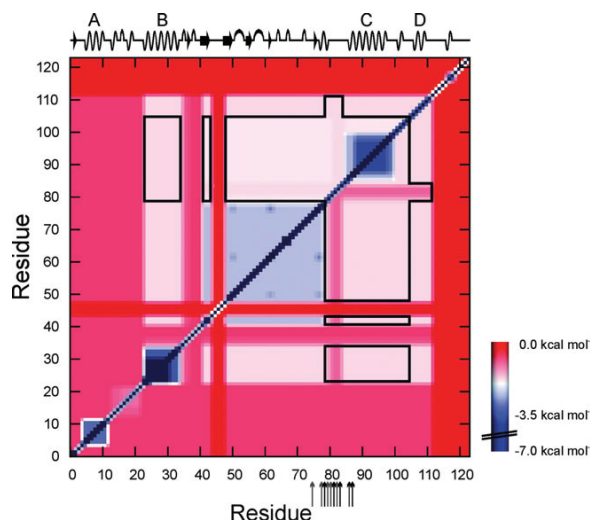


Figure 6. Stability map for α -lactalbumin. In the upper half of the matrix, stability information for the calcium-bound protein is shown; the lower half shows the protein in the ion unbound state. Red colors indicate pairs of residues where no or only a weak rigid contact exists. In contrast, blue colors indicate strong rigid contacts. Arrows highlight residues in the vicinity of the calcium binding site. The black frames indicate regions that are affected by removing the calcium ion. The secondary structure profile of α -lactalbumin is given at the top. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

vary in the noncovalent bond network and, hence, can lead to a different outcome of the rigidity analysis as observed by us^[26] and others.^[45] In general, this problem can be overcome by analyzing an ensemble of constraint networks, for example, generated from a conformational ensemble obtained by MD simulation.^[26,39,40] However, we note that using a single structure does not provide a limitation for the current study because here we only compare indices with respect to each other.

A related approach is provided by the DCM.^[31,67,68] Here, an ensemble of constraint topologies is generated by considering mean-field probabilities of hydrogen bonds and torsion constraints in a Monte-Carlo sampling. Average stability characteristics are then computed by performing a FIRST analysis on each constraint topology in the ensemble. Note that DCM requires knowledge of experimentally determined heat capacity curves for a protein-specific parameterization of the model.^[8,69]

We will now analyze the relation between, and the scope and limitations of indices used in connection with CNA and compare their informative value.

Relations among indices

The main purpose of applying global indices is to identify (a) phase transition point(s) where a molecule switches from a rigid to a flexible state. This is relevant for analyzing the thermostability of proteins^[6,7,37] or changes in a protein's global stability on binding of a ligand^[26] (see Table 2). Monitoring the floppy mode density Φ provides a measure for the intrinsic

Table 2. Application of global and local indices for explaining and interpreting experimental findings.

Application	Experimental data	Index type	Reference
Analyzing thermostability of and identifying 'weak-spots' in biomolecules.	Either optimal growth temperatures of the organism or experimentally determined melting temperatures of the biomolecules. 'Weak-spots' were compared with identified folding cores from H/D exchange NMR studies as well as mutation experiments.	Rigidity order parameter P_{∞} Cluster configuration entropy H Rigidity order parameter $P_{\infty}^{[a]}$ Cluster configuration entropy H FIRST dilution plots ^[b] Largest rigid cluster propensity P_{irc}	Radestock and Gohlke ^{6,7} , Rathi <i>et al.</i> ³⁹ Rader ³⁷
Identification of folding cores in biomolecules.	Folding cores as predicted by H/D exchange NMR experiments.	Rigidity order parameter $P_{\infty}^{[a]}$ FIRST dilution plots ^[b] FIRST dilution plots ^[b]	Rader <i>et al.</i> ⁶¹ Hespenheide <i>et al.</i> ⁷⁰
Analyzing the loss of structural stability in biomolecules.	Compared to the unfolding behavior of network glasses upon melting.	Floppy mode density Φ	Rader <i>et al.</i> ²⁴
Analyzing the flexibility of substrate-binding regions in enzymes.	Comparison of different structural information as well as thermal mobility (B-factor) determined by X-ray crystallography.	Stability maps rc_{ij}	Radestock and Gohlke ⁶
Analyzing the flexibility in proteins as well as RNA structures.	Thermal mobility (B-factor) determined by X-ray crystallography.	FIRST flexibility index f_i FIRST flexibility index g_i	Jacobs <i>et al.</i> ²³ , Fulle and Gohlke ^{38,71,72} Tan and Rader ⁶⁴
Analyzing changes in protein flexibility upon protein-protein complex formation.	Thermal mobility (B-factor) determined by X-ray crystallography and atomic fluctuations by MD simulations.	FIRST flexibility index f_i	Gohlke <i>et al.</i> ²⁶

[a] The author used the notations X_C and X_{irc} that match the definition of the type 2 rigidity order parameter P_{∞} [b] FIRST dilution plots are graphical representations of the rigid cluster decomposition along the thermal unfolding simulation. The percolation index p_i and the rigidity index r_i are measures that allow a numerical interpretation of the dilution plots.

flexibility along the unfolding trajectory. Alternatively, indices have been adapted from percolation theory by analyzing the microstructure of the network. These indices are derived from properties of the set of rigid clusters generated along the unfolding trajectory.^[6] The rigidity order parameter P_{∞} considers the extremum, that is, the largest rigid cluster, at a certain network state. In contrast, the largest rigid cluster is excluded in the case of the mean rigid cluster size S ; therefore, S monitors the size distribution of smaller rigid clusters. The cluster configuration entropy H is a morphological descriptor of heterogeneity in networks.^[62] Compared to the original implementation $H_{\text{type } 1}$, the alternative $H_{\text{type } 2}$ is more sensitive with respect to transition points that occur later in the unfolding trajectory, that is, when the network is already largely flexible. Both indices S and H monitor the complexity of the network; phrased differently, they measure the degree of disorder in the realization of a given network state.

The main purpose of applying local indices is to monitor the location and distribution of structurally weak or strong parts in biomolecules. This is relevant for guiding protein engineering efforts aimed at identifying unfolding nuclei (structurally weak parts) that, when mutated, may lead to an increase in the thermostability (Table 2).^[6,7] These indices can either reflect structural stability on a per-residue basis or characterize correlations of stability between pairs of residues. The FIRST flexibility index g_i belongs to the first class and is a local analog of the global index Φ in regions of collective modes. The index monitors how degrees of freedom and redundant con-

straints, respectively, are distributed throughout the network. It is the only index presented here that is computed from a static network state. The percolation index p_i is a local analog of P_{∞} and monitors the percolation behavior of the giant percolating cluster on a microscopic level. The rigidity index r_i is a generalization of p_i and monitors the transition when a residue segregates from any rigid cluster. Stability maps belong to the second class and characterize the correlation of stability in biomolecules; hence, they itemize the information provided by r_i in that they reflect the microscopic stabilities between all pairs of residues in the network. The relationship between global and local indices is shown in Figure 7.

Ambiguity in index definitions

To date, at least 14 indices have been introduced in the literature with sometimes overlapping or identical definitions. Rader *et al.*^[61] have used a global order parameter X_C with the goal to describe the percolation behavior in biomolecules. This index definition matches type 2 of the rigidity order parameter P_{∞} ; P_{∞} had been introduced before by Stauffer.^[60] Likewise, a parameter that considers the fraction of the network belonging to the percolating rigid cluster was used in the work of Chubynsky and Thorpe.^[55] The definition of this parameter matches type 1 of P_{∞} . We thus recommend using the original P_{∞} -based notation in future studies for clarity. The local indices p_i and r_i introduced in this study can be seen as "envelopes" of spikes in FIRST dilution plots.^[70] FIRST dilutions plots

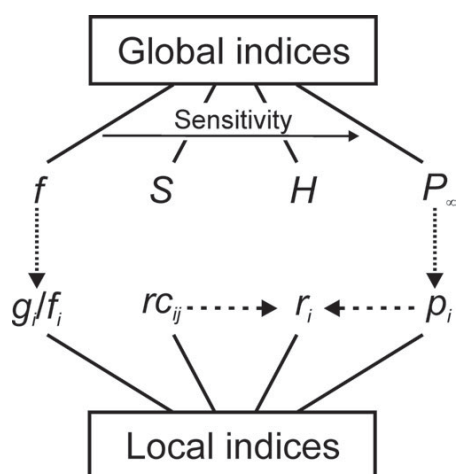


Figure 7. Relationship between global and local indices. Fine dashed arrows indicate 'local analog of global count' relationship. Coarsely dashed arrows indicate a 'generalization' relationship.

have been used as graphical representations of the rigid cluster decomposition along a thermal unfolding simulation so far. The local indices p_i and r_i allow exploiting the information from a rigid cluster decomposition in a quantitative manner (Fig. 5), for example, to calculate the stability change of a biomolecule on ligand binding or mutation. With a similar aim, Rader introduced a local index P_{lrc} for the projection of the percolation behavior on a microscopic level.^[37] This index monitors the propensity of a residue for being part of the largest rigid cluster along an unfolding trajectory. Rader's definition is related to the percolation index p_i defined here for the first time. In our opinion, analyzing the giant percolating cluster as in p_i instead of the actual largest rigid cluster as in P_{lrc} is more appropriate for identifying that part of a protein from where rigidity starts to propagate through the network.

Supporting Information, Table S1 also contains four index definitions used in the DCM.^[9] The global DCM flexibility index θ is the average of the number of independent degrees of freedom F over the DCM ensemble. The local DCM flexibility index ϑ is the ensemble average of the local density of floppy modes and redundant constraints^[8] and, therefore, related to the FIRST flexibility index f_i .^[8] A second local index P_R has been introduced for quantifying backbone flexibility by monitoring the probability whether backbone dihedral angles (ϕ, ψ) are rotatable over the ensemble.^[68] This definition is related to the rigidity index r_i that monitors when a bond segregates from a rigid cluster during the thermal unfolding. Finally, cooperativity correlation plots are provided by DCM^[68,69] that quantify the correlated stability for pairs of residues in terms of rotatable dihedral backbone angles. Related to this, stability maps rc_{ij} monitor the correlated stability between pairs of residues according to how long a rigid contact exists along the unfolding trajectory. Note that, although index definitions used within the frameworks of CNA or DCM are related, the way they are derived substan-

tially differs: in that CNA monitors changes in the constraint network along a thermal unfolding trajectory, whereas DCM performs Monte-Carlo sampling of network topologies at a fixed temperature. We thus recommend using the respective CNA- or DCM-based index notations in future studies to make these differences clear.

Informative value, scope, and limitations of global indices

We applied four global indices for analyzing the unfolding trajectory of α -lactalbumin, the floppy mode density Φ , the mean rigid cluster size S , the rigidity order parameter P_∞ and the cluster configuration entropy H . In general, all global indices were able to identify transition points in agreement with results from visual inspection of the unfolding trajectory. However, the indices dramatically differ in terms of the sensitivity, which leads to different numbers of transitions being identified. Monitoring Φ allows identifying those transition points where the structural features in the network change dramatically. As such, Φ identifies the second transition when the largest rigid cluster stops dominating the network but fails at identifying the other two transitions. The index S identifies the first and second transitions and, thus, detects the beginning of the network collapse; however, it fails at identifying the last transition. The indices P_∞ type 1 and 2 detect all three transition points identified visually; the type 2 index additionally identifies two more transitions highlighting the collapse of last rigid fragments. As for index H , type 1 and 2 are complementary: $H_{type\ 1}$ identifies the earlier transition points when the network is dominated by the largest rigid cluster, whereas $H_{type\ 2}$ identifies the later transition points when the network consists of multiple smaller rigid clusters.

The loss of stability in biomolecules during thermal unfolding is a multistep process.^[7] This is because biomolecules have a hierarchical organization in terms of structural stability as opposed to network glasses. This has implications for index sensitivity (early vs. late transitions) and the application domain of indices (e.g., folding core detection vs. thermostability analysis). The index Φ is particularly insensitive for detecting any transition points other than the one associated with the rigidity percolation threshold because the change in the number of independent degrees of freedom markedly changes only at this point.^[36] Accordingly, when applied for analyzing the (un-)folding of mainly globular biomolecules, Φ only provides a description of a two state nature of this transition, that is, of the biomolecule switching from a globally stable to a flexible state. Φ has been used for analyzing the loss of structural stability on protein unfolding and for comparing a protein's unfolding behavior to that of network glasses on melting by Rader et al.^[24] (Table 2). In contrast, indices S and H , based on the cluster size distribution of the network, show an increased sensitivity with respect to the detection of additional transition points. This is because removing a single constraint can lead to the collapse of a rigid region, which can strongly affect the cluster size distribution even if the network is already in the floppy state. Both indices S and $H_{type\ 1}$ are most sensitive for earlier transitions in unfolding and, hence, best

applied for analyzing the loss of the rigidity percolation in biomolecules. Phrased differently, $H_{\text{type } 1}$ is preferably applied for detecting when the first fragments segregate from the largest rigid cluster in a network. In contrast, $H_{\text{type } 2}$ is most sensitive for later transitions in unfolding and, hence, best applied for analyzing the decay of the largest rigid fragments.^[7] Consequently, $H_{\text{type } 2}$ was used to analyze the shift in the melting temperature of 20 pairs of orthologous proteins from mesophilic and thermophilic organisms^[6,7] and five citrate synthase structures that cover a wide range of thermostability^[39] (Table 2).

Finally, P_{∞} shows relatively pronounced steps also in between the transition points due to monitoring the decay of the giant or largest rigid cluster and, thus, is even more sensitive to network changes than S or H . This makes P_{∞} better suited for a detailed analysis of the loss of stability in a biomolecule, for example, when monitoring the segregation of secondary structure units from the core^[38] rather than for identifying transition points in an automated fashion, for example, as needed when computing melting points of a protein over an ensemble of structures.^[6,7] Along these lines, P_{∞} and the related index X_C were used as absolute measures to identify key amino acids that are important for the stability of rhodopsin^[61] and as relative measures to analyze the global shift in stability on ligand binding to HIV-1 gp120^[64] and the thermostability of rubredoxin structures^[37] (Table 2). Notably, homologous proteins have P_{∞} curves of a very similar shape,^[6,7] reflecting the evolutionary conservation of flexibility and rigidity in proteins and providing direct evidence for the hypothesis of corresponding states for orthologs from meso- and thermophilic organisms.^[6]

Besides providing qualitative information that allow for the successful prediction of folding cores^[61] and weak spots,^[6,7,37,39] some of the indices have also been used to connect with quantitative experimentally measurable data. In particular, this holds for (changes in) the stability of proteins measured in terms of melting temperatures (T_m) or, more indirectly, in terms of optimal growth temperatures (T_{og}) of the respective organisms. DCM index definitions describe the Landau free energy $G(T, \theta)$ as a function of the temperature T and a global flexibility order parameter θ and thus directly relate to a protein's thermostability.^[8,68] Likewise, Radestock and Gohlke found a linear relationship between computed phase transition temperatures and experimental T_m or T_{og} ,^[6,7] which was also applied in subsequent studies.^[37,39]

Informative value, scope, and limitations of local indices

We applied the four local indices FIRST flexibility index g_i , percolation index p_i , rigidity index r_i , and stability maps rc_{ij} for analyzing the unfolding trajectory of α -lactalbumin. The indices p_i and r_i are introduced here for the first time: p_i reflects the stability of the giant percolating cluster along the unfolding trajectory, whereas r_i monitors the rigid-to-flexible transition of bonds in the whole network. As such, p_i reflects the hierarchical organization of the giant percolating cluster and, hence, is useful for identifying how rigidity starts propagating through

the network. Furthermore, because p_i provides a microscopic view of the percolation behavior, the index allows tracing how structural changes (e.g., a mutation or the removal of a ligand or an ion) affect this (macroscopic) behavior. In contrast, r_i maps rigidity and flexibility at a more local scale, as demonstrated for the comparison of r_i with protection factors from H/D exchange experiments in the molten globule state of the protein (Fig. 4c). This comparison reveals that the structurally most stable regions are also those with the highest protection factors.^[65] Such regions have been interpreted as folding cores in this context.^[70] Thus, r_i may be useful for detecting residues that are part of folding cores in biomolecules. The complementarity of information provided by p_i and r_i is markedly demonstrated also when perturbing the constraint network as in the case of removing the calcium ion from α -lactalbumin (Figs. 5a and 5b): p_i monitors the stabilizing effect of the calcium ion on the giant percolating cluster (which includes helix C in the presence of the ion), whereas r_i detects no influence of the ion on the (local) stability of the helix. Notably, in both cases long-range changes in flexibility and rigidity are detected despite perturbing the network only by removing short-range constraints between the ion and its surroundings. Therefore, both indices should be particularly useful for detecting changes in a biomolecule's flexibility on ligand binding that may exert an entropic effect and for identifying adequate sites for mutations to increase the thermostability of a biomolecule. As an alternative to r_i , the related largest rigid cluster propensity P_{lrc} introduced by Rader et al.^[37] (Supporting Information, Table S1) has been used to identify the mechanically most stable residues in rubredoxin; these results agree well with results from H/D experiments (Table 2). The authors concluded that using this index is hence appropriate to analyze unfolding pathways in proteins. Finally, stability maps are applied to monitor correlations in structural stability for each pair of residues in a biomolecule; these maps have been successfully used to demonstrate that the distribution of functionally important flexible regions is conserved between meso- and thermophilic orthologs when considering the appropriate working temperatures of the enzymes (Table 2).^[6]

Indices to monitor the distribution of structural stability in biomolecules should be well behaved. In particular, one expects that, when constraints are removed from a network, the network's stability either remains unchanged or decreases. This relation is reflected by both p_i and r_i on removal of the calcium ion, as was also to be expected from the definitions of these indices. The FIRST flexibility index g_i did not behave as expected; however (Fig. 3), on removal of constraints, g_i decreased in some regions of collective modes and overconstrained structural parts. An explanation is given in Figure 8. g_i relates the number of independently rotatable bonds to the number of potentially rotatable bonds in collective modes (number of redundant constraints to the maximal number of redundant constraints in overconstrained regions), with both the numerator and denominator evaluated for that particular region. Breaking up this region into two parts by removing some constraints will in general lead to a disproportionate change in the numerator and denominator values. As a

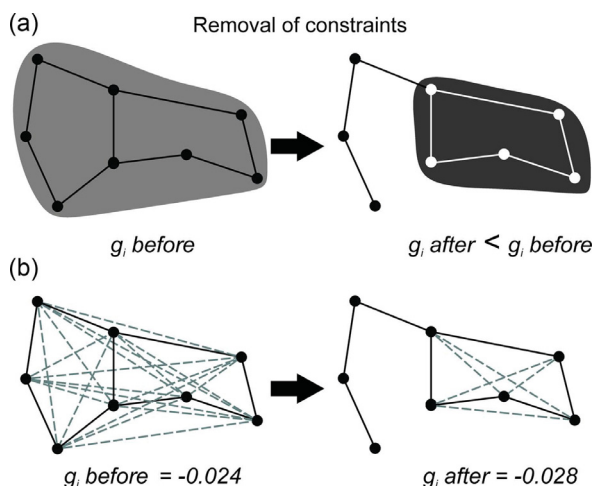


Figure 8. Example illustrating the unexpected behavior of the FIRST flexibility index g_i on removal of constraints. a) Schematic view of the collapse of a rigid cluster into a smaller one and a dangling end. b) A worked example of a) with two network states before and after the removal of constraints. Each edge represents five bars that connect atoms in a body-bar network. Solid edges represent regular bonds. Dashed edges represent additional bonds for a fully connected network. Applying eq. (6) to both network states leads to $g_{\text{before}} = -0.024$ and $g_{\text{after}} = -0.028$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

consequence, the removal of constraints does not necessarily lead to an increase of g_i . Thus, g_i (and the flexibility index f_i of previous FIRST versions) can be applied for analyzing absolute local flexibility and rigidity characteristics, as done in studies on proteins^[23,64] and RNA^[71] and when investigating the statics of the ribosomal exit tunnel^[38,72] (Table 2). However, we strongly advise against using g_i and f_i for analyzing changes in these characteristics, for example, when comparing networks that differ with respect to a bound ligand or the number of noncovalent interactions due to a temperature influence. Furthermore, note that even for analyzing absolute flexibility and rigidity characteristics it comes as a drawback that g_i tends to zero for large overconstrained regions [eq. (6)], which makes a comparison of such regions difficult. f_i does not suffer from that drawback.

Local indices describe intrinsic biomolecular flexibility and rigidity on a microscopic level. Hence, these indices can be compared with quantitative experimentally measurable data on biomolecular mobility such as B-factors determined by X-ray crystallography or S^2 order parameters and H/D exchange information determined by NMR. Although poor correlations have been obtained if correlating those experimental quantities with each other, it has been shown that DCM indices correlate well with all three of them.^[8] The FIRST index f_i was quantitatively compared to B-factor information of proteins, which showed a good agreement of both measures (Table 2).^[23] Similarly, indices p_i and r_i can be used to quantitatively assess (differences in) the stability of biomolecules as suggested in Figure 5. In fact, differences in p_i or r_i values summed over all residues correlate well with experimentally

determined stability changes in the case of wildtype eglin C and 11 of its mutants (Pfleger and Gohlke, unpublished results).

In conclusion, in this study, we presented concise definitions for four global and four local indices for describing stability characteristics in biomolecules. Three index definitions were introduced and applied to analyze a biomolecule's stability for the first time. Showcase analyses of the thermal unfolding of α -lactalbumin demonstrated the scope and limitations, and the informative value of each index. This allowed us to provide guidelines for future studies suggesting which of these indices could best be used for analyzing, understanding, and quantifying structural features that are important for protein stability and function. Finally, we made suggestions for proper index notations in future studies to prevent the misinterpretation and to facilitate the comparison of results obtained from flexibility and rigidity analyses.

Acknowledgments

We are grateful to Doris L. Klein and Prakash C. Rath (Heinrich-Heine-University, Düsseldorf) for fruitful discussions. The CNA program for calculating global and local flexibility indices is available from the authors upon request. CNA functions as a front- and back-end to the FIRST program; the latter is available from <http://flexweb.asu.edu/software/>.

Keywords: percolation theory · protein stability · phase transition · folding core · protein engineering · drug design

How to cite this article: C. Pfleger, S. Radestock, E. Schmidt, H. Gohlke, *J. Comput. Chem.* **2013**, *34*, 220–233. DOI: 10.1002/jcc.23122

Additional Supporting Information may be found in the online version of this article.

- [1] K. Henzler-Wildman, D. Kern, *Nature* **2007**, *450*, 964.
- [2] T. J. Kamerzell, C. R. Middaugh, *J. Pharm. Sci.* **2008**, *97*, 3494.
- [3] R. Sterner, E. Brunner, In *Thermophiles: Biology and Technology at High Temperatures*; F. Robb, G. Antranikian, D. Grogan, Driessen, A., Eds.; CRC Press: London, New York, **2008**, pp 25–38.
- [4] G. N. Somero, *Annu. Rev. Ecol. Syst.* **1978**, *9*, 1.
- [5] R. Jaenicke, G. Böhm, *Curr. Opin. Struct. Biol.* **1998**, *8*, 738.
- [6] S. Radestock, H. Gohlke, *Proteins* **2011**, *79*, 1089.
- [7] S. Radestock, H. Gohlke, *Eng. Life Sci.* **2008**, *8*, 507.
- [8] D. R. Livesay, S. Dallakyan, G. G. Wood, D. J. Jacobs, *FEBS Lett.* **2004**, *576*, 468.
- [9] J. M. Mottonen, M. Xu, D. J. Jacobs, D. R. Livesay, *Proteins* **2009**, *75*, 610.
- [10] A. Metz, C. Pfleger, H. Kopitz, S. Pfeiffer-Marek, K. H. Baringhaus, H. Gohlke, *J. Chem. Inf. Model* **2011**, *52*, 120.
- [11] S. Fulle, N. A. Christ, E. Kestner, H. Gohlke, *J. Chem. Inf. Model.* **2010**, *50*, 1489.
- [12] S. Fulle, H. Gohlke, In *Methods in Molecular Biology—Computational Drug Discovery and Design*; Baron, R., Ed.; Humana Press: New York, **2012**, pp 75–91.
- [13] D. Georlette, V. Blaise, T. Collins, S. D'Amico, E. Gratia, A. Hoyoux, J. C. Marx, G. Sonan, G. Feller, C. Gerday, *FEMS Microbiol. Rev.* **2004**, *28*, 25.
- [14] R. Abseher, L. Horstink, C. W. Hilbers, M. Nilges, *Proteins* **1998**, *31*, 370.
- [15] B. Halle, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 1274.

- [16] N. V. Dokholyan, L. Li, F. Ding, E. I. Shakhnovich, *Proc. Natl. Acad. Sci. USA* **2002**, 99, 8637.
- [17] M. Vendruscolo, N. V. Dokholyan, E. Paci, M. Karplus, *Phys. Rev. E* **2002**, 65, 1.
- [18] C. Böde, I. A. Kovács, M. S. Szalay, R. Palotai, T. Korcsmáros, P. Csermely, *FEBS Lett.* **2007**, 581, 2776.
- [19] L. H. Greene, V. A. Higman, *J. Mol. Biol.* **2003**, 334, 781.
- [20] J. Heringa, P. Argos, *J. Mol. Biol.* **1991**, 220, 151.
- [21] J. Heringa, P. Argos, M. R. Egmond, J. Devlieg, *Protein Eng.* **1995**, 8, 21.
- [22] B. Hendrickson, *SIAM J. Comput.* **1992**, 21, 65.
- [23] D. J. Jacobs, A. J. Rader, L. A. Kuhn, M. F. Thorpe, *Proteins* **2001**, 44, 150.
- [24] A. J. Rader, B. M. Hespeneide, L. A. Kuhn, M. F. Thorpe, *Proc. Natl. Acad. Sci. USA* **2002**, 99, 3540.
- [25] H. Gohlke, M. F. Thorpe, *Biophys. J.* **2006**, 91, 2115.
- [26] H. Gohlke, L. A. Kuhn, D. A. Case, *Proteins* **2004**, 56, 322.
- [27] D. J. Jacobs, B. Hendrickson, *J. Comput. Phys.* **1997**, 137, 346.
- [28] D. J. Jacobs, M. F. Thorpe, *Phys. Rev. Lett.* **1995**, 75, 4051.
- [29] D. J. Jacobs, *J. Phys. Chem. A* **1998**, 31, 6653.
- [30] N. Katoh, S. Tanigawa, *Discrete Comput. Geom.* **2011**, 45, 647.
- [31] D. J. Jacobs, S. Dallakyan, G. G. Wood, A. Heckathorne, *Phys. Rev. E* **2003**, 68, 061109.
- [32] M. Lei, M. I. Zavodszky, L. A. Kuhn, M. F. Thorpe, *J. Comput. Chem.* **2004**, 25, 1133.
- [33] S. Wells, S. Menor, B. Hespeneide, M. F. Thorpe, *Phys. Biol.* **2005**, 2, S127–S136.
- [34] A. Ahmed, H. Gohlke, *Proteins* **2006**, 63, 1038.
- [35] A. Ahmed, S. Villinger, H. Gohlke, *Proteins* **2010**, 78, 3341.
- [36] M. F. Thorpe, *J. Non-Cryst. Solids* **1983**, 57, 355.
- [37] A. Rader, *J. Phys. Biol.* **2010**, 7, 016002.
- [38] S. Fulle, H. Gohlke, *J. Mol. Biol.* **2009**, 387, 502.
- [39] P. C. Rath, S. Radestock, H. Gohlke, *J. Biotechnol.* **2012**, 159, 135.
- [40] P. C. Rath, C. Pfleger, S. Fulle, D. L. Klein, H. Gohlke, In *Statics of Biomacromolecules*; Comba, P., Ed.; Wiley-VCH: Weinheim, **2011**, pp 281–299.
- [41] J. Ren, D. I. Stuart, K. R. Acharya, *J. Biol. Chem.* **1993**, 268, 19292.
- [42] P. E. Bourne, K. J. Adress, W. F. Bluhm, L. Chen, N. Deshpande, Z. K. Feng, W. Fleri, R. Green, J. C. Merino-Ott, W. Townsend-Merino, H. Weissig, J. Westbrook, H. M. Berman, *Nucleic Acids Res.* **2004**, 32, D223.
- [43] R. W. W. Hoof, G. Vriend, C. Sander, E. E. Abola, *Nature* **1996**, 381, 272.
- [44] J. M. Word, S. C. Lovell, J. S. Richardson, D. C. Richardson, *J. Mol. Biol.* **1999**, 285, 1735.
- [45] T. Mamonova, B. Hespeneide, R. Straub, M. F. Thorpe, M. Kurnikova, *Phys. Biol.* **2005**, 2, S137.
- [46] B. I. Dahiyat, D. B. Gordon, S. L. Mayo, *Protein Sci.* **1997**, 6, 1333.
- [47] B. Folch, M. Rooman, Y. Dehouck, *J. Chem. Inf. Model* **2008**, 48, 119.
- [48] P. L. Privalov, S. J. Gill, *Adv. Protein Chem.* **1988**, 39, 191.
- [49] K. Kuwajima, Y. Harushima, S. Sugai, *Int. J. Pept. Protein Res.* **1986**, 27, 18.
- [50] E. A. Permyakov, V. V. Yarmolenko, L. P. Kalinichenko, L. A. Morozova, E. A. Burstein, *Biochem. Biophys. Res. Commun.* **1981**, 100, 191.
- [51] T. Segawa, S. Sugai, *J. Biochem.* **1983**, 93, 1321.
- [52] S. J. Prestrelski, D. M. Byler, M. P. Thompson, *Biochemistry* **1991**, 30, 8797.
- [53] The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.
- [53] Y. Cai, M. F. Thorpe, *Phys. Rev. B* **1989**, 40, 10535.
- [55] M. V. Chubynsky, M. F. Thorpe, *Curr. Opin. Solid State Mater. Sci.* **2001**, 5, 525.
- [56] H. He, M. F. Thorpe, *Phys. Rev. Lett.* **1985**, 54, 2107.
- [57] P. L. Privalov, *J. Mol. Biol.* **1996**, 258, 707.
- [58] C. A. Angell, In *Hydration Processes in Biology*; M.-C. Bellissent-Funel, Ed.; IOS Press: Amsterdam, **1999**, pp 127–139.
- [59] D. Stauffer, A. Aharony, *Introduction To Percolation Theory*; Taylor and Francis: London, **1994**.
- [60] D. Stauffer, *Phys. Rep.* **1979**, 54, 1.
- [61] A. J. Rader, G. Anderson, B. Isin, H. G. Khorana, I. Bahar, J. Klein-Seetharaman, *Proc. Natl. Acad. Sci. USA* **2004**, 101, 7246.
- [62] C. Andraud, A. Beghdadi, J. Lafait, *Physica A* **1994**, 207, 208.
- [63] I. R. Tsang, I. J. Tsang, *Phys. Rev. E* **1999**, 60, 2684.
- [64] H. P. Tan, A. J. Rader, *Proteins* **2009**, 74, 881.
- [65] B. A. Schulman, C. Redfield, Z. Peng, C. M. Dobson, P. Kim, *J. Mol. Biol.* **1995**, 253, 651.
- [66] E. D. Chrysina, K. Brew, K. R. Acharya, *J. Biotechnol.* **2000**, 275, 37021.
- [67] D. R. Livesay, D. H. Huynh, S. Dallakyan, D. J. Jacobs, *Chem. Cent. J.* **2008**, 2, 17.
- [68] D. R. Livesay, D. J. Jacobs, *Proteins* **2006**, 62, 130.
- [69] D. J. Jacobs, S. Dallakyan, *Biophys. J.* **2005**, 88, 903.
- [70] B. M. Hespeneide, A. J. Rader, M. F. Thorpe, L. A. Kuhn, *J. Mol. Graph. Model* **2002**, 21, 195.
- [71] S. Fulle, H. Gohlke, *Biophys. J.* **2008**, 94, 4202.
- [72] S. Fulle, H. Gohlke, *Methods* **2009**, 49, 181.

Received: 3 May 2012
 Revised: 26 August 2012
 Accepted: 28 August 2012
 Published online on 25 September 2012

13.3 Publication I – Supporting Information

Global and Local Indices for Characterizing Biomolecular Flexibility and Rigidity

Pfleger, C., Radestock, S., Schmidt, E., Gohlke, H.

J. Comput. Chem. (2013), 34, 220–233

Supporting Information

Global and local indices for characterizing biomolecular flexibility and rigidity

Christopher Pflieger[†], Sebastian Radestock[†], Elena Schmidt[‡], Holger Gohlke^{†*}

[†]Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine-University, Düsseldorf, Germany and [‡]Department of Biological Sciences, Goethe-University, Frankfurt am Main, Germany

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

Table SI: Definitions of global and local indices for characterizing biomolecular flexibility and rigidity.

Index Name/ Character	Index type	Description/Equation	Informative value/ boundaries	First publication/ application	Available in
Mean rigid cluster size S	Global	$S = \frac{1}{n} \sum_i^n s_i$ <p> n = Number of clusters except the largest rigid cluster s_i = Number of atoms in cluster i </p>	S monitors the macroscopic property associated with the rigid cluster size distribution of a network along a thermal unfolding simulation. S is zero as long as one rigid cluster dominates the whole network or if all rigid clusters have vanished.	Introduced by Stauffer and Aharony ¹ .	Constraint Network Analysis (CNA) ²
Rigidity order parameter P_∞	Global	$P_\infty = \frac{N_{gpc/lrc}}{N}$ <p> $N_{gpc/lrc}$ = Number of atoms in the giant percolating cluster ("type 1") or the largest rigid cluster ("type 2") N = Total number of atoms </p>	P_∞ monitors the fraction of the network belonging to the giant percolating cluster (type 1) or the actual largest rigid cluster (type 2) along a thermal unfolding simulation. P_∞ is close to one in the rigid phase and zero in the floppy phase.	Introduced by Stauffer ^{1,3} and applied to network glasses ⁴ and proteins ⁵⁻⁸ .	Constraint Network Analysis (CNA) ²
Cluster configuration entropy H	Global	$H = - \sum_s w_s \ln w_s$ $w_s = \frac{s^k n_s}{\sum_s s^k n_s}$ <p> s = Cluster of size s n_s = Number of clusters of size s, normalized by the total number of atoms in the network w_s = Probability that an atom belongs to a s-cluster $k = 1$ ("type 1") or 2 ("type 2") </p>	H monitors the degree of disorder in the realization of a given network state along a thermal unfolding simulation. H is zero in the rigid phase, and non-zero in the floppy phase. $H_{type 1}$ is most sensitive for earlier transitions in unfolding; $H_{type 2}$ is most sensitive for later transitions in unfolding.	Introduced by Andraud <i>et al.</i> ⁹ and applied to complex networks ¹⁰ and proteins ^{6,7,11} .	Constraint Network Analysis (CNA) ²

Table S1 continued.

Floppy mode density ϕ	Global	$\Phi = \frac{F}{6N - 6}$ $F = \text{Number of floppy modes}$ $N = \text{Number of atoms}$	ϕ monitors the rigid to flexible transition of networks by counting the floppy modes that are associated with dihedral rotations in a network.	Introduced by Thorpe ¹² and applied to proteins by Rader <i>et al.</i> ¹³ .	FIRST
DCM flexibility index θ	Global	$\theta \equiv \frac{\langle F \rangle}{N}$ $\langle F \rangle = \text{Average number of floppy modes}$ $N = \text{Number of residues in the protein}$	Characterizes global flexibility in an ensemble of network states. θ is zero in the rigid and large in the floppy state. For native-like proteins, $\theta \approx 1.0 \pm 0.4$.	Introduced and applied to proteins by Livesay <i>et al.</i> ¹⁴ .	Distance Constraint Model (DCM) ¹⁵
Percolation index p_i	Local	$p_i = \min \{ E_{cut} \mid A_{i1} \wedge A_{i2} \in c_{gpc}(E_{cut}) \}$ $p_i = \text{Percolation index of covalent bond } i$ $E_{cut} = \text{Hydrogen bond cutoff that defines a network state}$ $c_{gpc} = \text{Giant percolating cluster } c_{gpc} \text{ for a network state } \sigma = f(E_{cut})$ $A_{ij} = \text{Atom } j \text{ at bond } i$	p monitors the percolation behavior of a biomolecule on a microscopic level. It is derived for each bond i by E_{cut} when the bond segregates from the giant percolating cluster along a thermal unfolding simulation.	Introduced here for the first time.	Constraint Network Analysis (CNA) ²
Largest rigid cluster propensity P_{lrc_i}	Local	$P_{lrc_i} = \frac{1}{n_i} \sum_l^{n_l} \rho_l^\sigma$ $P_{lrc_i} = \text{Largest rigid cluster propensity of residue } i$ $n_i = \text{Number of steps in the bond dilution process}$ $\rho_i^l = \text{Residue } i, \text{ either part of the largest rigid cluster } (\rho = 1) \text{ or not } (\rho = 0) \text{ for a network state } \sigma.$	P_{lrc} monitors the propensity of a residue i of being part of the largest rigid cluster along a thermal unfolding simulation.	Introduced and applied to proteins by Rader ^{1,11,16} .	Not available

Table S1 continued.

Rigidity index r_i	Local	$r_i = \min \{ E_{cut} \mid \exists c \in C^{E_{cut}} : A_{i1} \wedge A_{i2} \in c \}$ r_i = Rigidity index of covalent bond i E_{cut} = Hydrogen bond cutoff that define a network state $C^{E_{cut}}$ = Set of rigid clusters for network state $\sigma = f(E_{cut})$ A_{ij} = Atom j at bond i	r monitors the transition from rigid to flexible for each residue. It is derived for each bond i by the E_{cut} value when the bond segregates from any rigid cluster along a thermal unfolding simulation.	Introduced here for the first time.	Constraint Network Analysis (CNA) ²
Backbone flexibility P_{R_i}	Local	$P_{R_i} = \frac{1}{k} \sum_{\sigma=1}^k \phi_i^{\sigma} \psi_i^{\sigma}$ P_{R_i} = Backbone flexibility of residue i k = Number of individual networks σ in an ensemble network $\phi_i^{\sigma} \psi_i^{\sigma}$ = Value is either “1” if the ϕ or ψ dihedral angle in residue i is rotatable, or “0” when both dihedral angles are locked	P_R monitors the probability that backbone dihedral angles (ϕ , ψ) are rotatable over the ensemble.	Introduced and applied to proteins by Livesay <i>et al.</i> ¹⁷ .	Distance Constraint Model (DCM) ¹⁵
DCM flexibility index ϑ_i	Local	$\vartheta_i = \frac{1}{k} \sum_{\sigma=1}^k (\rho_{F_i}^{\sigma} - \rho_{R_i}^{\sigma})$ ϑ_i = DCM flexibility index of region i k = Number of individual networks σ in an ensemble of networks ρ_{F_i} = Density of floppy modes in the i th region ρ_{R_i} = Density of redundant constraints in the i th region	ϑ monitors the average distance between the density of floppy modes and redundant constraints over an ensemble.	Introduced and applied to proteins by Livesay <i>et al.</i> ¹⁴ .	Distance Constraint Model (DCM) ¹⁵

Table S1 continued.

FIRST flexibility index g_i	Local	$g_i = \begin{cases} +1.0 & \text{dangling end} \\ \frac{F_j}{6E_j - B_j} & \text{flexible region} \\ 0.0 & \text{isostatic region} \\ -\frac{C_k - (6V_k - 6)}{6(V_k)(V_k - 1) - (6V_k - 6)} & \text{overconstrained region} \end{cases}$ <p> g_i = Flexibility index of the ith bond F_j = Number of floppy modes in the jth region E_j = Number of edges representing flexible bonds in the jth region B_j = Total number of bars in the jth region C_k = Total number of constraints in the kth region V_k = Number of atoms in the kth region </p>	g_i/f_i quantify the degree of flexibility/rigidity by counting the amount and spatial distribution of floppy modes respectively constraints.	Applied to proteins by Tan and Rader ⁸ .	FIRST
ProFlex/FIRST flexibility index f_i	Local	$f_i = \begin{cases} \frac{F_j}{H_j} & \text{flexible region} \\ 0.0 & \text{isostatic region} \\ -\frac{R_k}{C_k} & \text{overconstrained region} \end{cases}$ <p> f_i = Flexibility index of the ith bond F_j = Number of floppy modes in the jth region H_j = Total number of potentially rotatable bonds in the jth region R_k = Number of redundant constraints in the kth region C_k = Number of central-force bonds in the kth region </p>		Introduced by Jacobs et al. ¹⁸ and applied to proteins and RNA structures ¹⁹⁻²⁴ .	ProFlex/FIRST

Table S1 continued.

Stability maps rc_{ij}	Local	$rc_{ij} = \min \{E_{cut} \mid \exists c \in C^{E_{cut}} : R_i \wedge R_j \in c\}$ rc_{ij} = Rigid contact between the i th and j th residue E_{cut} = Hydrogen bond cutoff that define a network state R_i, R_j = Pair of residues i and j that is represented by their C_α atom $C^{E_{cut}}$ = Set of rigid clusters for network state $\sigma = f(E_{cut})$	<p>“Rigid contacts” between two residues, represented by their C_α atoms. A rigid contact exists if two residues belong to the same rigid cluster.</p>	Introduced and applied to proteins by Radestock and Gohlke ⁷ .	Constraint Network Analysis (CNA) ²
Cooperativity correlation plots	Local	$\frac{1}{k} \sum_{\sigma=1}^k \{1 \mid \exists m \in M^\sigma : R_i^\sigma \wedge R_j^\sigma \in m\}$ 0.0 $\frac{1}{k} \sum_{\sigma=1}^k \{1 \mid \exists c \in C^\sigma : R_i^\sigma \wedge R_j^\sigma \in c\}$ k M^σ C^σ R_i, R_j C_α atoms	<p>Cooperativity correlation plots monitor the statistical pairwise coupling in P_{R_i} information about which regions are correlated.</p> <p>Flexibly correlated residues Non-correlated residues Rigidly correlated residues</p> <p>= Number of individual networks in an ensemble network = Set of collective modes for a network state σ = Set of rigid cluster for a network state σ = Pair of residues i and j that is represented by their C_α atoms</p>	Introduced and applied to proteins by Livesay and Jacobs ^{17,25} .	Distance Constraint Model (DCM) ¹⁵

References in Supporting Information

1. Stauffer, D., Aharony, A. Introduction To Percolation Theory; Taylor and Francis: London, 1994.
2. Rath, P. C., Pfeiffer, C., Fulle, S., Klein, D.L., Gohlke, H. In Statics of biomacromolecules; Comba, P., Ed.; Wiley-VCH: Weinheim, 2011, p 281-299.
3. Stauffer, D. Phys Rep 1979, 54(1), 1-74.
4. Chubynsky, M. V.; Thorpe, M. F. Curr Opin Solid State Mater Sci 2001, 5(6), 525-532.
5. Rader, A. J.; Anderson, G.; Isin, B.; Khorana, H. G.; Bahar, I.; Klein-Seetharaman, J. Proc Natl Acad Sci USA 2004, 101(19), 7246-7251.
6. Radestock, S.; Gohlke, H. Eng Life Sci 2008, 8(5), 507-522.
7. Radestock, S.; Gohlke, H. Proteins 2010, 79(4), 1089-1110.
8. Tan, H. P.; Rader, A. J. Proteins 2009, 74(4), 881-894.
9. Andraud, C.; Beghdadi, A.; Lafait, J. Physica A 1994, 207(1-3), 208-212.
10. Tsang, I. R.; Tsang, I. J. Phys Rev E 1999, 60(3), 2684-2698.
11. Rader, A. J. Phys Biol 2010, 7(1), 016002.
12. Thorpe, M. F. J Non-Cryst Solids 1983, 57(3), 355-370.
13. Rader, A. J.; Hespenheide, B. M.; Kuhn, L. A.; Thorpe, M. F. Proc Natl Acad Sci USA 2002, 99(6), 3540-3545.
14. Livesay, D. R.; Dallakyan, S.; Wood, G. G.; Jacobs, D. J. FEBS Lett 2004, 576(3), 468-476.
15. Jacobs, D. J.; Dallakyan, S.; Wood, G. G.; Heckathorne, A. Phys Rev E 2003, 68(6), 061109.
16. Rader, A. J.; Brown, S. M. Mol Biosyst 2011, 7(2), 464-471.
17. Livesay, D. R.; Jacobs, D. J. Proteins 2006, 62(1), 130-143.
18. Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. Proteins 2001, 44(2), 150-165.
19. Hespenheide, B. M.; Rader, A. J.; Thorpe, M. F.; Kuhn, L. A. J Mol Graph Model 2002, 21(3), 195-207.
20. Gohlke, H.; Kuhn, L. A.; Case, D. A. Proteins 2004, 56(2), 322-337.
21. Mamonova, T.; Hespenheide, B.; Straub, R.; Thorpe, M. F.; Kurnikova, M. Phys Biol 2005, 2(4), S137-147.
22. Fulle, S.; Gohlke, H. Biophys J 2008, 94(11), 4202-4219.
23. Fulle, S.; Gohlke, H. Methods 2009, 49(2), 181-188.
24. Fulle, S.; Gohlke, H. J Mol Biol 2009, 387(2), 502-517.
25. Jacobs, D. J.; Livesay, D. R.; Hules, J.; Tasayco, M. L. J Mol Biol 2006, 358(3), 882-904.

13.4 Publication II

Efficient and Robust Analysis of Biomacromolecular Flexibility Using Ensembles of Network Topologies Based on Fuzzy Noncovalent Constraints

Pfleger, C., Gohlke, H.

Structure (2013), 21, 1725–1734

Author contribution to the publications:

My contribution to this publication was developing the ENT^{FNC} approach. To this end, I performed MD simulations on HEWL to derive definitions for fuzzy noncovalent constraints, parameterized and validated the ENT^{FNC} approach on several levels (II) on the HEWL dataset as well as (III) already existing external data. This results in a contribution of **70%** to this publication.

Cover article:



Efficient and Robust Analysis of Biomacromolecular Flexibility Using Ensembles of Network Topologies Based on Fuzzy Noncovalent Constraints

Christopher Pfleger¹ and Holger Gohlke^{1,*}

¹Mathematisch-Naturwissenschaftliche Fakultät, Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany

*Correspondence: gohlke@uni-duesseldorf.de
<http://dx.doi.org/10.1016/j.str.2013.07.012>

SUMMARY

We describe an approach (ENT^{FNC}) for performing rigidity analyses of biomacromolecules on ensembles of network topologies (ENT) generated from a single input structure. The ENT is based on fuzzy noncovalent constraints, which considers thermal fluctuations of biomacromolecules without actually sampling conformations. Definitions for fuzzy noncovalent constraints were derived from persistency data from molecular dynamics (MD) simulations. A very good agreement between local flexibility and rigidity characteristics from ENT^{FNC} and MD simulations-generated ensembles is found. Regarding global characteristics, convincing results were obtained when relative thermostabilities of citrate synthase and lipase A structures were computed. The ENT^{FNC} approach significantly improves the robustness of rigidity analyses, is highly efficient, and does not require a protein-specific parameterization. Its low computational demand makes it especially valuable for the analysis of large data sets, e.g., for data-driven protein engineering.

INTRODUCTION

Biomacromolecules are composed of flexible and rigid regions. This stability heterogeneity allows biomacromolecules to fulfill their diverse functional roles (Teague, 2003). Hence, a precise knowledge of flexibility and rigidity characteristics of a biomacromolecule is a prerequisite for understanding its function and valuable information for rational protein engineering and structure-based ligand design. Flexible and rigid regions in biomacromolecules can be determined with experimental methods (Henzler-Wildman and Kern, 2007; Cozzini et al., 2008; Sterner and Brunner, 2008; Bernadó, 2010; Kleckner and Foster, 2011; Hammel, 2012) and computational approaches, including molecular dynamics (MD) simulations (Young et al., 2001; Dodson and Verma, 2006; Cozzini et al., 2008) and approaches based on connectivity networks (Heringa and Argos, 1991; Dokholyan et al., 2002; Halle, 2002; Vendruscolo et al., 2002; Greene and Higman, 2003; Böde et al., 2007).

In a different approach, protein structures are modeled as constraint networks, which are analyzed by applying concepts based on rigidity theory. Here, atoms are represented as bodies and covalent and noncovalent interactions (hydrogen bonds, salt bridges, and hydrophobic tethers) as sets of bars (constraints; Jacobs et al., 2001; Rader et al., 2002). Initially, each body has six degrees of freedom (Whiteley, 2005). However, potential motions are constrained by the bars connecting the bodies. Once the network is constructed, the pebble game algorithm (Jacobs and Thorpe, 1995), implemented in the FIRST software, identifies flexible and rigid regions from the number and spatial distribution of the degrees of freedom ("constraint counting"). This analysis only takes seconds for a protein of ~300 residues. The theory underlying this approach is rigorous (Katoh and Tanigawa, 2011). Results from rigidity analyses have been successfully compared to results from experiments and those from other computational approaches (Hespenheide et al., 2002; Jacobs et al., 2003; Gohlke et al., 2004; Rader and Bahar, 2004; Livesay and Jacobs, 2006; Radestock and Gohlke, 2008; Fulle and Gohlke, 2009a).

Rather than analyzing "static" networks, several studies analyzed "perturbed" networks in which hydrogen bond constraints are sequentially removed, that way simulating the thermal unfolding of a biomacromolecule (Rader et al., 2002; Radestock and Gohlke, 2008, 2011; Rader, 2009; Rath et al., 2012). Hydrogen bonds are removed in the order of increasing strength according to a hydrogen bond energy E_{HB} (see Supplemental Experimental Procedures available online; Dahiyat et al., 1997). During the thermal unfolding simulation, a phase transition occurs at which the network loses the ability to carry stress; the transition point is referred to as rigidity percolation threshold (Thorpe, 1983) and has been related to the thermostability of proteins (Radestock and Gohlke, 2008, 2011; Rath et al., 2012).

Rigidity analyses are sensitive with respect to the input structural information (Gohlke et al., 2004; Mamonova et al., 2005). Two reasons account for this: (1) biomacromolecules have a soft matter-like character (Zaccai, 2000), i.e., noncovalent interactions frequently break and (re-)form ("flickering") such that the number and distribution of constraints in the networks vary; and (2) biomacromolecules are generally marginally stable (Taverna and Goldstein, 2002), i.e., their network state is close to the rigidity percolation threshold. As an overall consequence, a few constraints more or less can result in a network either being largely rigid or flexible.



The sensitivity problem can be overcome by performing rigidity analysis on an ensemble of conformations, e.g., generated by MD simulations (Gohlke et al., 2004; Rath et al., 2012). However, this compromises the efficiency of the rigidity analysis. To overcome this drawback, an ensemble of network topologies (ENT) can be generated from a single input structure by simulating the flickering of noncovalent constraints rather than the motions of the atoms. This idea has been pioneered in the distance constraint model (DCM; Jacobs et al., 2003; Livesay et al., 2004; Jacobs and Dalakyan, 2005) and the virtual pebble game (VPG) approach (González et al., 2012). While conceptually appealing, a downside of the DCM approach is that it requires experimental data for a protein-specific parameterization of the model. A downside of the VPG approach is that it is less accurate at the rigidity percolation threshold; analyzing network states of biomacromolecules around this threshold is particularly interesting, however.

In this study, we present an approach that performs rigidity analyses on an ENT generated from a single input structure (ENT^{FNC}) by using fuzzy noncovalent constraints (FNC). As such, the number and distribution of noncovalent constraints are modulated by random components within certain ranges, thus simulating thermal fluctuations of a biomacromolecule. The approach significantly improves the robustness of rigidity analyses, is highly efficient, and does not require a protein-specific parameterization. We analyzed MD simulations of hen egg white lysozyme (HEWL) structures as to the persistence of noncovalent bonds. From the analysis, we developed definitions for fuzzy hydrogen bond, salt bridge, and hydrophobic constraints. We validated the approach by comparison to rigidity analyses performed on single network topologies (SNT) of HEWL structures as well as on ensembles of HEWL conformations generated by MD simulations (ENT^{MD}). Furthermore, we demonstrate that our ENT^{FNC} approach is transferable to other protein systems. The ENT^{FNC} approach has been implemented into the CNA software (Pfleger et al., 2013a) and is available via the CNA web server (Krüger et al., 2013).

Theory

The FNC model consists of two parts related to the modeling of hydrogen bonds (including salt bridges) and hydrophobic tethers. Parameters of the model are derived from ENT^{MD} of HEWL. Values of the parameters are reported in the Results; here, we detail the theory underlying the FNC model.

Part Ia

To account for the flickering of hydrogen bonds (Zaccai, 2000), we determined the persistence characteristics of these interactions along MD trajectories. We did so for hydrogen bonds and salt bridges separately, and we distinguished hydrogen bonds in different secondary structure elements (α helices, 3_{10} helices, β sheets, and loop regions) following previous work (Stickle et al., 1992; Mamonova et al., 2005; Kieseritzky et al., 2006; Almond et al., 2007). From this, we derived the probability $p(HB, t)$ with which a hydrogen bond (salt bridge) of type t found in the input structure will be present across the ensemble of generated network topologies.

Part Ib

We next addressed that the hydrogen bond energy E_{HB} determines the order with which hydrogen bonds are removed in a thermal unfolding simulation. This order may strongly determine

the computed global and local stability characteristics. E_{HB} is computed by a simplified energy function (see Supplemental Experimental Procedures; Dahiyat et al., 1997). When analyzing a SNT, thermal motions of atoms are neglected that may influence E_{HB} and, hence, the order of hydrogen bond removal. To account for this effect, Gaussian white noise is added to $E_{HB,i}^{initial}$ computed for a hydrogen bond i from the single input structure (Equation 1).

$$E_{HB,i} = E_{HB,i}^{initial} + \mathcal{N}(0, SD_{HB,t}) \quad (\text{Equation 1})$$

The white noise is sampled from a Gaussian distribution $\mathcal{N}(0, SD_{HB,t})$ with a mean of zero and a standard deviation ($SD_{HB,t}$) that depends on the type t of the hydrogen bond; $SD_{HB,t}$ is determined from analyzing hydrogen bond energies in ENT^{MD}. Sampling from a Gaussian distribution follows the rationale that the shape of the energy well of a hydrogen bond can be fitted by a harmonic approximation (Leach, 2001) and that fluctuations about the minimum of a quadratic function show a Gaussian distribution (Levy and Karplus, 1979).

Part II

Regarding hydrophobic interactions, we wanted to model that these interactions are less specific than polar ones (Rose and Wolfenden, 1993). To do so, we developed a fuzzy constraint representation in which tethers between closer atoms are included with a higher probability in a network topology than those between atoms further apart; more specifically: (1) tethers between atoms that are in van der Waals contact d_{vdW} (van der Waals radii: C: 1.7 Å, S: 1.8 Å) are always included; (2) tethers between atoms that are further apart than $d_{vdW} + D_{max}$ are never included; here, $D_{max} = 1.5$ Å because this value equates to half of the distance between the contact minimum and the solvent-separated minimum in a potential of mean force of hydrophobic solutes (Pratt and Chandler, 1977); and (3) for distances in between, the probability for a tether to be included is computed from Equation 2, which approaches these two extremes.

$$p(d_{ij}) = e^{-\frac{1}{2} \left(\frac{(d_{ij} - d_{vdW})^2}{D_{cut}^2} \right)} \quad (\text{Equation 2})$$

$p(d_{ij})$ is a Gaussian with a squared distance dependency, d_{ij} is the distance between two atoms i and j , and D_{cut} determines the full width at half maximum of the Gaussian. Gaussians have been applied successfully for modeling the strength of pairwise interactions between hydrophobic atoms (Crivelli et al., 2002; Huey et al., 2007; Forli and Olson, 2012). Preliminary tests have shown that it is advantageous to favor hydrophobic tethers at shorter distances; this is accounted for by the squared distance dependency.

With this FNC model, in five steps, the ENT^{FNC} is generated from a single input structure, and global and local stability characteristics are analyzed (Figure 1):

- (1) An initial network topology is generated from the input structure.
- (2) Information about noncovalent constraints is extracted. Hydrogen bonds (including salt bridges) with

$$E_{HB}^{initial} < 0 \text{ kcal/mol}$$

and hydrophobic tethers between C and/or S atoms with a distance $< d_{vdW} + D_{max}$ are identified. The secondary

Structure

Fuzzy Noncovalent Constraints in Rigidity Analysis

CNA

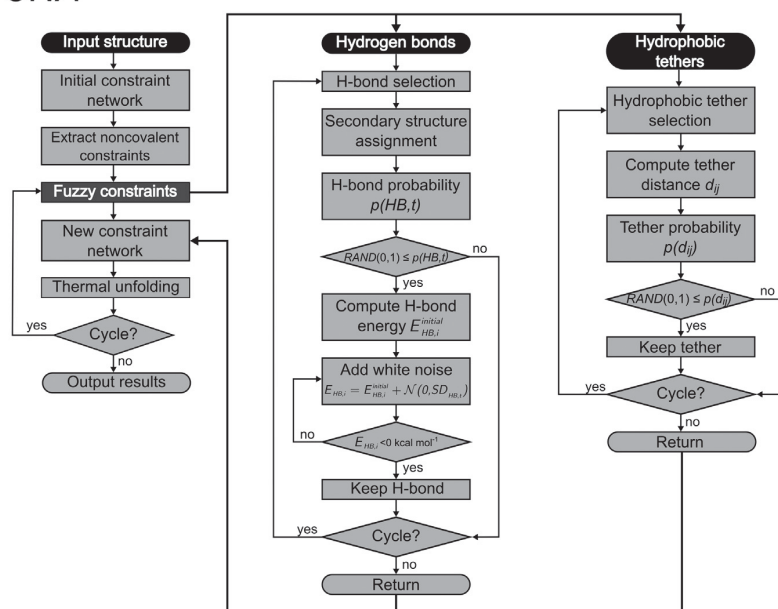


Figure 1. Work Flow of the ENT^{FNC} Approach

The ENT^{FNC} approach has been integrated into the CNA software package. $RAND(0,1)$ draws a random number with equal probability from the range $[0, 1]$. See Theory for further details.

Data Bank [PDB] IDs: 1LYO, 1VDP, 1F10, 1HSX, 1LSE, and 1LYS); and (2) a single rigid cluster dominates the system and contains between 57% and 72% of all atoms (PDB IDs: 2C8O, 1LSF, 3LZT, and 193L). These differences in the RCD results originate from differences in the number and distribution of noncovalent constraints in the network topologies: Network topologies with more constraints result in the “single rigid cluster” RCD, and the largest difference in the number of constraints across all ten networks is 41 (22%; Table S2).

To quantify the results of the rigidity analysis, the C_α atom-based rigidity index r_i was computed by CNA for each HEWL structure (Figures 2B and 2D). r_i maps flexibility and rigidity characteristics within a network topology by monitoring when a covalent bond segregates from a rigid cluster during a thermal unfolding simulation (Pfleger et al., 2013b). Figure 2B reveals that the local stability characteristics are only moderately consistent in helices A, B, D, and E (SD ~ 0.8 kcal/mol) and vary strongly in the beta sheet region C (SD up to 2.5 kcal/mol). Particularly large SDs are observed for the “spikes” at residues D52, W62, N65, T69, and I78, which reveal regions that are highly stabilized by interactions to hydrophobic atoms (Figure S4). Accordingly, only 11% of the HEWL residues show r_i values that differ across all ten curves by < 1.0 kcal/mol; the maximal difference is found for residue M105 (6.7 kcal/mol) in helix E. Note that differences of 1.0 kcal/mol can already lead to misinterpretations of results from rigidity analysis considering that significant differences in the (relative) thermostability of proteins have been mapped to energy differences of the same magnitude (Radestock and Gohlke, 2008, 2011; Rathi et al., 2012). The overall picture does not change when energy-minimized structures are used as input (Figure S1), although the SDs drop to values of ~ 0.5 (0.9) kcal/mol in the case of helices A, B, and D (helix E and region C). Two “spikes” still prevail. Accordingly, only 30% of the HEWL residues show r_i values that differ across all ten curves by < 1.0 kcal/mol. Overall, these results demonstrate a high sensitivity of the rigidity analysis with respect to the input structure. This is remarkable because the structural deviation of the ten HEWL structures is only slightly larger than the uncertainty in the structure determination.

structure a hydrogen bond or salt bridge is involved in is assigned from the input structure using DSSP (Joosten et al., 2011).

- (3) The number and distribution of noncovalent constraints is modified according to the definitions of FNC.
- (4) A network topology is built.
- (5) Global and local stability characteristics are computed by the CNA software.

Steps 3–5 are repeated until a user-specified number of networks is generated over which the global and local stability characteristics are averaged.

RESULTS

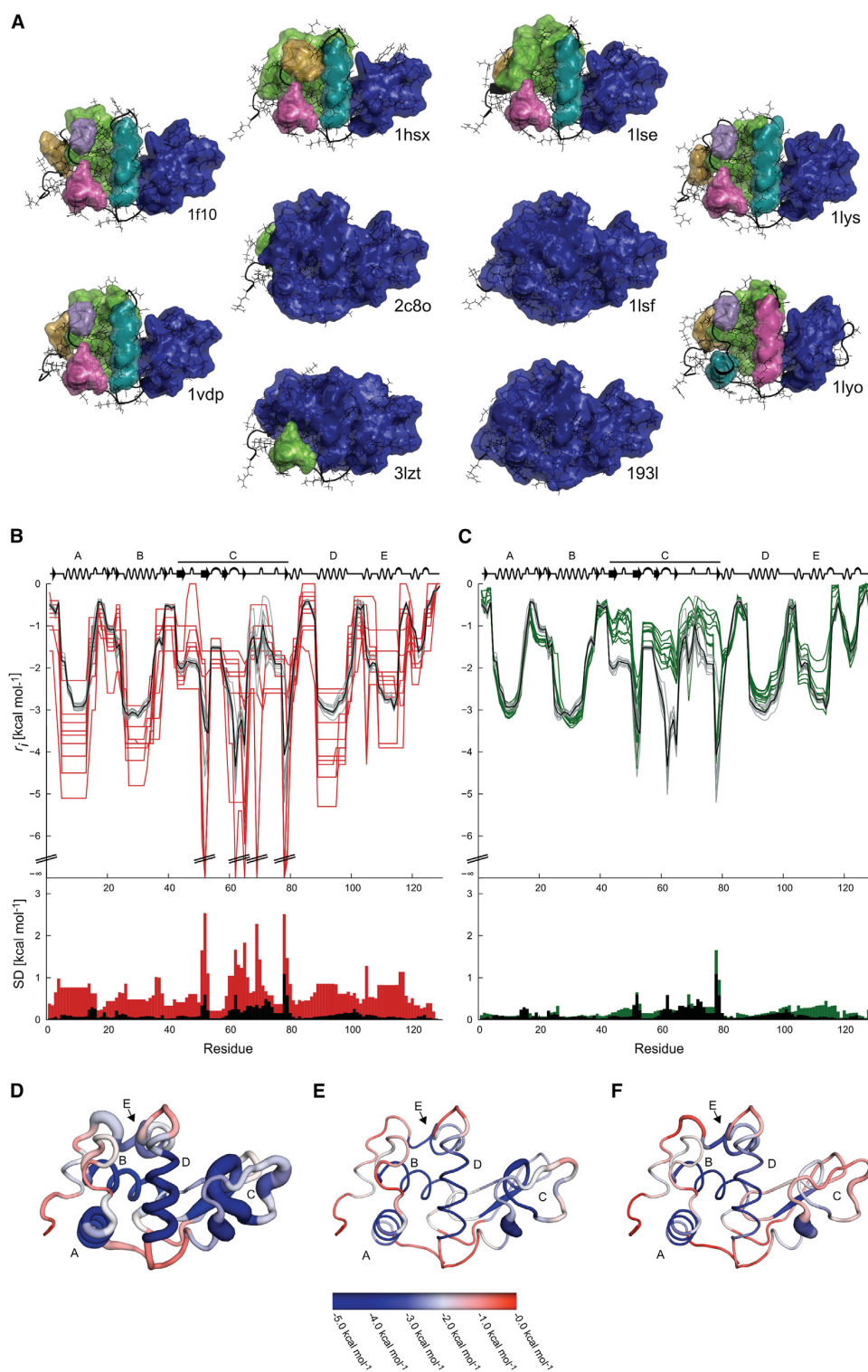
Rigidity Analyses Are Highly Sensitive with Respect to the Input Structure

To illustrate to what extent the results of a rigidity analysis depend on the chosen input structure, we computed rigid cluster decompositions (RCD) using FIRST for network topologies derived from ten HEWL crystal structures (SNT approach; Figure 2A). All of these structures (see Supplemental Experimental Procedures) have been resolved from 0.93 Å to 1.93 Å. Structures of a similar quality have been used in FIRST analyses before (Gohlke et al., 2004; Radestock and Gohlke, 2008; Fulle and Gohlke, 2009b; Rathi et al., 2012). The structures have been selected from a set of 38 HEWL structures with the aim to avoid structural redundancy (see Supplemental Experimental Procedures; Table S1). This resulted in mutual root-mean-square deviations (rmsd) of the C_α atoms (of all atoms) of at most 0.86 Å (1.85 Å). The RCD results fall into two classes: (1) between 43% and 50% of all atoms are part of a rigid cluster with the remaining ones being located in flexible regions (Protein

flexibility and rigidity characteristics within a network topology by monitoring when a covalent bond segregates from a rigid cluster during a thermal unfolding simulation (Pfleger et al., 2013b). Figure 2B reveals that the local stability characteristics are only moderately consistent in helices A, B, D, and E (SD ~ 0.8 kcal/mol) and vary strongly in the beta sheet region C (SD up to 2.5 kcal/mol). Particularly large SDs are observed for the “spikes” at residues D52, W62, N65, T69, and I78, which reveal regions that are highly stabilized by interactions to hydrophobic atoms (Figure S4). Accordingly, only 11% of the HEWL residues show r_i values that differ across all ten curves by < 1.0 kcal/mol; the maximal difference is found for residue M105 (6.7 kcal/mol) in helix E. Note that differences of 1.0 kcal/mol can already lead to misinterpretations of results from rigidity analysis considering that significant differences in the (relative) thermostability of proteins have been mapped to energy differences of the same magnitude (Radestock and Gohlke, 2008, 2011; Rathi et al., 2012). The overall picture does not change when energy-minimized structures are used as input (Figure S1), although the SDs drop to values of ~ 0.5 (0.9) kcal/mol in the case of helices A, B, and D (helix E and region C). Two “spikes” still prevail. Accordingly, only 30% of the HEWL residues show r_i values that differ across all ten curves by < 1.0 kcal/mol. Overall, these results demonstrate a high sensitivity of the rigidity analysis with respect to the input structure. This is remarkable because the structural deviation of the ten HEWL structures is only slightly larger than the uncertainty in the structure determination.

Results of Rigidity Analyses Averaged over ENT^{MD} Are Starting Structure Independent

Previous studies have pointed to the benefit of averaging results from rigidity analysis over ensembles of conformations from MD simulations in terms of a much decreased sensitivity with



(legend on next page)

Structure

Fuzzy Noncovalent Constraints in Rigidity Analysis

respect to the input structure (Gohlke et al., 2004; Rath et al., 2012). To provide benchmark results for subsequent analyses, we generated MD trajectories of 300 ns length starting from each of the ten HEWL structures (Table S1). Our simulation setup and a simulation length of that order provide an accurate representation of HEWL dynamics (Koller et al., 2008). We then repeated the rigidity analyses by CNA, averaging over 1,500 conformations extracted from each of the trajectories (ENT^{MD} approach). As the main outcome, the averaged local stability characteristics are much more consistent (Figures 2B and 2E) than if SNT were analyzed, with the SD of all MD averages < 0.3 kcal/mol in general and < 0.6 kcal/mol in region C. The sole exception concerns residue I78 with SD = 1.1 kcal/mol. Accordingly, 90% of the HEWL residues show r_i values that differ across all 10 curves by < 1.0 kcal/mol; the maximal difference is found at residue I78 in region C (3.3 kcal/mol). Additionally, the stable regions identified in all ensembles for residues Y53, W62–N65, and I78 are in very good agreement with protection factors determined by H/D experiments for HEWL (Radford et al., 1992). In contrast, the stable regions are only identified in five of the ten HEWL structures when using the SNT approach (Figure S4). Additionally, analyzing structures simulated at 300 K leads to larger r_i values (indicating a higher flexibility) for most of the regions of HEWL than in the case of the HEWL crystal structures (Figures 2B and 2D). This finding is important if results from rigidity analysis on biomacromolecules are to be compared to experimental data obtained at room temperature. In summary, averaging over ENT^{MD} leads to robust, i.e., starting structure-independent, results of rigidity analyses.

Parameterizing Fuzzy Noncovalent Constraints Using Data on Breaking and (Re-)Forming of Noncovalent Interactions from MD Simulations

For parameterizing the FNC model, we first analyzed the persistence characteristics of noncovalent bonds during the MD simulations of the ten HEWL structures. Using definitions of noncovalent constraints as in FIRST, we monitored hydrogen bonds (including salt bridges) and hydrophobic tethers. The results show a bimodal distribution of the persistence characteristics of hydrogen bonds, with about 77% persisting for < 60 ns (i.e., 20% of the trajectory length) and 7% being stable almost across the entire simulation (persistence time > 90% of the trajectory length; Figure S2A). In contrast, the majority of hydrophobic tethers (89%) persist only for < 10% of the trajectory length (Figure S2B).

We furthermore analyzed whether hydrogen bonds are more persistent in secondary structure elements such as α helices, 3_{10} helices, or β sheets than in loop regions. Secondary structure

Table 1. Type-Dependent Probabilities and Standard Deviations of Hydrogen Bond Energies

Type of Hydrogen Bond	$p(\text{HB}, t)^{a,b}$	$SD_{\text{HB}, t}^c$
α Helix (1 \rightarrow 5)	0.8	2.0 ± 0.02
3_{10} Helix (1 \rightarrow 4) ^d	0.6	1.2 ± 0.03
β Sheet	0.8	1.5 ± 0.04
sp ² -sp ²	0.4	1.6 ± 0.02
sp ² -sp ³	0.3	2.0 ± 0.05
sp ³ -sp ²	0.5	1.6 ± 0.04
sp ³ -sp ³	0.5	1.5 ± 0.03
Salt bridge	0.8	0.7 ± 0.04

^aHydrogen bond energies were computed from geometric parameters (Dahiyat et al., 1997).

^bProbability with which a hydrogen bond (salt bridge) will be present in generated network topologies.

^cSD and SEM in kcal/mol.

^dIncluding hydrogen bonds in β turns.

elements were identified by DSSP (Joosten et al., 2011) in each conformation extracted from the MD trajectories. The analysis reveals that >85% of the backbone hydrogen bonds in α helices, 3_{10} helices, and β sheets persist in > 80% of the trajectory length (data not shown), in agreement with previous studies (Stickle et al., 1992; Kieseritzky et al., 2006; Almond et al., 2007). In contrast, hydrogen bonds between charged groups (salt bridges) are only present in about 20% of the extracted conformations (data not shown), again in agreement with previous findings (Mamonova et al., 2005).

To determine to what extent the energy E_{HB} of a hydrogen bond fluctuates, hydrogen bonds with a persistence of >10% of the trajectory length were analyzed. We distinguished between backbone hydrogen bonds of α helices, 3_{10} helices (including β turns), and β sheets. For all other polar interactions, we distinguished hydrogen bonds from salt bridges and further classified hydrogen bonds with respect to the hybridization state of the donor and acceptor atoms (sp²-sp², sp²-sp³, sp³-sp², and sp³-sp³). $SD_{\text{HB}, t}$ (Equation 1) was then calculated from the energies of all hydrogen bonds of type t found in all conformations extracted from the ten independent MD trajectories. Table 1 shows that $SD_{\text{HB}, t}$ is type-dependent. The largest fluctuations are found for backbone hydrogen bonds in α helices and hydrogen bonds involving sp²-sp³ hybridized donor and acceptor atoms. The lowest fluctuation is found for salt bridges (0.7 kcal/mol). The standard error of the mean (SEM) of $SD_{\text{HB}, t}$ is estimated from ten fluctuation values originating from each of the ten independent MD trajectories. The SEM is

Figure 2. Local Stability Characteristics of HEWL

(A) Rigid cluster decompositions using the FIRST program (Jacobs et al., 2001) obtained with a cutoff of the hydrogen bond energy $E_{\text{HB}} = -1.0$ kcal/mol and a hydrophobic tether distance cutoff $D_{\text{cut}} = 0.25$ Å. Rigid clusters are depicted as uniformly colored bodies with the largest rigid cluster in blue.
 (B) Rigidity index r_i for the SNT analyses of the ten HEWL structures (red), r_i curves for the ten ENT^{MD} analyses of HEWL (gray), and the average over all ENT^{MD} analyses (black). The histogram below shows the standard deviation of the r_i s across the crystal structures and the MD ensembles, respectively.
 (C) Rigidity index r_i for the ENT^{FNC} analyses of the ten HEWL structures (green). The histogram below shows the standard deviation of the r_i s. For comparison, the results from the ENT^{MD} analyses are depicted again.
 (D–F) The mean r_i values and standard deviations from the SNT analyses (D), ENT^{MD} analyses (E), and ENT^{FNC} analyses (F) are mapped onto a HEWL structure. The colors show the r_i values and the diameter of the putty plot the standard deviation at each residue position. The diameter is scaled with respect to the maximum SD of all three analyses.
 See also Table S2 and Figures S1 and S4.

≤ 0.05 kcal/mol for each hydrogen bond type, i.e., it is $< 10\%$ even in the case of the lowest fluctuation found for salt bridges.

We next used these results for parameterizing the flickering of hydrogen bonds (including salt bridges) in the FNC model (Part Ia in Theory). Some backbone hydrogen bonds are particularly important for the stability of secondary structures and have a pronounced persistence along a MD trajectory. As such, we required backbone hydrogen bonds in α helices ($1 \rightarrow 5$) and β sheets to be present in 80% of the generated network topologies (i.e., $p(HB, t) = 0.8$). Preliminary tests showed that backbone hydrogen bonds in 3_{10} helices ($1 \rightarrow 4$) should be included in 60% of the generated network topologies. Hydrogen bonds in β turns are treated as those in 3_{10} helices (Baker and Hubbard, 1984). All other sp^2 - sp^2 , sp^2 - sp^3 , sp^3 - sp^2 , and sp^3 - sp^3 hydrogen bonds have been found to be less persistent and, thus, are included in 40% (sp^2 - sp^2), 30% (sp^2 - sp^3), and 50% (sp^3 - sp^2 , sp^3 - sp^3) of the generated network topologies. In total, this leads to average numbers of hydrogen bonds in the generated network topologies that differ by $< 4\%$ from the average values of the MD ensembles (Figure S3A). Although hydrogen bonds between charged groups (salt bridges) reveal a low persistence along the MD trajectories, we required that these interactions be present in 80% of the generated network topologies. This still leads to generally lower numbers of salt bridges in the generated network topologies compared to MD results (Figure S3A). However, the difference in the absolute numbers amounts to only two to three salt bridges (i.e., $\sim 5\%$ with respect to all polar interactions) for the HEWL system.

Second, we addressed the effect of thermal motions on computed hydrogen bond energies E_{HB} (Part Ib in Theory) by applying the $SD_{HB, t}$ (Table 1) in Equation 1. Only hydrogen bonds with $E_{HB} < 0$ kcal/mol are included in a new network. Varying E_{HB} that way yields rearranged orders in which hydrogen bonds are removed during a thermal unfolding simulation. Kendall's τ coefficient reveals that the orders in 1,500 networks generated that way are independent from the order in the underlying crystal structure (SNT versus ENT^{FNC} in Table S3; the mean \pm SEM over all ten cases is $\tau = 0.029 \pm 0.013$). The same result is obtained if the order of hydrogen bonds is compared between networks generated from 1,500 conformations extracted from MD trajectories and the underlying crystal structure, respectively (SNT versus ENT^{MD} in Table S3; $\tau = 0.019 \pm 0.007$). Finally, the pairwise comparison of the orders of hydrogen bonds in 1,500 networks from ENT^{FNC} versus 1,500 networks from ENT^{MD} also reveals nonexistent correlations on average (Table S3; $\tau = 0.033 \pm 0.013$). However, for all ten independent MD simulations, for $> 97\%$ of the networks extracted from the MD trajectory at least one network from ENT^{FNC} is found where the orders of hydrogen bonds significantly ($p < 0.01$) correlate; in these cases $\tau > 0.15$ (Figure S3B). This observation is notable in that it already suggests that ensembles generated from either sampling scheme should lead to similar results in thermal unfolding simulations.

Third, we addressed the less specific character of hydrophobic tethers by favoring tethers at shorter distances over those at longer distances (Part II in Theory). In Equation 2, D_{cut} was set to 0.25 Å as used in the SNT and ENT^{MD} approaches. With these settings, the average numbers of hydrophobic tethers in networks generated by the ENT^{FNC} approach differ by $< 14\%$ from those found in networks generated from MD trajectories (Figure S3A).

In summary, our definitions of FNCs for polar interactions and hydrophobic tethers yield noncovalent constraints in ENT^{FNC} derived from single crystal structures that agree very well in terms of ensemble properties with noncovalent bonds identified in structures from MD simulations.

Averaged Results from ENT^{FNC} Agree Almost Perfectly with Those from ENT^{MD}

For validation, we applied the ENT^{FNC} approach on each of the ten energy minimized HEWL structures. Results from CNA were averaged over 1,500 network topologies each. The results of the ENT^{FNC} analyses (Figures 2C and 2F) are considerably more consistent than if a SNT was analyzed, with the SD of all ensemble averages < 1.0 kcal/mol except for residue I78 with $SD = 1.9$ kcal/mol. Accordingly, 88% of the HEWL residues show r_i values that differ across all ten curves by < 1.0 kcal/mol, which is in remarkable agreement with ENT^{MD} results. The largest differences are observed in region C and helix E. For testing the sensitivity of these results on $SD_{HB, t}$ in Equation 1, we repeated the above calculations, once setting the SD to $SD_{HB, t} + SEM_t$ and once to $SD_{HB, t} - SEM_t$ (Table 1). In each case, the results are within the uncertainty of the calculations obtained with $SD_{HB, t}$ across all ten HEWL systems (data not shown). This demonstrates that the ENT^{FNC} results are robust with respect to variations in $SD_{HB, t}$. Compared to the ENT^{MD} results, the identification of known stable regions (Y53, W62–N65, and I78) is less pronounced; still, these regions reveal the highest stability characteristics in region C. Additionally, analyzing ENT^{FNC} topologies leads to larger r_i values (indicating a higher flexibility) for most of the regions of HEWL than in the case of the SNT from the HEWL structures (Figures 2B and 2D). Except for region C, these r_i values are nearly identical to those derived from ENT^{MD} topologies. In summary, averaging over an ENT^{FNC} leads to robust, i.e., less starting structure-dependent rigidity analyses, as observed for the ENT^{MD} approach. The local flexibility and rigidity characteristics also agree almost perfectly with those from ENT^{MD} analyses in terms of the magnitudes of the r_i values.

Validation of the ENT^{FNC} Approach on External Data Sets

We next applied the ENT^{FNC} approach on an external data set not used for the parameterization of the FNC. We investigated five citrate synthase (CS) structures from different organisms with respect to their global stability characteristics by means of thermal unfolding simulations (Rathi et al., 2012). The organisms differ in their optimal growth temperatures (T_{og}), which range from 310.2 K to 373.2 K (Table S4). Four of the CS structures are crystal structures; the fifth (TsCS) has been generated by homology modeling. We generated ENT^{FNC} with 1,500 topologies for each CS structure. As a reference, ENT^{MD} with 1,500 conformations were extracted from MD trajectories of 30 ns length. Additionally, we analyzed either the crystal structures/homology model (SNT) or these structures after energy minimization (SNT^{min}). Phase transitions temperatures T_p were computed from the change of the cluster configuration entropy H_{type2} along the thermal unfolding simulation (Table S4; Pfleger et al., 2013b), making use of a linear relationship parameterized on melting temperatures of pairs of homologs from meso- and thermophilic organisms (Radestock and Gohlke, 2011). Often, melting temperatures are estimated from T_{og} values by assuming

Structure

Fuzzy Noncovalent Constraints in Rigidity Analysis

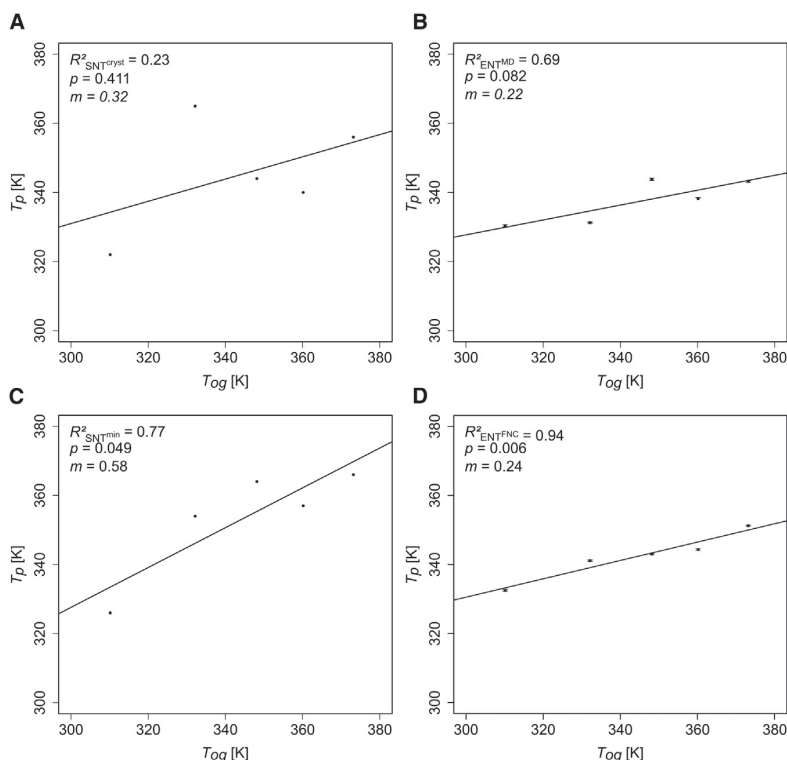


Figure 3. Correlations between Predicted T_p and Optimal Growth Temperatures T_{og}
(A–D) Correlations between predicted T_p and optimal growth temperatures T_{og} from SNT (A), ENT^{MD} (B), SNT^{min} (C), and ENT^{FNC} (D) analyses of five different CS structures. Error bars in (B) and (D) show the SEM. Least squares fit lines have been added.

See also Table S4.

largest slope $m_{SNT^{min}} = 0.58$. In view of this, the phase transition temperatures determined from thermal unfolding simulations should be considered relative values only (Radestock and Gohlke, 2011). Still, the temperatures are very helpful, e.g., when it comes to comparing the thermostability of two or more homologous proteins or the stability of a wild-type with its mutant (Radestock and Gohlke, 2008, 2011; Rathi et al., 2012).

All CS are structurally highly similar (rmsd of the C_α atoms: 1.22–2.32 Å) but differ strongly in the pairwise sequence identities (~20% to ~60%). In contrast, for another data set currently under investigation in our group (P.C. Rathi, H.G., unpublished data), mutants with improved thermostability have been generated from a wild-type lipase A (Ahmad et al.,

2008; Ahmad and Rao, 2009), resulting in high pairwise sequence identities of the 14 structures (> 93%). This allows us to cite some additional, yet preliminary results. When considering 12 of the 14 structures (the structures with the lowest and highest thermostabilities are treated as outliers), T_p computed by the ENT^{FNC} approach correlate fairly with experimentally determined melting temperatures ($R^2 = 0.52$, $p = 0.008$, $m_{ENT^{single}} = 0.28$). In our opinion, this is a remarkable result as to the predictive power of the ENT^{FNC} approach because the sequences of some of the mutant pairs differ by only one amino acid.

that they are ~25 K higher (Dehouck et al., 2008; Radestock and Gohlke, 2008, 2011); that way, they also provide a relationship between T_p and T_{og} . In the case of analyzing ENT^{FNC} and ENT^{MD} , T_p is averaged over ensembles of 1,500 networks; in both cases, the SEM of T_p is < 0.5 K (Table S4).

No correlation ($R^2 = 0.23$, $p = 0.411$) between T_p and T_{og} values is found if SNT were used (Figure 3A); the T_p of the homology model is largely over predicted. If CNA is performed on ENT^{MD} , a fair correlation is found ($R^2 = 0.69$, $p = 0.082$; Figure 3B), demonstrating a successful structural refinement of the homology model. However, the CS structures are not correctly ranked with respect to their T_{og} values. The SNT^{min} yield a good correlation ($R^2 = 0.77$, $p = 0.049$; Figure 3C), again largely due to a better prediction for the homology model. Still, the ranking of the CS structures is not perfect. The best correlation is obtained if the ENT^{FNC} approach is pursued ($R^2 = 0.94$, $p = 0.006$; Figure 3D). Now, all CS structures are correctly ranked with respect to their T_{og} values.

The comparison of the results from the SNT, SNT^{min} , ENT^{MD} , and ENT^{FNC} approaches reveals an influence of the treatment of the CS structures on the slope of the correlation lines. The ensemble-based approaches lead to lower slopes ($m_{ENT^{MD}} = 0.22$, $m_{ENT^{FNC}} = 0.24$) than if the crystal structures/homology model are analyzed ($m_{SNT} = 0.32$). “Heating” structures to 300 K thus seems to obliterate differences in the thermostability. Apparently, the ENT^{FNC} approach implicitly captures part of this temperature effect. The opposite is observed if the structures are “cooled” as in the case of SNT^{min} , resulting in the

DISCUSSION

We introduced the ENT^{FNC} approach with the aim to improve the robustness of rigidity analyses of biomacromolecules while preserving their computational efficiency. To this end, we provided definitions for FNC based on persistency data of noncovalent bonds derived from MD simulations. With the FNC, an ENT is generated from a single input structure over which results from rigidity analyses are averaged. Thus, by mimicking the flickering of noncovalent bonds, the ENT^{FNC} approach allows performing rigidity analyses on ensembles of network topologies rather than on ensembles of conformations.

The approach was validated at several levels. As to the definitions of fuzzy hydrogen bonds and salt bridges, the network topologies generated with FNC differ by at most 5% from those generated from MD ensembles in terms of the average number of polar interactions. The definition of fuzzy hydrophobic

tethers leads to a related difference of at most 14%. The higher difference in the latter case may reflect that hydrophobic interactions are less specific than polar ones (Mamonova et al., 2005). Furthermore, for almost all networks extracted from an MD trajectory at least one network from ENT^{FNC} is found where the sequences with which hydrogen bonds are removed from a network during a thermal unfolding simulation significantly correlate.

At the next level, we compared results from rigidity analyses on ENT^{FNC} to those on ENT^{MD}; in both cases, the same ten HEWL structures were used as input for the ENT^{FNC} approach or as starting structures for the MD simulations. Remarkably, averaging over an ENT^{FNC} leads to robust, almost starting structure-independent rigidity analyses, as found for the ENT^{MD} approach. Furthermore, local flexibility and rigidity characteristics determined for the ENT^{FNC} agree almost perfectly with those from the ENT^{MD} approach in terms of the magnitudes of the r_i . These findings indirectly confirm the appropriateness of the FNC definitions. Furthermore, they suggest that the ENT^{FNC} approach is viable for overcoming the problem of the sensitivity of rigidity analyses with respect to the input structure.

Finally, we tested the ENT^{FNC} approach for computing relative thermostabilities on data sets of CS and lipase A structures. In both cases, convincing results were obtained. These results are encouraging, for two reasons: (1) both protein systems were not used in the course of parameterizing the FNC. Hence, this demonstrates the transferability of the ENT^{FNC} approach; and (2) both protein systems strongly differ in terms of the extent of the sequence similarity among the members. Yet, even for the series of sequentially highly similar lipase A proteins, the ENT^{FNC} approach shows a high predictive power. This indicates that, while the ENT^{FNC} approach is largely insensitive with respect to small conformational changes of input structures, it remains sensitive enough to pick up effects on the thermostability due to small sequential variations.

A major advantage of the ENT^{FNC} approach over the ENT^{MD} approach is the computational efficiency. As such, the MD simulations for ENT^{MD} required ~16 (8) days of computing time per HEWL (CS) structure on a single NVIDIA Tesla M2070 GPU. Clearly, that way the computational efficiency of a rigidity analysis is compromised. In contrast, the ENT^{FNC} approach only required ~6 min (~2 hr) per HEWL (CS) structure for energy minimization and ~2 hr (~19 hr) per HEWL (CS) structure for the ENT^{FNC} analyses. This amounts to a speed up of a factor of ~4,000 (~100) in the case of HEWL (CS). With that, the ENT^{FNC} approach seems well suited for application in large-scale studies on proteins, e.g., for predicting the effects of site-saturation mutagenesis on thermostability.

As a downside, the ENT^{FNC} approach only mimics the flickering of noncovalent bonds for a given conformation of the biomacromolecule. In contrast, changes in the network due to conformational changes of the biomacromolecule—as additionally detected by the ENT^{MD} approach—will be missed. Thus, the ENT^{FNC} approach is prone to fail if such conformational changes have a determining influence on the biomacromolecule's stability or function. In turn, the ENT^{FNC} approach should be most suitable for comparing biomolecular systems where major conformational changes are not expected. This should be given when comparing homologous proteins (Radestock and Gohlke,

2008, 2011) or wild-type and mutant proteins, making the ENT^{FNC} approach well applicable for data-driven protein engineering. Likewise, the ENT^{FNC} approach could be applied for estimating the influence of ligand molecules on biomolecular stability (Gohlke et al., 2004) if the ligand binding is not accompanied by a large induced fit. In that respect, it is encouraging to note that HEWL structures within one MD trajectory differed by up to 3.3 Å C_α rmsd, yet, local flexibility and rigidity characteristics determined by the ENT^{FNC} approach agree almost perfectly with those from the ENT^{MD} approach.

A few approaches exist that are similar in spirit to the ENT^{FNC} approach. The DCM approach generates an ENT by considering mean-field probabilities of hydrogen bond and torsion constraints in a Monte Carlo sampling. Average stability characteristics are then calculated by rigidity analyses on each topology in the ensemble. While conceptually appealing, a downside of the DCM approach is that it requires experimental data for a protein-specific parameterization of the model (Jacobs et al., 2003; Livesay et al., 2004; Jacobs and Dallakyan, 2005). Recently, the VPG has been introduced, which provides ensemble averaged descriptions of a biomacromolecule's flexibility and rigidity without having to sample multiple network topologies (González et al., 2012). While it is highly efficient, the VPG suppresses fluctuations of network rigidity and, hence, tends to be less accurate at the rigidity percolation threshold where most such fluctuations occur (Gonzalez et al., 2011). This is a drawback when analyzing biomacromolecules considering that they are generally marginally stable (Taverna and Goldstein, 2002), i.e., their network state is close to the rigidity percolation threshold. As a further development, the VPG-x approach improves the accuracy of the description of network rigidity by combining the original VPG with a statistical sampling approach albeit at the cost of losing VPG's efficiency. As all approaches sample over an ENT either directly or indirectly, DCM, VPG-x, and ENT^{FNC} belong to the same computational complexity class. Regarding the representation of noncovalent constraints in these approaches, we see it as an advantage that the FNC defined in this study have been parameterized based on data from state-of-the-art MD simulations. Thus, the definitions should implicitly include solvation and temperature effects. Furthermore, no protein-specific information was used; rather, the definitions are based on hybridization states, atom types, and secondary structure and thus are transferable to other protein systems.

In summary, the ENT^{FNC} approach introduced here has been demonstrated to be a viable approximation to the ENT^{MD} approach for performing ensemble-based rigidity analyses on biomacromolecules in a computationally efficient manner. Our results position the ENT^{FNC} approach for linking biomolecular structure, flexibility, (thermo-)stability, and/or function for large-scale data sets of systems where only limited conformational changes occur. The ENT^{FNC} approach should thus be a valuable complement to the existing approaches for biomolecular rigidity analysis.

EXPERIMENTAL PROCEDURES

Details on the structure preparation of the HEWL and CS systems, the setup and execution of MD simulations of the HEWL and CS systems, and the computation of global and local stability characteristics in the case of the

Structure

Fuzzy Noncovalent Constraints in Rigidity Analysis

SNT and ENT^{MD} approaches are given in the [Supplemental Experimental Procedures](#).

SUPPLEMENTAL INFORMATION

Supplemental information includes Supplemental Experimental Procedures, four figures, and four tables and can be found with this article online at <http://dx.doi.org/10.1016/j.str.2013.07.012>.

ACKNOWLEDGMENTS

We thank Doris L. Klein, Nadine Homeyer, Daniel Mulnaes, and Britta Nisius for fruitful discussions. We thank Prakash Chandra Rath for providing the dataset of lipase A structures. H.G. acknowledges insightful conversations with Ileana Streinu, Meera Sitharam, and Mike Thorpe.

Received: April 27, 2013

Revised: July 4, 2013

Accepted: July 17, 2013

Published: August 29, 2013

REFERENCES

- Ahmad, S., and Rao, N.M. (2009). Thermally denatured state determines refolding in lipase: mutational analysis. *Protein Sci.* **18**, 1183–1196.
- Ahmad, S., Kamal, M.Z., Sankaranarayanan, R., and Rao, N.M. (2008). Thermostable *Bacillus subtilis* lipases: in vitro evolution and structural insight. *J. Mol. Biol.* **381**, 324–340.
- Almond, A., Blundell, C.D., Higman, V.A., MacKerell, A.D., and Day, A.J. (2007). Using molecular dynamics simulations to provide new insights into protein structure on the nanosecond timescale: Comparison with experimental data and biological inferences for the hyaluronan-binding link module of TSG-6. *J. Chem. Theory Comput.* **3**, 1–16.
- Baker, E.N., and Hubbard, R.E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.
- Bernadó, P. (2010). Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering. *Eur. Biophys. J.* **39**, 769–780.
- Böde, C., Kovács, I.A., Szalay, M.S., Palotai, R., Korcsmáros, T., and Cserehely, P. (2007). Network analysis of protein dynamics. *FEBS Lett.* **581**, 2776–2782.
- Cozzini, P., Kellogg, G.E., Spyraakis, F., Abraham, D.J., Costantino, G., Emerson, A., Fanelli, F., Gohlke, H., Kuhn, L.A., Morris, G.M., et al. (2008). Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **51**, 6237–6255.
- Crivelli, S., Eskow, E., Bader, B., Lamberti, V., Byrd, R., Schnabel, R., and Head-Gordon, T. (2002). A physical approach to protein structure prediction. *Biophys. J.* **82**, 36–49.
- Dahiyat, B.I., Gordon, D.B., and Mayo, S.L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337.
- Dehouck, Y., Folch, B., and Rooman, M. (2008). Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity. *Protein Eng. Des. Sel.* **21**, 275–278.
- Dodson, G., and Verma, C.S. (2006). Protein flexibility: its role in structure and mechanism revealed by molecular simulations. *Cell. Mol. Life Sci.* **63**, 207–219.
- Dokholyan, N.V., Li, L., Ding, F., and Shakhnovich, E.I. (2002). Topological determinants of protein folding. *Proc. Natl. Acad. Sci. USA* **99**, 8637–8641.
- Forli, S., and Olson, A.J. (2012). A force field with discrete displaceable waters and desolvation entropy for hydrated ligand docking. *J. Med. Chem.* **55**, 623–638.
- Fulle, S., and Gohlke, H. (2009a). Statics of the ribosomal exit tunnel: implications for cotranslational peptide folding, elongation regulation, and antibiotics binding. *J. Mol. Biol.* **387**, 502–517.
- Fulle, S., and Gohlke, H. (2009b). Constraint counting on RNA structures: linking flexibility and function. *Methods* **49**, 181–188.
- Gohlke, H., Kuhn, L.A., and Case, D.A. (2004). Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* **56**, 322–337.
- Gonzalez, L.C., Livesay, D.R., and Jacobs, D.J. (2011). Improving protein flexibility predictions by combining statistical sampling with a mean-field virtual pebble game. *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 294–298.
- González, L.C., Wang, H., Livesay, D.R., and Jacobs, D.J. (2012). Calculating ensemble averaged descriptions of protein rigidity without sampling. *PLoS ONE* **7**, e29176.
- Greene, L.H., and Higman, V.A. (2003). Uncovering network systems within protein structures. *J. Mol. Biol.* **334**, 781–791.
- Halle, B. (2002). Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA* **99**, 1274–1279.
- Hammel, M. (2012). Validation of macromolecular flexibility in solution by small-angle X-ray scattering (SAXS). *Eur. Biophys. J.* **41**, 789–799.
- Henzler-Wildman, K., and Kern, D. (2007). Dynamic personalities of proteins. *Nature* **450**, 964–972.
- Heringa, J., and Argos, P. (1991). Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol.* **220**, 151–171.
- Hespenheide, B.M., Rader, A.J., Thorpe, M.F., and Kuhn, L.A. (2002). Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graph. Model.* **21**, 195–207.
- Huey, R., Morris, G.M., Olson, A.J., and Goodsell, D.S. (2007). A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **28**, 1145–1152.
- Jacobs, D.J., and Thorpe, M.F. (1995). Generic rigidity percolation: The pebble game. *Phys. Rev. Lett.* **75**, 4051–4054.
- Jacobs, D.J., and Dallakyan, S. (2005). Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys. J.* **88**, 903–915.
- Jacobs, D.J., Rader, A.J., Kuhn, L.A., and Thorpe, M.F. (2001). Protein flexibility predictions using graph theory. *Proteins* **44**, 150–165.
- Jacobs, D.J., Dallakyan, S., Wood, G.G., and Heckathorne, A. (2003). Network rigidity at finite temperature: Relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **68**.
- Joosten, R.P., te Beek, T.A., Krieger, E., Hekkelman, M.L., Hooft, R.W., Schneider, R., Sander, C., and Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39**(Database issue), D411–D419.
- Katoh, N., and Tanigawa, S. (2011). A proof of the molecular conjecture. *Discrete Comput. Geom.* **45**, 647–700.
- Kieseritzky, G., Morra, G., and Knapp, E.W. (2006). Stability and fluctuations of amide hydrogen bonds in a bacterial cytochrome c: a molecular dynamics study. *J. Biol. Inorg. Chem.* **11**, 26–40.
- Kleckner, I.R., and Foster, M.P. (2011). An introduction to NMR-based approaches for measuring protein dynamics. *Biochim. Biophys. Acta* **1814**, 942–968.
- Koller, A.N., Schwalbe, H., and Gohlke, H. (2008). Starting structure dependence of NMR order parameters derived from MD simulations: implications for judging force-field quality. *Biophys. J.* **95**, L04–L06.
- Krüger, D.M., Rath, P.C., Pfleger, C., and Gohlke, H. (2013). CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo)stability, and function. *Nucleic Acids Res.* **41**, W340–W348.
- Leach, A.R. (2001). *Molecular Modelling: Principles and Applications*, Second Edition (Harlow: Prentice Hall).
- Levy, R.M., and Karplus, M. (1979). Vibrational approach to the dynamics of an alpha-helix. *Biopolymers* **18**, 2465–2495.

- Livesay, D.R., and Jacobs, D.J. (2006). Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins* 62, 130–143.
- Livesay, D.R., Dallakyan, S., Wood, G.G., and Jacobs, D.J. (2004). A flexible approach for understanding protein stability. *FEBS Lett.* 576, 468–476.
- Mamonova, T., Hespenheide, B., Straub, R., Thorpe, M.F., and Kurnikova, M. (2005). Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.* 2, S137–S147.
- Pfleger, C., Rath, P.C., Klein, D.L., Radestock, S., and Gohlke, H. (2013a). Constraint Network Analysis (CNA): a Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model.* 53, 1007–1015.
- Pfleger, C., Radestock, S., Schmidt, E., and Gohlke, H. (2013b). Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.* 34, 220–233.
- Pratt, L.R., and Chandler, D. (1977). Theory of hydrophobic effect. *J. Chem. Phys.* 67, 3683–3704.
- Rader, A.J. (2009). Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys. Biol.* 7, 16002.
- Rader, A.J., and Bahar, I. (2004). Folding core predictions from network models of proteins. *Polymer (Guildf.)* 45, 659–668.
- Rader, A.J., Hespenheide, B.M., Kuhn, L.A., and Thorpe, M.F. (2002). Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. USA* 99, 3540–3545.
- Radestock, S., and Gohlke, H. (2008). Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.* 8, 507–522.
- Radestock, S., and Gohlke, H. (2011). Protein rigidity and thermophilic adaptation. *Proteins* 79, 1089–1108.
- Radford, S.E., Buck, M., Topping, K.D., Dobson, C.M., and Evans, P.A. (1992). Hydrogen exchange in native and denatured states of hen egg-white lysozyme. *Proteins* 14, 237–248.
- Rath, P.C., Radestock, S., and Gohlke, H. (2012). Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* 159, 135–144.
- Rose, G.D., and Wolfenden, R. (1993). Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu. Rev. Biophys. Biomol. Struct.* 22, 381–415.
- Stern, R., and Brunner, E. (2008). The relationship between catalytic activity, structural flexibility and conformational stability as deduced from the analysis of mesophilic-thermophilic enzyme pairs and protein engineering studies. In *Thermophiles: Biology and Technology at High Temperatures*, F. Robb, G. Antranikian, D. Grogan, and A. Driessen, eds. (London, New York: CRC Press), pp. 25–38.
- Stickle, D.F., Presta, L.G., Dill, K.A., and Rose, G.D. (1992). Hydrogen bonding in globular proteins. *J. Mol. Biol.* 226, 1143–1159.
- Taverna, D.M., and Goldstein, R.A. (2002). Why are proteins marginally stable? *Proteins* 46, 105–109.
- Teague, S.J. (2003). Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* 2, 527–541.
- Thorpe, M.F. (1983). Continuous deformations in random networks. *J. Non-Cryst. Solids* 57, 355–370.
- Vendruscolo, M., Dokholyan, N.V., Paci, E., and Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E: Stat., Nonlinear. Soft Matter Physiol.* 65, 1–4.
- Whiteley, W. (2005). Counting out to the flexibility of molecules. *Phys. Biol.* 2, S116–S126.
- Young, M.A., Gonfloni, S., Superti-Furga, G., Roux, B., and Kuriyan, J. (2001). Dynamic coupling between the SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation. *Cell* 105, 115–126.
- Zaccai, G. (2000). How soft is a protein? A protein dynamics force constant measured by neutron scattering. *Science* 288, 1604–1607.

13.5 Publication II – Supporting Information

Efficient and Robust Analysis of Biomacromolecular Flexibility Using Ensembles of Network Topologies Based on Fuzzy Noncovalent Constraints

Pfleger, C., Gohlke, H.

Structure (2013), 21, 1-10

Structure, Volume 21

Supplemental Information

Efficient and Robust Analysis of Biomacromolecular Flexibility Using Ensembles of Network Topologies Based on Fuzzy Noncovalent Constraints

Christopher Pflieger and Holger Gohlke

Inventory of Supplemental Information

Figure S1: Local stability characteristics of energy minimized HEWL structures. Referred to section '*Rigidity analyses are highly sensitive with respect to the input structure*' and related to Figure 2.

Figure S2: Persistence of non-covalent interactions in conformational ensembles. Referred to section '*Parameterizing fuzzy noncovalent constraints using data on breaking and (re-)forming of noncovalent interactions from MD simulations*'.

Figure S3: Non-covalent constraints in ENT^{FNC} , SNT, and ENT^{MD} for the 10 HEWL structures. Referred to section '*Parameterizing fuzzy noncovalent constraints using data on breaking and (re-)forming of noncovalent interactions from MD simulations*'.

Figure S4: Rigidity indices r_i depicted separately for the 10 HEWL structure. Related to Figure 2.

Table S1: Clustering of HEWL crystal structures based on the pairwise all-atom RMSD. Referred to the sections '*Rigidity analyses are highly sensitive with respect to the input structure*' and '*Results of rigidity analyses averaged over ENT^{MD} are starting structure-independent*'.

Table S2: Results from the rigid cluster decomposition and number of noncovalent constraints for the networks of the ten representative HEWL structures taken from the PDB. Related to Figure 2A.

Table S3: Correlation of the ordering of hydrogen bonds by energy. Referred to the section '*Parameterizing fuzzy noncovalent constraints using data on breaking and (re-)forming of noncovalent interactions from MD simulations*'.

Table S4: Optimal growth temperatures T_{og} and computed phase transition temperatures T_p . Related to Figure 3.

Description of the HEWL and CS dataset, referred to the sections '*Rigidity analyses are highly sensitive with respect to the input structure*' and '*Validation of the ENT^{FNC} approach on external data sets*'.

Description of the MD simulation protocol used to generate ensembles of HEWL and CS structures, referred to the sections '*Quantifying global and local stability characteristics in the case of the SNT and ENT^{MD} approaches*', '*Parameterizing fuzzy noncovalent constraints using data on breaking and (re-)forming of noncovalent interactions from MD simulations*' and '*Validation of the ENT^{FNC} approach on external data sets*'.

Description of how global and local stability characteristics are quantified in the case of SNT and ENT^{MD} approaches, referred to the sections '*Rigidity analyses are highly sensitive with respect to the input structure*' and '*Results of rigidity analyses averaged over ENT^{MD} are starting structure-independent*'.

Description of the function used to compute the energy of polar interactions, referred to the sections '*Introduction*' and '*Theory*'.

Supplemental Data

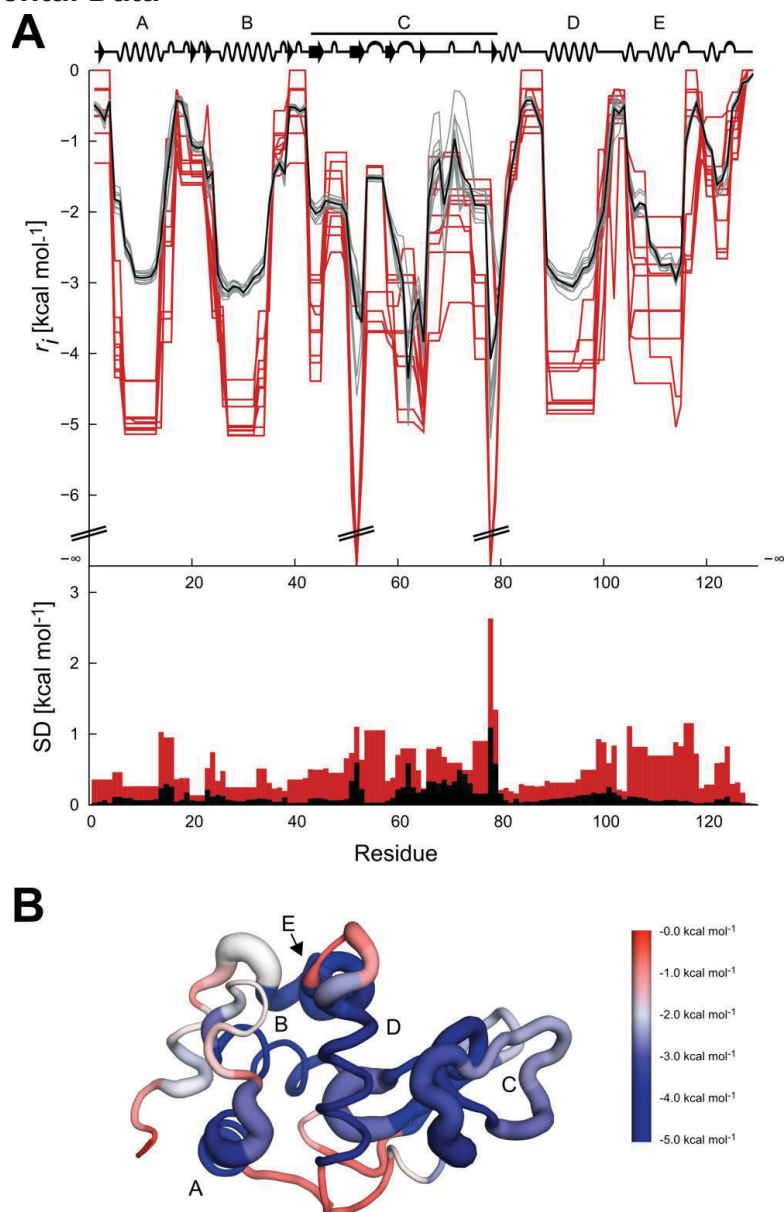


Figure S1: Local stability characteristics of energy minimized HEWL structures.

(A) Rigidity index r_i curves for the SNT analyses of the ten energy minimized HEWL crystal structures (red), r_i curves for the ten ENT^{MD} analyses of HEWL (gray), and the average over all ENT^{MD} analyses (black). The histogram below shows the standard deviation of the r_i 's across the crystal structures and the MD ensembles, respectively. The results from the ENT^{MD} analyses are depicted again from Figure 2 for comparison. (B) The mean r_i values and standard deviations are mapped on a HEWL structure.

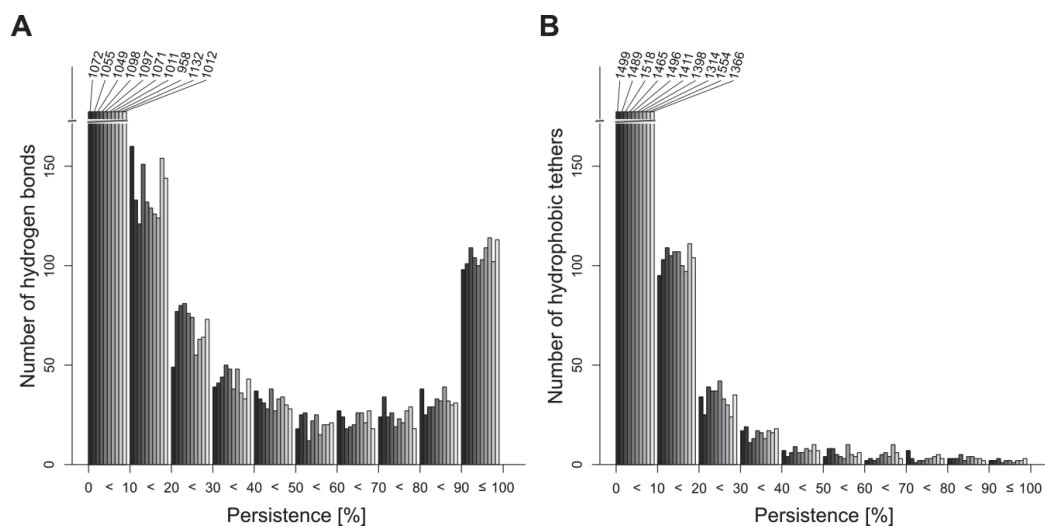


Figure S2: Persistence of non-covalent interactions in conformational ensembles. Conformations are extracted from MD trajectories of ten different HEWL structures. The persistence is shown for (A) hydrogen bonds (including salt bridges) and (B) hydrophobic tethers.

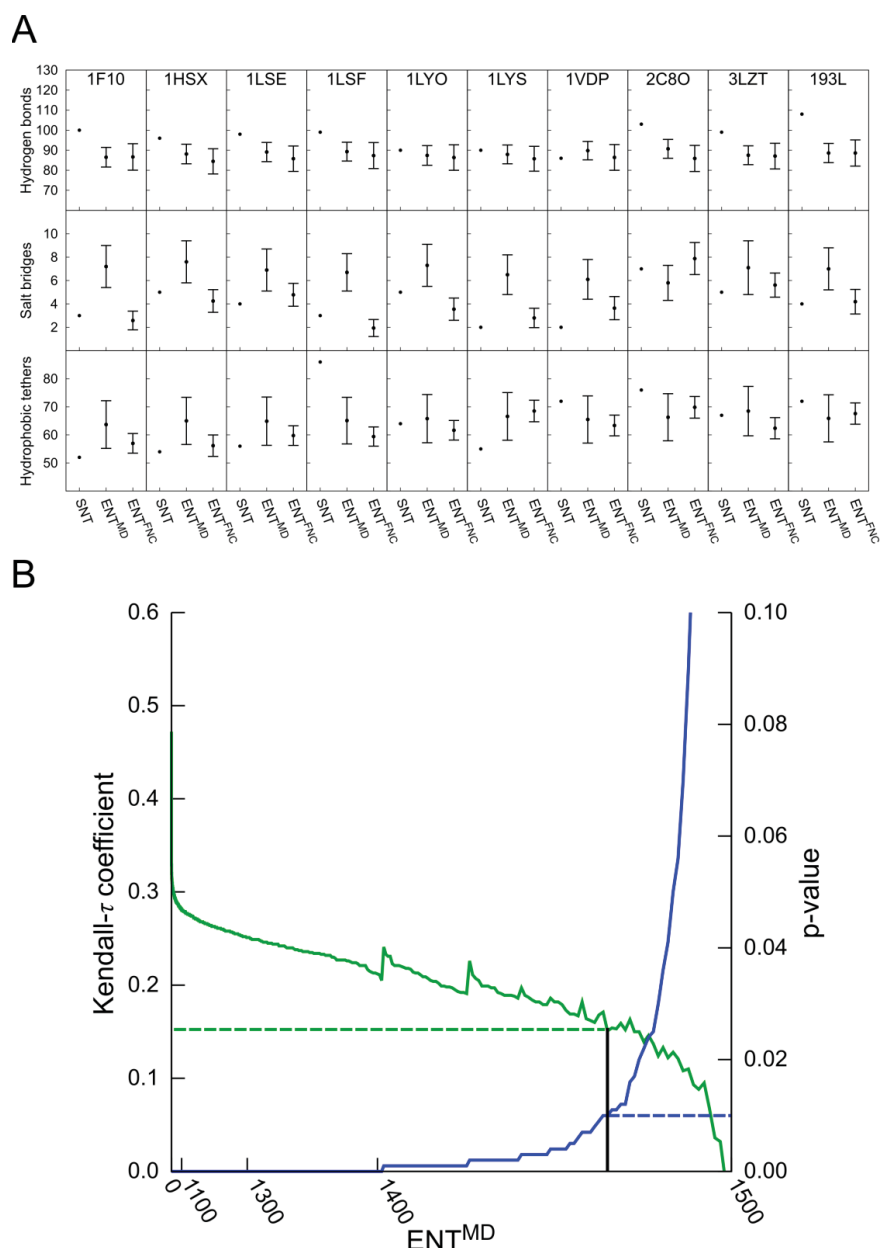


Figure S3: Noncovalent constraints in ENT^{FNC} , SNT, and ENT^{MD} for the 10 HEWL structures. (A) The number of hydrogen bonds, salt bridges, and hydrophobic tethers as sampled by the ENT^{FNC} approach, in the SNT, and ENT^{MD} for the respective HEWL structure. In the case of ENT^{FNC} and ENT^{MD} the average \pm the standard deviation is given. (B) Kendall- τ coefficients (green) and associated p -values (blue) showing the respective best correlation between orders of hydrogen bonds in a network from ENT^{MD} and one from ENT^{FNC} for the HEWL structure 1hsx. A network from either ENT approach can only be used once for a pairwise comparison. The dashed blue line marks $p \leq 0.01$; the dashed green line marks $\tau > 0.15$. The p -values are sorted according to the magnitude; the abscissa has an exponential scale.

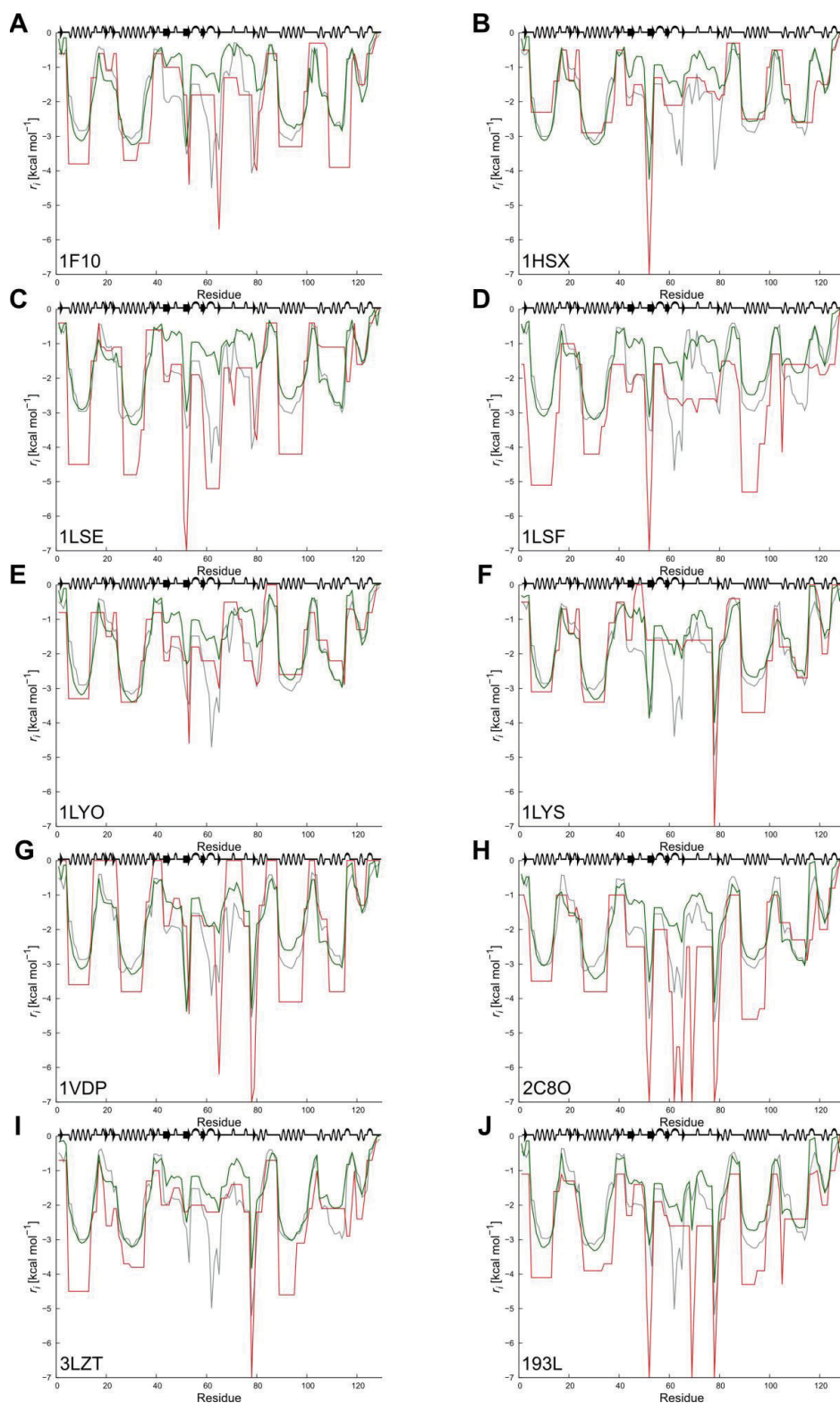


Figure S4: Rigidity indices r_i (see Figure 2) depicted separately for the 10 HEWL structure. (A-J) r_i curves of the SNT (red), ENT^{MD} (gray), and ENT^{FNC} (green).

Table S1: Clustering of HEWL crystal structures based on the pairwise all-atom RMSD.¹

Cluster id	PDB id
1	1bwj, 1lza, 193l , 194l, 1bwi, 1bvz, 1bwh
2	1lsd, 1lse
3	1lsa, 1lsb, 1lsf
4	1lyo , 4lyo, 1lsc, 2cds, 1vds, 1vdt
5	2c8o , 2c8p, 1iee, 1w6z
6	1f10 , 1hsw
7	1f0w, 1hsx , 1jpo, 1vdq, 1ved, 1wtm, 1bgi, 1wtn
8	1lma, 1vdp , 5lym
9	1lys
10	2lzt, 3lzt

¹ Cluster representatives are marked in bold.**Table S2:** RCD results and number of noncovalent constraints for the networks of the ten representative HEWL structures taken from the PDB.

PDB id	Rigid clusters ¹	Percentage of atoms in rigid clusters ¹	Hydrogen bonds/salt bridges	Hydrophobic tethers
1lyo	6	49.6	95	64
1vdp	6	49.5	88	72
1f10	6	45.6	103	52
1hsx	5	48.3	101	54
1lse	5	49.0	102	56
1lys	6	42.9	92	55
2c8o	2	67.5	110	76
1lsf	1	71.6	102	86
3lzt	2	56.8	104	67
193l	1	71.0	112	72

¹ Rigid clusters are only counted if their size ≥ 20 atoms.

Table S3: Correlation of the ordering of hydrogen bonds by energy.

PDB id	SNT vs. ENT ^{FNC}		SNT vs. ENT ^{MD}		ENT ^{MD} vs. ENT ^{FNC}	
	τ^1	FWHM ²	τ^1	FWHM ²	τ^1	FWHM ²
193l	0.034	0.170	0.016	0.166	0.020	0.192
1lse	0.025	0.181	0.032	0.163	0.058	0.199
1lsf	0.048	0.170	0.018	0.152	0.023	0.187
1lyo	0.014	0.193	0.018	0.152	0.037	0.193
2c8o	0.039	0.173	0.026	0.157	0.038	0.190
1f10	0.012	0.162	0.016	0.159	0.040	0.195
1hsx	0.010	0.177	0.004	0.158	0.042	0.194
1vdp	0.034	0.178	0.019	0.163	0.021	0.189
1lys	0.034	0.169	0.017	0.168	0.019	0.197
3lzt	0.042	0.168	0.023	0.157	0.036	0.190

¹ Kendall's τ coefficients.² Full width at half maximum for the distributions of τ coefficients.**Table S4:** Optimal growth temperatures T_{og} and computed phase transition temperatures T_p .

PDB id's	T_{og}^1	$T_{p(\text{cryst})}^2$	$T_{p(\text{min})}^2$	$T_{p(\text{MD})}^3$	$T_{p(\text{FNC})}^3$
3enj	310.2	322	326	330.3±0.3	336.3±0.4
TsCS ⁴	332.2	365	354	331.3±0.2	342.4±0.2
1iom	348.2	344	364	343.8±0.3	344.0±0.2
1o7x	360.2	340	357	338.2±0.3	346.3±0.3
2ibp	373.2	356	366	343.2±0.3	352.9±0.2

¹ Optimal growth temperature T_{og} in K (Darland et al., 1970; Oshima and Imahori, 1974; Zillig et al., 1980; Volkl et al., 1993; Bell et al., 2002; Boutz et al., 2007; Larson et al., 2009).² T_p values of SNT and SNT^{min} in K.³ T_p values and standard error of ENT^{MD} and ENT^{FNC} in K.⁴ Homology model of the open CS using PDB id 2r26 as template.

Supplemental Experimental Procedures

Structure preparation

Dataset of HEWL structures: In order to analyze the persistence characteristics of noncovalent interactions and for evaluating the ENT^{FNC} approach, a set of 10 HEWL structures was used. They were filtered out of an initial dataset of 38 HEWL structures determined by X-ray crystallography to a resolution below 2.5 Å with the aim to reduce structural redundancy. All structures were taken from the Protein Data Base (Berman et al., 2000): 193l, 194l, 1bgi, 1bvx, 1bwh, 1bwi, 1bwj, 1f0w, 1f10, 1hsw, 1hsx, 1iee, 1jpo, 1lma, 1lsa, 1lsb, 1lsc, 1lsd, 1lse, 1lsf, 1lyo, 1lys, 1lza, 1vdp, 1vdq, 1vds, 1vdt, 1ved, 1w6z, 1wtm, 1wtn, 2c8o, 2c8p, 2cds, 2lzt, 3lzt, 4lyo, and 5lym. To filter the structures, a clustering was carried out on the entire dataset based on the pairwise all-atom rmsd according to Ward's method as implemented in the hclust module of R (R Development Core Team, 2008). This resulted in 10 clusters (see Table S1). Out of each cluster the structure with the smallest pairwise rmsd to all other cluster members was selected. If a cluster contained only two structures, the structure with the better resolution was taken. The resulting cluster representatives were then used for MD simulations and rigidity analyses. For this, they were checked for good structural quality using the PDBREPORT database (Hooft et al., 1996), hydrogens were added by the REDUCE program (Word et al., 1999), and, where necessary, Asn, Gln, or His side-chains were flipped.

Dataset of CS structures: In order to apply the ENT^{FNC} approach to an external data set not used in the parameterization, five orthologous CS structures in the *apo* ("open") form were used. These structures had already been used in a previous study investigating thermostabilizing mutations (Rathi et al., 2012). Each protein originates from a different organism with living temperatures T_{og} in the range from 37°C to 100°C (*Sus scrofa*, 37°C, PDB id 3enj; *Thermoplasma acidophilum*, 59°C, TsCS2r26; *Thermus thermophilus*, 75°C, PDB id 1iom; *Sulfolobus solfataricus*, 87°C, PDB id 1o7x; *Pyrobaculum aerophilum*, 100°C, PDB id 2ibp). In the case of TsCS, no open CS structure is available. Hence, a homology model of the open form had been created using the closed (PDB id 2r26) and the open conformations (PDB id 1o7x) as templates (Rathi et al., 2012).

For the SNT approach, structures were either used "as is" or minimized by at most 5000 steps of conjugate gradient minimization or until the root-mean-square gradient of the energy was $< 1.0 \cdot 10^{-4}$ kcal mol⁻¹ Å⁻¹. The energy minimization was carried out with Amber11 using the Cornell *et al.* force field (Wang et al., 2000) with modifications for proteins (ff99SB) (Hornak et al., 2006) and the GB^{OBC} generalized Born model (Onufriev et al., 2004).

Molecular dynamics simulations

MD simulations of HEWL were carried out with the AMBER 11 package of molecular simulation programs using the GPU accelerated version of PMEMD (Case et al., 2010). The Cornell *et al.* force field (Wang et al., 2000) with modifications for proteins (ff99SB) (Hornak et al., 2006) was employed. The structures were solvated in a truncated octahedron of TIP3P water (Jorgensen et al., 1983) such that the distance

between the boundary of the box and the closest solute atom was at least 11 Å. Periodic boundary conditions were applied using the particle mesh Ewald (PME) method (Darden et al., 1993) to treat long-range electrostatic interactions. Bond lengths involving bonds to hydrogen atoms were constrained by SHAKE (Ryckaert et al., 1977; Miyamoto and Kollman, 1992). The time step for all MD simulations was 2 fs, and a direct-space non-bonded cutoff of 8 Å was applied. First, the solvent was minimized for 250 steps by using the steepest descent method followed by conjugate gradient minimization of 50 steps. Subsequently, the same approach was used to minimize the entire system including the protein. Afterwards, the system was heated from 100 K to 300 K using canonical ensemble (NVT) MD, and the solvent density was adjusted using isothermal-isobaric ensemble (NPT) MD. Positional restraints applied during equilibration were reduced in a stepwise manner over 50 ps followed by 50 ps of unrestrained canonical ensemble (NVT) MD at 300 K with a time constant of 2 ps for heat bath coupling. Each simulation ran for 300 ns, and coordinates were saved at 200 ps intervals to obtain 10 individual ensembles of 1500 conformations each.

The simulation protocol for CS is described elsewhere (Rathi et al., 2012). For the present study, we extended the trajectories from 10 ns to 30 ns. Coordinates were saved at 20 ps intervals to obtain five individual ensembles of 1500 conformations each.

Quantifying global and local stability characteristics in the case of the SNT and ENT^{MD} approaches

As input for CNA solely the biomacromolecule was used; water molecules and buffer ions were removed. The network of covalent and noncovalent (hydrogen bonds including salt bridges and hydrophobic tethers) constraints was constructed with the FIRST software (version 6.2) (Jacobs et al., 2001). The strength of each hydrogen bond (including salt bridges) was assigned by the energy E_{HB} computed by FIRST (Dahiyat et al., 1997). Hydrophobic interactions between carbon or sulfur atoms were taken into account if the distance between these atoms was less than the sum of their van der Waals radii (C: 1.7 Å, S: 1.8 Å) plus $D_{cut} = 0.25$ Å (Rader et al., 2002). Global and local stability characteristics were computed by the CNA software (Pfleger et al., 2013a) along a thermal unfolding trajectory by means of the cluster configuration entropy H_{type2} and the rigidity index r_i as described in ref. (Pfleger et al., 2013b). For the ENT^{MD} approach, these results were averaged over the ensemble of conformations generated by MD simulations.

Energy of polar interactions

The energy of polar interactions is computed by two functions that have been adapted from ref. (Dahiyat et al., 1997) (see also ref. (Jacobs et al., 2001) and FIRST 6.2.1 user guide available on <http://flexweb.asu.edu/>), one for uncharged hydrogen bonds (eq. S1) and one for salt bridges (eq. S2). In the case of uncharged hydrogen bonds the energy is computed by a distance-dependent and an angle-dependent term. In the case of salt bridges only a distance-dependent term is used. R_0 is the equilibrium distance, and V_0 is the well-depth of the interaction. The angle term varies depending on the hybridization state of the donor and acceptor atoms; θ is the angle between donor-hydrogen-acceptor; ϕ is the angle between hydrogen-acceptor-base

atom bonded to the acceptor; ϕ is the torsion angle between the normals of two planes defined by the sp^2 centers. If ϕ is less than 90° , the supplement of the angle will be used.

$$E_{HB,uncharged} = V_0 \left\{ 5 \left(\frac{R_0}{R} \right)^{12} - 6 \left(\frac{R_0}{R} \right)^{10} \right\} F(\theta, \phi, \varphi) \quad (S1)$$

$$V_0 = 8 \text{ kcal mol}^{-1}$$

$$R_0 = 2.8 \text{ \AA}$$

$$sp^3 \text{ donor} - sp^3 \text{ acceptor: } F = \cos^2 \theta e^{-(\pi-\theta)^6} \cos^2(\phi - 109.5)$$

$$sp^3 \text{ donor} - sp^2 \text{ acceptor: } F = \cos^2 \theta e^{-(\pi-\theta)^6} \cos^2 \phi$$

$$sp^2 \text{ donor} - sp^3 \text{ acceptor: } F = \cos^4 \theta e^{-2(\pi-\theta)^6}$$

$$sp^2 \text{ donor} - sp^2 \text{ acceptor: } F = \cos^2 \theta e^{-(\pi-\theta)^6} \cos^2(\max[\phi, \varphi])$$

$$E_{HB,charged} = V_0 \left\{ 5 \left(\frac{R_0}{R+x} \right)^{12} - 6 \left(\frac{R_0}{R+x} \right)^{10} \right\} \quad (S2)$$

$$V_0 = 10 \text{ kcal mol}^{-1}$$

$$R_0 = 3.2 \text{ \AA}$$

$$x = 0.375 \text{ \AA}$$

Supplemental References

Darland, G., Brock, T.D., Samsonof, W., and Conti, S.F. (1970). Thermophilic, Acidophilic Mycoplasma Isolated from a Coal Refuse Pile. *Science* 170, 1416-1418.

Oshima, T., and Imahori, K. (1974). Description of *Thermus-Thermophilus* (Yoshida and Oshima) Comb-Nov, a Nonsporulating Thermophilic Bacterium from a Japanese Thermal Spa. *Int. J. Syst. Bacteriol.* 24, 102-112.

Zillig, W., Stetter, K.O., Wunderl, S., Schulz, W., Priess, H., and Scholz, I. (1980). The *Sulfolobus-Caldariella* Group - Taxonomy on the Basis of the Structure of DNA-Dependent Rna-Polymerases. *Arch. Microbiol.* 125, 259-269.

Volkl, P., Huber, R., Drobner, E., Rachel, R., Burggraf, S., Trincone, A., and Stetter, K.O. (1993). *Pyrobaculum-Aerophilum* Sp-Nov, a Novel Nitrate-Reducing Hyperthermophilic Archaeum. *Appl. Environ. Microbiol.* 59, 2918-2926.

Bell, G.S., Russell, R.J.M., Connaris, H., Hough, D.W., Danson, M.J., and Taylor, G.L. (2002). Stepwise adaptations of citrate synthase to survival at life's extremes - From psychrophile to hyperthermophile. *Eur. J. Biochem.* 269, 6250-6260.

Boutz, D.R., Cascio, D., Whitelegge, J., Perry, L.J., and Yeates, T.O. (2007). Discovery of a thermophilic protein complex stabilized by topologically interlinked chains. *J. Mol. Biol.* 368, 1332-1344.

Larson, S.B., Day, J.S., Nguyen, C., Cudney, R., and McPherson, A. (2009). Structure of pig heart citrate synthase at 1.78 angstrom resolution. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* 65, 430-434.

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

R Development Core Team (2008). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

Hooft, R.W.W., Vriend, G., Sander, C., and Abola, E.E. (1996). Errors in protein structures. *Nature* 381, 272.

Word, J.M., Lovell, S.C., Richardson, J.S., and Richardson, D.C. (1999). Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* 285, 1735-1747.

Rathi, P.C., Radestock, S., and Gohlke, H. (2012). Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* 159, 135-144.

Wang, J.M., Cieplak, P., and Kollman, P.A. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* *21*, 1049-1074.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* *65*, 712-725.

Onufriev, A., Bashford, D., and Case, D.A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* *55*, 383-394.

Case, D.A., Darden, T.A., Cheatham, I., T.E., Simmerling, C.L., Wang, J., Duke, R.E., Luo, R., Walker, R.C., Zhang, W., Merz, K.M., *et al.* (2010). AMBER 11 (University of California, San Francisco.).

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

Darden, T., York, D., and Pedersen, L. (1993). Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* *98*, 10089-10092.

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

Miyamoto, S., and Kollman, P.A. (1992). Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. *J. Comput. Chem.* *13*, 952-962.

Jacobs, D.J., Rader, A.J., Kuhn, L.A., and Thorpe, M.F. (2001). Protein flexibility predictions using graph theory. *Proteins* *44*, 150-165.

Dahiyat, B.I., Gordon, D.B., and Mayo, S.L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.* *6*, 1333-1337.

Rader, A.J., Hespenheide, B.M., Kuhn, L.A., and Thorpe, M.F. (2002). Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. U. S. A.* *99*, 3540-3545.

Pfleger, C., Rathi, P.C., Klein, D.L., Radestock, S., and Gohlke, H. (2013a). Constraint Network Analysis (CNA): A Python software package for efficiently linking biomolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model.* *53*, 1007-1015.

Pfleger, C., Radestock, S., Schmidt, E., and Gohlke, H. (2013b). Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.* *34*, 220-233.

13.6 Publication III

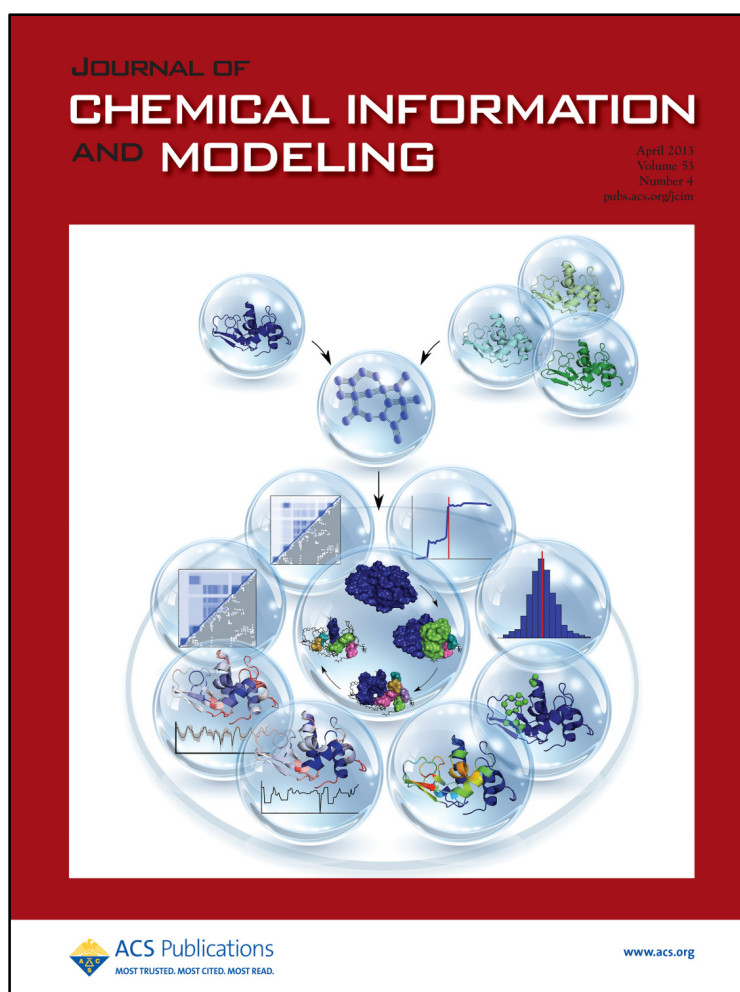
Constraint Network Analysis (CNA): A Python Software Package for Efficiently Linking Biomacromolecular Structure, Flexibility, (Thermo)Stability, and Function

§Pfleger, C., §Rathi, P.C., Klein, D., Radestock, S., Gohlke, H.

J. Chem. Inf. Model. (2013), 53, 1007-1015

Author contribution to the publications:

My contribution to this publication was developing the *pyFIRST* interface module, implementing global and local indices as described in publication II, and providing a robust procedure for the treatment of ligand molecules in rigidity analysis. This results in a contribution of **35%** to this publication.

Cover article:

§ Both authors contributed equally to this work

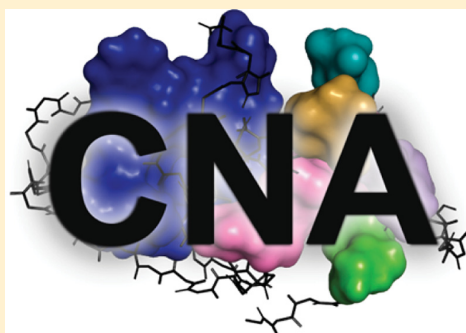
Constraint Network Analysis (CNA): A Python Software Package for Efficiently Linking Biomacromolecular Structure, Flexibility, (Thermo-)Stability, and Function

Christopher Pfleger,[‡] Prakash Chandra Rath,† Doris L. Klein, Sebastian Radestock,† and Holger Gohlke*

Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine-University, Universitätsstr. 1, 40225, Düsseldorf, Germany

ABSTRACT: For deriving maximal advantage from information on biomacromolecular flexibility and rigidity, results from rigidity analyses must be linked to biologically relevant characteristics of a structure. Here, we describe the Python-based software package Constraint Network Analysis (CNA) developed for this task. CNA functions as a front- and backend to the graph-based rigidity analysis software FIRST. CNA goes beyond the mere identification of flexible and rigid regions in a biomacromolecule in that it (I) provides a refined modeling of thermal unfolding simulations that also considers the temperature-dependence of hydrophobic tethers, (II) allows performing rigidity analyses on ensembles of network topologies, either generated from structural ensembles or by using the concept of fuzzy noncovalent constraints, and (III) computes a set of global and local indices for quantifying biomacromolecular stability.

This leads to more robust results from rigidity analyses and extends the application domain of rigidity analyses in that phase transition points ("melting points") and unfolding nuclei ("structural weak spots") are determined automatically. Furthermore, CNA robustly handles small-molecule ligands in general. Such advancements are important for applying rigidity analysis to data-driven protein engineering and for estimating the influence of ligand molecules on biomacromolecular stability. CNA maintains the efficiency of FIRST such that the analysis of a single protein structure takes a few seconds for systems of several hundred residues on a single core. These features make CNA an interesting tool for linking biomacromolecular structure, flexibility, (thermo-)stability, and function. CNA is available from <http://cpclab.uni-duesseldorf.de/software> for nonprofit organizations.



INTRODUCTION

The concepts of biomacromolecular flexibility and its opposite, rigidity, are crucial for understanding the relationship between biomacromolecular structure, (thermo-)stability, and function. In the field of statics, flexibility and rigidity denote the possibility (or impossibility) of internal motion but are not associated with information about directions and magnitudes of movements. Identifying and modulating the heterogeneous composition of biomacromolecules in terms of flexible and rigid regions is becoming increasingly important for successful protein engineering and rational drug-design.^{1–5} Several computational approaches have been developed that identify flexible and rigid regions by either determining spatial variations in the local packing density⁶ or representing and analyzing a structure as a connectivity network of interacting atoms or residues.^{7–12} The approaches benefit from being computationally highly efficient. A related concept has been introduced by Jacobs et al.¹³ Here, biomacromolecules were initially represented as bond-bending networks in which each atom has three degrees of freedom representing the dimensions of motion in 3-space. In later versions, the equivalent body-bar representation is used where atoms are modeled as bodies with six degrees of freedom.^{13–15} By adding constraints (representing covalent and noncovalent bonds in a biomacromolecular

context) between the bodies, internal motions become restricted. Each constraint is modeled as a set of bars, and each bar removes one degree of freedom. According to the type of interaction, the number of bars varies in that stronger interactions are modeled with a higher number of bars than weaker ones. Noncovalent interactions such as hydrogen bonds, salt bridges, hydrophobic tethers, and stacking interactions contribute most to the biomacromolecular stability; hence, these interactions are modeled as constraints in addition to covalent bonds. Once the network is constructed, the Pebble Game algorithm, available within the FIRST (Floppy Inclusions and Rigid Substructure Topography) software, efficiently decomposes the network into rigid clusters and flexible hinge regions from the number and spatial distribution of bond-rotational degrees of freedom.^{16,17} A rigid region is a collection of interlocked bonds allowing no relative motion of the bodies. Such a region can either be overconstrained, if it has redundant constraints, or is isostatically rigid. In a flexible region, dihedral rotation is not locked in by other bonds. The theory underlying this approach is rigorous¹⁸ and has been applied in different areas of biomacromolecular research.^{5,19–35}

Received: January 20, 2013

Published: March 21, 2013

We developed the command-line Python-based software package Constraint Network Analysis (CNA) for analyzing structural features of biomacromolecules that are important for the molecule's stability. CNA functions as a front- and backend to the FIRST software and allows (I) setting up a variety of constraint network representations for analysis by FIRST, (II) processing the results obtained from FIRST, and (III) calculating seven indices for quantifying biomacromolecular stability, both globally and locally.³⁶ As to the latter, the indices are calculated by monitoring changes of the network stability along a thermal unfolding simulation. The thermal unfolding is simulated by consecutively removing hydrogen bond (including salt bridge) constraints from the network with increasing temperature. Thermal unfolding simulations have been successfully applied in several studies on proteins, RNAs, and the ribosome in order to understand how flexibility and rigidity is linked to biomacromolecular stability and function.^{4,5,14,19,28,31,34,35,37}

CNA goes beyond the mere identification of flexible and rigid regions in a biomacromolecular structure in that it allows linking results from constraint network analysis to biologically relevant characteristics of a structure. This is key for deriving maximal advantage from information on biomacromolecular flexibility and rigidity. Here, we describe the design and implementation of the CNA software package. We then demonstrate its application scope in a showcase example on Hen Egg White Lysozyme (HEWL) structures. The CNA software package is available under an academic license from <http://cpclab.uni-duesseldorf.de/software>.

METHODS AND IMPLEMENTATION

General Overview. The CNA software package allows three different types of rigidity analysis: (I) based on a *single network topology* generated from a single input structure, (II) based on an *ensemble of network topologies* generated from a conformational ensemble provided as input,^{21,35} and (III) based on an *ensemble of network topologies* generated from a single input structure by considering fuzzy noncovalent constraints (FNC) (C. Pfleger, H. Gohlke, to be published elsewhere). The last variant mimics that noncovalent constraints thermally break and reform even in the native state of a biomacromolecule.³⁸ In short, we developed a system-independent parametrization of fuzzy noncovalent constraints by analyzing the atom type and location-dependent persistence characteristics of noncovalent constraints (hydrogen bonds, salt-bridges, and hydrophobic tethers) during MD simulations. With this, the number and distribution of noncovalent constraints are modulated by random components within certain ranges, simulating thermal fluctuations of a biomacromolecule without actually moving atoms. In the related distance constraint model (DCM), ensembles of network topologies are generated considering mean-field probabilities of hydrogen bond and torsion constraints in a Monte Carlo sampling.^{20,39} Average stability characteristics are then calculated by constraint counting on each topology in the ensemble.⁴⁰ As a downside, the DCM approach requires experimental data for a system-specific parametrization of the model.

The analysis of a *single network topology* by CNA consists of the following steps. Initially, a constraint network is generated from the input structure by placing covalent and noncovalent constraints according to rules described in refs 13–15. Next, a thermal unfolding simulation is carried out by sequentially removing noncovalent constraints from the network (see

section Thermal Unfolding Simulation for details). For each network during the simulation, a rigidity analysis by FIRST is performed and then post-processed to calculate global and local indices to characterize biomacromolecular flexibility and rigidity. The workflow of the software is illustrated in Figure 1. In the case of analyzing an *ensemble of network topologies*, these steps are repeated for each network, and the results are averaged over the ensemble.

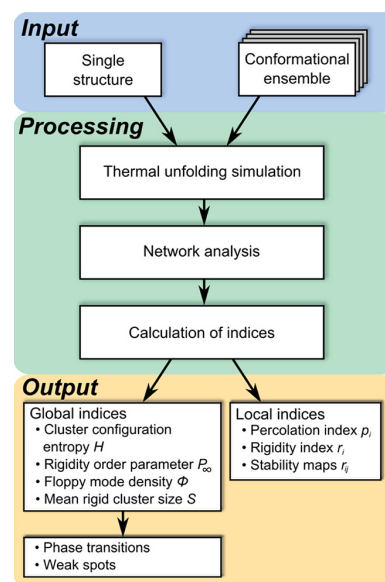


Figure 1. Schematic workflow of the CNA software.

Upon running a thermal unfolding simulation (a) phase transition(s) can be identified at which the network changes from mainly rigid to flexible. For this, the change in the global indices is monitored during the simulation. Four different global indices are implemented in CNA. They monitor (I) the normalized number of independent internal degrees of freedom (floppy mode density, Φ), (II) the fraction of the network belonging to a rigid component (rigidity order parameter, P_∞), (III) the degree of disorder in the network (cluster configuration entropy, H), and the rigid cluster size distribution (mean rigid cluster size, S). In addition, CNA calculates three local indices that characterize the flexibility and rigidity at the bond level: (I) the percolation index p_i monitors the percolation behavior of a biomacromolecule on a microscopic level and thus allows the identification of the hierarchical organization of the giant percolating cluster during a thermal unfolding simulation, (II) the rigidity index r_i monitors when a bond segregates from a rigid cluster, (III) a stability map is a two-dimensional itemization of the rigidity index r_i and is derived by identifying “rigid contacts” between two residues. Exact definitions of these indices and guidelines for when to use them are given in ref 36. Furthermore, the CNA software identifies unfolding nuclei, i.e., those residues that break apart from the giant cluster at the phase transition point.^{4,28,35} The unfolding nuclei can be considered weak spots in the structure; accordingly, this knowledge can be exploited in data-driven protein engineering to focus on residues that are highly likely to improve thermostability upon mutation.

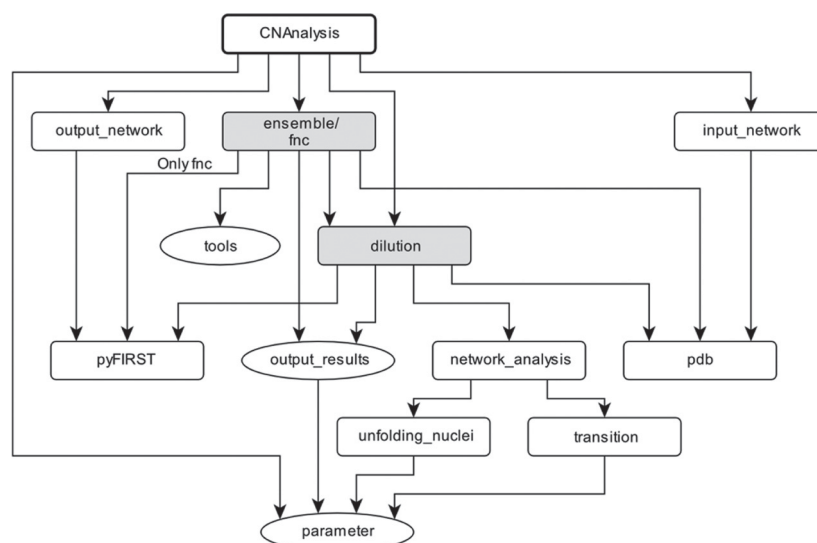


Figure 2. Hierarchical structure of the CNA software. All modules that contain (a) class definition(s) are shown in rectangles. The core module *CNAAnalysis* is highlighted by a bold frame. Modules colored in gray contain the simulation methods for analyzing a single network topology and an ensemble of network topologies. Modules that solely contain methods are shown as ellipses. An arrow indicates the call of a module by another module.

CNA is implemented as a Python-based software package making use of an object-oriented design (Figure 2). Third party software is required for full functionality (Table 1): (I) The

Table 1. External Software Needed by the CNA Software

name	version	description and use
Python	2.73	Python interpreter used by the CNA software package
Biopython	1.58	for reading PDB files using the Bio.PDB package and parsing results from the DSSP program using the Bio.DSSP package
NumPy	1.6.1	for statistical analyses
SciPy	0.11.0	for statistical analyses
Open Babel	2.3.1	for identifying the connectivity and bond orders of ligand molecules
DSSP		for computing secondary structure information that is required by the FNC approach
SWIG	2.0.8	for compiling the <i>pyFIRST</i> interface module

Biopython package⁴¹ is needed to parse input PDB files and provides information on secondary structure from a DSSP analysis.^{42,43} (II) For statistical analysis and detecting the phase transitions, the Numpy⁴⁴ and SciPy⁴⁵ extensions for Python are required. (III) The Open Babel^{46,47} Python-bindings are required to determine the bond order of small-molecule ligands. To facilitate the installation of the CNA software package, the third party software is provided with the CNA source tree except for the DSSP program, which is available at <http://swift.cmbi.ru.nl/gv/dssp/>. The CNA source tree also contains a comprehensive documentation detailing the installation and usage of the software and a suite of test cases to check the validity of the installation. CNA is a command-line based software that is called by the shell script *CNA.sh*. A “--help” argument lists all available options and required arguments, their descriptions, default values, and the range of allowed values. An erroneous argument set for an option produces an informative error message. The CNA software has

been successfully tested on Debian, OpenSuse, and CentOS Linux platforms.

Constraint Network Analysis Is the Core Module. The *CNAAnalysis* module is the core of the CNA software. The *CNAAnalysis* module consists of a single class *ConstraintNetworkAnalysis*. Upon creating an instance of the type *ConstraintNetworkAnalysis*, it (I) parses the command line options that specify the analysis type, (II) checks whether the values of the command line arguments conform to the desired data-type, and (III) performs the requested analysis. Depending on the type of analysis, the *ConstraintNetworkAnalysis* instance creates an instance of the class *Dilution* if the analysis of a *single network topology* is requested. Otherwise, it creates an instance of the class *Fnc* or *Ensemble*, which then creates instances of the class *Dilution* for each network of the ensemble. The command line options provided by the user are checked for validity by the module *Parameter*; this module also contains default values for the options and internal constants.

PyFIRST as an Interface. We developed the *pyFIRST* interface module to directly access the functionality of the FIRST software (available at <http://flexweb.asu.edu>) within the Python environment of CNA. The interface module was implemented using the SWIG (Simplified Wrapper and Interface Generator) software tool (<http://www.swig.org/>).⁴⁸ SWIG automatically generates a wrapper code for C/C++ programs that then acts as an interface for other high level programming languages such as Python. The SWIG interface file is written in C++ and contains a single class *pyFIRST*. The class contains methods that are later on accessible within the Python environment of CNA. Upon instantiating a *pyFIRST* object, a data structure is generated that represents the constraint network topology of the input structure. Additionally, the *pyFIRST* object provides methods that are used to (I) read constraint information (covalent bonds, hydrogen bonds, salt bridges, hydrophobic tethers, and stacking interactions) from the network topology, (II) remove constraints from the network with respect to all or a certain type of constraints, and

(III) perform a rigid cluster decomposition. Finally, methods are available that return warnings issued by FIRST when initializing the data structure for the constraint network topology. Note that the *pyFIRST* interface module has been written such that it can be used in any other Python-based application requiring a rigidity analysis by FIRST, thus providing a general Python interface to FIRST.

Structural Information as Input. Single or multiple (in case of a conformational ensemble) input structures for CNA must be in PDB format.⁴⁹ Although the validity of the input structure(s) is checked upon creating an instance of the class PDB, we recommend subjecting only complete structures without missing residues or atoms. Hydrogen atoms must be present, too, because otherwise the identification of hydrogen bond and salt bridge constraints cannot be performed. Ligand molecules, if present, are extracted from the input structure and, subsequently, analyzed to determine the bond order by means of Open Babel.^{46,47} The last step requires the presence of hydrogen atoms at the ligand. All identified rotatable bonds (single bonds) are then modeled by five bars, whereas nonrotatable bonds (double, triple, amide, and aromatic bonds) are modeled by six bars.¹⁵ Finally, the covalent constraint information for the ligand is merged with the covalent and noncovalent constraint network of the biomacromolecule also generating noncovalent constraints between them. Ions, water, and buffer molecules are handled by FIRST. If an NMR structure is used as input, only the first model is considered. Furthermore, Amber-conform residue names (HIE, HID, HIP, and CYX) are replaced by standard residue names (HIS and CYS) in order to allow the use of PDB structures extracted from molecular dynamics (MD) trajectories created by the Amber software.⁵⁰ In the case of a conformational ensemble, a PDB object is instantiated for each conformation. Apart from checking the validity of and preparing the input structure, the PDB class provides several functions that can be used to work with the structure in terms of getting single atom and residue objects, finding neighbor residues within a certain distance cutoff, and writing out structures (including biomacromolecules and ligand molecules) in the PDB format.

Accessing the Network Topology. The *output_network* and *input_network* modules of CNA contain the OutputNetwork and InputNetwork class definitions. Upon instantiating an object, these classes are used to write and read the constraint network topology of a single structure or of each conformation of an ensemble. This is particularly useful for adding user-defined constraints that are not identified automatically, for example, constraints between ions and protein atoms. In the file containing the constraint network topology, each entry of a covalent constraint contains the identifiers of the involved atoms and number of bars of the constraint. For constraints representing hydrophobic or stacking interactions, in addition to the atom identifiers, the distance between the atoms is given plus an indicator whether the constraint occurs within a protein or between protein and ligand. For hydrogen bond and salt bridge constraints, the energy and type of interaction is written instead of the distance and indicator. This file can be modified and used as input for CNA again. In this case, user-defined constraints will overwrite constraint information identified from the input structure(s).

Thermal Unfolding Simulation. The thermal unfolding simulation allows analyzing changes in the network stability upon removing hydrogen bond (including salt bridge)

constraints from the network.^{4,14,28} To do so, the energy of a hydrogen bond E_{HB} is determined by an empirical energy function.⁵¹ Then, during the thermal unfolding simulation,^{4,28} intermediate networks σ are created such that hydrogen bonds with an energy $E_{HB} > E_{cut}(\sigma)$ are removed from the network.⁵¹ This follows the idea that stronger hydrogen bonds will break at higher temperatures than weaker ones. By means of an empirically determined linear function, E_{cut} can be related to a temperature T .²⁸

Consequently, the simulation mimics a rise in the temperature by analyzing a range of networks having many hydrogen bonds (equivalent to low temperatures) to having few hydrogen bonds (equivalent to high temperatures). Note that the temperatures should be considered relative values only because the absolute values may depend on the size and architecture of the analyzed protein.⁴ Still, the temperatures are very helpful, for example, when it comes to comparing the thermostability of two or more homologous proteins or the stability of a wild-type with its mutant.^{4,28,35} An alternative concept grounded in mean-field theory directly connects network rigidity and absolute temperature; while appealing, it requires experimental data for a system-specific parametrization.^{20,40} Each of the intermediate networks σ is then subjected to rigidity analysis by FIRST. While the principal idea of the thermal unfolding simulation has been adapted from the FIRST software,¹³ the method implemented here allows for additional settings that are not available in the FIRST implementation. These include specifying the energy range and step-size for removing hydrogen bonds. Furthermore, a modified method has been implemented that also considers the temperature dependence of hydrophobic tethers along the thermal unfolding simulation.³⁵ This approach follows the idea that hydrophobic interactions become stronger with increasing T .^{52,53} Accordingly, more hydrophobic tethers are added to the network by linearly increasing the distance cutoff for including hydrophobic tethers $D_{cut}(\sigma)$ from a starting value of 0.25 Å at 300 K to an ending value of 0.40 Å at 420 K. Doing this has been shown to improve thermostability predictions of citrate synthases.³⁵

The thermal unfolding simulation is done by the *dilution* module containing the Dilution class. Upon instantiating an object of this class, the object creates new intermediate networks σ and passes the networks through FIRST by instantiating a *pyFirst* object. Subsequently the module *networkAnalysis* is used to calculate the global and local indices (see section Analyzing the Results from the Rigidity Analysis). Via the global indices, phase transition(s) are identified by an object of the class Transitions. Finally, unfolding nuclei are identified by an object of the class UnfoldingNuclei.

Analyzing the Results from the Rigidity Analysis. The *network_analysis* module comprises in total four classes that process the results from the FIRST rigidity analysis. The main class NetworkAnalysis contains methods to calculate the size and size distribution of rigid clusters and to identify the actual largest rigid cluster as well as the giant percolating cluster of the network. The giant percolating cluster is the largest rigid cluster present at the highest E_{cut} value (i.e., at the lowest temperature) with all constraints in place. During the thermal unfolding simulation, the melting of the giant percolating cluster is monitored, and the largest rigid subcluster of the previous giant percolating cluster becomes the new giant percolating cluster of the present network state σ . Subsequently, the NetworkAnalysis object is passed to three classes for calculating the global and

local indices called GlobalIndices, LocalIndices, and LocalStabilityMaps.

The class GlobalIndices contains all methods that are required to calculate the floppy mode density Φ , the rigidity order parameter P_∞ , the cluster configuration entropy H , and the mean rigid cluster size S .³⁶ Apart from this, the class GlobalIndices also instantiates objects of the classes Transitions and UnfoldingNuclei that are required for the identification of phase transition points and unfolding nuclei of the structure. For identifying phase transition points, two methods have been implemented that make use of the data of the global indices: fitting of a mono/double sigmoid curve and interpolating with a smoothed spline. By default, phase transition points are identified by the double sigmoid curve.³⁵ However, the user can choose as an option that Akaike's information criterion⁵⁴ be used to identify whether a mono or double sigmoid curve gives better fitting results. Finally, if more than two phase transitions are expected or shall be identified, interpolation with the smoothed spline is recommended. Multiple transitions can occur in multimeric proteins. The transition point is then identified for each global index as the point at which the maximal rigidity loss occurs in the structure. Occasionally, a Transitions object does not return a transition point; this occurs if no "sharp" transition can be detected or if multiple transitions with comparable rigidity losses are present.

The class LocalIndices is used to calculate the percolation index p_i and the rigidity index r_i . Both reflect structural stability on a per-residue basis³⁶ and, thus, can be used to identify the location and distribution of structurally weak or strong parts in biomacromolecules. Finally, the class LocalStabilityMaps is used to calculate the two-dimensional itemization of the rigidity index r_i , the stability map, and a so-called "neighbor stability map", where values of the stability map of residue pairs separated by more than 5 Å are masked. That way, the latter map provides useful information about the stability of neighboring residues only, which can be used for focusing on short-range weak and strong connections within a biomacromolecule.

Writing the Analysis Results. The module *output_results* is used to write results files containing information about global and local indices, phase transition points, and unfolding nuclei. For a phase transition point, the hydrogen bond energy cutoff E_{cut} and the respective temperature are listed. Unfolding nuclei are written out as a text file and PDB file; in the latter, the B-factor column is used to record whether or not a residue is an unfolding nucleus by setting the values to one or zero. If the analysis is performed on an *ensemble of network topologies*, an additional file summarizing the average local indices and standard deviations is written. Similarly, for the phase transition points, mean, median, and standard error are provided in addition. Furthermore, the percentage of network topologies in which a residue is predicted to be an unfolding nucleus is recorded.

Showcase Example: Flexibility Characteristics of HEWL. In a showcase example, we applied the CNA software to a HEWL structure. We show the results for two analysis types, analyzing a single network topology derived from a single input structure (PDB ID: 3LZT) and analyzing an ensemble of network topologies derived from a conformational ensemble. The conformational ensemble was generated by extracting 1500 conformations from a trajectory of 300 ns length obtained by MD simulations starting from an X-ray structure of HEWL (PDB ID: 3LZT). The MD simulation was carried out in

explicit solvent at 300 K with the AMBER 11 package of molecular simulation programs.⁵⁰ The detailed simulation protocol is described elsewhere (C. Pfleger, H. Gohlke, to be published elsewhere). Water molecules were removed from each conformation before the ensemble was subjected to CNA. Analyzing a single network topology took about 40 s, and the ensemble of 1500 conformations required ~11 h on a single-core workstation computer, which demonstrates the computational efficiency of CNA and FIRST.

Snapshots from the thermal unfolding simulation of the single input structure are depicted in Figure 3. They show the

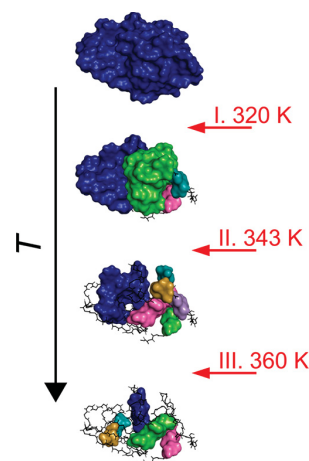


Figure 3. Rigid cluster decompositions along the thermal unfolding simulation of the showcase example HEWL. Rigid clusters are shown as uniformly colored bodies connected by flexible hinge regions (black). The roman numbers relate to three major steps of rigidity loss.

loss of rigidity in terms of the decay of rigid clusters with increasing temperature. The first transition relates to the beginning of the collapse of the giant rigid cluster, which occurs in the interface region of the α - and β -domains. At this state, the network is dominated by two large rigid components. During the next transition, the rigid cluster covering the α -domain collapses, and the helical elements remain as single rigid clusters. Finally, during the last transition, the rigid cluster covering the β -domain collapses, and nearly the whole system becomes flexible. The results from the thermal unfolding simulation agree, in reverse order, with the "fast track" folding pathway described in refs 55 and 56. Here, both domains of HEWL fold concurrently but with a slight preference to initially form native contacts in the β -domain.⁵⁷ Alternatively, a "slow track" folding reaction of HEWL has been described,^{56,58,59} in which the majority of the protein molecules populate an intermediate state with persistent structures in only the α -domain.⁵⁷ Still, parts of the α -domain need to unfold again to enable the subsequent folding of the β -domain.

As an example for a global index, the cluster configuration entropy H is shown, which monitors the loss of network stability during the thermal unfolding simulation. In the analysis of the single network topology (Figure 4a), an early phase transition at 319 K indicates the beginning decay of structural stability, with most of the network still being captured in rigid clusters. The dominant phase transition at 343 K then refers to the point at which the network loses its ability to carry stress

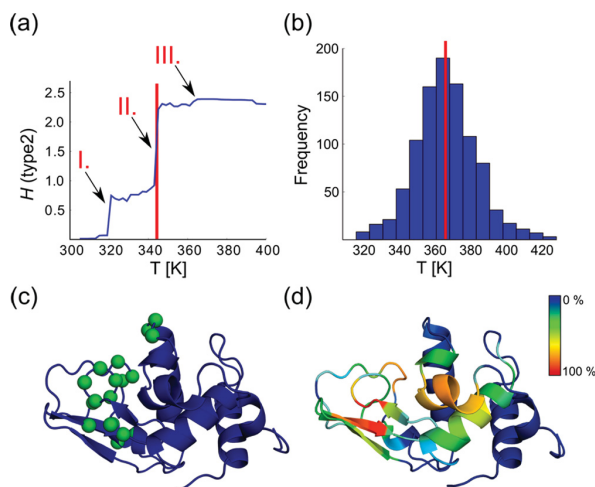


Figure 4. (a) Cluster configuration entropy H (type 2) derived from the single network topology. The entropy is plotted as a function of the temperature, and the roman numbers correspond to the three major steps depicted in Figure 3. The phase transition automatically identified by CNA is marked by the red vertical line. (b) Frequency distribution of phase transitions identified from analyzing the ensemble of network topologies. The median is marked with a red vertical line. (c) Weak spot detection for the single network topology. Green spheres highlight the identified weak spot residues in the HEWL structure. (d) Weak spot detection over the ensemble of network topologies. For depicting the probability of being a weak spot, each residue is colored according to a color scale ranging from blue (low probability) to red (high probability).

and, hence, corresponds to the folded–unfolded transition. The last transition indicates the loss of the remaining rigid components. In the case of analyzing the ensemble of network topologies, the frequency distribution of the identified phase transition points is shown (Figure 4b). From this, a median transition temperature of 358 K is revealed, which is 15 K higher than the dominant phase transition point identified from analyzing a single network topology. Note that, in general, phase transitions identified using a single input structure can be different from ensemble results, as shown in a previous study on citrate synthase.³⁵ We thus recommend performing CNA analyses on ensembles of network topologies, in particular, when quantitative results are desired. At the transition point, unfolding nuclei are identified (Figure 4c). Almost all unfolding nuclei are located in the β -domain of HEWL, which disintegrates at the dominant phase transition (Figures 3 and 4c). Furthermore, for the ensemble of network topologies, the probability of a residue being found as an unfolding nucleus over the entire ensemble is provided (Figure 4d). The higher this probability the more likely will it be that rigidifying this residue will improve protein stability. The ensemble results are more detailed than the ones from the single structure in that now unfolding nuclei are not only located in the β -domain but also in helix B, which agrees with the view that this helix plays a crucial role in stabilizing the tertiary structure of HEWL.⁶⁰

As for local indices, we exemplarily show the rigidity index r_i , which characterizes the stability of the HEWL structure down to the bond level (Figure 5a, b). As such, r_i monitors the point when a residue segregates from a rigid cluster along the thermal unfolding simulation: the lower r_i the longer is a residue part of a rigid cluster. Secondary structure elements are generally

found to be more stable than loop regions. Furthermore, averaging r_i values over the ensemble of network topologies leads to a smoother r_i curve and to the spike located at residue 78 becoming less pronounced than in the case of analyzing the single network topology. The spike reveals a region that is highly stabilized by hydrophobic interactions; these regions only melt at a late stage of the thermal unfolding simulations. Notably, the stable regions identified for residues 53 and 62–65 are in very good agreement with those identified by high protection factors in H/D experiments for the native and denatured states of HEWL.⁶¹ During the catalytic cycle, HEWL undergoes a reorientation of the α - and β -domains due to a bending movement around a central hinge region.⁶² Along these lines, the identified flexible hinge regions (Figure 5a, b) are in agreement with those suggested by McCammon et al.⁶² and coincide with results obtained from Gaussian network models and MD simulations.^{60,63} Such a decomposition into rigid clusters and flexible regions is used as a first step in a normal mode-based geometric simulation approach (NMSim) working on a coarse-grained protein representation.⁶⁴ With this, stereochemically and energetically favorable conformations of HEWL were generated previously.⁶⁴

As yet another local index, stability maps rc_{ij} are two-dimensional itemizations of the r_i and report when a “rigid contact” between two residues of the network vanishes during the thermal unfolding simulation. The upper triangles of Figure 5c and d show the stability maps for the single network topology and the ensemble of network topologies, respectively. Again, blocks of stable contacts are pronounced for secondary structures elements. In contrast, very weak contacts are identified for residues 81–87 that partially form a 3_{10} helix. This is in agreement with results from NMR experiments that reveal a disordered structure of this region.⁶⁵ The lower triangles of Figure 5c and d show a modification of the stability map that highlights solely those residue pairs with a “rigid contact” where the residues are within a distance of 5 Å. This map is referred to as “neighbor stability map”. Accordingly, a rigid contact in such a map that melts early in the thermal unfolding simulation is a prominent target for rigidification and, hence, for improving protein stability.

CONCLUSIONS

In recent years, there has been encouraging progress in characterizing the flexibility and rigidity of biomacromolecules down to the residue level by graph theoretical approaches. However, for deriving maximal advantage from information on biomacromolecular flexibility and rigidity, results from rigidity analyses must be linked to biologically relevant characteristics of a structure, such as (thermo-)stability and function. This provided the incentive for us to develop the CNA software package presented here. CNA functions as a front- and backend to the FIRST software and allows setting up a variety of constraint network representations, processing the results obtained from FIRST, and calculating global and local indices for quantifying biomacromolecular stability.

Thus, while CNA relies on FIRST as a core engine, it goes beyond the mere identification of flexible and rigid regions in a biomacromolecular structure. Major advancements in that respect include (I) a refined modeling of thermal unfolding simulations that considers the temperature-dependence of hydrophobic tethers, (II) the ability to perform rigidity analyses on ensembles of network topologies, either generated from structural ensembles provided as input or by using the concept

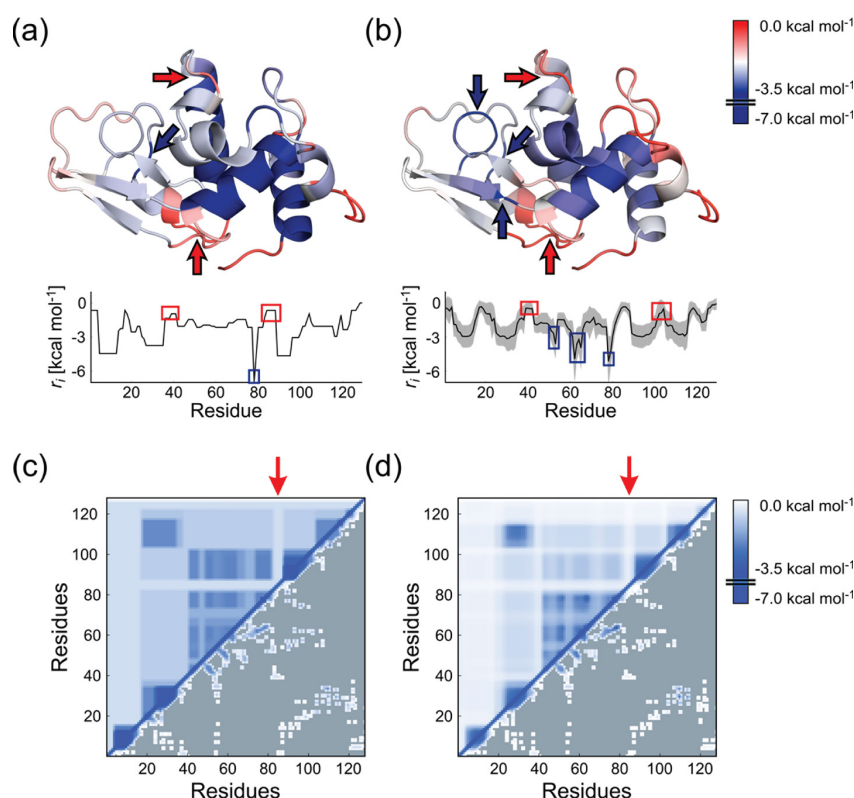


Figure 5. (a) Rigidity index r_i determined by analyzing the single network topology and (b) ensemble of network topologies plotted against a residue identifier and color coded onto the structure (range of color code: red (flexible) to blue (rigid)). In addition, the plot in (b) shows the standard deviation as a gray area. Blue rectangles and blue arrows in panels (a) and (b) highlight structurally stable regions for which high protection factors have been determined by H/D experiments. Red rectangles and red arrows in panels (a) and (b) highlight structurally flexible regions that are associated with hinge regions of HEWL. Stability maps (upper triangle) and neighbor stability maps (lower triangle) determined by analyzing the single network topology (c) and the ensemble of network topologies (d). The color depicts how stably two residues are connected and ranges from white (low stability) to blue (high stability). Red arrows highlight regions that reveal a disordered structure in NMR experiments. Gray areas in the neighbor stability map are displayed when residues are more than 5 Å away from each other.

of fuzzy noncovalent constraints, and (III) computing a set of global and local indices for characterizing biomacromolecular flexibility and rigidity, three of which have been introduced only recently by us.⁵⁶ The advancements allow (I) modeling in a more detailed manner the thermal unfolding of biomacromolecules, (II) obtaining more robust results from rigidity analyses due to a reduced sensitivity to the structural input, and (III) extending the application domain of rigidity analyses in that phase transition points (“melting points”) and unfolding nuclei (“structural weak spots”) are determined automatically. Such advancements are important for data-driven protein engineering, for example, for identifying structural parts that influence protein thermostability.²⁸ Furthermore, CNA robustly handles small-molecule ligands in general. This is important when it comes to estimating the influence of ligands on biomacromolecular stability, for example, for probing signal transmission across a protein structure for understanding and predicting “dynamic allostery”⁶⁶ and in assessing (changes in) flexibility characteristics of binding sites and interface regions.⁶⁷ How CNA can be applied in that respect has been demonstrated in a showcase example on HEWL.

CNA maintains the efficiency of FIRST. This has been achieved by linking CNA and FIRST via the *pyFIRST* interface module, minimizing the I/O overhead. The analysis of a single

protein structure by CNA usually takes only a few seconds for systems of several hundred residues on a single core. The runtime for analyses of ensembles of network topologies, which is in the order of hours currently, could be further reduced given that processing individual members of such an ensemble is trivially parallelizable. Finally, the hierarchical design of the software makes CNA highly adaptable and extensible, for example, by adding new index definitions.

Overall, we believe that these unique features make CNA an interesting tool for linking biomacromolecular structure, flexibility, (thermo-)stability, and function.

AUTHOR INFORMATION

Corresponding Author

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

Present Address

[†]Elsevier Information Systems GmbH, Frankfurt am Main, Germany.

Author Contributions

[‡]These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to the Ministry of Innovation, Science, and Research of North Rhine-Westphalia and Heinrich-Heine-University Düsseldorf for a scholarship to PCR within the CLIB-Graduate Cluster Industrial Biotechnology. We are grateful to Daniel Mulnaes (Heinrich-Heine-University, Düsseldorf) for proofreading the manuscript. CNA is available from <http://cpclab.uni-duesseldorf.de/software> for nonprofit organizations.

■ ABBREVIATIONS

CNA, Constraint Network Analysis; FIRST, Floppy Inclusions and Rigid Substructure Topography; HEWL, Hen Egg White Lysozyme; FNC, Fuzzy Noncovalent Constraints; PDB, Protein Data Bank; DSSP, Define Secondary Structure of Proteins

■ REFERENCES

- (1) Ahmed, A.; Kazemi, S.; Gohlke, H. Protein flexibility and mobility in structure-based drug design. *Front. Drug Des. Discovery* **2007**, *3*, 455–476.
- (2) Heal, J. W.; Jimenez-Roldan, J. E.; Wells, S. A.; Freedman, R. B.; Romer, R. A. Inhibition of HIV-1 protease: The rigidity perspective. *Bioinformatics* **2012**, *28*, 350–357.
- (3) Jagodzinski, F.; Hardy, J.; Streinu, I. Using rigidity analysis to probe mutation-induced structural changes in proteins. *J. Bioinf. Comput. Biol.* **2012**, *10*.
- (4) Radestock, S.; Gohlke, H. Protein rigidity and thermophilic adaptation. *Proteins* **2011**, *79*, 1089–1108.
- (5) Tan, H. P.; Rader, A. J. Identification of putative, stable binding regions through flexibility analysis of HIV-1 gp120. *Proteins* **2009**, *74*, 881–894.
- (6) Halle, B. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1274–1279.
- (7) Dokholyan, N. V.; Li, L.; Ding, F.; Shakhnovich, E. I. Topological determinants of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8637–8641.
- (8) Vendruscolo, M.; Dokholyan, N. V.; Paci, E.; Karplus, M. Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E* **2002**, *65*, 1–4.
- (9) Böde, C.; Kovács, I. A.; Szalay, M. S.; Palotai, R.; Korcsmáros, T.; Csermely, P. Network analysis of protein dynamics. *FEBS Lett.* **2007**, *581*, 2776–2782.
- (10) Greene, L. H.; Higman, V. A. Uncovering network systems within protein structures. *J. Mol. Biol.* **2003**, *334*, 781–791.
- (11) Heringa, J.; Argos, P. Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol.* **1991**, *220*, 151–171.
- (12) Heringa, J.; Argos, P.; Egmond, M. R.; Devlieg, J. Increasing thermal stability of subtilisin from mutations suggested by strongly interacting side-chain clusters. *Protein Eng.* **1995**, *8*, 21–30.
- (13) Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins* **2001**, *44*, 150–165.
- (14) Rader, A. J.; Hespeneide, B. M.; Kuhn, L. A.; Thorpe, M. F. Protein unfolding: Rigidity lost. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 3540–3545.
- (15) Whiteley, W. Counting out to the flexibility of molecules. *Phys. Biol.* **2005**, *2*, S116–S126.
- (16) Jacobs, D. J.; Thorpe, M. F. Generic rigidity percolation: The pebble game. *Phys. Rev. Lett.* **1995**, *75*, 4051–4054.
- (17) Jacobs, D. J.; Hendrickson, B. An algorithm for two-dimensional rigidity percolation: The pebble game. *J. Comput. Phys.* **1997**, *137*, 346–365.
- (18) Katoh, N.; Tanigawa, S. A proof of the molecular conjecture. *Discrete Comput. Geom.* **2011**, *45*, 647–700.
- (19) Hespeneide, B. M.; Rader, A. J.; Thorpe, M. F.; Kuhn, L. A. Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graphics Modell.* **2002**, *21*, 195–207.
- (20) Jacobs, D. J.; Dallakyan, S.; Wood, G. G.; Heckathorne, A. Network rigidity at finite temperature: Relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *Phys. Rev. E* **2003**, *68*.
- (21) Gohlke, H.; Kuhn, L. A.; Case, D. A. Change in protein flexibility upon complex formation: Analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* **2004**, *56*, 322–337.
- (22) Rader, A. J.; Anderson, G.; Isin, B.; Khorana, H. G.; Bahar, I.; Klein-Seetharaman, J. Identification of core amino acids stabilizing rhodopsin. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7246–7251.
- (23) Rader, A. J.; Bahar, I. Folding core predictions from network models of proteins. *Polymer* **2004**, *45*, 659–668.
- (24) Mamonova, T.; Hespeneide, B.; Straub, R.; Thorpe, M. F.; Kurnikova, M. Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.* **2005**, *2*, S137–S147.
- (25) Wells, S.; Menor, S.; Hespeneide, B. M.; Thorpe, M. F. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* **2005**, *2*, S127–S136.
- (26) Livesay, D. R.; Jacobs, D. J. Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins* **2006**, *62*, 130–143.
- (27) Ahmed, A.; Gohlke, H. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. *Proteins* **2006**, *63*, 1038–1051.
- (28) Radestock, S.; Gohlke, H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.* **2008**, *8*, 507–522.
- (29) Fulle, S.; Gohlke, H. Analyzing the flexibility of RNA structures by constraint counting. *Biophys. J.* **2008**, *94*, 4202–4219.
- (30) Fulle, S.; Gohlke, H. Constraint counting on RNA structures: Linking flexibility and function. *Methods* **2009**, *49*, 181–188.
- (31) Fulle, S.; Gohlke, H. Statics of the ribosomal exit tunnel: Implications for cotranslational peptide folding, elongation regulation, and antibiotics binding. *J. Mol. Biol.* **2009**, *387*, 502–517.
- (32) Fulle, S.; Christ, N. A.; Kestner, E.; Gohlke, H. HIV-1 TAR RNA spontaneously undergoes relevant apo-to-holo conformational transitions in molecular dynamics and constrained geometrical simulations. *J. Chem. Inf. Model.* **2010**, *50*, 1489–1501.
- (33) Mottonen, J. M.; Jacobs, D. J.; Livesay, D. R. Allosteric response is both conserved and variable across three CheY orthologs. *Biophys. J.* **2010**, *99*, 2245–2254.
- (34) Rader, A. J. Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys. Biol.* **2010**, *7*, 016002.
- (35) Rathi, P. C.; Radestock, S.; Gohlke, H. Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* **2012**, *159*, 135–144.
- (36) Pfleger, C.; Radestock, S.; Schmidt, E.; Gohlke, H. Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.* **2013**, *34*, 220–233.
- (37) Wells, S. A.; Jimenez-Roldan, J. E.; Romer, R. A. Comparative analysis of rigidity across protein families. *Phys. Biol.* **2009**, *6*.
- (38) Zaccai, G. Biochemistry - How soft is a protein? A protein dynamics force constant measured by neutron scattering. *Science* **2000**, *288*, 1604–1607.
- (39) Jacobs, D. J.; Dallakyan, S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys. J.* **2005**, *88*, 903–915.
- (40) Gonzalez, L. C.; Wang, H.; Livesay, D. R.; Jacobs, D. J. Calculating ensemble averaged descriptions of protein rigidity without sampling. *PLoS One* **2012**, *7*.
- (41) Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.
- (42) Joosten, R. P.; Beek, T. A. H. T.; Krieger, E.; Hekkelman, M. L.; Hoof, R. W. W.; Schneider, R.; Sander, C.; Vriend, G. A series of PDB

- related databases for everyday needs. *Nucleic Acids Res.* **2011**, *39*, D411–D419.
- (43) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
- (44) Ascher, D.; Dubois, P. F.; Hinsén, K.; Hugunin, J.; Oliphant, T. *Numerical Python*, 2001.
- (45) Jones, E.; Oliphant, T.; Peterson, P. *SciPy: Open Source Scientific tools for Python*, 2001.
- (46) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*.
- (47) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: A Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2008**, *2*.
- (48) Beazley, D. M. Automated scientific software scripting with SWIG. *Future Gener. Comput. Syst.* **2003**, *19*, 599–609.
- (49) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (50) Case, D.A.; T. A. D., Cheatham, T.E.; , III, Simmerling, C.L.; Wang, J.; Duke, R.E.; Luo, R.; Walker, R.C.; Zhang, W.; Merz, K.M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; I. Kolossváry, Wong, K.F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S.R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D.R.; Mathews, D.H.; Seetin, M.G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P.A. *AMBER 11*; University of California: San Francisco, 2010.
- (51) Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6*, 1333–1337.
- (52) Privalov, P. L.; Gill, S. J. Stability of protein–structure and hydrophobic interaction. *Adv. Protein Chem.* **1988**, *39*, 191–234.
- (53) Schellman, J. A. Temperature, stability, and the hydrophobic interaction. *Biophys. J.* **1997**, *73*, 2960–2964.
- (54) Burnham, K. P.; Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information–Theoretic Approach*, 2. ed.; Springer: New York, 2002; pp XXVI, 488 S.
- (55) Radford, S. E.; Dobson, C. M.; Evans, P. A. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* **1992**, *358*, 302–307.
- (56) Matagne, A.; Radford, S. E.; Dobson, C. M. Fast and slow tracks in lysozyme folding: Insight into the role of domains in the folding process. *J. Mol. Biol.* **1997**, *267*, 1068–1074.
- (57) Dinner, A. R.; Sali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* **2000**, *25*, 331–339.
- (58) Kiefhaber, T. Kinetic traps in lysozyme folding. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9029–9033.
- (59) Wildegger, G.; Kiefhaber, T. Three-state model for lysozyme folding: Triangular folding mechanism with an energetically trapped intermediate. *J. Mol. Biol.* **1997**, *270*, 294–304.
- (60) Haliloglu, T.; Bahar, I. Structure-based analysis of protein dynamics: Comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. *Proteins* **1999**, *37*, 654–667.
- (61) Radford, S. E.; Buck, M.; Topping, K. D.; Dobson, C. M.; Evans, P. A. Hydrogen-exchange in native and denatured states of hen egg-white lysozyme. *Proteins* **1992**, *14*, 237–248.
- (62) McCammon, J. A.; Gelin, B. R.; Karplus, M.; Wolynes, P. G. The hinge-bending mode in lysozyme. *Nature* **1976**, *262*, 325–6.
- (63) Kohn, J. E.; Afonine, P. V.; Ruscio, J. Z.; Adams, P. D.; Head-Gordon, T. Evidence of functional protein dynamics from X-ray crystallographic ensembles. *PLoS Comput. Biol.* **2010**, *6*.
- (64) Ahmed, A.; Rippmann, F.; Barnickel, G.; Gohlke, H. A normal mode-based geometric simulation approach for exploring biologically relevant conformational transitions in proteins. *J. Chem. Inf. Model.* **2011**, *51*, 1604–1622.
- (65) Smith, L. J.; Sutcliffe, M. J.; Redfield, C.; Dobson, C. M. Structure of hen lysozyme in solution. *J. Mol. Biol.* **1993**, *229*, 930–944.
- (66) Tzeng, S. R.; Kalodimos, C. G. Dynamic activation of an allosteric regulatory protein. *Nature* **2009**, *462*, 368–U139.
- (67) Metz, A.; Pfeleger, C.; Kopitz, H.; Pfeiffer-Marek, S.; Baringhaus, K. H.; Gohlke, H. Hot spots and transient pockets: Predicting the determinants of small-molecule binding to a protein–protein interface. *J. Chem. Inf. Model.* **2011**, *52*, 120–133.
- (68) Krüger, D. M.; Rath, P. C.; Pfeleger, C.; Gohlke, H. CNA Web server: Rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo)stability, and function. *Nucleic Acids Res.* **2013**, DOI: 10.1093/nar/gkt292.

NOTE ADDED IN PROOF

A CNA Web server is available at <http://cpclab.uni-duesseldorf.de/cna/>.⁶⁸

NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on April 8, 2013, with an error in reference 68. The corrected version was published to the Web on April 9, 2013.

13.7 Publication IV

Allosteric Coupling deduced from Altered Rigidity Percolation in Biomacromolecules

Pfleger, C., Minges, A.R.M., Gohlke, H.

Submitted manuscript (2014)

Author contribution to the publications:

My contribution to this publication was developing the ensemble-based perturbation approach with the aim to investigate allosteric signaling and regulation. This results in a contribution of **70%** to this publication.

Abstract

Allostery is the coupling between different sites in biomacromolecules in which an impact at one site causes an effect at a distant site. Over the last decades, aspects of dynamics extended the classical view of conformational allostery. This type of dynamic allostery can happen in the absence of conformational changes, and thus, is difficult to deduce from static X-ray structures. Molecular dynamics simulations are widespread used to identify correlated dynamics in biomacromolecules but require an accurate separation of noise from signal. Hence, we introduced an ensemble-based perturbation approach for analyzing dynamic allostery in order to quantify cooperativity free energies and allosteric communication. To this end, the approach monitors altered mechanical stability by perturbing constraint network topologies of biomacromolecules. The approach was validated by mutational perturbations on eglin c, which showed a good correlation between predicted and experimental free energies of destabilization and almost perfectly reproduced a continuous pathway of dynamically coupled residues as found in NMR experiments. In case studies on PTP1B and LFA-1, we found long-range effects of altered rigidity, connecting the allosteric and the orthosteric sites in both systems. Finally, the predicted free energy of cooperativity in LFA-1 is in agreement with the underlying mechanism of negative cooperativity in LFA-1. Our results emphasize that the perturbation approach provides an exciting opportunity to consider entropic effects in analyzing allosteric regulation.

Introduction

Allostery is the process by which biomacromolecules transmit the effect of binding at one site to another, often distal, functional site (1). Allosteric regulation occurs in enzyme activation (2), metabolism regulation (3), transcription control (4,5), and signal transduction (6). The widespread exploitation of allostery by nature thus renders a quantitative and predictive description of allostery fundamental for understanding biological processes at the molecular level and beyond (7,8). Likewise, such a description is expected to have a major impact in the field of drug design (9).

The first two models for allostery (10,11) dominated the field for decades and centered on the importance of conformational change between two well-defined structural end states (12). However, both models are phenomenological (1,12). For gaining insights into how a biomacromolecular structure enables allosteric communication, further model developments were required (13,14) that led to the “structural view” of allosteric mechanism (15-17). This view culminates in some approaches positing the existence of conserved pathways of allosteric signal transmission between the sites (1). Accordingly, a number of computational techniques, including mapping of residue networks by evolutionary, topology, and simulation analyses, have been developed for identifying such pathways at the atomistic level (16,18,19).

The Monod-Wyman-Changeux model posits an equilibrium shift between the two structural end states upon ligand binding (10), and the Koshland-Nemethy-Filmer model involves an induced fit of a binding site (20) enabled by the inherent flexibility of proteins (11). Thus, both models farsightedly touched on the dynamic nature of biomacromolecules *in the context of conformational changes*. The role of changes in protein dynamics for allosteric communication *even in the absence of conformational changes* was proposed by Cooper and Dryden (21). Based on a statistical thermodynamics analysis of ligand binding, they showed that allosteric communication can arise out of changes in frequencies and amplitudes of biomacromolecular thermal fluctuations and that this “dynamic allostery” is primarily an entropy effect. This view of allostery has gained much attention since (22-24), fueled by the development of experimental techniques that allow distinguishing between the two aspects of the role of dynamics in transitions at the residue level (25-27) and a theoretical underpinning in terms of the free energy landscape of a biomacromolecule (28-32). According to this framework, all possible conformations of a protein are sampled, and binding of an allosteric ligand shifts the population by stabilizing a certain conformation. Based on this, the ensemble allosteric model (33) proposes that allosteric mechanisms may be more statistical, and less deterministic, than the classical models suggests (1), and that it is the relative stability of the different biomacromolecular states that determines the type and magnitude of coupling between both sites (1,33).

Here, we integrate an ensemble-based perturbation approach with the analysis of biomacromolecular statics to construct a model of dynamic allostery. We apply our model to two systems encompassing ligand-based K- and V-type allostery as well as a system showing allosteric changes in dynamics from (surface) mutations that do not affect structure. We introduce a free energy quantity based on the mechanical stability of the biomacromolecule and show that this quantity predicts coupling in all three systems in a manner consistent with experiment. By construction, our model excludes any conformational changes of the biomacromolecule upon perturbations. Hence, the observed allosteric communication must arise from changes in the width of the distributions of biomacromolecular states only and so is entropic in nature. We discuss that the model is able to describe the effect of tunable sensitivity of a biomacromolecular ensemble.

Overall strategy and theory

Cooper and Dryden stressed the aspect of dynamics as a carrier for allosteric signaling between distant sites (21). A quantity of cooperativity is the allosteric free energy ($\Delta\Delta G$), defined as the difference between the free energies of the first and second binding process. By separating the free energies into its entropy and enthalpy contributions, they concluded that allosteric cooperativity effects could be primarily entropic, and thus, happen in the absence of conformational changes. It follows that, in positive cooperativity, binding of the first ligand results in a major loss of entropy, and thereby lowers the entropic costs for a subsequent binding process. In contrast, if binding of the first ligand does not quench all dynamics in the biomacromolecule, the major loss of entropy can occur during the second binding process, resulting in negative cooperativity (22).

In the context of statics, biomacromolecules can be decomposed into flexible and rigid regions by determining the number and spatial distribution of the internal independent degrees of freedom (floppy modes) (34,35). Stiffening the network due to ligand binding cause a loss of floppy modes and, hence, can lower the entropic costs of a further binding process (Figure 1A). Because floppy modes are delocalized, i.e. they represent collective motions in a certain region, they are related to low frequency modes of motion within a system. In turn, high frequency modes are neglected. However, it has been noted that, despite their local characteristics, high-frequency modes can contribute to allosteric signaling as well (36).

Overall strategy: Following this formulation of entropy-driven cooperativity, rigidity analysis should be particularly useful for detecting altered biomacromolecular stability upon ligand binding that may exert an entropic effect (37,38). For this, a biomacromolecule is modeled as a constraint network. Here, atoms are represented as bodies and covalent and noncovalent interactions (hydrogen bonds, salt bridges, and hydrophobic tethers) as sets of bars (constraints) (37,39). A fast combinatorial algorithm, the pebble game (40), then counts the bond rotational degrees of freedom in the constraint network. The theory underlying this approach is rigorous (41), and results from rigidity analyses have been successfully compared with those from experiments and other computational approaches (42-48).

Rigidification due to ligand binding can percolate through the network of interactions within the biomacromolecule and affects distant site(s) if the biomacromolecule is poised to this before perturbation (37,40). Our perturbation approach compares ensembles from the ground and perturbed states: First, the ground-state ensemble is extracted from an MD trajectory. The perturbed ensemble is then obtained by removing the covalent and non-covalent constraints associated to the allosteric effector from each network topology of the ground-state ensemble. As any changes in biomolecular conformation are excluded, changes in the biomacromolecular stability must arise solely from the perturbation in the network topology. The results from rigidity analysis are post-processed to predict putative residues involved in allosteric signaling and/or the type and magnitude of coupling by monitoring long-range altered rigidity.

Free energy quantity derived from mechanical stability: Biomacromolecules usually display a hierarchy of structural stability reflecting the modularity of their structure. In order to identify this hierarchy, rigidity analysis is performed by Constraint Network Analysis (CNA) such that a variety of network topologies obtained by removing non-covalent constraints is analyzed. To this end, hydrogen bond constraints (including salt bridges) are removed from the network in the order of increasing strength (39,47,49). For each threshold value E_{cut} at state σ a new network is generated where all hydrogen bonds with an energy $E_{HB} > E_{cut}(\sigma)$ are removed from the network (Figure 1 B). The hydrogen bond energy E_{HB} is determined from an empirical energy function (50). From these analyses, stability maps rc_{ij} are derived that identify pairs of residues that are flexibly or rigidly correlated across the

structure. More specifically, rc_{ij} maps are derived by identifying “rigid contacts” between residue pairs $R_{\{i,j\}}$ where a rigid contact exists if two residues belong to the same rigid cluster c of the set of rigid clusters $C^{E_{cut}}$ and are separated by $\leq 5 \text{ \AA}$ (eq. 1).

$$rc_{ij} = \min\{E_{cut} \mid \exists c \in C^{E_{cut}} : R_i \wedge R_j \in c\}. \quad (1)$$

That way, a contact’s stability relates to the microscopic stability in the network and, taken together, the microscopic stabilities of all residue–residue contacts display the hierarchy of stability in the biomacromolecule. The sum over all rigid contacts then yields a measure for the mechanical energy of the system (eq. 2)

$$E_{CNA} = \sum_i^n \sum_{j>i}^n rc_{ij} \quad (2).$$

We note that (the average of) E_{CNA} has been used recently as a proxy for the melting enthalpy of a protein and correlates with the protein’s melting temperature (P. Rathi, K.E. Jaeger, H. Gohlke, unpublished results), as expected from the work of Robertson and Murphy (51).

The difference in the mechanical energy $\Delta E_{CNA} = E_{CNA,perturbed} - E_{CNA,ground}$ reflects the change in biomacromolecular stability due to the presence of the ligand. As the perturbation of a ground state network by removal of covalent and noncovalent constraints from the ligand is localized and small (usually about 0.7 – 2.0% of the total number of constraints in a network topology is removed), the ensembles of ground and perturbed states should highly overlap. This warrants applying a free energy perturbation approach to compute a mechanical perturbation free energy ΔG_{CNA} (perturbed \rightarrow ground state) according to eq. 3

$$\Delta G_{CNA} = -k_B T \cdot \ln \left\langle \exp \left(- \frac{\Delta E_{CNA}}{k_B T} \right) \right\rangle_{ground} \quad (3)$$

where the product of the Boltzmann constant k_B and temperature T is 0.596 kcal mol⁻¹ at 300 K and $\langle \dots \rangle_{ground}$ denotes averaging over all states in the ground state ensemble.

Cooperative free energy: For two binding events at sites s1 and s2 of a biomacromolecule, the type and magnitude of cooperativity can be deduced from the cooperative free energy $\Delta\Delta G$ (eq. 4)

$$\Delta\Delta G = \Delta G_{s1/s2} - (\Delta G_{s1} + \Delta G_{s2}) \quad (4).$$

In dynamic allostery, according to Cooper and Dryden (21), if $\Delta\Delta G < 0$, the major loss of entropy already happens after the first binding process und favors further binding, i.e. both sites are coupled by positive cooperativity. Otherwise, if $\Delta\Delta G > 0$, the major loss of entropy must occur during the second binding step, and thus, both sites are coupled by negative cooperativity.

Identification of pathways for allosteric signaling: While the mechanical perturbation free energy is a global quantity, a per-residue decomposition of the mechanical energy is required to identify those residues that contribute most to the allosteric signaling. A per-residue decomposition of the free energy is obtained by following a linear response approach according to eq. 5

$$\Delta G_{i,CNA} = \alpha \left(\langle E_{i,CNA}^{perturbed} \rangle - \langle E_{i,CNA}^{ground} \rangle \right) \quad (5)$$

where $E_{i,CNA}$ is the mechanical energy of all pairwise “rigid contacts” related to residue i (eq. 6)

$$E_{i,CNA} = \frac{1}{2} \sum_{j \neq i}^n r c_{ij} . \quad (6)$$

Based on this per-residue free energy it should be possible to examine how much each residue is affected upon perturbing the ground state and how residues influence each other that are spatially separated. From this, we want to probe whether pathways are constituted of a few residues that are strongly coupled (“small pathway”) or multiple residues (“broad pathway”) that are weakly coupled. The latter question relates to the point of network stability and robustness of the allosteric coupling pathway in view of potential mutations, whereby less detrimental effects are expected for the “broad pathway” variant.

Results

Allosteric effects of mutations on eglin c dynamics

The small serine protease inhibitor eglin c is an ideal test case for validation as experimental data is available (I) from chemical denaturation experiments providing free energies of unfolding for 13 single and double mutants (see Table 1) and (II) from NMR relaxation experiments for these mutants revealing continuous pathways of dynamically coupled residues (52,53).

Comparison between mechanical perturbation free energies and experimental free energies of stability: For validation, we calculated the mechanical perturbation free energies (eq. 3) of 13 single and double mutants. Each *in silico* mutation were done by cutting the respective amino acid side chain of the wild-type between the C α and C β atoms, and hence, mutating the residue to alanine. As reference, we used free energies of unfolding between wild-type and mutants from chemical denaturation experiments (52, 53) ranging from -0.16 kcal mol⁻¹ to -3.74 kcal mol⁻¹. The computed and experimental free energies yield a good and significant correlation ($R^2 = 0.80$, 95% confidence interval: $0.43 < R^2 < 0.94$, $p = 0.0001$) (Figure 2A). The V63A mutant, though destabilizing eglin c, was not considered in the correlation analysis because no altered rigidity upon *in silico* perturbation was found. This behavior is in agreement with findings from NMR experiments according to which the mutation V63A results only in a few changes in eglin c dynamics (52).

Comparison between computed and experimental coupling free energies: Next, we analyzed the coupling between distant sites of six double mutants. None of the mutation sites makes any direct interactions to another site. While almost all double mutants show sizeable destabilization effects (Table 1), experimental coupling free energies are either very small (< 0.27 kcal mol⁻¹) and/or just above the experimental error except for the double mutation V18A/V54A, where the coupling free energy indicates a negative coupling (52). Likewise, computed coupling free energies (eq. 4) do not differ significantly from zero in all six cases.

Identification of dynamically coupled residues: NMR studies on the mutants V14A, V34A, and V54A showed continuous pathways of dynamically coupled residues up to 10 Å apart from the site of mutation (52, 53). All three mutants have no significant effect on the tertiary structure of eglin c, and hence, the response appears to be purely dynamic driven (52). Because the V34A mutant shows the most far-reaching dynamic response in the NMR experiment, we focused on this mutant. Upon *in silico* perturbation the network topologies lose on average ~2 (5.6% of all) non-covalent constraints associated to hydrophobic tethers. The per-residue free energy $\Delta G_{i,CNA}$ (eq. 5) was calculated from differences between stability maps of the ground (wild-type) and perturbed (mutant) state (Figure 2B). Almost all residues in the vicinity of the mutation site felt the perturbation but only a few residues show larger $\Delta G_{i,CNA}$'s. The largest per-residue $\Delta G_{i,CNA}$'s was found for residues that make native contacts with V34; these are residues Y35, F36, V52, R53, and V54 with $\Delta G_{i,CNA}$'s ≥ 0.67 kcal mol⁻¹. When considering only those residues with $\Delta G_{i,CNA} > 0.2$ kcal mol⁻¹ (i.e., with a $\Delta G_{i,CNA}$ value of at least 24% of the maximal $\Delta G_{i,CNA}$), a contiguous pathway of coupled residues was revealed (Figure 2C). Striking examples for the long-range character of altered rigidity are residues V62 and V63, which are 15.9 Å (13.3 Å) apart from the site of mutation. Notably, both residues also show altered dynamics in NMR experiments (Figure 2D). On the opposite site of eglin c, the connection with the reactive binding loop via residue V43 (13.4 Å) is missed: None of the loop residues show significant non-zero $\Delta G_{i,CNA}$'s. However, the strong influence observed for R53 may indicate a forwarding effect on the reactive binding loop.

Overall, the identified pathway is in good agreement with results from NMR experiments (Figure 2C and D): When phrased as a binary classification problem, our results allow discriminating between pathway and non-pathway residues with a sensitivity of ~73%, a specificity of ~87%, and an accuracy of ~85%. Additionally, we analyzed the mutants V54A and V14A. In the case of V14A, no altered rigidity was observed by our perturbation approach, but this mutant is also reported as having the lowest impact on dynamics in the NMR experiments (53). In the case of V54A, we found a similar contiguous pathway as for V34A (Figure S1). While the sensitivity is lower than for V34A (~56%), a specificity of ~86% and an accuracy of ~82% demonstrate that we can particularly reliably identify non-pathway residues.

Probing V- and K-type allosteric mechanisms in PTP1B and LFA-1

The perturbation approach was next applied to PTP1B and LFA-1. The human PTP1B is part of the signaling transduction cascade leading to the phosphorylation of the insulin receptor (54). The LFA-1 domain is part of β_2 -integrin and binds to intercellular adhesion molecules (ICAM) (55). Both systems possess different allosteric mechanisms: PTP1B possesses a V-type mechanism, i.e. the allosteric effector antagonizes the enzyme activity but does not change the affinity for substrate binding. In contrast, LFA-1 possesses a K-type mechanism. Here, the allosteric effector inhibits the complex formation of LFA-1 and ICAM molecules. Both systems involve small (< 2.8 Å rmsd overall; < 1.7 Å rmsd in either the allosteric or orthosteric binding sites) conformational changes during the allosteric regulation. The ground state ensembles were extracted from MD trajectories starting from the effector bound structures of PTP1B and LFA-1. Network topologies were then *in silico* perturbed by removing the allosteric effector from each network.

Probing the allosteric mechanism in PTP1B: PTP1B is composed of an active site, a neighboring non-catalytic site (together referred to as orthosteric site) (56), and an allosteric site which is ~20 Å apart from the orthosteric site (54). During allosteric regulation, PTP1B undergoes two conformational changes: The first change is located at helix $\alpha 7^*$ next to the allosteric site, and the second one is located at the catalytically important WPD loop (residues T177-P185). Importantly, none of the residues in the allosteric site directly interacts with the WPD loop. Binding of the allosteric effector prevents the closure of the WPD loop as well as interactions between helix $\alpha 7^*$ and the two helices $\alpha 3$ and $\alpha 6$ (54). Consequently, the structure of PTP1B is trapped in the inactive state (Figure 3B).

Upon *in silico* perturbation, the network topologies lose on average ~4 (1.2% of all) hydrogen bond constraints and ~17 (14% of all) hydrophobic tether constraints. The per-residue $\Delta G_{i,CNA}$'s (eq. 5) of altered rigidity show that almost all residues felt the *in silico* perturbation but that in many cases the effects are very small or just above the standard error (SEM = 0.09 kcal mol⁻¹) (Figure 3A). As in the case of eglin c, we only considered residues with $\Delta G_{i,CNA} > 0.2$ kcal mol⁻¹ further, which keeps ~33% of the residues in PTP1B. Strong altered rigidity was found for the region enclosing helices $\alpha 3$, $\alpha 4$, and $\alpha 6$. In particular, L192 in helix $\alpha 3$, which is part of the allosteric site, shows the largest $\Delta G_{i,CNA}$ of 2.08 kcal mol⁻¹. By mapping those residues with $\Delta G_{i,CNA} > 0.2$ kcal mol⁻¹ onto the structure of PTP1B, we found residues up to ~21 Å apart from the allosteric site being influenced by effector binding, demonstrating the long-range effects of altered rigidity in PTP1B (Figure 3B). Remarkably, affected residues are not dispersed but rather show a distinct connection to active site residues (Y20, C215-S222, R254, G259, I261, Q262, Q266) and residues in the hinge region of the WPD loop (T177-W179, V184, P185). To ensure that results from altered rigidity are non-trivial, we calculated the packing density for each residue averaged over the ground state ensemble. From this, no correlation was observed between $\Delta G_{i,CNA}$ values and the packing density (Figure S3A).

So far, the analysis was based on a visual inspection of $\Delta G_{i,CNA}$ values mapped onto the protein structure. To determine to what extent a certain region is affected upon *in silico* perturbation in a more formalized way, we applied methods from the field of graph analysis to probe the allosteric communication in PTP1B. To this end, differences between stability maps of the ground and perturbed ensembles were represented by a stress-minimized graph G . In this graph, nodes are representing residues with $\Delta G_{i,CNA} > 0.2$ kcal mol⁻¹ and edges the pairwise mechanical perturbation energy ΔE_{ij} (eq. 6, see Material & Methods) between residues i and j . The stress-minimized graph was analyzed for residues that show a high allosteric communication and, hence, are most likely involved in allosteric signaling. Centrality indices are common measures to analyze the information flow through a graph and were already successfully applied to identify allosteric signaling (57-59). Here, we calculated the betweenness centrality for each edge within the graph (see Material & Methods). Remarkably, only a few edges have larger centrality values while most of the edges are close to zero (0.5% of the edges are significantly larger than zero) (Figure S4A). By considering only edges with centrality values significantly larger than zero, we identified two continuous pathways with more than two residues. Both pathways originate from the allosteric site and then spread through the graph (Figure 3C). The first pathway originates at residue G277 in the allosteric site and heads to helix $\alpha 7^*$. More striking is the second pathway, which originates at residue L192 in helix $\alpha 3$ and then branches at position F269 (helix $\alpha 6$) to W179 in the WPD loop (Figure 3D) and via T224 to residues G220, I261, and Q266 in the orthosteric site (Figure 3E). Notably, the identification of residue F191 is encouraging because F191 in the inactive state occupies a site that W179 must enter during the closure of the WPD loop (54). Although the affinity for substrate binding is not affected by binding of the allosteric inhibitor to PTP1B, the identification of residues in the orthosteric site is encouraging because it could indicate an interruption of a catalytically important network of residues. As such, Q266 is involved in a water-mediated hydrogen bonding network, which is a prerequisite for formation of an intermediate state during catalysis (60).

Probing the allosteric mechanism in LFA-1: The tertiary structure of LFA-1 adopts a Rossmann fold composed of a central β -sheet surrounded by seven α -helices. A metal ion-independent adhesion site (MIDAS, further referred to as orthosteric site) at the top of LFA-1 coordinates a magnesium ion, which is required for binding of ICAM (61). The activation of LFA-1 involves a conformational change of the C-terminal helix $\alpha 7$, which is neighboring the allosteric site (62). Binding of an allosteric effector holds the structure in the inactive state, and thus, inhibits the complex formation of LFA-1 and ICAM molecules (63). Remarkably, residues in the orthosteric site show only minor structural differences between the ICAM-1 bound intermediate state (PDB ID 1MQ8) and inactive state (PDB ID 3BQM) (heavy atom rmsd = 1.7 Å).

Upon *in silico* perturbation, the network topologies lose on average ~ 0.5 (0.3% of all) hydrogen bond constraints and ~ 8 (7% of all) hydrophobic tether constraints. The results from the predicted per-residue $\Delta G_{i,CNA}$ (Figure 4A) indicates that almost all residues felt the perturbation but only 38% have $\Delta G_{i,CNA} > 0.2$ kcal mol⁻¹. By mapping only residues with $\Delta G_{i,CNA} > 0.2$ kcal mol⁻¹ onto the structure of LFA-1, we found that the β -sheets (except for $\beta 2'$) in the core region show a strong altered rigidity (Figure 4B). Notably, the opposite region enclosing the helices $\alpha 1$, $\alpha 2$, $\alpha 3$, $\alpha 4$, $\alpha 5$, $\alpha 6$ was less altered upon *in silico* perturbation, which is in agreement with findings from NMR experiments (62). Again, $\Delta G_{i,CNA}$'s derived from altered rigidity do not correlate with the packing density of residues, indicating that the derived result is not trivial (Figure S3B). M140 in the orthosteric site is among the most distant residues with altered rigidity (~ 18 Å) and part of a network of residues that are important for the protein-protein interaction with ICAM (64). Besides M140, we also found the orthosteric site residues D137, S139, L142, and Q143. Again, this finding demonstrates

the long-range effect of altered rigidity in biomacromolecules. Finally, DrugScore^{PPI} was applied to probe the importance of residues in the orthosteric site for the protein-protein complex formation (Figure S4) (65). To this end, we used the crystal structure of LFA-1 in complex with ICAM-1 (PDB ID 1mq8). DrugScore^{PPI} detects L205 as *warm spot* residue, which has also been reported to abolish ICAM binding in mutagenic analyses (66) but does not show a significant change in the predicted $\Delta G_{i,CNA}$ values from the perturbation approach. Notably, the second largest contribution to protein-protein binding calculated by DrugScore^{PPI} is for M140. Even though mutagenesis analysis on M140A shows no effect on the binding affinity of ICAM-1 (66), the overall importance of this residue for ICAM-1 binding has been reported (64).

To probe the allosteric signaling more formally, we computed the centrality betweenness for each edge of the stress minimized graph shown in Figure 4C. When considering only edges with centrality values significantly larger than zero, only a few edges (0.4 %) remain left (Figure S2B). This results in a condensed representation of the graph (Figure 4C) from which two pathways spread from the allosteric site. The first pathway encloses residues L295 and L289 and heads to the functionally important hinge region between the sheet $\beta 4$ and helix $\alpha 7$ (Figure 4D). The second pathway encloses I237 and I150 and ends at residue L142 in the orthosteric site (Figure 4C and F). Notably, between the allosteric and orthosteric sites there is a “buffer zone” of at least two residues (Figure 4C). From there, L142 is connected to the other orthosteric site residues D137, S139, M140, and Q143.

Negative cooperativity in LFA-1 deduced from rigidity analyses

Next, we applied the perturbation approach to probe the underlying type and magnitude of the cooperativity in LFA-1. Following the formulation of Cooper and Dryden, in positive cooperativity the allosteric effector lowers the entropic costs for the subsequent substrate binding by stabilizing the biomacromolecule (Figure 1A). In negative cooperativity, binding of an allosteric effector quenches not all dynamics and, hence, the major loss of entropy must occur upon substrate binding. The latter case is the type of mechanism that would agree with experimental findings of a negative cooperativity by inhibiting the complex formation of LFA-1 and ICAM (67, 68).

To probe the negative cooperativity in LFA-1, we used an MD trajectory starting from a modelled structure of LFA-1 in complex with both the allosteric effector (BQM) and ICAM-1 (see Material and Methods). The modelled complex is stable over the course of the trajectory (heavy atom rmsd < 4.8 Å) (Figure S5). A structural ensemble of LFA-1_{BQM/ICAM-1} was extracted from the trajectory, which was used to generate perturbed ensembles of LFA-1_{apo}, LFA-1_{BQM}, and LFA-1_{ICAM-1}. The cooperative free energy (eq. 4) in LFA-1 was calculated from comparing the ensembles of LFA-1_{BQM/ICAM-1}, LFA-1_{BQM}, and LFA-1_{ICAM-1} against the ensemble of LFA-1_{apo}. For LFA-1_{BQM} and LFA-1_{ICAM-1}, the mechanical perturbation free energy ΔG_{CNA} (eq. 3) of altered rigidity is ~ 0.6 kcal mol⁻¹ (BQM) and ~ 8.9 kcal mol⁻¹ (ICAM-1), respectively. The ~ 15 times smaller free energy upon BQM binding shows that the allosteric effector does not drastically lower the entropic burden for the subsequent ICMA-1 binding. Still, the mechanical perturbation free energy in LFA-1_{BQM/ICAM-1} is ~ 12.7 kcal mol⁻¹. This results in a cooperative free energy $\Delta\Delta G$ of 3.2 kcal mol⁻¹. Accordingly, the entropic burden of ICAM-1 binding is increased if BQM is already bound to LFA-1, which correctly points to a mechanism of negative cooperativity in LFA-1.

Discussion

In this study, we presented an ensemble-based perturbation approach for analyzing dynamic allostery by tracking altered rigidity in biomacromolecules. To identify altered rigidity characteristics we applied *in silico* perturbations on constraint network topologies. To this end, covalent and non-covalent constraints associated to the allosteric effector are removed from the network topology. By way of construction of this model, the conformation of the biomacromolecule remains unchanged, and solely the perturbations in the network topology are leading to changes in biomacromolecular stability.

The perturbation approach was validated on eglin c at two levels. First, we compared predicted mechanical perturbation free energies with free energies of destabilization from chemical denaturation experiments of 13 single- and double-mutants. The obtained good correlation between predicted and experimental free energies is convincing due to two reasons: (I) changes in the network topologies upon *in silico* perturbations are relatively small and (II) predicted free energies of all *in silico* mutants are based on one conformational ensemble extracted from the MD trajectory starting from the wild-type structure of eglin c. Our findings show that the perturbation approach is sensitive enough to pick up effects of single mutations that lead to changes of the free energy of destabilization by $< 1.6 \text{ kcal mol}^{-1}$.

At the second level, we tested the perturbation approach to reproduce experimentally determined pathways of dynamically coupled residues in eglin c. To compute per-residue free energies, we used a linearized free energy expression following a linear response approach. Encouragingly, the predicted pathway, as reported by NMR experiments (52, 53), is reproduced for $\sim 85\%$ ($\sim 82\%$) of the involved residues for the mutant V34A (V54A). Remarkably, residues on the pathways are up to $\sim 16 \text{ \AA}$ apart from the site of mutation, which demonstrates a long-range percolation of altered rigidity in biomacromolecules. Only for the mutant V14A could no pathway be detected; however, this mutant also revealed the weakest impact on dynamics as reported by NMR experiments (53).

Next, we applied the perturbation approach on PTP1B and LFA-1. Both systems show small conformational changes during allosteric regulation. By comparing the active and inactive state of PTP1B, these conformational changes manifest in the catalytically important WPD loop and helix $\alpha 7^*$ close to the allosteric site. In LFA-1, the allosteric effector prevents the movement of helix $\alpha 7$, which is required for binding of ICAM molecules to LFA-1. However, the orthosteric site shows only minor conformational changes upon binding to ICAM-1 as found in an intermediate state of LFA-1 in complex with ICAM-1. For both systems, long-range altered rigidity was found upon *in silico* perturbation by removing the allosteric effector from the network topologies. Even though multiple residues (“broad pathway”) felt the perturbation at the allosteric site, we found a connection to residues in the orthosteric site as well as to other functionally important regions. For performing the pathway identification in a more formal way, we computed the allosteric communication within a graph representation of the residue pairwise altered rigidity. From this, we obtained condensed (“small pathway”) and continuous pathways that originate from the allosteric site. In PTP1B, the pathway connects the allosteric site to the WPD loop and to catalytically important residues in the orthosteric site. In LFA-1, the analysis discloses two pronounced pathways: One that heads to a hinge region, which is required for the movement of helix $\alpha 7$ upon activation and another that heads to the orthosteric site.

Finally, the computed cooperative free energy for binding of both the allosteric effector and the orthosteric ligand to LFA-1 was positive in agreement with negative cooperativity observed experimentally for LFA-1.

Contrary to other studies, in the introduced perturbation approach the signal-to-noise ratio is apparently high enough such that elaborate statistical filtering techniques, as required for analyzing changes in correlated motions upon binding of an allosteric effector (69), do not

need to be applied here. Furthermore, the approach does not have to be specifically parameterized for the protein to be investigated, in contrast to related work based on the Distance Constraint Model (70, 71). Finally, ensemble-based analyses instead of analyses of single structures (48) ensure robust results and are unbiased by conformational effects from direct comparison of the active and inactive states (72). A downside of the perturbation approach is the need for a computationally expensive MD simulation beforehand to generate an ensemble as structural input for CNA. Carrying out the MD simulation of PTP1B required ~14 days of computing time on a single NVIDIA Tesla M2070 GPU. Clearly, that way the computational efficiency of CNA is compromised; a single CNA run required 8 h on a workstation computer.

In summary, we introduced an ensemble-based perturbation approach to analyze dynamic allostery by way of computing mechanical perturbation free energies, cooperative free energies, pathways of allosteric signaling. The *in silico* perturbations solely affect the network topologies, which are the foundation for the rigidity analysis, without actually moving atoms. In all three tested systems (eglin c, PTP1B, and LFA-1), long-range effects of altered rigidity were found despite only small changes in the network topologies. This demonstrates that the perturbation approach is sensitive to pick up effects related to allosteric regulation. In PTP1B and LFA-1, altered rigidity is observed along continuous pathways of residues connecting the allosteric and the orthosteric site.

Materials and Methods

Structure preparation

In this study we used eglin c, the protein tyrosine phosphatase 1B (PTP1B), and the lymphocyte function-associated antigen 1 (LFA-1) to probe the allosteric mechanism by the perturbation approach. In the case of PTP1B and LFA-1, the binding of allosteric effectors lead to negative cooperativity.

Eglin c: The perturbation approach was initially validated on the small protease inhibitor eglin c. The wild-type structure of eglin c taken from the Protein Data Bank (PDB ID 1cse) (73) was used as starting structure for the MD simulations (see SI). To this end, we extracted the structure of eglin c from the complex with subtilisin. The ensemble of the wild-type structure was generated by extracting 1500 conformations from an MD trajectory of 300 ns length. Ensembles of mutants (Table 1) were then generated by *in silico* mutational perturbation of the wild-type ensemble. To this end, side chains of the residues were cut between the C α and the C β position and missing hydrogen atoms were added by using the tleap program (which is part of the AmberTools software package (74)).

Protein tyrosine phosphatase 1B (PTP1B): The starting structure for the MD simulation was taken from PDB ID 1t4j. The crystal structure is in complex with the allosteric effector FRJ (54). The helix $\alpha 7^*$ which is not resolved in the crystal structure was modeled using the MODELLER program (75). As template, we used the phosphotyrosine-bound structure of PTP1B (PDB ID 1pty). The partial charges of the allosteric effector FRJ were generated according to the RESP procedure (76, 77). The ensemble of the ground state was generated by extracting 1500 conformations from the MD trajectory of 300 ns length. All water molecules and ions were removed except for water molecules that make water-mediated hydrogen bonds between FRJ and PTP1B. The perturbed ensemble was then generated by *in silico* perturbation, i.e. removing the allosteric effector and water molecules (if present), of the ground state ensemble.

Lymphocyte function-associated antigen 1 (LFA-1): The starting structure for the MD simulation was taken from PDB ID 3bqm. The LFA-1 structure is in complex with the allosteric effector BQM. Partial charges of BQM were generated according to the RESP procedure (76, 77). The ensemble of the ground state was generated by extracting 1500 conformations from the MD trajectory of 300 ns length. An ensemble of the perturbed state of LFA-1 was generated by removing the allosteric effector from the ground state ensemble. All water molecules and ions were removed. In order to probe the allosteric cooperativity in LFA-1 we modelled a complex structure of LFA-1_{BQM/ICAM-1}. To improve the stability of the modelled complex during the MD simulation we used the intermediate state of LFA-1_{ICMA-1} (PDB ID 1mq8) even though a better resolved crystal structure in the high affinity state (PDB ID 1t0p) is available. To model the complex, the structure of LFA-1 in complex with ICAM-1 was aligned on PDB ID 3bqm. Afterwards, coordinates of ICAM-1 and the magnesium ion buried in the metal-ion-dependent adhesion site (MIDAS) were merged with PDB ID 3bqm.

Allosteric response deduced from rigidity analysis

Rigidity analyses were performed using the CNA software package (78). CNA was applied on ensembles of network topologies generated from conformational ensembles provided as input (see structure preparation). Average stability characteristics are calculated by constraint counting on each topology in the ensemble. Each network of covalent and noncovalent (hydrogen bonds including salt bridges and hydrophobic tethers) constraints was constructed with the FIRST software (version 6.2), which is part of the CNA software package (37). The

strength of hydrogen bonds (including salt bridges) were assigned by the energy E_{HB} computed by FIRST (50). Hydrophobic interactions between carbon or sulfur atoms were taken into account if the distance between these atoms was less than the sum of their van der Waals radii (C: 1.7 Å, S: 1.8 Å) plus $D_{cut} = 0.25$ Å (39). Fluor atoms of CF_3 groups in BQM are modelled as carbons to consider the increased bulky character of CF_3 groups (79).

Allosteric signaling from network analysis: Stability maps (80) as derived from rigidity analysis were used to generate a undirected graph representation. In such a graph, nodes represent the residues and edges represent the pairwise mechanical perturbation energy ΔE_{ij} between residue i and j by eq. 6

$$\Delta E_{ij} = \langle rc_{ij}^{perturbed} \rangle - \langle rc_{ij}^{ground} \rangle \quad (\text{eq. 6})$$

where $\langle rc_{ij} \rangle$ is the average strength of a rigid contact (eq. 1) between the two residues i and j . From the graph representation, we calculated centrality indices to quantify the relative importance of each site within the graph, e.g. the information flow through a graph. For the graph construction and calculation of centrality indices, we used the NetworkX extension for Python (81). The underlying algorithm of the betweenness centrality for each edges is documented in detail in refs. (82, 83). The betweenness centrality is defined as the sum of the fraction of all-pairs shortest paths that pass through the edge e (eq. 7)

$$C_B(i, j) = \frac{2}{n(n-1)} \sum_{i, j \in V} \frac{\sigma(i, j | e)}{\sigma(i, j)} \quad (\text{eq. 7})$$

The centrality is normalized where n is the number of nodes in the graph. $\sigma(i, j)$ is the number of shortest paths and $\sigma(i, j | e)$ is the number of those that pass through the edge e . For visualization of the graph, we used the software Visone (84). The graph embedding is minimized based on the pairwise $C\alpha$ distance in the PDB structure used as input for the perturbation approach.

Acknowledgments

We thank D. Mulnaes for fruitful discussions.

Tables

Table 1: Free energies from chemical denaturation experiments and the mechanical perturbation approach of mutants in eglin c.

Mutation	Chemical denaturation ^{1,2}		Rigidity analysis ^{3,4}	
	$\Delta\Delta G$	$\Delta\Delta\Delta G$	ΔG_{CNA}	$\Delta\Delta G_{CNA}$
V18A/V54A	3.74	0.95	1.24	-0.07
V63A	3.20	---	--- ⁷	---
V34A/V62A	2.54	0.25	0.90	-0.04
V54A/V62A	2.54 ⁵	0.26	1.02	-0.05
V14A/V18A	2.53 ⁵	0.27	0.62	<0.01
V18A/V34A	2.50	0.02	1.12	-0.06
L27A/V54A	1.98 ⁵	0.22	0.80	<0.01
V54A	1.59 ⁵	---	0.79	---
V34A	1.27	---	0.66	---
V18A	1.21 ⁵	---	0.52	---
V62A	1.02 ⁵	---	0.28	---
V14A	1.04 ⁶	---	0.10	---
L27A	0.16 ⁶	---	0.01	---

¹ Experimental $\Delta\Delta G$ of unfolding and thermodynamic couplings $\Delta\Delta\Delta G$ taken from ref. (52, 53). We considered positive free energies of destabilization instead of reported negative values.

² The standard error in $\Delta\Delta G$ is 0.13 kcal mol⁻¹ and in $\Delta\Delta\Delta G$ is 0.19 kcal mol⁻¹ as reported in ref. (52).

³ Predicted ΔG_{CNA} (eq. 3) relative to the WT and cooperativity $\Delta\Delta G_{CNA}$ (eq. 4); in kcal mol⁻¹.

⁴ The standard error in ΔG_{CNA} is 0.03 kcal mol⁻¹ and $\Delta\Delta G_{CNA}$ is 0.04 kcal mol⁻¹.

⁵ Updated standard error from ref. (53).

⁶ The standard error in $\Delta\Delta G$ is 0.12 kcal mol⁻¹, as reported in ref. (53).

⁷ Mutant shows no altered rigidity.

Figure captions

Figure 1:(A) Illustration of positive and negative cooperativity deduced from altered rigidity. The vertical dashed line indicates the rigidity percolation threshold where the biomacromolecule switches from a flexible to a rigid state. (B) Rigid cluster decomposition during the bond dilution process of LFA-1. The four states show the decay of rigidity where at each state all hydrogen bonds with an energy $E_{HB} > E_{cut}$ are removed from the network.

Figure 2:*In silico* mutation perturbation in eglin c. (A) Correlation between predicted ΔG_{CNA} 's (eq. 3) and $\Delta\Delta G$'s from chemical denaturation experiments (52, 53). The SEM of the predicted (horizontal) and experimentally determined (vertical) free energies are exemplarily shown for V18A/V54A. The mutant V63A (red) is considered an outlier because it shows no altered rigidity upon *in silico* perturbation and is reported to show only small changes in the dynamics by NMR experiments (52). (B) Differences between stability maps of the ground (wild-type) and perturbed (mutant) state upon V34A mutation. The difference map shows the pairwise mechanical perturbation energy ΔE_{ij} (eq. 6) between residue i and j and the attached histogram the per-residue $\Delta G_{i,CNA}$ (eq. 5). The dashed line at 0.2 kcal mol⁻¹ indicates a threshold at which we assume residues to be prone for allosteric regulation. Residues with $\Delta G_{i,CNA}$ above the threshold are mapped on the structure of eglin c (D) and are in good agreement with a continuous dynamic network derived from NMR experiments (E) (52). The site of mutation (V34) is shown in red (D and E), blue colors reflect predicted $\Delta G_{i,CNA}$ values in (D), and white residues in (E) are those that lack NMR data on changes in the dynamics but are included in the network as suggested by the authors of ref. (52).

Figure 3: Probing the allosteric mechanism in PTP1B. (A) Differences between stability maps of the ground (effector bound) and perturbed (*apo*) state of PTP1B. The attached histogram shows the per-residue $\Delta G_{i,CNA}$ (eq. 5) with a threshold of 0.2 kcal mol⁻¹ (dashed lines). Secondary structure as well as allosteric (red) and orthosteric (green) site information are attached at the top and right. (B) Residues with $\Delta G_{i,CNA} > 0.2$ kcal mol⁻¹ are mapped on the structure of PTP1B. The color indicates if a residue is part of the allosteric (red), orthosteric (green), WPD loop (orange), or any other site (blue). Darker colors indicate greater $\Delta G_{i,CNA}$ values. The same information is represented as a stress-minimized graph where nodes represent residues and edges the pairwise mechanical perturbation energy ΔE_{ij} (eq. 6) as in (A). The graph embedding is minimized based on the pairwise C α distance in PDB ID 1T4J. Highlighted residues show the dominant pathways for allosteric communication through the graph as derived from calculating the betweenness centrality (eq. 7). One pathway is mapped on the structure and branches at position F269 to connect the allosteric site with the hinge region of the WPD loop (D, orange) and residues in the orthosteric site (E, green). Additionally, (E) shows the aligned closed structure of PTP1B (black, PDB id 1pty).

Figure 4: Probing the allosteric mechanism in LFA. (A) Differences between stability maps of the ground (effector bound) and perturbed (*apo*) state of LFA-1.. The attached histogram shows the per-residue $\Delta G_{i,CNA}$ (eq. 5) with a threshold of 0.2 kcal mol⁻¹ (dashed lines). Secondary structure as well as allosteric (red) and orthosteric (green) site information are shown at the top and right. (B) Residues above the threshold are mapped on the structure of LFA-1. The color indicates if a residue is part of the allosteric (red), orthosteric (green), or any other site (blue). The same information is represented as a stress minimized-graph where nodes represent the residues and edges the pairwise mechanical perturbation energy ΔE_{ij} (eq. 6) as in (A). The graph embedding is minimized based on the pairwise C α distance in

PDB ID 3BQM. Highlighted residues illustrate the dominant allosteric pathway through the graph as calculated from the betweenness centrality (eq. 7). Two pathways are mapped on LFA-1: (D) pathway between the allosteric site and the hinge region between strand $\beta 4$ and helix $\alpha 7$ (blue); (E) pathway between the allosteric and the orthosteric site (green).

Figures

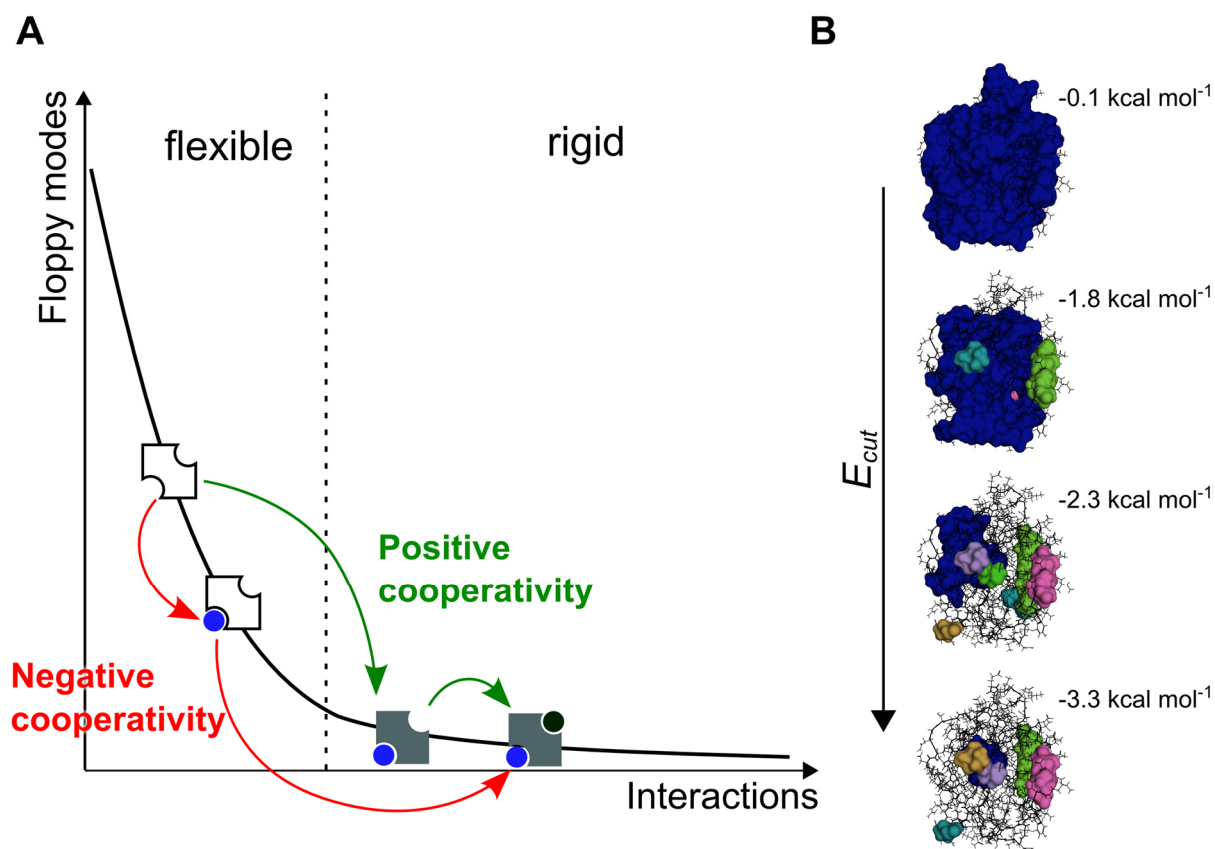


Figure 1

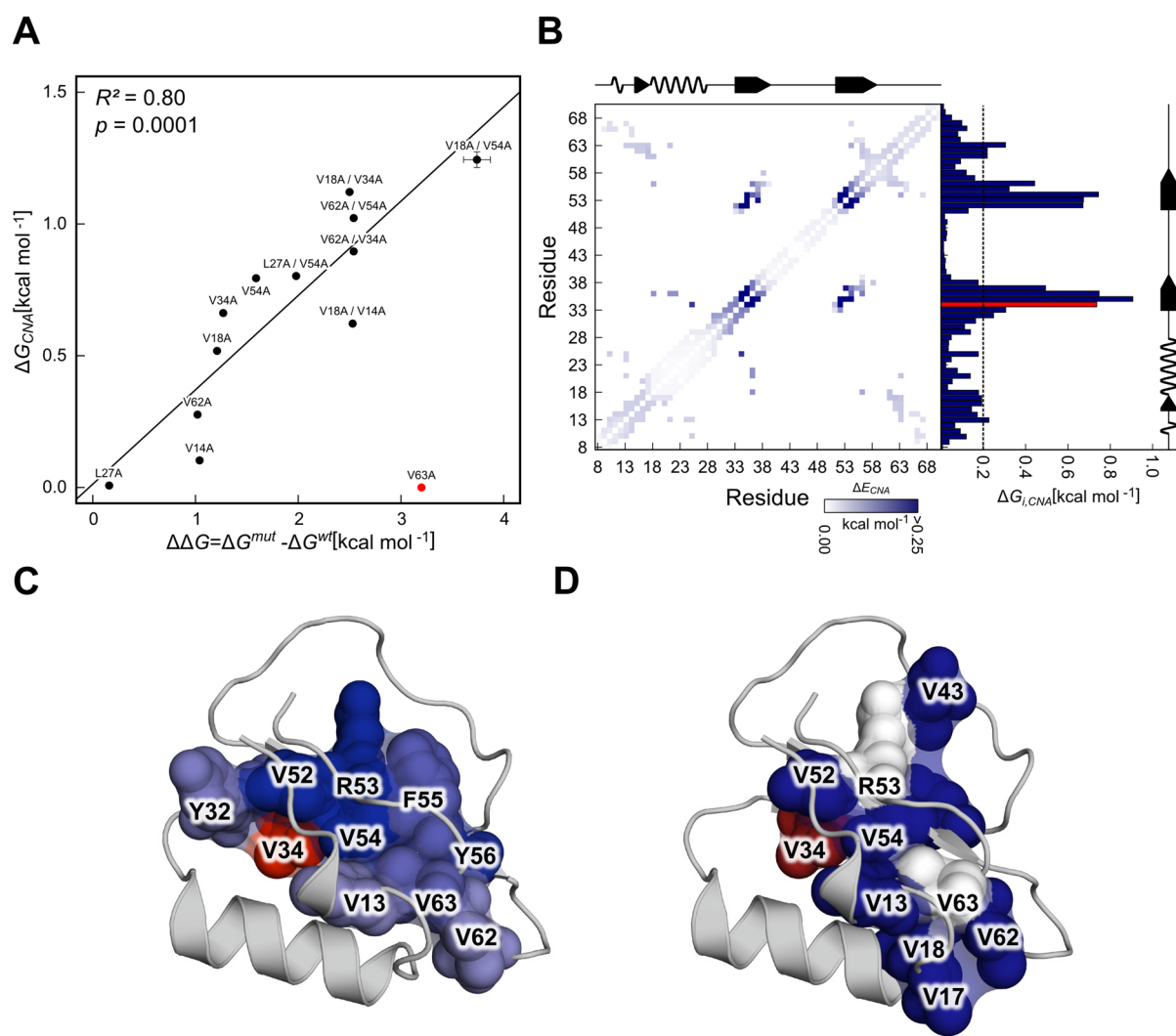


Figure 2

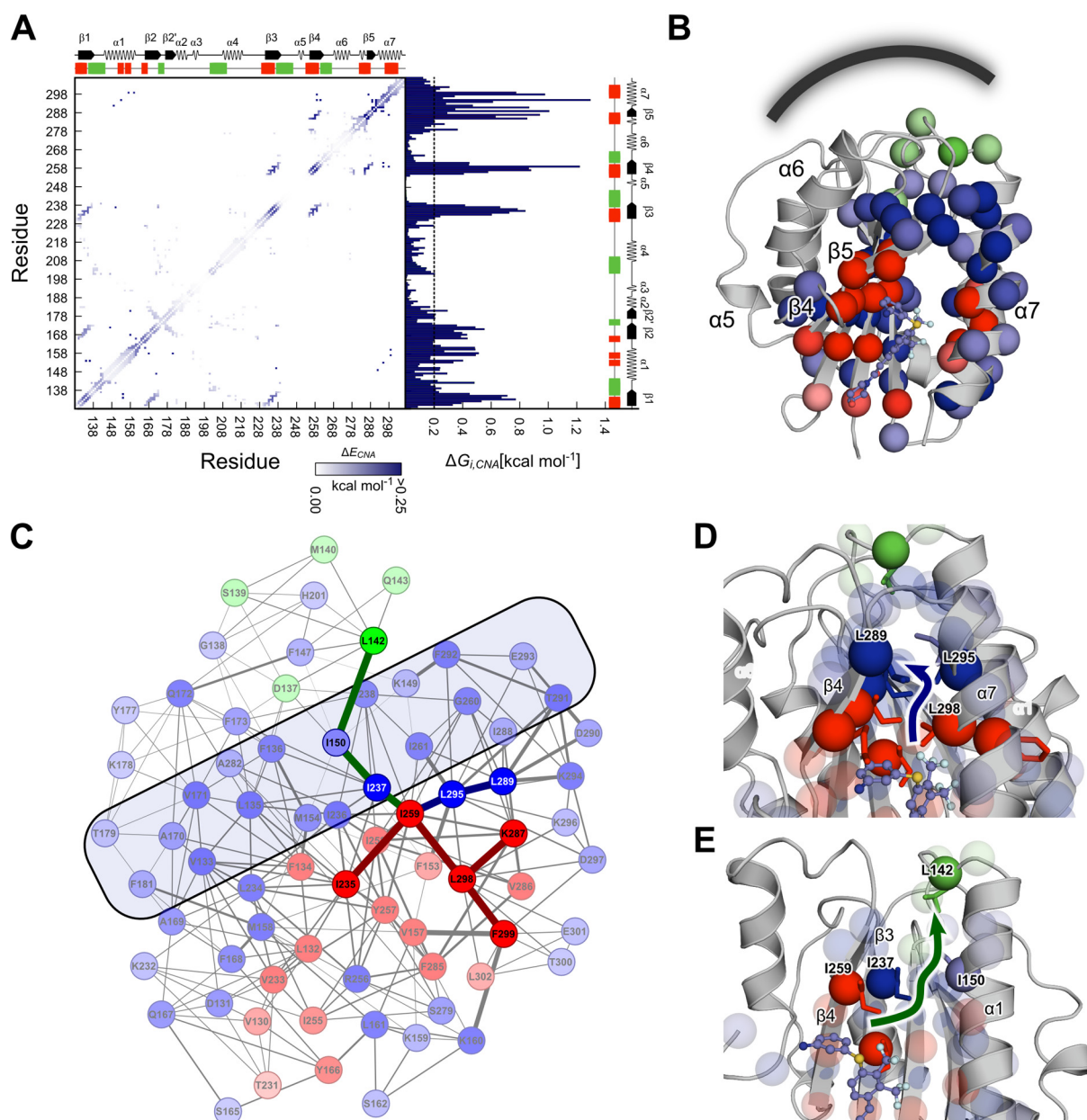


Figure 4

References

1. Motlagh HN, Wrabl JO, Li J, Hilser VJ (2014) The ensemble nature of allostery. *Nature* 508(7496):331-339.
2. Jeffrey PD, *et al.* (1995) Mechanism of CDK activation revealed by the structure of a cyclin-CDK2 complex. *Nature* 376(6538):313-320.
3. Blackmore NJ, *et al.* (2013) Three sites and you are out: ternary synergistic allostery controls aromatic amino acid biosynthesis in *Mycobacterium tuberculosis*. *J. Mol. Biol.* 425(9):1582-1592.
4. Bruning JB, *et al.* (2010) Coupling of receptor conformation and ligand orientation determine graded activity. *Nat. Chem. Biol.* 6(11):837-843.
5. Srinivasan S, *et al.* (2013) Ligand-binding dynamics rewire cellular signaling via estrogen receptor- α . *Nat. Chem. Biol.* 9(5):326-+.
6. Jura N, *et al.* (2011) Catalytic control in the EGF Receptor and Its connection to general kinase regulatory mechanisms. *Mol Cell* 42(1):9-22.
7. Freiburger LA, *et al.* (2011) Competing allosteric mechanisms modulate substrate binding in a dimeric enzyme. *Nat. Struct. Mol. Biol.* 18(3):288-U270.
8. Nussinov R, Tsai CJ, Ma B (2013) The underappreciated role of allostery in the cellular network. *Annu. Rev. Biophys.* 42:169-189.
9. van Westen GJ, Gaulton A, Overington JP (2014) Chemical, target, and bioactive properties of allosteric modulation. *PLoS Comput. Biol.* 10(4):e1003559.
10. Monod J, Wyman J, Changeux JP (1965) On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* 12:88-118.
11. Koshland DE, Jr., Nemethy G, Filmer D (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* 5(1):365-385.
12. Cui Q, Karplus M (2008) Allostery and cooperativity revisited. *Protein Sci.* 17(8):1295-1307.
13. Perutz MF (1970) Stereochemistry of cooperative effects in haemoglobin. *Nature* 228(5273):726-&.
14. Changeux JP, Edelstein SJ (2005) Allosteric mechanisms of signal transduction. *Science* 308(5727):1424-1428.
15. Freire E (1999) The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme. *Proc. Natl. Acad. Sci. U. S. A.* 96(18):10118-10122.
16. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286(5438):295-299.
17. Tang S, *et al.* (2007) Predicting allosteric communication in myosin via a pathway of conserved residues. *J. Mol. Biol.* 373(5):1361-1373.
18. Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* 10(1):59-69.
19. Feher VA, Durrant JD, Van Wart AT, Amaro RE (2014) Computational approaches to mapping allosteric pathways. *Curr. Opin. Struct. Biol.* 25:98-103.
20. Koshland DE (1959) Enzyme flexibility and enzyme action. *J Cell Comp Physiol* 54:245-258.
21. Cooper A, Dryden DT (1984) Allostery without conformational change. A plausible model. *Eur. Biophys. J.* 11(2):103-109.
22. Kern D, Zuiderweg ER (2003) The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.* 13(6):748-757.

23. Popovych N, Sun SJ, Ebricht RH, Kalodimos CG (2006) Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.* 13(9):831-838.
24. Tzeng SR, Kalodimos CG (2009) Dynamic activation of an allosteric regulatory protein. *Nature* 462(7271):368-372.
25. Sekhar A, Kay LE (2013) NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc. Natl. Acad. Sci. U. S. A.* 110(32):12867-12874.
26. Wand AJ (2013) The dark energy of proteins comes to light: conformational entropy and its role in protein function revealed by NMR relaxation. *Curr. Opin. Struct. Biol.* 23(1):75-81.
27. Manley G, Rivalta I, Loria JP (2013) Solution NMR and computational methods for understanding protein allostery. *J. Phys. Chem. B* 117(11):3063-3073.
28. Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* 254(5038):1598-1603.
29. Kumar S, Ma BY, Tsai CJ, Sinha N, Nussinov R (2000) Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Sci.* 9(1):10-19.
30. McMahon B, Frauenfelder H, Austin R, Chu K, Groves JT (2001) The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Biophys. J.* 80(1):286a-286a.
31. Onuchic JN, LutheySchulten Z, Wolynes PG (1997) Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545-600.
32. Kar G, Keskin O, Gursoy A, Nussinov R (2010) Allostery and population shift in drug discovery. *Curr. Opin. Pharmacol.* 10(6):715-722.
33. Hilser VJ, Thompson EB (2007) Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl. Acad. Sci. U. S. A.* 104(20):8311-8315.
34. Thorpe MF (1983) Continuous deformations in random networks. *J Non-Cryst Solids* 57(3):355-370.
35. Duxbury PM, Jacobs DJ, Thorpe MF, Moukarzel C (1999) Floppy modes and the free energy: Rigidity and connectivity percolation on Bethe lattices. *Physical Review E* 59(2):2084-2092.
36. Hawkins RJ, McLeish TCB (2006) Coupling of global and local vibrational modes in dynamic allostery of proteins. *Biophys. J.* 91(6):2055-2062.
37. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) Protein flexibility predictions using graph theory. *Proteins* 44(2):150-165.
38. Gohlke H, Kiel C, Case DA (2003) Insights into protein-protein binding by binding free energy calculation and free energy decomposition using a generalized born model. *J. Mol. Biol.* 330(4):891-913.
39. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. U. S. A.* 99(6):3540-3545.
40. Jacobs DJ, Thorpe MF (1995) Generic rigidity percolation: The pebble game. *Phys. Rev. Lett.* 75(22):4051-4054.
41. Katoh N, Tanigawa S (2011) A proof of the molecular conjecture. *Discrete Comput. Geom.* 45(4):647-700.
42. Hespenheide BM, Rader AJ, Thorpe MF, Kuhn LA (2002) Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graphics Modell.* 21(3):195-207.
43. Jacobs DJ, Dallakyan S, Wood GG, Heckathorne A (2003) Network rigidity at finite temperature: relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* 68(6).

44. Gohlke H, Kuhn LA, Case DA (2004) Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* 56(2):322-337.
45. Rader AJ, *et al.* (2004) Identification of core amino acids stabilizing rhodopsin. *Proc. Natl. Acad. Sci. U. S. A.* 101(19):7246-7251.
46. Livesay DR, Jacobs DJ (2006) Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins* 62(1):130-143.
47. Radestock S, Gohlke H (2008) Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.* 8(5):507-522.
48. Fulle S, Gohlke H (2009) Statics of the ribosomal exit tunnel: implications for cotranslational peptide folding, elongation regulation, and antibiotics binding. *J. Mol. Biol.* 387(2):502-517.
49. Radestock S, Gohlke H (2011) Protein rigidity and thermophilic adaptation. *Proteins* 79(4):1089-1108.
50. Dahiyat BI, Gordon DB, Mayo SL (1997) Automated design of the surface positions of protein helices. *Protein Sci.* 6(6):1333-1337.
51. Robertson AD, Murphy KP (1997) Protein structure and the energetics of protein stability. *Chem. Rev.* 97(5):1251-1267.
52. Clarkson MW, Gilmore SA, Edgell MH, Lee AL (2006) Dynamic coupling and allosteric behavior in a nonallosteric protein. *Biochemistry* 45(25):7693-7699.
53. Clarkson MW, Lee AL (2004) Long-range dynamic effects of point mutations propagate through side chains in the serine protease inhibitor eglin c. *Biochemistry* 43(39):12448-12458.
54. Wiesmann C, *et al.* (2004) Allosteric inhibition of protein tyrosine phosphatase 1B. *Nat. Struct. Mol. Biol.* 11(8):730-737.
55. Lu CF, Takagi J, Springer TA (2001) Association of the membrane proximal regions of the alpha and beta subunit cytoplasmic domains constrains an integrin in the inactive state. *J. Biol. Chem.* 276(18):14642-14648.
56. Puius YA, *et al.* (1997) Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: A paradigm for inhibitor design. *Proceedings of the National Academy of Sciences of the United States of America* 94(25):13420-13425.
57. Blacklock K, Verkhivker GM (2014) Computational modeling of allosteric regulation in the Hsp90 chaperones: A statistical ensemble analysis of protein structure networks and allosteric communications. *PLoS Comput. Biol.* 10(6).
58. Pandini A, Fornili A, Fraternali F, Kleinjung J (2012) Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J.* 26(2):868-881.
59. Daily MD, Upadhyaya TJ, Gray JJ (2008) Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins* 71(1):455-466.
60. Pannifer AD, Flint AJ, Tonks NK, Barford D (1998) Visualization of the cysteinyl-phosphate intermediate of a protein-tyrosine phosphatase by x-ray crystallography. *J. Biol. Chem.* 273(17):10454-10462.
61. Stanley P, Hogg N (1998) The I domain of integrin LFA-1 interacts with ICAM-1 domain 1 at residue Glu-34 but not Gln-73. *J. Biol. Chem.* 273(6):3358-3362.
62. Huth JR, *et al.* (2000) NMR and mutagenesis evidence for an I domain allosteric site that regulates lymphocyte function-associated antigen 1 ligand binding. *Proc. Natl. Acad. Sci. U. S. A.* 97(10):5231-5236.
63. Kallen J, *et al.* (1999) Structural basis for LFA-1 inhibition upon lovastatin binding to the CD11a I-domain. *J. Mol. Biol.* 292(1):1-9.

-
64. Huang CC, Springer TA (1995) A binding interface on the I-domain of lymphocyte function-associated antigen-1 (LFA-1) required for specific interaction with intercellular-adhesion molecule-1 (ICAM-1). *J. Biol. Chem.* 270(32):19008-19016.
 65. Kruger DM, Gohlke H (2010) DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. *Nucleic Acids Res.* 38(Web Server issue):W480-486.
 66. Edwards CP, Fisher KL, Presta LG, Bodary SC (1998) Mapping the intercellular adhesion molecule-1 and -2 binding site on the inserted domain of leukocyte function-associated antigen-1. *J. Biol. Chem.* 273(44):28937-28944.
 67. Welzenbach K, Hommel U, Weitz-Schmidt G (2002) Small molecule inhibitors induce conformational changes in the I domain and the I-like domain of lymphocyte function-associated antigen-1 - Molecular insights into integrin inhibition. *J. Biol. Chem.* 277(12):10590-10598.
 68. Weitz-Schmidt G, Welzenbach K, Dawson J, Kallen J (2004) Improved lymphocyte function-associated antigen-1 (LFA-1) inhibition by statin derivatives - Molecular basis determined by X-ray analysis and monitoring of LFA-1 conformational changes in vitro and ex vivo. *J. Biol. Chem.* 279(45):46764-46771.
 69. McClendon CL, Friedland G, Mobley DL, Amirkhani H, Jacobson MP (2009) Quantifying correlations between allosteric sites in thermodynamic ensembles. *J. Chem. Theory Comput.* 5(9):2486-2502.
 70. Mottonen JM, Jacobs DJ, Livesay DR (2010) Allosteric response is both conserved and variable across three CheY orthologs. *Biophys. J.* 99(7):2245-2254.
 71. Verma D, Jacobs DJ, Livesay DR (2012) Changes in lysozyme flexibility upon mutation are frequent, large and long-ranged. *PLoS Comput. Biol.* 8(3):e1002409.
 72. Rader AJ, Brown SM (2011) Correlating allostery with rigidity. *Mol. Biosyst.* 7(2):464-471.
 73. Bode W, Papamokos E, Musil D (1987) The high-resolution X-Ray crystal-structure of the complex formed between subtilisin carlsberg and eglin-C, an elastase Inhibitor from the Leech *Hirudo-Medicinalis* - structural-analysis, subtilisin structure and interface geometry .2. *Eur J Biochem* 166(3):673-692.
 74. D.A. Case TAD, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, P.A. Kollman (2010) AMBER 11University of California, San Francisco.).
 75. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234(3):779-815.
 76. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. *J. Comput. Chem.* 25(9):1157-1174.
 77. Cieplak P, Cornell WD, Bayly C, Kollman PA (1995) Application of the multimolecule and multiconformational resp methodology to biopolymers - charge derivation for DNA, RNA, and proteins. *J. Comput. Chem.* 16(11):1357-1377.
 78. Pfleger C, Rath PC, Klein DL, Radestock S, Gohlke H (2013) Constraint Network Analysis (CNA): a Python software package for efficiently linking biomolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model.*
 79. Bégué J-P, Bonnet-Delpon D (2008) *Bioorganic and medicinal chemistry of fluorine* (John Wiley & Sons, Hoboken, N.J) pp XVII, 365 S.
 80. Pfleger C, Radestock S, Schmidt E, Gohlke H (2013) Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.* 34(3):220-233.

81. Hagberg AA, Schult DAS, P. J. (2008) Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science conference*, eds Varoquaux G, Vaught T, & Millman J, pp 11-15.
82. Brandes U (2008) On variants of shortest-path betweenness centrality and their generic computation. *Soc Networks* 30(2):136-145.
83. Brandes U (2001) A faster algorithm for betweenness centrality. *J Math Sociol* 25(2):163-177.
84. Brandes U, Wagner D (2004) Analysis and visualization of social networks. *Graph Drawing Software, Mathematics and Visualization*, eds Jünger M & Mutzel P (Springer Berlin Heidelberg, Berlin ;Heidelberg), pp 321-340.

13.8 Publication IV – Supporting Information

Allosteric Coupling deduced from Altered Rigidity Percolation in Biomacromolecules

Pfleger, C., Minges, A.R.M., Gohlke, H.

Submitted manuscript (2014)

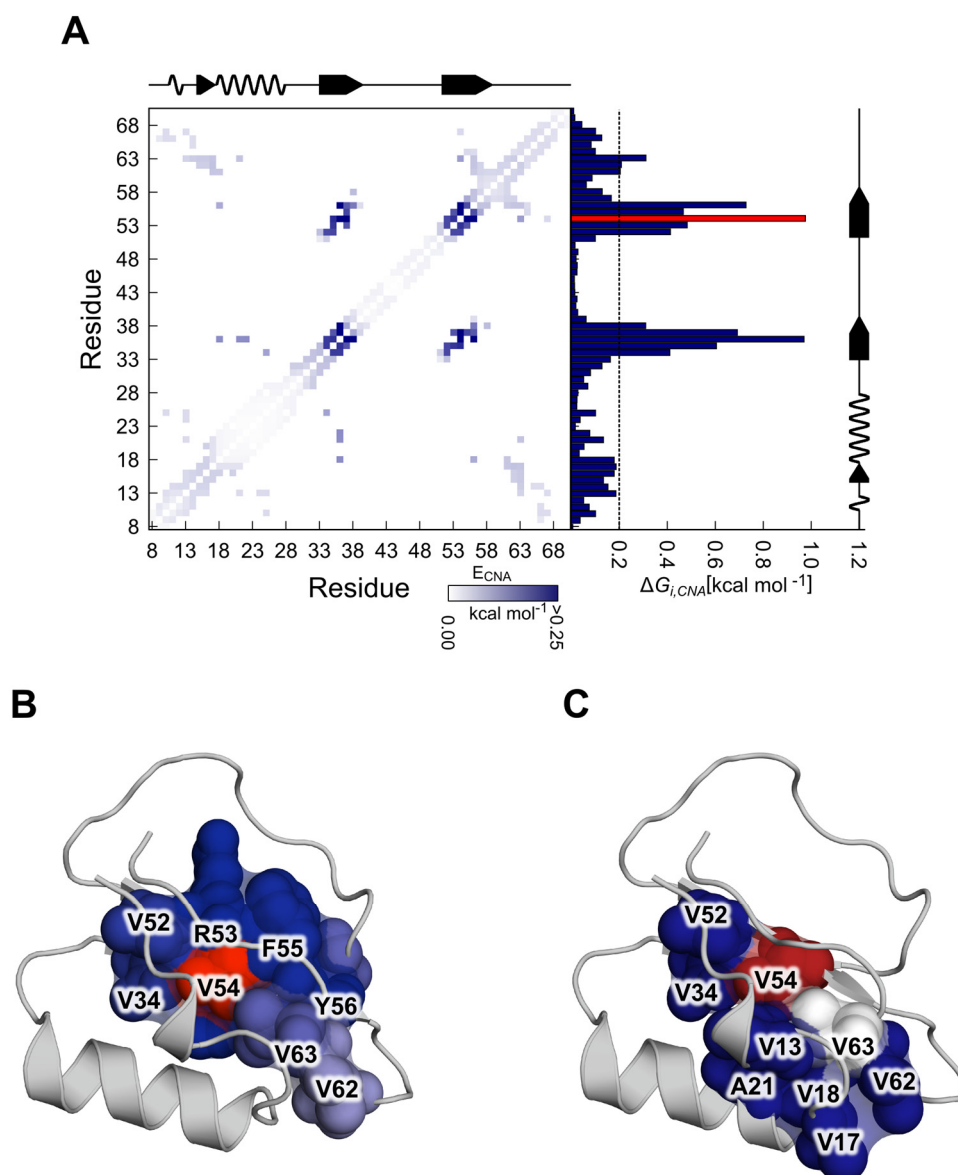


Figure S1: In silico mutational perturbation in eglin c. (A) Differences between stability maps of the ground (wild-type) and perturbed (mutant) state upon V54A mutation. The attached histogram shows the per-residue $\Delta G_{i,CNA}$ (eq. 5 in the main text). The dashed line at 0.2 kcal mol^{-1} indicates a threshold of about 20% of the maximal $\Delta G_{i,CNA}$ found. Residues with $\Delta G_{i,CNA}$ above the threshold are mapped onto the structure of eglin c (B) and are in good agreement with a continuous dynamic network of residues from NMR experiments of the same mutant (C) (40). The site of mutation (V54) is shown in red (B and C), blue colors reflect predicted $\Delta G_{i,CNA}$ values in (B), and white residues in (C) are those that lack NMR data on changes in the dynamics but are included in the network as suggested by the authors of ref. (61).

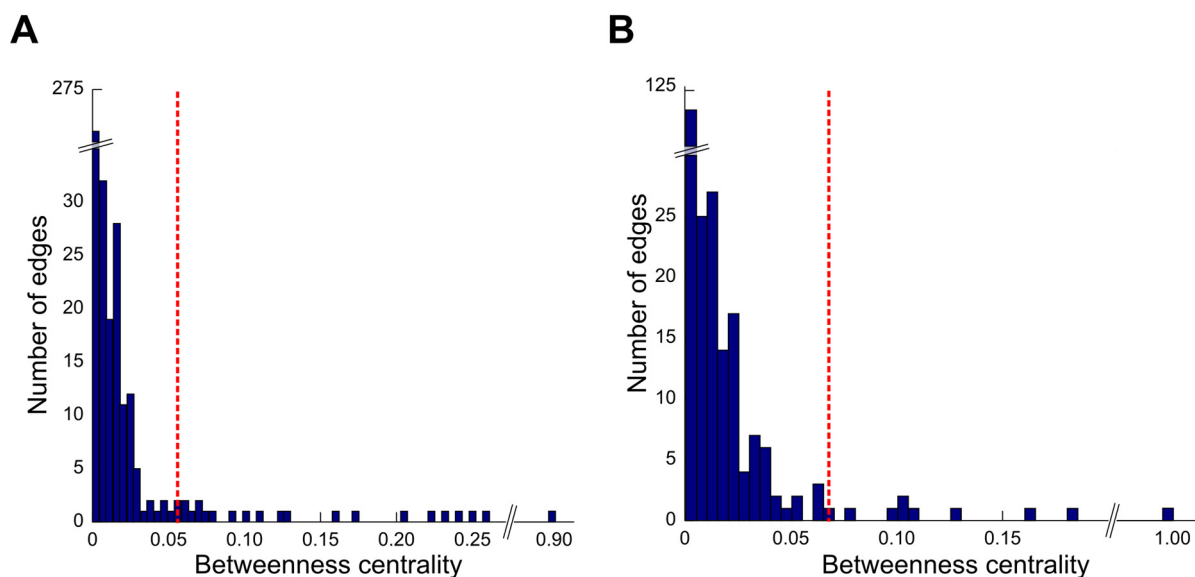


Figure S2: Distribution of the betweenness centrality for each edge of PTP1B (A) and LFA-1 (B). The vertical dashed line indicates a threshold for the betweenness centrality of edges, which are significantly larger than zero, and thus, contribute markedly to allosteric communication.

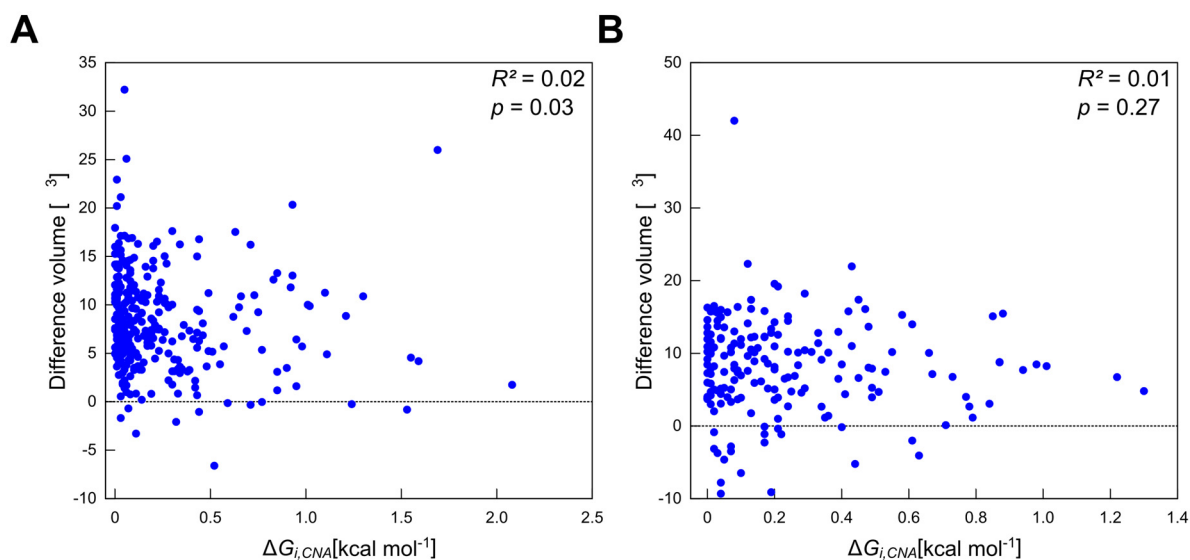


Figure S3: Correlation analysis between per-residue Voronoi volumes and mechanical perturbation free energies of PTP1B (A) and LFA-1 (B).

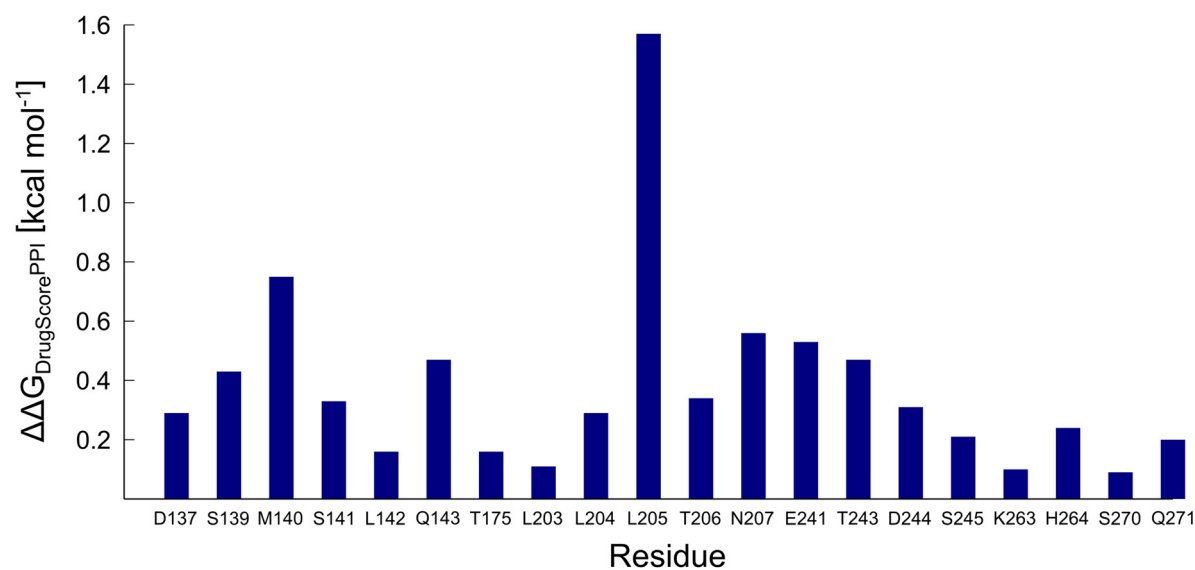


Figure S4: Free energies for residue-to-Aa mutations in the protein-protein interface of LFA-1 in complex with ICAM-1 computed from DrugScore^{PPI} (1).

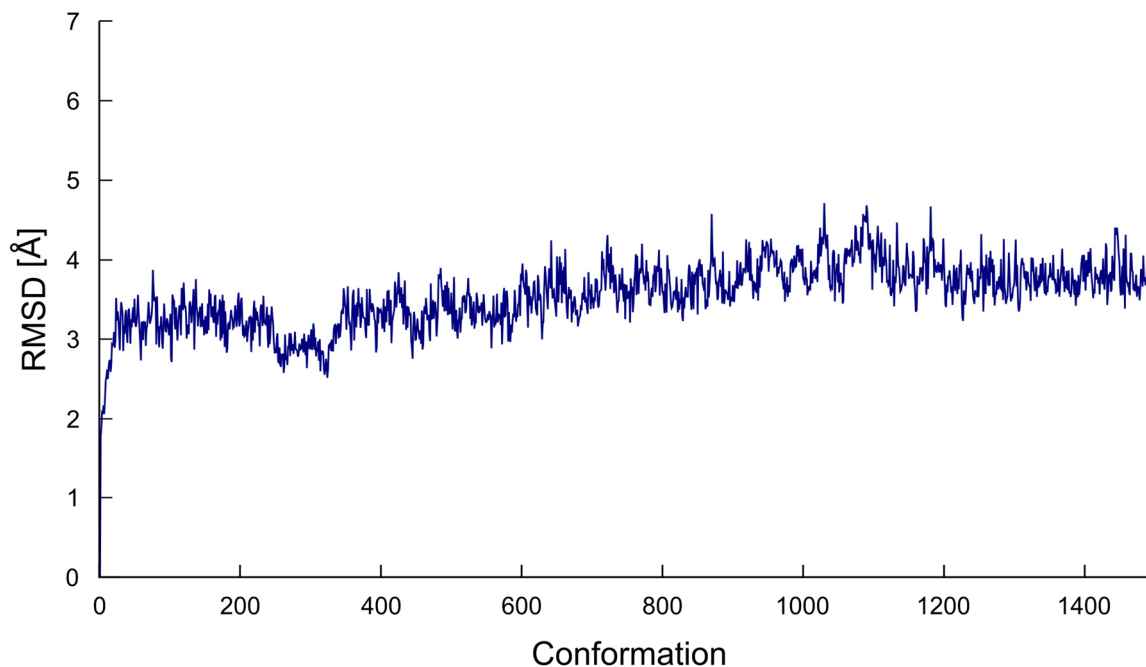


Figure S5: Heavy atom rmsd of conformations extracted from a 300ns MD trajectory of the modelled complex of LFA-1, BQM, and ICAM-1.

Supplemental Experimental Procedures

Molecular dynamics simulations

All MD simulations were carried out with the AMBER 11 package of molecular simulation programs using the GPU accelerated version of PMEMD.(2) The Cornell *et al.* force field (3) with modifications for proteins (ff99SB)(4) was employed. The structures were solvated in a truncated octahedron of TIP3P water(5) such that the distance between the boundary of the box and the closest solute atom was at least 11 Å. Periodic boundary conditions were applied using the particle mesh Ewald (PME) method (6) to treat long-range electrostatic interactions. Bond lengths involving bonds to hydrogen atoms were constrained by SHAKE.(7, 8) The time step for all MD simulations was 2 fs, and a direct-space non-bonded cutoff of 8 Å was applied. First, the solvent was minimized for 250 steps by using the steepest descent method followed by conjugate gradient minimization of 50 steps. Subsequently, the same approach was used to minimize the entire system including the protein. Afterwards, the system was heated from 100 K to 300 K using canonical ensemble (NVT) MD, and the solvent density was adjusted using isothermal-isobaric ensemble (NPT) MD. Positional restraints applied during equilibration were reduced in a stepwise manner over 50 ps followed by 50 ps of unrestrained canonical ensemble (NVT) MD at 300 K with a time constant of 2 ps for heat bath coupling. Each simulation ran for 300 ns, and coordinates were saved at 200 ps intervals to an ensemble of 1500 conformations.

Voronoi volumes of proteins

We calculated the Voronoi volumes (9, 10) for each atom in PTP1B and LFA-1 applying an algorithm implemented in C-language (11). For each residue, the individual atom volumes were summed. The method was applied on ensembles of 1500 conformations generated by MD simulations starting from the allosteric inhibitor bound state of PTP1B and LFA-1. Because Voronoi volumes cannot be computed for surface exposed residues, we retained water molecules that are within a distance cutoff of 3.5 Å to the protein. The resultant average volumes were then used in a correlation analysis between altered rigidity characteristics and per-residue volumes of PTP1B (Figure S3 A) and LFA-1 (Figure S3 B).

Supplemental References

1. Kruger DM, Gohlke H (2010) DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. *Nucleic Acids Res.* 38(Web Server issue):W480-486.
2. D.A. Case TAD, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, P.A. Kollman (2010) AMBER 11 University of California, San Francisco.).
3. Wang JM, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* 21(12):1049-1074.
4. Hornak V, *et al.* (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* 65(3):712-725.
5. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* 79(2):926-935.
6. Darden T, York D, Pedersen L (1993) Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* 98(12):10089-10092.
7. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* 23(3):327-341.
8. Miyamoto S, Kollman PA (1992) Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. *J. Comput. Chem.* 13(8):952-962.
9. Gerstein M, Tsai J, Levitt M (1995) The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J. Mol. Biol.* 249(5):955-966.
10. Harpaz Y, Gerstein M, Chothia C (1994) Volume changes on protein folding. *Structure* 2(7):641-649.
11. Richards FM (1985) Calculation of Molecular Volumes and Areas for Structures of Known Geometry. *Methods Enzymol* 115:440-464.

13.9 Publication V

Pocket-Space Maps to Identify Novel Binding-Site Conformations in Proteins

Craig, I.R., Pfleger, C., Gohlke, H., Essex, J.W., Spiegel, K.

J. Chem. Inf. Model. (2011), 51, 2666–2679

Author contribution to the publications:

My contribution to this publication was developing the PocketAnalyzer program. To this end, I re-implemented an initial Python-based version of the underlying pocket detection algorithm, validated the program and performed a parameter optimization for pocket detection. This results in a contribution of **30%** to this publication.


Pocket-Space Maps To Identify Novel Binding-Site Conformations in Proteins

Ian R. Craig,^{*,ll†} Christopher Pfleger,[‡] Holger Gohlke,[‡] Jonathan W. Essex,[§] and Katrin Spiegel[†]

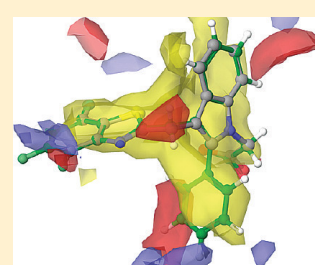
[†]Novartis Institutes for Biomedical Research, Wimblehurst Road, Horsham, West Sussex, RH12 5AB, U.K.

[‡]Mathematisch-Naturwissenschaftliche Fakultät, Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany

[§]School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K.

 Supporting Information

ABSTRACT: The identification of novel binding-site conformations can greatly assist the progress of structure-based ligand design projects. Diverse pocket shapes drive medicinal chemistry to explore a broader chemical space and thus present additional opportunities to overcome key drug discovery issues such as potency, selectivity, toxicity, and pharmacokinetics. We report a new automated approach to diverse pocket selection, PocketAnalyzer^{PCA}, which applies principal component analysis and clustering to the output of a grid-based pocket detection algorithm. Since the approach works directly with pocket shape descriptors, it is free from some of the problems hampering methods that are based on proxy shape descriptors, e.g. a set of atomic positional coordinates. The approach is technically straightforward and allows simultaneous analysis of mutants, isoforms, and protein structures derived from multiple sources with different residue numbering schemes. The PocketAnalyzer^{PCA} approach is illustrated by the compilation of diverse sets of pocket shapes for aldose reductase and viral neuraminidase. In both cases this allows identification of novel computationally derived binding-site conformations that are yet to be observed crystallographically. Indeed, known inhibitors capable of exploiting these novel binding-site conformations are subsequently identified, thereby demonstrating the utility of PocketAnalyzer^{PCA} for rationalizing and improving the understanding of the molecular basis of protein–ligand interaction and bioactivity. A Python program implementing the PocketAnalyzer^{PCA} approach is available for download under an open-source license (<http://sourceforge.net/projects/papca/> or <http://cpclab.uni-duesseldorf.de/downloads>).



INTRODUCTION

Advances in experimental and computational methodology, and improvements in hardware and instrumentation, have rapidly increased the rate at which molecular structures of proteins can be generated.¹ Multiple experimentally determined structures are often now available for a particular protein of interest, and it is also more frequently feasible to derive a large set of protein conformations using computational methods. This is driving structure-based drug design to move beyond a “one target, one structure” perspective to account for and embrace the flexibility of proteins.^{2–4} Since medicinal chemists can use protein structures to guide their ligand design toward the formation of specific protein–ligand interactions, a diverse set of protein conformations presents an opportunity to explore chemical space more widely. This freedom increases the chances that key drug discovery issues such as potency, selectivity, ADME (absorption, distribution, metabolism, and excretion), and toxicity can be overcome. In the past, computational methods have confirmed the existence of different pocket shapes a posteriori, such as the additional subpocket in the HIV-integrase active site, exploited by the first marketed HIV-integrase inhibitor.^{5,6} The goal of structure based computational chemistry is to identify druggable pocket shapes beforehand and guide chemistry to exploit alternative pocket shapes.

The analysis and interpretation of large volumes of protein structural information can be a lengthy process if visual inspection is required in order to detect and confirm diverse and potentially novel pocket shapes. Alternatively, computational approaches to diverse pocket selection have been devised that typically analyze Molecular Dynamics (MD) trajectories and often involve some form of clustering to bundle similar structures together, followed by selection of just one representative structure from each group.⁷ Some form of dimensionality reduction is also common, Principal Component Analysis (PCA)⁸ being a prominent example. These techniques provide a characterization of the dominant deformation modes of the binding pocket and map out the pocket conformational space.

Of critical importance is the set of coordinates to which techniques such as clustering and PCA are applied. The set of Cartesian coordinates for all active site C_α atoms can be a poor proxy for local binding pocket conformation because protein structures with similar C_α conformations may still have very different side-chain conformations and will thus incorporate very different binding pocket shapes. This can lead to dissimilar pocket conformations being clustered or mapped together, and to novel pocket shapes being missed. Thus, methods for select-

Received: April 13, 2011

Published: September 12, 2011

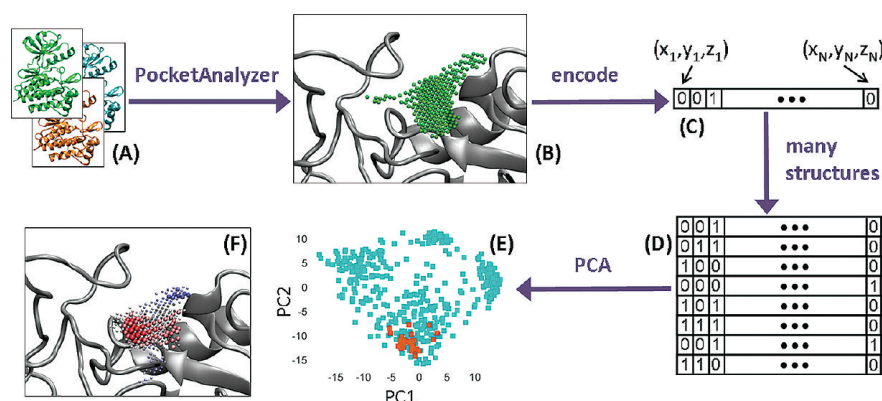


Figure 1. A graphical summary of the PocketAnalyzer^{PCA} approach. The PocketAnalyzer pocket detection algorithm is applied to each protein structure in the set (A). For each structure, the resulting pocket shape (B) is encoded as a row vector of integers (C). Merging the row vectors from each protein structure produces the pocket shape matrix (D). Projecting the row vectors onto principal components derived from the pocket shape matrix generates a map of the pocket conformational space (E). The principal components themselves describe the dominant changes in pocket shape within the set of protein conformations (F).

ing a diverse set of pocket conformations should instead use descriptors that account for all the atoms lining the pocket. An obvious choice is the set of Cartesian positions for all atoms bordering the binding site.^{7,9} However, collating this set of descriptors can be technically awkward when comparing protein structures with different residue numberings and/or atom orderings. It also precludes the inclusion of protein structures with mutations of one or more binding pocket residues or structures of more than one isoform. Furthermore, there can be considerable ambiguity in defining the set of “binding site atoms”, particularly for flexible proteins.

In this paper, we address the problem of diverse pocket selection from an alternative perspective by working directly with pocket shape descriptors rather than a set of proxy coordinates. In particular, we develop a procedure that reduces a large collection of structures of the same protein to a subset that retains a number of substantially distinct binding pocket conformations. The approach is based on applying PCA and clustering to the output of a grid-based pocket detection algorithm. The PCA yields (a) principal component (PC) eigenvectors, which reveal the dominant deformation modes of the pocket, and (b) PC projections (“scores”), which provide characterization and visualization of the pocket conformational distribution (PCD). Clustering of the PCD then results in an all-atom approach to diverse pocket selection that is free from the problems hampering methods based on atomic coordinates. As our method builds upon the PocketAnalyzer pocket detection code developed in the Gohlke group, we call it PocketAnalyzer^{PCA}. Although itself unpublished, the original PocketAnalyzer approach implements a variant of the LIGSITE algorithm.^{10,11} Minor differences between PocketAnalyzer and LIGSITE are described below. The approach is technically straightforward and allows simultaneous analysis of mutants, isoforms, and homologous proteins. Furthermore, although we focus on the identification of novel pocket conformations from MD simulations, our procedure is applicable to any source of atomistic protein structural information, and all combinations thereof.

Here, we apply PocketAnalyzer^{PCA} to two proteins that exhibit moderate but significant active site flexibility, namely aldose reductase and neuraminidase. Both proteins have been extensively studied, resulting in a well-characterized set of binding pocket

conformations with which to compare the output of our approach to diverse pocket selection. Since the PocketAnalyzer^{PCA} technique is potentially applicable to problems beyond diverse pocket selection, the paper closes by highlighting some directions for future work.

MATERIALS AND METHODS

PocketAnalyzer^{PCA}. *Outline.* We first give a short overview of the PocketAnalyzer^{PCA} approach (Figure 1), before dealing with the methodological details in more depth. In the first step, a grid-based pocket detection algorithm is applied to an ensemble of protein structures. The identified pockets from each single protein structure are represented as a row vector of integers, each encoding the inclusion (“1”) or exclusion (“0”) of a particular grid point. The row vector describes the pocket shape corresponding to a specific conformation of the protein. Since the same set of grid-points is used to analyze each protein structure, the row vectors can be merged to produce a pocket shape matrix. Each column of this pocket shape matrix then represents the varying inclusion/exclusion of a particular grid-point in the pockets of an ensemble of protein structures.

The pocket shape matrix is subjected to Principal Component Analysis (PCA).¹² This results in (a) a set of PC eigenvectors, (b) a set of eigenvalues that correspond to the variance along each PC, and (c) the projections (or “scores”) of each protein structure along each PC. The high-variance PCs describe the dominant changes in pocket conformation within the given set of protein structures. The scores characterize the distribution of pocket shapes along the PCs, providing a map of the pocket conformational space. Analysis and comparison of these pocket conformational distributions (PCDs) may provide a useful perspective on a variety of questions related to molecular recognition and protein dynamics. Of particular relevance to medicinal chemistry efforts is the rephrasing of the diverse pocket selection problem in terms of finding a small number of protein structures whose pockets nevertheless provide coverage of the significant regions of the PCD. In this work a clustering of the PCD is used to achieve this aim.

Pocket Detection. The pocket detection algorithm implemented within the PocketAnalyzer code is a variant of the LIGSITE

algorithm¹⁰ very similar to that described by Stahl and co-workers.¹¹ As such, a grid map is defined across the protein where each grid point must meet a number of criteria in order to be included in the pocket. First, the grid point should not be within the van der Waals radius of any protein atom. Second, it must be sufficiently enclosed within the protein structure. This degree of buriedness is assessed by scanning away from the grid point along fourteen vectors (positive/negative in the x , y , and z axes and the four cube diagonals) and counting the number of directions in which protein atoms are encountered within a distance of 10 Å. A grid point is excluded if this count is less than the user-defined threshold *dob* (degree of buriedness). Third, the grid point must be surrounded by a certain number of other well-buried neighboring grid points, determined by the parameter *mn**b* (minimal number of neighbors). Fourth, after clustering of the grid points meeting the preceding criteria, the point must belong to a cluster of size greater than the parameter *mcs* (minimal cluster size). The default values for the parameters are *dob* = 11, *mn**b* = 15, and *mcs* = 50 at a grid-spacing of 0.8 Å.

The parameters can be used to optimize the pocket detection with respect to the protein system under investigation. Solvent-exposed pockets require a lower degree of buriedness threshold (*dob*) to adjust to the generally lower enclosure of grid-points in an open binding site. Smaller pockets might only be identified by setting a lower minimal cluster size threshold (*mcs*). The number of neighbors (*mn**b*) affects the shape of the identified pockets. Increasing this value leads to pockets with a more globular shape, whereas lower values result in more disperse and filamentous pocket shapes, e.g. (sub)pockets that are connected by a tunnel. In this work the aldose reductase analysis used the default parameter settings, whereas the values have been modified for the neuraminidase analysis. In particular, the *dob* threshold parameter was reduced from the default value of 11 to 9 to account for the large and solvent-exposed binding pocket. As this then leads to disperse pocket shapes in the case of neuraminidase, the *mn**b* value has been increased from the default of 15 to 18. Other parameters were kept at their default values. As for the grid-spacing, changing this value within the boundaries of 0.5 to 1.5 Å does not grossly affect the identification of pockets. We thus chose a grid spacing of 0.8 Å, equivalent to approximately one-half of a C–C bond distance, in order to detect pockets reliably with moderate computational effort.

Note that two differences between the LIGSITE/Stahl algorithm and the current one are that PocketAnalyzer: (a) does not ignore hydrogen atoms and (b) does not add a tolerance of 0.8 Å to the van der Waals radii. Regarding the former, in this work the Protein Preparation Wizard in *Maestro*¹³ was used to add hydrogens to the crystal structures. The same utility was also used to optimize the resulting hydrogen-bonding networks.

Structural Alignment. A crucial aspect of the PocketAnalyzer^{PCA} approach is that all the protein structures are analyzed on the same set of grid points (currently initialized using the first structure in the list). For this to be meaningful, the protein structures must be aligned before submission to the pocket detection algorithm. Here, protein structures were aligned to minimize the rmsd between the Cartesian coordinates of the C_α atoms of a specific set of active site residues (defined below). As the structural alignment depends on the set of atoms chosen,¹⁴ the question arises as to how sensitive the PocketAnalyzer^{PCA} approach is to a change in alignment. Notably, changing from an all C_α alignment to an active site C_α alignment caused little change in the resulting PCs and PCDs for large sets of protein structures (see

Supporting Information Figure S1): although the PC scores of a few individual structures sometimes changed more significantly, the spectrum of pocket shapes returned by the clustering process generally remained the same. This demonstrates that the PocketAnalyzer^{PCA} approach is empirically rather insensitive to a switch between two sensible alignments.

For ALR, C_α atoms of the following manually selected active site residues were used for the alignment: 1ADS: W20, H46, V47, Y48, K77, W79, C80, H110, W111, T113, G114, F121, F122, L124, V130, N160, Q183, Y209, W219, V297, C298, L300, L301, S302, C303, H306, and Y309. For neuraminidase, the C_α atoms of all residues were used for the alignment.

Principal Component Analysis. PCA is used to reduce the dimensionality of the pocket shape matrix. Since all matrix elements are of the same type (i.e., binary integers representing grid-point inclusion/exclusion) the covariance matrix is directly diagonalized rather than first normalizing to the correlation matrix. Importantly, the PCA does not have to use all the grid-point inclusion/exclusion row vectors. Some can be withheld and then projected onto the resulting principal components. This provides a mechanism to investigate how well one set of structures (e.g., derived crystallographically) overlap or cover the PCD derived from another set (e.g., derived from a MD simulation).

To minimize the number of descriptors used by the PCA, only grid points that are included in the pocket of interest for at least one of the protein conformations are considered. For typical binding pockets only a few hundred grid-points meet this criterion, and the resulting PCA takes less than a minute for a few hundred protein structures. For both aldose reductase and neuraminidase the 'pocket of interest' in this work is the active (i.e., catalytic) site.

Clustering. Clustering is used to derive a small subset of protein structures whose pockets nevertheless provide coverage of the significant regions of the pocket conformational distribution. In particular, the CLARA algorithm¹⁵ implemented in the R package *cluster*¹⁶ is applied to group the protein structures according to their scores along the PCs. Retaining only the representative protein structure from each cluster leaves a subset of protein conformations corresponding to a diverse selection of pocket shapes. In CLARA the cluster representative is the medoid, i.e. the member of the cluster with minimal average Euclidean distance to the other members of the cluster. The PocketAnalyzer^{PCA} program performs this clustering via an external call to R.

The PocketAnalyzer^{PCA} program allows clustering of pocket shapes either on the original grid-point inclusion/exclusion descriptors or on the scores (i.e., projections) along a user-defined number of the PCs. Data sets with a steep scree plot, which shows the variance along the PC versus the PC index, could be meaningfully clustered using only a few PCs. In contrast, using all of the PCs is equivalent to clustering on the original grid-point variables. In between these extremes, based on their relatively flat scree plots, the examples presented here employ a CLARA clustering of the first 50 PC scores. Since this accounts for 83% of the variance in the aldose reductase data set, and 63% for neuraminidase, this corresponds to clustering on all pocket shape changes except small and/or infrequently observed ones.

Druggability Testing. SiteMap is used to assess the druggability of the cluster representatives.¹⁷ The DScore druggability metric computed by SiteMap is a linear function of three pocket properties: size, enclosure (akin to an average degree of buriedness), and hydrophilicity. These descriptors are calculated for each pocket using a grid-based approach similar to the algorithm adopted by

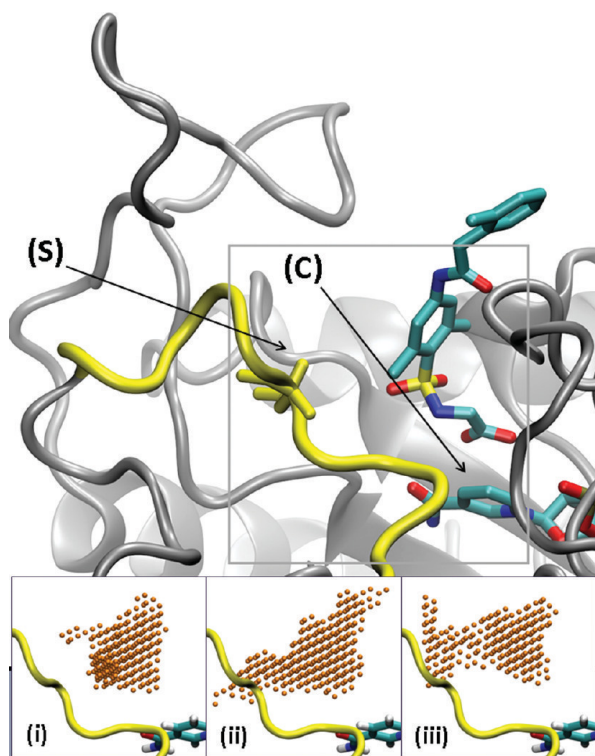


Figure 2. The large figure shows the active site of aldose reductase in an *apo*-like conformation (PDB 1EL3). The protein backbone is represented in gray ribbon with a more flexible section highlighted in yellow, as is the torsionally labile Leu300 residue. The selectivity pocket and the catalytic site are labeled (S) and (C) respectively. The ligand (IDD384) is shown right-of-center, and part of the NADP⁺ cofactor is visible in the lower right-hand corner. For reference, the gray box indicates the area displayed in subsequent figures. The smaller figures below illustrate pocket shapes characterized by PocketAnalyzer for three representative ALR crystal structures: (i) 1ADS (*apo* conformation), (ii) 1PWL (*zenarestat*), and (iii) 2FZD (*tolrestat*).

PocketAnalyzer. More details on SiteMap and DScore can be found in the literature.¹⁷ Many other methods for characterizing the physicochemical properties and druggability of protein pockets are available.^{18–21}

Data Sets. *Aldose Reductase.* To provide a reference for active site conformations generated by MD simulations, a set of eight crystallographic ALR protein structures was assembled with the help of the CavBase module of Relibase+.^{22,23} The ALR structures show two main active site conformations: an *apo* conformation in which the selectivity pocket illustrated in Figure 2 is closed (PDB codes 1ADS, 1EL3) and a group of conformations in which the selectivity pocket is open. Named after the ligands bound to them,²⁴ the latter are *tolrestat* (2FZD and 2FZB), *zenarestat* (1IEI and 1PWL), and *IDD594* (1US0 and 2R24). Note that (a) all ALR residue numberings in this work are those of the 1ADS structure, (b) the ligand in the 1PWL structure is actually *minalrestat* not *zenarestat*, and (c) the *apo* conformation was termed ‘*holo*’ in some previous work.²⁴

It is natural to speculate whether any known ALR inhibitors without a crystallographically confirmed binding mode might in fact occupy the novel pocket conformations identified by PocketAnalyzer^{PCA} (see Results section). To address this issue,

inhibitors with structural similarity to a ligand with an available crystallographic binding mode were aligned to that template using the substructure-based ligand alignment method available in ICM.²⁵ Five crystallographic binding modes were initially chosen as templates: *tolrestat* (from PDB code 2FZB), *zenarestat* (1IEI), *IDD384* (1EL3), *minalrestat* (1PWL), and *IDD594* (1US0). The first three bind the catalytic site of aldose reductase via an α -aminoacid substructure, *minalrestat* binds via a cyclic imide substructure, and *IDD594* via an α -hydroxyacid substructure. A set of 395 ALR inhibitors with $IC_{50} < 1 \mu M$ was downloaded from the bindingDB,²⁶ of which 66 matched the substructure of at least one of the templates (see Supporting Information, List S9). Since at least 36 of the 66 substructure-matched bindingDB compounds were obvious derivatives of *lidorestat*,²⁷ its crystallographic binding mode from PDB structure 1Z3N was added as a sixth alignment template.

Neuraminidase. The set of crystallographic reference structures of neuraminidase N1 consists of an *apo* structure (2HTY), the oseltamivir bound structure in the closed 150-loop conformation (2HU4), the oseltamivir bound structure in the open 150-loop conformation (2HU0), the zanamivir bound *holo* structure (3B7E), a second *apo* structure (3BEQ), and the oseltamivir resistant H274Y mutant in complex with zanamivir and oseltamivir (3CKZ, 3CL0). To this set of neuraminidase N1 protein conformations, two structures of neuraminidase N8 are added in which the protein is bound to oseltamivir with the 150 loop in the open and closed positions (2HT7, 2HT8). All neuraminidase residue numberings in this work correspond to those of the 2HTY structure.

As for ALR, structural alignments to crystallographic binding-mode templates were used to predict whether known N1 inhibitors might bind to any of the novel pocket shapes identified by the PocketAnalyzer^{PCA} protocol. A set of 64 N1 inhibitors with $IC_{50} < 1 \mu M$ were downloaded from the bindingDB.²⁶ All but one of these matched the substructure of oseltamivir. In consequence, the only crystallographic binding-mode used was that of oseltamivir from PDB entry 2HU4.

MD Simulations. To generate computationally derived ensembles of ALR and neuraminidase structures, MD simulations were performed using the Amber 9 package.^{28,29}

Protocol for Aldose Reductase. Three trajectories were initiated from the 1US0 crystal structure of ALR in complex with the carboxylic-acid type inhibitor *IDD594*. This initial structure was chosen due to its high resolution (0.66 Å) and to be consistent with previous simulations.²⁴ All crystallographic waters, citrate ions, and the ligand were deleted. The NADP⁺ cofactor was retained. Side-chain tautomerization and protonation states were assigned with the help of the Protein Preparation Wizard in *Maestro*.¹³ Hydrogens were deleted in *Maestro* before being added back in AMBER-compatible format using the *tleap* module. The Amber ff03 force-field³⁰ was used for the protein, and the Ryde parametrization was used for the cofactor.³¹ The *tleap* module was also used to solvate the protein in a rectangular box of TIP3P water molecules, with a minimal distance between the solute and the boundary of the box of 11 Å.³² Two potassium ions were then added to ensure overall charge neutrality. This resulted in a simulation cell with dimensions 81 × 70 × 81 Å³ and a total of 37299 atoms.

One equilibration run was used to prepare all three trajectories, beginning with a two-step minimization approach. First, the protein and cofactor were held fixed while the solvent was minimized using 500 steps of the steepest descent method

followed by conjugate gradient minimization. In a second round of minimization, using the same approach, the restraints on the protein atoms were relaxed. The entire system was then heated from 5 to 300 K over 200 ps at constant volume. In this and all following simulations, a Langevin thermostat³³ with a collision frequency of 5 ps^{-1} was employed to regulate the temperature, and a time step of 1 fs was used. The SHAKE algorithm³⁴ was used throughout to constrain the lengths of bonds involving hydrogen, and the particle mesh Ewald method³⁵ was employed to treat long-range electrostatic interactions. The nonbonded cutoff was set at 12 Å.

The equilibration was completed with a second stage of dynamics, this time at constant pressure for 800 ps at a fixed temperature of 300 K. The pressure was regulated to a reference of 1 bar using AMBER's weak-coupling barostat³⁶ with a relaxation time of 4 ps.²⁹ Three production simulations were then launched from the equilibrated structure. These trajectories differed only in the sequence of random numbers used by the thermostat. Each of the simulations ran for 20 ns, and coordinates were saved at 500 ps intervals to give three sequences of 40 computationally derived protein conformations.

Protocol for Neuraminidase. To generate a computationally derived ensemble of neuraminidase conformations, MD simulations were initiated from the 2.5 Å resolution *apo* structure (PDB 2HTY) and the 2.4 Å resolution *holo* structure in complex with oseltamivir (PDB 2HU4). Topologies compatible with the Amber 9 program were prepared as outlined for aldose reductase. The calcium ion was retained as it is structurally important.³⁷ Force field parameters for oseltamivir have been assigned using the generalized amber force field (GAFF).³⁸ The atom types, charges, and prep-files for oseltamivir are included in the Supporting Information (Section S2). To achieve charge neutrality, three potassium ions were added. The systems were solvated with TIP3P water molecules as described above for ALR. This resulted in a final box size of $80 \times 81 \times 81 \text{ Å}^3$ and a total of 43201 atoms for the *apo* simulation and a box of size $79 \times 82 \times 82 \text{ Å}^3$ and a total of 43236 atoms for the two *holo* simulations.

The neuraminidase structure features eight disulfide bridges, which are expected to stabilize the protein framework. The minimization and equilibration protocol was shortened with respect to aldose reductase: in a first step, the entire system was minimized without restraints by 5000 steps of steepest descent. In a second step, the temperature was gradually increased during 60 ps to 300 K using the Andersen temperature coupling scheme, randomizing velocities every 1000 steps, and a time step of 2 fs. Harmonic restraints of $5 \text{ kcal mol}^{-1} \text{ Å}^{-2}$ were applied to the protein backbone. The density of the system was allowed to adjust at the same time, again using AMBER's weak coupling barostat.³⁶ Finally, the harmonic restraints were released, and the system was allowed to relax during another 100 ps before starting the production run. Throughout all MD simulations, the SHAKE algorithm³⁴ and the particle mesh Ewald method³⁵ were used. The nonbonded cutoff was set at 12 Å. The two *holo* simulations differed only in the sequence of random numbers used by the thermostat. Each of the simulations ran for 20 ns, and coordinates were saved at 10 ps intervals to give three sequences of 200 computationally derived protein conformations. A greater number of conformations was extracted from the neuraminidase simulations than from the ALR trajectories to compensate for the higher mobility of the protein in the former.

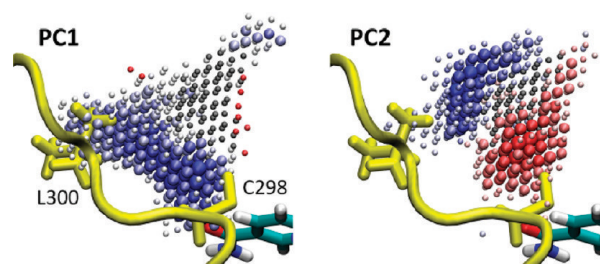


Figure 3. The highest-variance principal components of the aldose reductase data set: PC1 (left) and PC2 (right). Each is a linear combination of the original grid-point descriptors. The size of the sphere representing each grid-point reflects the absolute magnitude of its coefficient in the linear combination: the largest spheres indicate the largest coefficients, i.e. the most important grid-points for that particular PC. Any two grid-points with the same color are positively correlated: according to that particular PC, they tend to be in the pocket shape simultaneously. Any two grid points with different colors are negatively correlated: they do not tend to be in the pocket shape simultaneously. The dark gray spheres indicate grid-points which are included in more than 80% of the pocket shapes. Also shown for reference are the side-chains of residues C298 and L300.

RESULTS

PocketAnalyzer^{PCA} Applied to Aldose Reductase. PocketAnalyzer^{PCA} was first applied to aldose reductase (ALR). As a target for the prevention of several diabetic complications, ALR has been the subject of numerous drug design studies.³⁹ Crystallographic and computational investigations have already identified a set of distinct ALR active site conformations.^{24,40–42} This makes ALR a useful benchmark system for diverse pocket selection, with the crystal structures revealing an active site of moderate flexibility driven more by changes in side-chain conformation rather than backbone rearrangement.

Some important components of the ALR active site are illustrated in Figure 2. Active site conformations identified by Klebe and co-workers^{24,40–42} differ principally in the region of the “selectivity” pocket. In contrast, the region around the catalytic residues on the opposite side of the active site is markedly rigid.^{24,43} The protein structure with PDB code 1ADS exemplifies the *apo* conformation in which the selectivity pocket is closed. The result is an active site of low volume mainly comprised of the residues surrounding the catalytic site and the NADP⁺ cofactor. A side-chain rotation of residue L300, and an associated adjustment of certain backbone torsional angles, opens the selectivity pocket in the *tolrestat*-, *zenarestat*-, and *IDD594*-bound conformations.²⁴ However, differences in the positioning of other side-chains, particularly those of residues W111, F115, F122, and C303, result in a distinct selectivity pocket conformation in each case.

Diverse Pocket Selection. The PocketAnalyzer^{PCA} diverse pocket selection protocol was used to analyze and explore the active site conformations visited in the three MD simulations of aldose reductase (see Materials and Methods). The principal aim was to compare the MD-derived binding-site conformations with those observed in the set of available ALR crystal structures and to thus identify novel computationally generated pocket shapes. As a first step, the protein conformations contained in the crystallographic and MD data sets were structurally aligned and processed with PocketAnalyzer. A Principal Component Analysis was then applied to the grid-point inclusion/exclusion row

vectors corresponding to the 120 computationally derived structures (3 trajectories, 40 frames from each). In contrast, the grid-point row vectors corresponding to the 8 crystallographic pocket conformations were not submitted to the PCA but were subsequently projected onto the PCs resulting from the analysis of the MD-derived protein structures. However, including the crystallographic pocket conformations in the PCA makes little difference to the resulting PCs and the PCD (see Supporting Information Figure S3).

The two highest-variance PCs (out of 120 PCs in total) are shown in Figure 3. The first, labeled PC1, describes an expansion of the pocket in the direction of the flexible L300 loop, simultaneously opening both the selectivity pocket and another sub-pocket near residue C298. In contrast, the positive and negative lobes of PC2 show alternate, i.e. mutually exclusive, expansion into different areas. According to PC2, pocket conformations which include the grid-points toward the selectivity pocket and L300 (blue) tend not to include certain grid-points around the C298 subpocket (red). Since these high-variance principal components produced by PocketAnalyzer^{PCA} characterize the dominant changes in pocket shape within the structural data set, they provide a novel and easily visualized perspective on protein flexibility that is directly relevant to ligand binding and ligand design. It is worth noting that, although PC1 and PC2 are the highest

variance principal components, in this case they are cumulatively responsible for only 21% of the data set's total variance. In general it may be necessary to consider more than two PCs to obtain a full picture of the possible pocket deformations.

PocketAnalyzer^{PCA} also generates a score for each pocket conformation along each PC by projecting the row vector of a pocket conformation onto the PC. These scores collectively characterize the pocket conformational distribution (PCD) and provide a map of the regions of pocket space explored in the current data set. Figure 4 exemplifies this with a plot of the scores along PC1 against those along PC2. Such plots enable comparisons to be made between the PCDs of different parts of the structural data set. For example, it is clear from Figure 4 that the MD simulations approach some of the crystallographic pocket conformations more closely than others.

A second observation from Figure 4 is that the three MD simulations explore different regions of pocket conformational space. In particular, the second trajectory visits an extensive region of pocket space that is neither covered by the other two trajectories nor the crystallographic structures. This comparison of their respective PCDs indicates that the conformational sampling is incomplete in these 20 ns MD simulations. However, combining multiple short trajectories increases the likelihood of achieving a better coverage of the pocket shape space.^{24,44}

To address the diverse pocket selection problem, the CLARA clustering algorithm was then applied to cluster the ALR pocket shapes according to their principal component scores. Here, ten clusters were identified based on the pockets' scores along the first 50 PCs, which are cumulatively responsible for 83% of the total variance in the MD data set (see Supporting Information Figure S4). Figure 5 shows the cluster representatives and indicates the number of members for each resulting cluster.

The ten cluster representatives span a diversity of pocket shapes. There are small, *apo*-like conformations such as clusters 2, 3, 5, 6, and 7. In contrast, the selectivity pocket is open in clusters 4, 8, and 9 and in fact adopts a different conformation in each case.²⁴ A collapsed pocket shape is apparent for cluster 10, whereas clusters 1 and 9 extend into the novel C298 subpocket that is involved in PC1 and PC2 (see Figure 3). This subpocket is of some interest because, unlike the selectivity pocket, it is not exploited by any of the ligands in the publicly available ALR crystal structures.

Notably, while at least two cluster representatives are drawn from each of the three MD simulations (two from the first, three from the second, and two from the third), three cluster

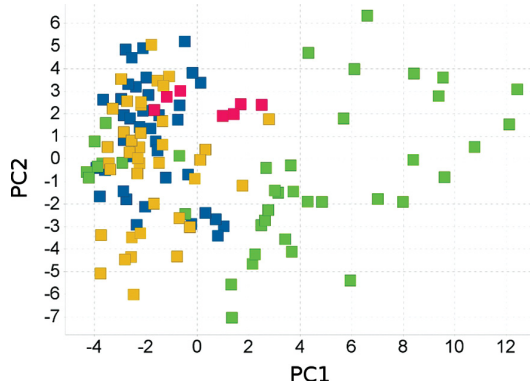


Figure 4. The pocket conformational distribution for the aldose reductase data set along PC1 and PC2. Pocket shapes derived from crystal structures are represented by red squares. Pocket shapes derived from the three trajectories MD1, MD2, and MD3 are colored blue, green, and yellow respectively.

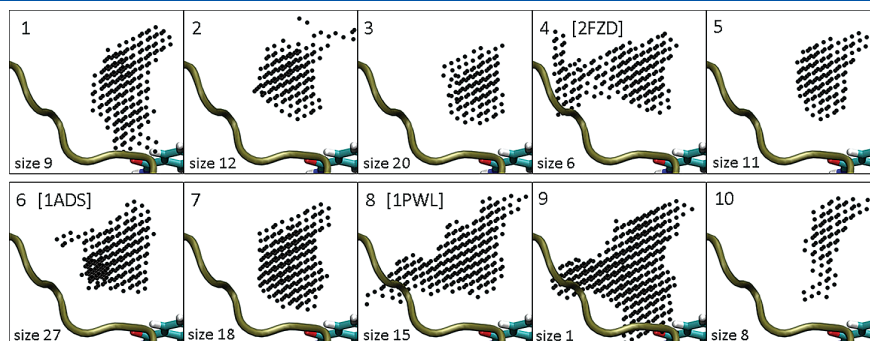


Figure 5. The diverse pocket selection for the active site of aldose reductase. The ten cluster representatives are shown along with the size of each cluster (number of members). Each representative is viewed from the same angle as Figure 2 but zoomed in to better discriminate between pocket shapes. For clarity, most of the protein is not shown — only the flexible L300 loop and the NADP⁺ cofactor are included for reference.

Table 1. Druggability Analysis of the Ten ALR Cluster Representatives

cluster representative	source	DScore	size ^a	enclosure ^b	hydrophilicity ^c
1	MD 1	1.02	112	0.66	0.95
2	MD 1	0.94	88	0.65	1.02
3	MD 3	0.90	78	0.69	1.06
4	crystal	1.08	117	0.66	0.79
5	MD 3	0.97	124	0.66	1.12
6	crystal	0.98	95	0.68	1.06
7	MD 2	1.06	87	0.66	0.64
8	crystal	1.16	103	0.76	0.73
9	MD2	1.19	127	0.77	0.65
10	MD2	0.79	58	0.75	1.16

^aNumber of grid-points. ^bFraction of radial rays which intersect the protein within a certain distance;¹⁷ ^cIn kcal mol⁻¹ (but see SiteMap reference¹⁷ for calibration).

representatives actually arise from crystallographically derived protein structures: number 4 is from 2FZD (*tolrestat*), number 6 is from 1ADS (*apo*), and number 8 is from 1PWL (*zenarestat*). The PocketAnalyzer^{PCA} diverse pocket selection has therefore automatically distinguished three of the four pocket conformations characterized by Sotriffer and co-workers.²⁴ The fourth crystallographic conformation (IDD594) is represented by structures 1US0 and 2R24 that are both members of cluster 8 along with 1PWL.

Within the context of the current clustering pattern, clusters which do not contain a single crystallographically derived pocket shape (which are all except clusters 4, 6, and 8) represent novel computationally derived ALR binding-site conformations. Undoubtedly, some of these “novel” conformations are in fact fairly similar to a pocket shape identified in one or other of the crystal structures. Others seem genuinely distinct, such as the collapsed pocket of cluster 10 or the pocket shapes expanding into the C298 subpocket as in cluster representatives 1 and 9.

Testing the Druggability of Novel Pocket Conformations. Although novel pocket conformations may indeed provide valuable opportunities for diversifying ligand design, the question arises as to whether any of these new computationally derived pocket shapes are realistically druggable. To address this issue, the druggability of all ten cluster representatives was analyzed with SiteMap¹⁷ (see Materials and Methods section). The results are displayed in Table 1.

A higher DScore druggability metric is intended to indicate a more druggable pocket conformation, with a threshold of DScore = 0.98 taken to delineate “druggable” sites.¹⁷ Pockets with DScore < 0.83 are assumed to be “undruggable”, and any in between are predicted to be “difficult”. In Table 1 the MD-derived pocket conformations cover a wider range of druggabilities than the crystallographic pocket shapes. This is not entirely surprising, given that this observation compares conformations of protein–ligand cocrystal structures with those from ligand-free MD simulations. The more druggable computationally derived pockets are cluster representatives 1, 7, and 9, each of which has a DScore > 1. Although cluster representative 7 is rather similar to *apo* conformations like 1ADS, the other two are more novel conformations that expand into the C298 subpocket. Cluster representative 9 is in fact much larger than is apparent

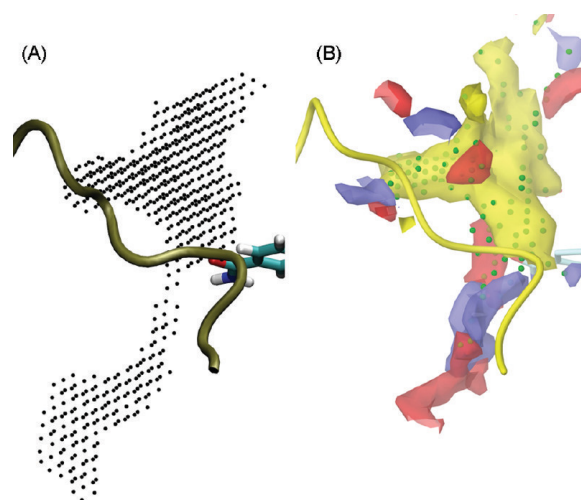


Figure 6. Full view of cluster representative 9 showing the open channel and the connection to a second cavity on the surface of the ALR protein structure. (A) The pocket shape identified by PocketAnalyzer. (B) The SiteMap analysis of the same protein structure, showing the hydrophobic (yellow), hydrogen-bond donor (blue), and hydrogen-bond acceptor (red) fields. SiteMap site points are represented as green spheres.

from Figure 5; its full extent is shown in Figure 6. It is formed by a twist of the protein backbone at residues C298 and A299, with associated changes in side-chain conformation of C298 and Y209. Smaller adjustments of side-chain torsional angles occur for residues W112 and N160. The net effect is the opening of a channel past C298, underneath the L300 loop, connecting the ALR binding site with a second cavity as shown in Figure 6. The high DScore of this pocket derives in part from its size and in part from its low hydrophilic character.

The predicted druggability of cluster representative 9 makes this computationally derived pocket conformation a potentially attractive target for ALR ligand design. Therefore, the pocket conformation was analyzed to determine if already known ALR inhibitors bind to this region. By comparing 52 publicly available ligand-bound wild-type ALR crystal structures, no ligand was found that entered the region of the novel C298 subpocket. This may suggest that in reality the cluster representative 9 pocket shape is rarely formed, or equivalently that it has an unfavorable free energy of formation. It seems plausible that the protein would have to adopt a relatively high energy conformation in order to open the C298 channel. Aspects such as protein strain and entropic changes are not explicitly incorporated into the current DScore druggability metric.

For the majority of known ALR inhibitors, however, there is no publicly available crystallographic information on the binding mode. Therefore, to address the question of whether these ligands might in fact occupy the C298 subpocket, those with structural similarity to a ligand with an available crystallographic binding mode were aligned to that template using the substructure-based ligand alignment method available in ICM.²⁵ The resulting ligand alignments provide predicted binding modes for a set of 66 substructure-matched ALR inhibitors (for details see Material and Methods).

Remarkably, the alignments of two lidorestat derivatives predict that they must bind to a pocket conformation similar to cluster

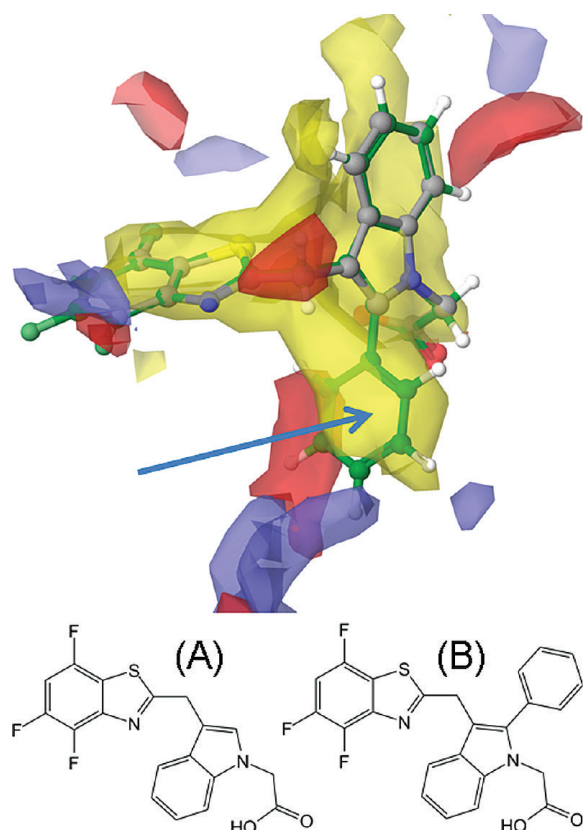


Figure 7. Ligand alignment of the 2-phenyl derivative of lidorestat (green; bindingDB monomerid = 16471) with the crystallographic binding mode of the parent compound from PDB structure 1Z3N (gray). The blue arrow indicates the occupation of the C298 subpocket by the 2-phenyl substituent. Lidorestat (A) and its 2-phenyl derivative (B) are sketched below the main figure. The SiteMap analysis of the cluster representative 9 is also shown in the same frame of reference following alignment of the associated protein structure with 1Z3N using the Align Binding Sites tool in Maestro.¹³

representative 9. The most striking example is shown in Figure 7, in which a phenyl substituent at the 2-position of the indole core of lidorestat is predicted to fill the hydrophobic part of the C298 subpocket. This phenyl-derivative of lidorestat has an IC_{50} of 100 nM in an *in vitro* ALR inhibition assay,²⁷ providing experimentally based evidence that the novel computationally derived pocket conformation identified by PocketAnalyzer^{PCA} is a plausible and druggable target for structure-based drug design and a potentially valuable opportunity to design novel ALR inhibitor chemotypes. Along these lines, since the phenyl-derivative of lidorestat also enters the more polar area toward the entrance to the channel, a hydrophilic meta or para substituent or the replacement of the phenyl ring with a nitrogen-containing heterocycle (e.g., pyridine, pyrrole, imidazole) might add further interactions with the protein.

PocketAnalyzer^{PCA} Applied to Neuraminidase. In a second example the PocketAnalyzer^{PCA} protocol was applied to viral neuraminidase subtype N1, which is a target for antiviral inhibitors useful in the treatment of influenza. Current drugs on the market include zanamivir and oseltamivir, both of which bind to the main catalytic (sialic acid-binding or SA) site of the enzyme. X-ray crystallography and extended MD simulations have revealed

significant conformational mobility in this region, which has not been observed for other neuraminidase subtypes.⁴⁵ In particular, the 150-loop can adopt an open or closed conformation; in its open conformation, it gives access to a second cavity, adjacent to the SA site. A third cavity in the neuraminidase binding site has been identified by means of MD simulations close to the 430-loop.⁷ In the following we therefore refer to (i) the SA-cavity where oseltamivir binds, (ii) the 150-cavity, and (iii) the 430-cavity. Even though no current drugs are known to bind to these latter two cavities, a technique called computational solvent mapping has identified low energy binding-sites for small molecular probes in both regions,⁴⁶ and a virtual screening campaign has identified several new chemotypes which could potentially target these sites.⁷

Recently, McCammon and co-workers have characterized the conformational mobility of the neuraminidase active site using extensive MD simulations.^{47,48} In their work, all-atom root-mean-square deviation-based (rmsd) clustering on a subset of 62 residues lining the active site was used to identify and compare clusters of pocket conformations.⁷ Here we show that similar results and insight about the binding site flexibility can be gained by applying the PocketAnalyzer^{PCA} diverse pocket selection protocol.

Diverse Pocket Selection. The PocketAnalyzer^{PCA} protocol was used to analyze and cluster snapshots extracted from three N1 MD simulations, including one *apo* and two *holo* simulations (see Materials and Methods). First, the aligned protein conformations are processed with the PocketAnalyzer algorithm and the grid-point inclusion/exclusion row vectors corresponding to the 600 computationally derived structures (3 trajectories, 200 frames from each) are fed into the PCA. The row vectors corresponding to nine crystallographic pocket conformations (see Material and Methods section) are withheld from the PCA but subsequently projected onto the resulting PCs. In Figure 8, the first two principal components, cumulatively responsible for 25% of the total variance, show that the most variable regions involve the 430-cavity and the 150-cavity, in agreement with the findings of Amaro et al.⁴⁷ The SA-cavity is more conserved, but in the *apo* simulation a side-chain movement of Arg152 causes some variation along a channel adjacent to the 150-loop. The PCD for the neuraminidase data set along PC1 and PC2 illustrates the different regions of pocket space explored by the three MD simulations (Supporting Information Figure S6).

To extract a representative set of pocket shapes we clustered the pockets calculated from the three MD simulations using their scores along the first 50 principal components, which are cumulatively responsible for 63% of the total variance in this data set (see Supporting Information Figure S5). This approach differs from the earlier rmsd-based clustering⁷ in that rather than clustering each simulation separately, all 600 snapshots were analyzed simultaneously. We furthermore decided to define only 10 clusters to facilitate further processing and visual inspection of the resulting representative structures. It would however be straightforward to define a larger number of clusters or to extract additional structures from a given cluster of interest.

An overlay of the MD snapshots corresponding to the cluster representatives is shown in Figure 9 and compared to the corresponding overlay of the crystallographic protein structures. The most striking difference between the experimental and computational ensembles is the much larger conformational variability in the 430-loop in the MD simulations. In previous simulations,⁴⁷ the most dramatic conformational changes have

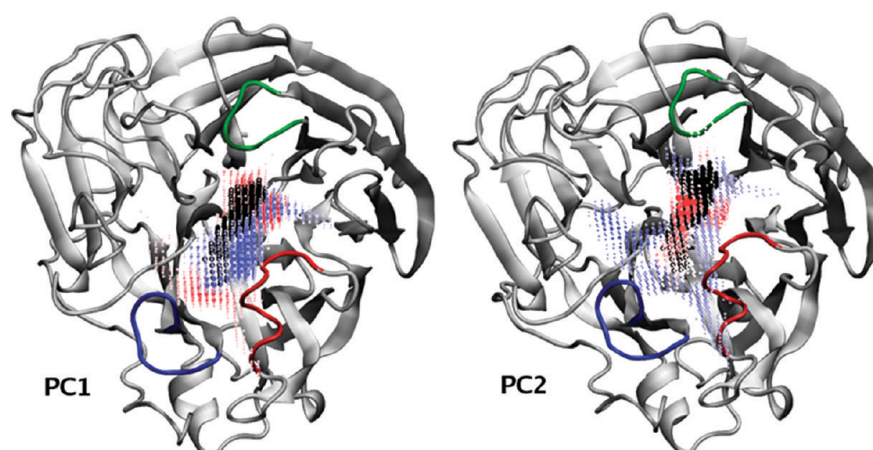


Figure 8. The highest variance principal components of the neuraminidase data set. The grid-points are sized and colored in the same way as in Figure 3. The 150-loop is shown in red, the 430-loop is shown in blue, and the SA-loop is shown in green. The latter comprises residues Pro245 to Ala250.

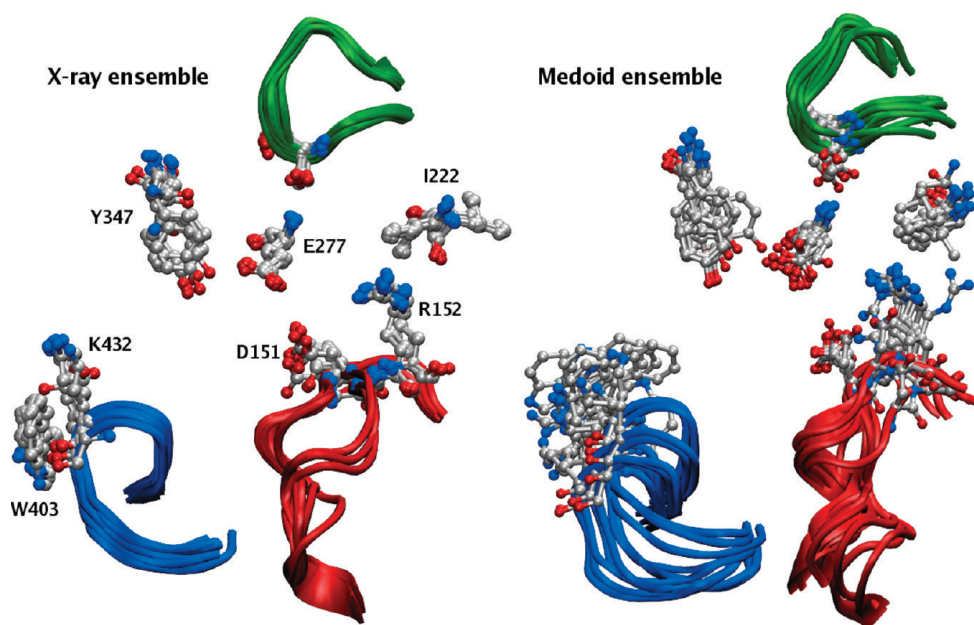


Figure 9. Left: Superposition of a representative set of *apo* and *holo* X-ray structures of N1 and N8 neuraminidases (see Material and Methods section). Right: Superposition of MD-derived cluster representatives selected by PocketAnalyzer^{PCA}. The 150-, 430-, and SA-loops are color-coded as in Figure 8. Selected active site residues are shown in ball and stick representation.

been observed in the *apo* simulations, with the 430-loop reaching a wide-open conformation that results in a 430-loop C_{α} rmsd of 4 Å and a markedly increased solvent accessible surface area. In our case, it is the *holo* simulations that display the larger variations in the 430-loop, with an average 430-loop C_{α} rmsd of 4 Å (Supporting Information Figure S7, middle and lower panels). The opening of the 430-loop is accompanied by a positional change of Lys432 and Trp403, which swap positions (Figure 9). The 150-loop is more stable in the *apo* and first *holo* simulations, whereas it transitions into a wide-open state in the second *holo* simulation, leading to a final 150-loop C_{α} rmsd of 4 Å. This is again in agreement with previous MD simulations and leads to the opening of the 150-cavity.

More subtle differences in pocket-shapes are due to side-chain movements and are apparent in the pocket shapes shown in Figure 10. Clusters representatives 2, 3, and 4 originate from the *apo* simulation and reflect the absence of the extensive protein–ligand hydrogen-bond and salt-bridge network that exists in the oseltamivir-bound *holo* simulations. In the latter, the positively charged exocyclic amino group on the ligand interacts with the conserved Glu119 and Asp151 of the 150-loop. In addition, the ligand's carboxylic acid functionality interacts with Arg292, Arg371, and Tyr347. In the absence of the ligand, Asp151 is more mobile and moves outward, leading to a larger central pocket shape. Cluster representatives 8, 9, and 10 expand significantly into the 150-cavity; they all originate from the second *holo*

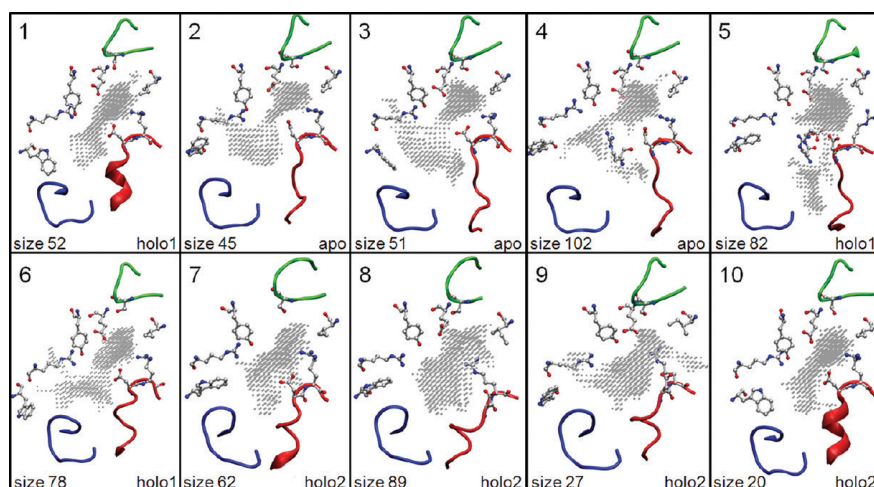


Figure 10. The diverse pocket selection for the neuraminidase data set. Only the 150 (red), 430 (blue), and SA-loop (green) are shown for clarity. A few residues lining the binding pocket are displayed in ball and stick representation in order to highlight side-chain movements that lead to changes in the pocket shape, such as Tyr347 and Asp151. The size of the cluster and source of its representative are indicated to the left and to the right of each pocket shape.

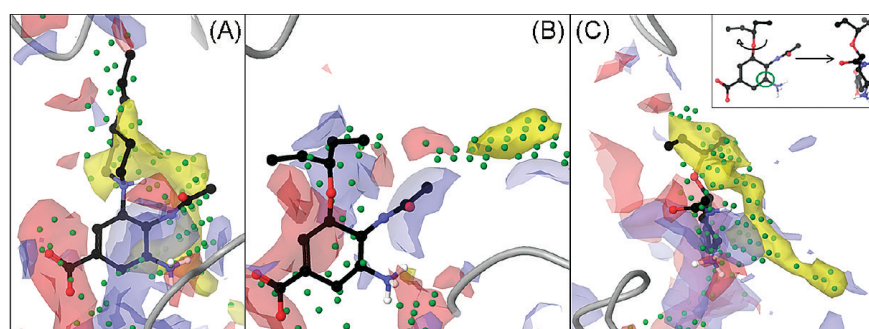


Figure 11. SiteMap analysis of three cluster representatives exhibiting small novel subpockets around the main SA-cavity. Fields are color-coded as for Figure 6. Green spheres represent the SiteMap site-points. For orientation, the SA-loop and part of the 150-loop are shown in gray ribbon at the top and bottom of each panel. (A) Cluster representative 6 showing the extension toward the SA-loop at the top of the figure. An oseltamivir derivative predicted to exploit this subpocket is superimposed (see text; bindingDB monomerid = 5261). (B) Cluster representative 4 showing the open channel and the hydrophobic patch (yellow) adjacent to the 150-loop. The channel runs horizontally across the middle of the figure. Oseltamivir is superimposed for reference. (C) Cluster representative 5 showing the opening of a hydrophobic tunnel in the floor of the catalytic site. This figure is rotated roughly 90° compared to the other panels, as illustrated for oseltamivir by the curly arrow in the inset. The green circle in the inset indicates the possible attachment point discussed in the text.

simulation which has the largest 150-loop rmsd. Pockets that extend into the 430-cavity were extracted from all three simulations, consider for example cluster representatives 1 (first *holo* simulation), 4 (*apo* simulation), and 7, 8, and 9 (second *holo* simulation).

To summarize the section up to this point, pocket shapes have been identified which correspond to: (a) the wide-open conformation of the 150-loop and (b) the wide-open conformation of the 430-loop. The PocketAnalyzer^{PCA} results thus broadly match the all-atom rmsd-based analysis of Cheng et al.⁷

Testing the Druggability of Novel Pocket Conformations. Of the ten pocket shapes in the diverse pocket selection in Figure 10, numbers 4, 5, and 6 were found to be of particular interest as novel (and potentially druggable) binding-site conformations of N1 neuraminidase. In making this selection, we have neglected pocket shapes that merely show expansion into the 150- and 430-cavities since similar binding-site conformations have been discovered and described in the earlier work discussed above.

Instead, we focus on more subtle variations caused by the opening of small but distinct subpockets around the main SA-cavity.

The first of these is exemplified by cluster representative 6 and involves an extension of the pocket toward the SA-loop (see Figure 10). This coincides with the creation of a strongly hydrophobic patch in this region of the protein, which is at least partially responsible for cluster representative 6 having a relatively high predicted druggability; its SiteMap DScore is 1.05 as opposed to a median DScore of 0.97 (minimum 0.96, maximum 1.04) among the crystallographic binding-site conformations (see Supporting Information Table S8).

As for aldose reductase, an obvious question is whether any experimentally confirmed inhibitors of N1 neuraminidase might exploit this SA-loop subpocket. As above, we used structural alignment of known inhibitors to crystallographically derived binding-modes of structurally similar ligands to address this issue (for details see Materials and Methods). The result-

ing alignments provide predicted binding modes for 64 substructure-matched derivatives of oseltamivir. Within this set of submicromolar inhibitors, 95% involve a variation of the hydrophobic ether substituent that is directed toward the SA-loop in the crystallographic binding-mode of oseltamivir. The substructure alignments indeed predict that five of the molecules occupy the novel SA-loop subpocket to some extent. One example is the 4-propylpiperidine derivative that is superimposed on the fields from the SiteMap analysis of cluster representative 6 in panel A of Figure 11. This molecule has an IC_{50} of 40 nM in an enzymatic N1 inhibition assay. Indeed, the report describing the synthesis of this compound also includes an illustration of a proprietary crystal structure of the related piperidine derivative.⁴⁹ This confirms the binding mode of these compounds and provides experimental support for plausibility of this novel binding-site conformation identified by PocketAnalyzer^{PCA}.

A second variation of the SA-cavity is exhibited by cluster representative 4 and is induced by a change in the conformation of the side-chain of Arg152. In particular, the guanidine group of Arg152 moves toward, and forms a salt bridge with, the side-chain of Glu277. This rearrangement opens a narrow channel that is adjacent to the 150-loop (Figure 11, panel B). Interestingly, this rotation of Arg152 side-chain would not be compatible with the binding modes of known SA-cavity ligands (e.g., oseltamivir and zanamivir) and, indeed, is only observed in the *apo* MD simulation. Furthermore, although the novel channel includes a moderately sized hydrophobic patch, the rearrangement leads to an overall decrease in depth and size of the catalytic binding site and hence a relatively low DScore of 0.92 (see Supporting Information Table S8). None of the substructure-matched oseltamivir derivatives are predicted to occupy this channel by the structural alignment approach discussed above. However, some support for the existence of this pocket conformation is provided by the earlier computational solvent mapping calculations.⁴⁶ In that work it was observed that a low energy cluster of chemical probe positions is found in the region of this channel.

Cluster representative 5 (Figure 11, panel C) exemplifies a third conformation of the SA cavity that, to the best of our knowledge, has not been previously observed in either experimental or computational studies. It involves the opening of a narrow but hydrophobic tunnel in the floor of the catalytic site. Again, none of the publicly available oseltamivir derivatives are predicted to enter this tunnel by the structural alignment approach. However, the tunnel is also present to some extent in cluster representatives 1, 2, 3, and 6, so it is a characteristic and recurrent feature of the simulations. The addition of several rather hydrophobic site-points results in high predicted druggability with a DScore of 1.08. Furthermore, the central cyclohexene ring of oseltamivir has an obvious attachment point for hydrophobic substituents that might access this tunnel (Figure 11). It will therefore be of interest to see if future medicinal chemistry studies report evidence of inhibitors binding to this novel pocket conformation.

CONCLUDING REMARKS

We have introduced the PocketAnalyzer^{PCA} methodology to address the problem of diverse pocket selection, i.e. how to reduce a large collection of structures of the same protein to a subset that retains a number of substantially distinct binding pocket conformations. Diverse pocket shapes drive medicinal chemistry to explore a broader chemical space and so present additional opportunities to overcome key drug discovery issues

such as potency, selectivity, toxicity, and pharmacokinetics. The identification of diverse pocket shapes and novel binding-site conformations can therefore greatly assist the progress of structure-based ligand design projects.

The PocketAnalyzer^{PCA} approach combines a grid-based pocket detection algorithm with PCA and clustering. The resulting principal component (PC) eigenvectors reveal the dominant binding-site deformation modes within an ensemble of protein structures, and the corresponding PC scores provide characterization and visualization of the pocket conformational distributions. From a methodological point of view, the PocketAnalyzer^{PCA} approach provides a novel and complementary perspective on protein dynamics that may prove particularly relevant for ligand binding and drug design. PocketAnalyzer^{PCA} was primarily envisioned as a tool for analyzing trajectories of protein conformations produced by MD simulations. However, the procedure is applicable to any source of atomistic protein structural information and also to combinations of structures from several such sources, as in the examples presented above. Therefore, PocketAnalyzer^{PCA} may be useful for exploring the increasing volume of experimentally derived structural information resulting from high-throughput crystallography and advances in NMR-based techniques.

A technically related approach to a different problem combines pocket detection and clustering to track the opening and closing of transient binding pockets in protein–protein interaction surfaces along MD trajectories.⁵⁰ This ePOS method analyses each in a sequence of MD frames using the PASS pocket detection algorithm⁵¹ and clusters the resulting pockets by the set of pocket lining atoms to define unique pockets and track their opening and closing as time progresses. The recently announced fpocket Web server uses a different pocket detection algorithm to perform a similar analysis.⁵² A precedent for combining grid-based pocket characterization with PCA is the GRID/CPCA approach of Kastenholz et al.^{53,54} However, rather than focusing on the analysis of many structures of the same protein (as here) the GRID/CPCA method is directed at comparing structures of different targets to derive insights that assist in improving compound selectivity.

When applied to aldose reductase, a protein with moderate binding-site flexibility and a well-characterized set of crystallographic binding-site conformations, PocketAnalyzer^{PCA} distinguishes three of the four crystallographically observed binding-site conformations previously reported by the Klebe group.²⁴ In addition, the approach identifies a number of distinct pocket shapes that have not been observed experimentally and which therefore represent novel computationally derived binding-site conformations. From a medicinal chemistry point of view, the most outstanding result is that one MD-derived pocket shape is particularly striking in its difference to the crystallographic conformations. A rotation of a short section of the protein backbone and accompanying adjustments in the positions of a few amino-acid side-chains open a channel connecting the active site with another pocket on the protein surface. Although the channel itself and the second pocket are rather polar, the ‘entrance’ to the channel forms a reasonably large hydrophobic subpocket, and SiteMap¹⁷ analysis predicted good druggability for this novel conformation of the ALR active site. Indeed, subsequent alignment of known ALR inhibitors to the crystallographic binding modes of structurally similar ALR ligands identified a derivative of lidorestat that is predicted to fill the novel hydrophobic subpocket with a phenyl ring. This compound has an IC_{50} of

100 nM in an *in vitro* ALR inhibition assay,²⁷ providing experimental evidence that the novel computational derived pocket conformation identified by PocketAnalyzer^{PCA} is a plausible and druggable target for structure-based drug design against ALR.

In a second example, the PocketAnalyzer^{PCA} approach is used to derive a diverse set of binding-site conformations from viral neuraminidase. Similarly to ALR, the binding-site flexibility of neuraminidase is reasonably well-established, for example as a result of MD simulations,⁴⁸ and a number of distinct binding-site conformations have been characterized crystallographically. The PocketAnalyzer^{PCA} diverse pocket approach was found to identify a qualitatively similar range of binding-site conformations as a previously reported atom-based rmsd clustering method,^{47,48} with the advantage of quickly highlighting conserved and variable regions in the pocket. The method also allows facile comparison of structures from different sources and direct visualization of differences in pocket shape rather than changes in proxy descriptors such as backbone and side-chain positions.

The N1 diverse pocket selection included three particularly interesting and novel subpockets adjacent to the main catalytic site of N1 neuraminidase. Alignment of known submicromolar N1 inhibitors to the crystallographic binding-mode of oseltamivir identified several molecules predicted to occupy the first of these subpockets, for example with a 4-propylpiperidine substituent. Indeed, the report describing the synthesis of this compound includes an illustration of a proprietary crystal structure of the unsubstituted piperidine derivative,⁴⁹ confirming the binding mode and the plausibility of this computationally derived binding-site conformation as a druggable target for N1 inhibition.

In addition to direct application to structure-based ligand design, the PocketAnalyzer^{PCA} protocol produces an ensemble of protein structures incorporating diverse and potentially novel pocket shapes that could be useful as input to numerous structure-based drug design methods.⁵⁵ For example, this would provide an effective way to account for protein flexibility in docking and virtual screening,^{56–59} receptor-based pharmacophore generation,³ and druggability analysis.⁶⁰ Applied in this way, PocketAnalyzer^{PCA} would be just one component in a larger drug discovery workflow, providing a rational approach to selecting an ensemble of protein conformations.^{61,62}

Moving beyond the specific application of PocketAnalyzer^{PCA} to diverse pocket selection, in the future the approach may be more broadly applied to address questions regarding the effect of various perturbations on pocket conformational distributions. For example, when coupled with the appropriate MD simulations, PocketAnalyzer^{PCA} may provide a useful perspective on the change in binding-site conformation induced by factors such as mutation, allosteric modulation, solvent pH, and post-translational modification.

■ ASSOCIATED CONTENT

■ **Supporting Information.** Additional PCD plots and PCA scree plots for the ALR and NA data sets, rmsd plots for all MD simulations, and details of the oseltamivir force field parametrization. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ianrcraig@gmail.com.

Present Addresses

^{||}BASF SE, GVC/C - A030, 67056 Ludwigshafen, Germany.

■ ACKNOWLEDGMENT

We are grateful to Teresa Jimenez Vaquero for providing an initial version of the PocketAnalyzer code. Financial support from the Education Office of Novartis Institutes for Biomedical Research is gratefully acknowledged (I.C.).

■ ABBREVIATIONS:

ALR, aldose reductase; NA, neuraminidase; SA, sialic acid; PCA, principal component analysis; PC, principal component; PCD, pocket conformational distribution; MD, molecular dynamics; GAFF, generalized amber force field; rmsd, root mean square deviation

■ REFERENCES

- (1) Dutta, S.; Burkhardt, K.; Young, J.; Swaminathan, G.; Matsuura, T.; Henrick, K.; Nakamura, H.; Berman, H. Data Deposition and Annotation at the Worldwide Protein Data Bank. *Mol. Biotechnol.* **2009**, *42*, 1–13.
- (2) Ahmed, A.; Kazemi, S.; Gohlke, H. Protein flexibility and mobility in structure-based drug design. *Front. Drug Des. Discovery* **2007**, *3*, 455–476.
- (3) Carlson, H. A.; Masukawa, K. M.; Rubins, K.; Bushman, F. D.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A. Developing a dynamic pharmacophore model for HIV-1 integrase. *J. Med. Chem.* **2000**, *43*, 2100–2114.
- (4) Cozzini, P.; Kellogg, G. E.; Spyridis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **2008**, *51*, 6237–6255.
- (5) Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A. Discovery of a novel binding trench in HIV integrase. *J. Med. Chem.* **2004**, *47*, 1879–1881.
- (6) Sotriffer, C. A.; Ni, H.; McCammon, J. A. Active site binding modes of HIV-1 integrase inhibitors. *J. Med. Chem.* **2000**, *43*, 4109–4117.
- (7) Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. *J. Med. Chem.* **2008**, *51*, 3878–3894.
- (8) Zhou, Z.; Madrid, M.; Evansek, J. D.; Madura, J. D. Effect of a bound non-nucleoside RT inhibitor on the dynamics of wild-type and mutant HIV-1 reverse transcriptase. *J. Am. Chem. Soc.* **2005**, *127*, 17253–17260.
- (9) Grant, B. J.; Rodrigues, A. P. C.; Elsayy, K. M.; McCammon, J. A.; Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **2006**, *22*, 2695–2696.
- (10) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Modell.* **1997**, *15*, 359–363.
- (11) Stahl, M.; Taroni, C.; Schneider, G. Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng.* **2000**, *13*, 83–88.
- (12) Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer: New York, 2002.
- (13) *Maestro*, version 9.1; Schrödinger, LLC: New York, 2010.
- (14) Godzik, A. The structural alignment between two proteins: Is there a unique answer? *Protein Sci.* **1996**, *5*, 1325–1338.
- (15) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley: New York, 1990.

- (16) Available at <http://cran.r-project.org/web/packages/cluster> (accessed September 27, 2011).
- (17) Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377–389.
- (18) Carl, N.; Konc, J.; Janezic, D. Protein surface conservation in binding sites. *J. Chem. Inf. Model.* **2008**, *48*, 1279–1286.
- (19) Carl, N.; Konc, J.; Vehar, B.; Janezic, D. Protein-Protein Binding Site Prediction by Local Structural Alignment. *J. Chem. Inf. Model.* **2010**, *50*, 1906–1913.
- (20) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.
- (21) Coleman, R. G.; Burr, M. A.; Souvaine, D. L.; Cheng, A. C. An intuitive approach to measuring protein surface curvature. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 1068–1074.
- (22) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607–620.
- (23) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387–406.
- (24) Sotriffer, C. A.; Kramer, O.; Klebe, G. Probing flexibility and “induced-fit” phenomena in aldose reductase by comparative crystal structure analysis and molecular dynamics simulations. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 52–66.
- (25) ICM, version 3.6; MolSoft, LCC; La Jolla, CA, 2010.
- (26) Liu, T. Q.; Lin, Y. M.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.
- (27) Van Zandt, M. C.; Jones, M. L.; Gunn, D. E.; Geraci, L. S.; Jones, J. H.; Sawicki, D. R.; Sredy, J.; Jacot, J. L.; DiCioccio, A. T.; Petrova, T.; Mitschler, A.; Podjarny, A. D. Discovery of 3-[(4,5,7-trifluorobenzothiazol-2-yl)methyl]indole-N-acetic acid (Lidorestat) and congeners as highly potent and selective inhibitors of aldose reductase for treatment of chronic diabetic complications. *J. Med. Chem.* **2005**, *48*, 3141–3152.
- (28) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- (29) Case, D. A.; Darden, T. A.; Cheatham, Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *Amber 9*; University of California: San Francisco, 2006.
- (30) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.
- (31) Bryce, R. AMBER parameter database. <http://www.pharmacy.manchester.ac.uk/bryce/amber> (accessed March 15, 2010).
- (32) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (33) Izaguirre, J. A.; Cattarelli, D. P.; Wozniak, J. M.; Skeel, R. D. Langevin stabilization of molecular dynamics. *J. Chem. Phys.* **2001**, *114*, 2090–2098.
- (34) Ryckaert, J. P.; Cicotti, G.; Berendsen, H. J. C. Numerical-Integration of Cartesian Equations of Motion of A System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (35) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald - An N. Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (36) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-Dynamics with Coupling to An External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (37) Lawrenz, M.; Wereszczynski, J.; Amaro, R.; Walker, R.; Roitberg, A.; McCammon, J. A. Impact of calcium on N1 influenza neuraminidase dynamics and binding free energy. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2523–2532.
- (38) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (39) Varkonyi, T.; Kempler, P. Diabetic neuropathy: new strategies for treatment. *Diabetes Obes. Metab.* **2008**, *10*, 99–108.
- (40) Steuber, H.; Zentgraf, M.; Gerlach, C.; Sotriffer, C. A.; Heine, A.; Klebe, G. Expect the unexpected or caveat for drug designers: Multiple structure determinations using aldose reductase crystals treated under varying soaking and co-crystallisation conditions. *J. Mol. Biol.* **2006**, *363*, 174–187.
- (41) Steuber, H.; Zentgraf, M.; Podjarny, A.; Heine, A.; Klebe, G. High-resolution crystal structure of aldose reductase complexed with the novel sulfonyl-pyridazinone inhibitor exhibiting an alternative active site anchoring group. *J. Mol. Biol.* **2006**, *356*, 45–56.
- (42) Steuber, H.; Zentgraf, M.; La Motta, C.; Sartini, S.; Heine, A.; Klebe, G. Evidence for a novel binding site conformer of aldose reductase in ligand-bound state. *J. Mol. Biol.* **2007**, *369*, 186–197.
- (43) Luque, L.; Freire, E. Structural stability of binding sites: Consequences for binding affinity and allosteric effects. *Proteins: Struct., Funct., Genet.* **2000**, *63*–71.
- (44) Caves, L. S. D.; Evanseck, J. D.; Karplus, M. Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin. *Protein Sci.* **1998**, *7*, 649–666.
- (45) Russell, R. J.; Haire, L. F.; Stevens, D. J.; Collins, P. J.; Lin, Y. P.; Blackburn, G. M.; Hay, A. J.; Gamblin, S. J.; Skehel, J. J. The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **2006**, *443*, 45–49.
- (46) Landon, M. R.; Amaro, R. E.; Baron, R.; Ngan, C. H.; Ozonoff, D.; McCammon, J. A.; Vajda, S. Novel druggable hot spots in avian influenza neuraminidase H5N1 revealed by computational solvent mapping of a reduced and representative receptor ensemble. *Chem. Biol. Drug Des.* **2008**, *71*, 106–116.
- (47) Amaro, R. E.; Minh, D. D. L.; Cheng, L. S.; Lindstrom, W. M.; Olson, A. J.; Lin, J. H.; Li, W. W.; McCammon, J. A. Remarkable loop flexibility in avian influenza N1 and its implications for antiviral drug design. *J. Am. Chem. Soc.* **2007**, *129*, 7764.
- (48) Amaro, R. E.; Xiaolin, C.; Ivaylo, I.; Dong, X.; McCammon, J. A. Characterizing Loop Dynamics and Ligand Recognition in Human and Avian Type Influenza Neuraminidases via Generalized Born Molecular Dynamics and End-Point Free Energy Calculations. *J. Am. Chem. Soc.* **2009**, *131*, 4702–4709.
- (49) Lew, W.; Wu, H. W.; Chen, X. W.; Graves, B. J.; Escarpe, P. A.; MacArthur, H. L.; Mendel, D. B.; Kim, C. U. Carbocyclic influenza neuraminidase inhibitors possessing a C-3-cyclic amine side chain: Synthesis and inhibitory activity. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1257–1260.
- (50) Eyrisch, S.; Helms, V. Transient pockets on protein surfaces involved in protein-protein interaction. *J. Med. Chem.* **2007**, *50*, 3457–3464.
- (51) Brady, G. P.; Stouten, P. F. W. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383–401.
- (52) Schmidtke, P.; Le Guilloux, V.; Maupetit, J.; Tuffery, P. fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* **2010**, *38*, W582–W589.
- (53) Afzelius, L.; Raubacher, F.; Karlen, A.; Jorgensen, F. S.; Andersson, T. B.; Masimirembwa, C. M.; Zamora, I. Structural analysis of CYP2C9 and CYP2C5 and an evaluation of commonly used molecular modeling techniques. *Drug Metab. Dispos.* **2004**, *32*, 1218–1229.
- (54) Kastenholtz, M. A.; Pastor, M.; Cruciani, G.; Haaksma, E. E. J.; Fox, T. GRID/CPA: A new computational tool to design selective ligands. *J. Med. Chem.* **2000**, *43*, 3033–3044.
- (55) Perot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A. C.; Villoutreix, B. O. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today* **2010**, *15*, 656–667.

- (56) Barril, X.; Morley, S. D. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.* **2005**, *48*, 4432–4443.
- (57) Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 511–524.
- (58) Huang, S. Y.; Zou, X. Q. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 399–421.
- (59) Virtanen, S. I.; Pentikainen, O. T. Efficient Virtual Screening Using Multiple Protein Conformations Described as Negative Images of the Ligand-Binding Site. *J. Chem. Inf. Model.* **2010**, *50*, 1005–1011.
- (60) Egner, U.; Hillig, R. C. A structural biology view of target drugability. *Expert Opin. Drug Discovery* **2008**, *3*, 391–401.
- (61) Bolstad, E. S. D.; Anderson, A. C. In pursuit of virtual lead optimization: Pruning ensembles of receptor structures for increased efficiency and accuracy during docking. *Proteins: Struct., Funct., Bioinf.* **2009**, *75*, 62–74.
- (62) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. *J. Chem. Inf. Model.* **2010**, *50*, 186–193.

13.10 Publication V – Supporting Information

Pocket-Space Maps to Identify Novel Binding-Site

Conformations in Proteins

Craig, I.R., Pfleger, C., Gohlke, H., Essex, J.W., Spiegel, K.

J. Chem. Inf. Model. (2011), 51, 2666–2679

Supporting Information:

Pocket-space maps to identify novel binding-site conformations in proteins

Ian R. Craig, Christopher Pflieger, Holger Gohlke, Jonathan W. Essex, and Katrin Spiegel

Contents

Figure S1. Sensitivity of PocketAnalyzer ^{PCA} principal component scores to a change in the protein structural alignment
Section S2. AMBER force field parameters for oseltamivir
Figure S3. PocketAnalyzer ^{PCA} analysis of the aldose reductase dataset including the crystallographic pocket shapes in the principal component analysis
Figure S4. PCA scree plot for the aldose reductase dataset
Figure S5. PCA scree plots for the neuraminidase dataset
Figure S6. Pocket conformational distribution for the neuraminidase dataset
Figure S7. RMSD plots for the neuraminidase dataset
Table S8. SiteMap analysis for neuraminidase dataset
List S9. SMILES strings of 395 submicromolar aldose reductase inhibitors

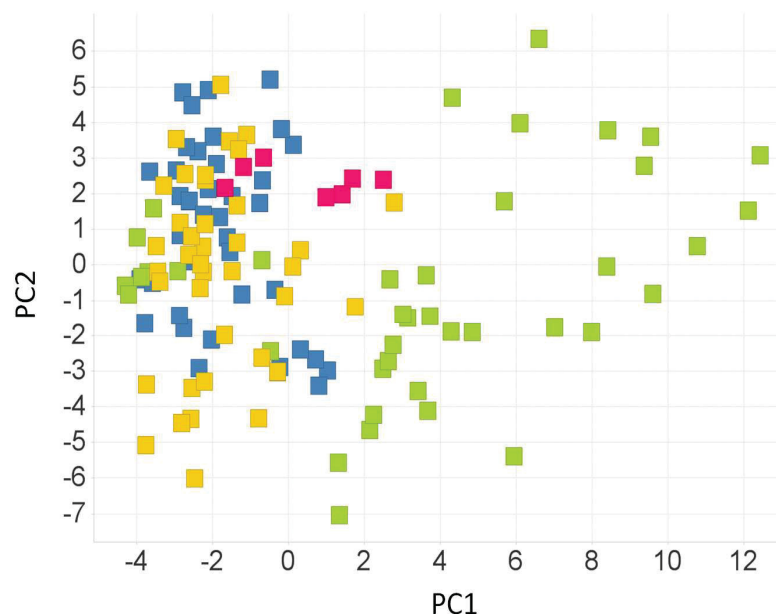
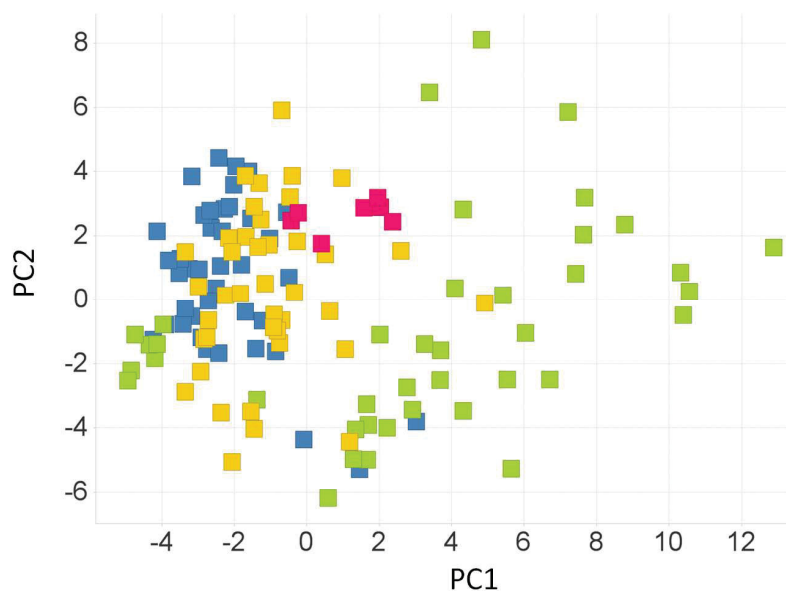


Figure S1. Pocket conformational distributions for the aldose reductase dataset along PC1 and PC2. Above: Protein structures aligned according to the C α positions of the set of active-site residues listed in the main text (this figure is identical to Figure 4 of the main text). Below: Protein structures aligned according to all C α positions. Other parameters as described in the main text. In both cases, pocket shapes derived from crystal structures are represented by red squares, whilst pocket shapes derived from the three trajectories MD1, MD2, and MD3 are coloured blue, green, and yellow respectively.

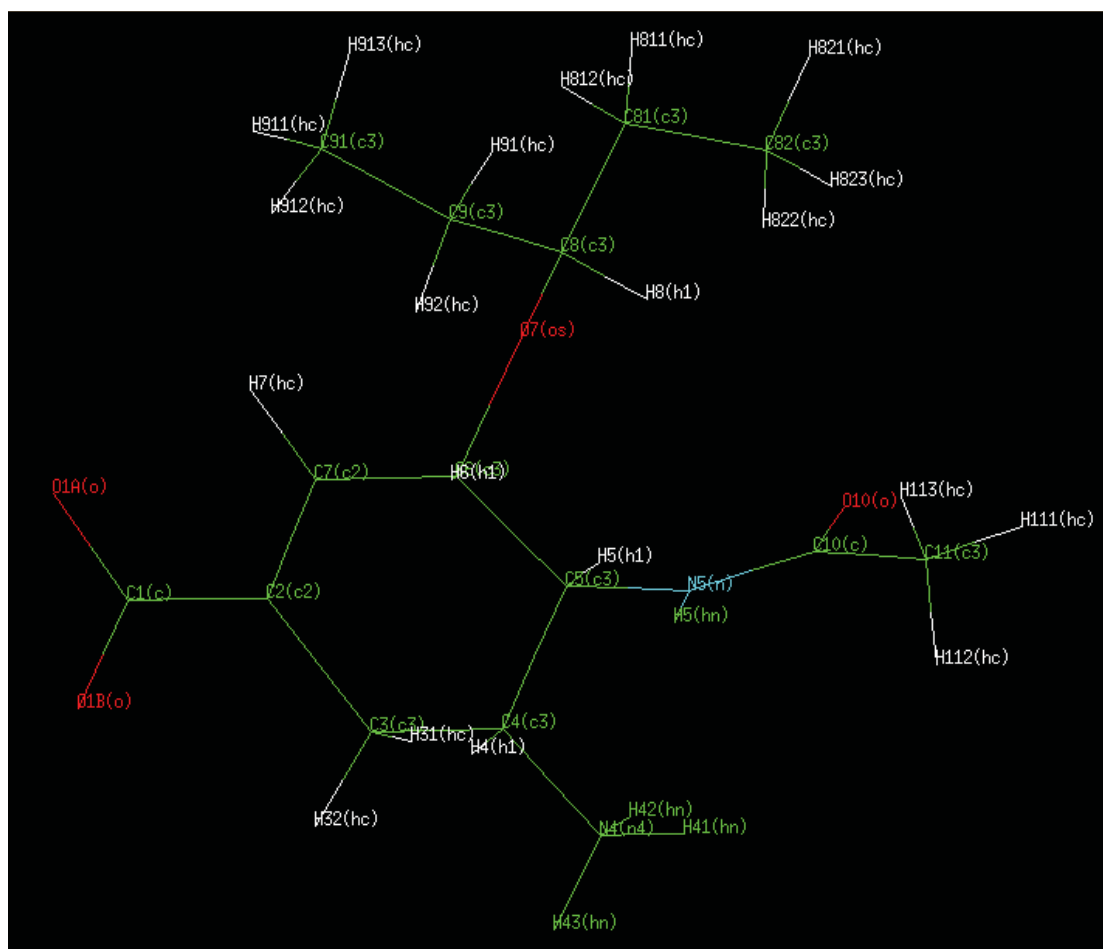


S2

Section S2. AMBER force field parameters for oseltamivir

Oseltamivir Parameterization:

Atom Name	gaff atom type	partial charge [e]
C1	c	0.326845
O1A	o	-0.590693
O1B	o	-0.590693
C2	c2	-0.154556
C3	c3	-0.078101
C4	c3	0.102548
C5	c3	0.036882
N5	n	-0.409405
C10	c	0.589654
O10	o	-0.555389
C11	c3	-0.124481
H111	hc	0.049439
H112	hc	0.049439
H113	hc	0.049439
H5	hn	0.256978
C6	c3	0.138702
C7	c2	-0.151065
H7	hc	0.111851
O7	os	-0.364698
C8	c3	0.113821
C9	c3	-0.100685
C91	c3	-0.144913
H911	hc	0.048477
H912	hc	0.048477
H913	hc	0.048477
H91	hc	0.051266
H92	hc	0.051266
C81	c3	-0.100685
C82	c3	-0.144913
H821	hc	0.048477
H822	hc	0.048477
H823	hc	0.048477
H811	hc	0.051266
H812	hc	0.051266
H8	h1	0.064078
H6	h1	0.065515
H5	h1	0.059495
N4	n4	-0.283955
H41	hn	0.371275
H42	hn	0.371275
H43	hn	0.371275
H4	h1	0.064443
H31	hc	0.052676
H32	hc	0.052676



Function and Bioinformatics **2010**, 78, 2523-2532). It has therefore been maintained in the simulation, using the parameterization of Aqvist adapted for the parm99 force field.

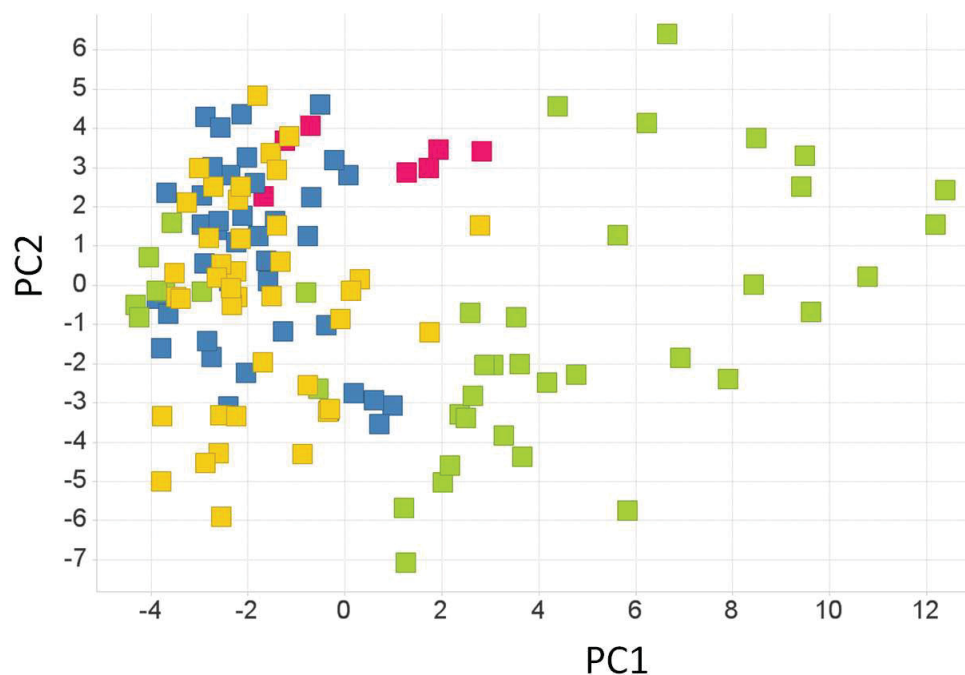


Figure S3. Pocket conformational distributions for the aldose reductase dataset along PC1 and PC2 when the crystallographic pocket shapes are included in the PCA. Other parameters as described in the main text. This figure should be compared to Figure 4 in the main text. In both cases, pocket shapes derived from crystal structures are represented by red squares, whilst pocket shapes derived from the three trajectories MD1, MD2, and MD3 are coloured blue, green, and yellow respectively.

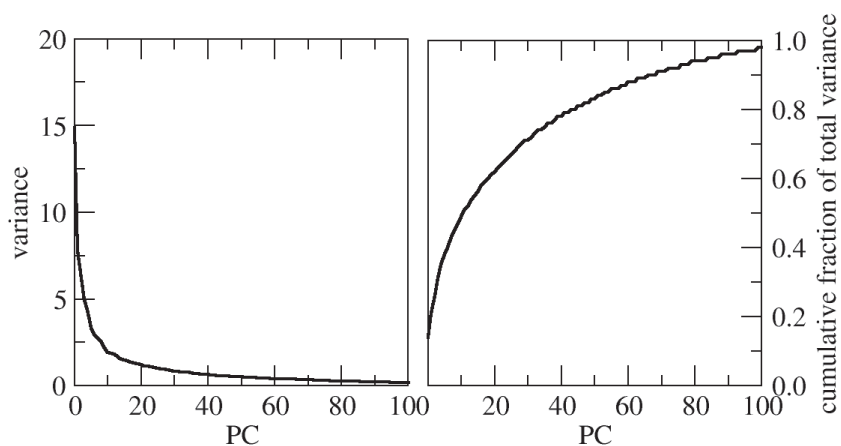


Figure S4. PCA scree plot for the aldose reductase dataset. The left panel plots the variance along each principal component (PC) as a function of PC index. The right panel plots the cumulative fraction of the total variance as a function of PC index.

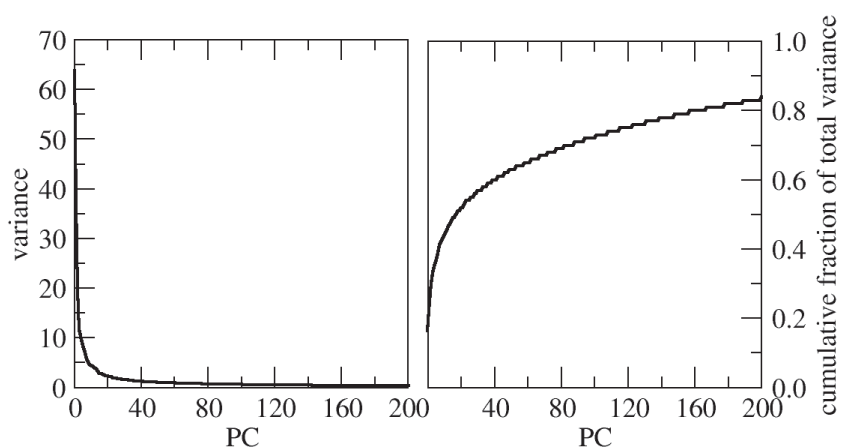


Figure S5. PCA scree plots for the neuraminidase dataset. The left panel plots the variance along each principal component (PC) as a function of PC index. The right panel plots the cumulative fraction of the total variance as a function of PC index.

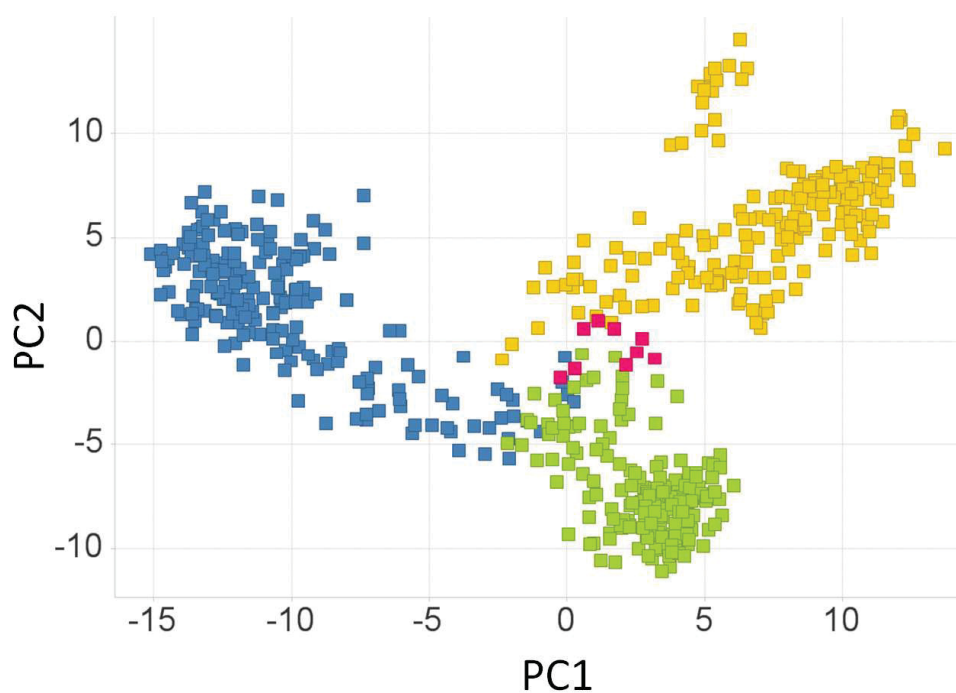


Figure S6. The pocket conformational distribution for the neuraminidase dataset along PC1 and PC2. Pocket shapes derived from crystal structures are represented by red squares. Pocket shapes derived from the *apo*, first *holo*, and second *holo* trajectories are coloured blue, green, and yellow respectively.

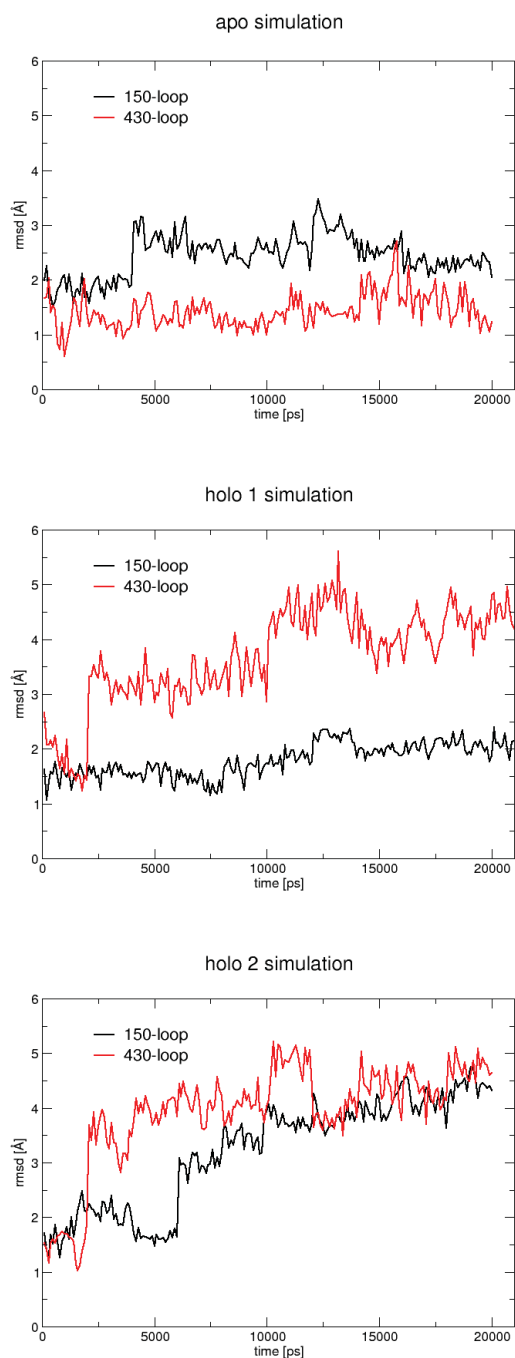


Figure S7. RMSD plots for the *apo* (upper), first *holo* (middle), and second *holo* (lower) neuraminidase trajectories. In each case, the black line shows the C α RMSD of the 150-loop (residues Asn146 to Pro154 inclusive) compared to the initial protein structure, and the red lines shows the C α RMSD of the 430-loop (residues Arg430 to Thr439 inclusive).

Table S8. SiteMap analysis for neuraminidase datasets. Upper: the ten cluster representatives derived from the MD simulations. Lower: the crystallographic protein structures.

Cluster representative	Source	DScore	Size^{a)}	Enclosure^{b)}	Hydrophilicity^{c)}
1	holo1	0.98	75	0.83	1.04
2	apo	1.00	115	0.75	1.19
3	apo	0.60	32	0.70	1.07
4	apo	0.92	86	0.66	1.05
5	holo1	1.08	127	0.81	1.07
6	holo1	1.05	117	0.75	1.05
7	holo2	0.76	78	0.79	1.65
8	holo2	1.05	123	0.79	1.13
9	holo2	1.05	176	0.68	0.90
10	holo2	1.01	214	0.74	1.14

Crystal structure	DScore	Size^{a)}	Enclosure^{b)}	Hydrophilicity^{c)}
2HTY	0.97	208	0.72	1.25
2HU4	0.96	145	0.80	1.42
2HU0	0.96	217	0.73	1.28
3B7E	1.00	120	0.77	1.23
3BEQ	0.96	176	0.75	1.32
3CKZ	0.98	162	0.76	1.28
3CL0	0.97	149	0.78	1.33
2HT7	0.97	216	0.76	1.29
2HT8	1.04	147	0.78	1.12

a) Number of grid-points; b) A ratio; c) In kcal mol⁻¹ (but see SiteMap reference for calibration)

List S9. BindingDB identifiers and SMILES strings of 395 submicromolar aldose reductase inhibitors

BindingDB ID	SMILES
BindingDB_7460	<chem>c1cc(c(cc1c2c(c(=O)c3c(cc(cc3o2)O)O)O)O)O</chem>
BindingDB_16233	<chem>c1cc(cc(c1)[N+](=O)[O-])c2nc(on2)CCCC(=O)[O-]</chem>
BindingDB_16239	<chem>c1cc(cc(c1)[N+](=O)[O-])CNC(=O)c2ccc(cc2OCC(=O)[O-])Cl</chem>
BindingDB_16240	<chem>c1ccc2c(c1)cc(c3c2C(=O)N(S3(=O)=O)CC(=O)[O-])C(=O)[O-]</chem>
BindingDB_16242	<chem>c1ccc2c(c1)cc(c3c2C(=O)N(S3(=O)=O)CC(=O)[O-])C(=O)OCC(=O)[O-]</chem>
BindingDB_16313	<chem>c1cc(c(cc1F)OCC(=O)[O-])C(=S)NCc2ccc(cc2F)Br</chem>
BindingDB_16314	<chem>CN(CC(=O)[O-])C(=S)c1cccc2c1ccc(c2C(F)(F)F)OC</chem>
BindingDB_16315	<chem>Cc1c2cc(ccc2oc1S(=O)(=O)c3ccc(=O)[nH]n3)Cl</chem>
BindingDB_16442	<chem>c1ccc(cc1)S(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16444	<chem>c1ccc(c(c1)S(=O)(=O)c2ccc(=O)[nH]n2)Cl</chem>
BindingDB_16445	<chem>c1ccc(c(c1)S(=O)(=O)c2ccc(=O)[nH]n2)Br</chem>
BindingDB_16449	<chem>c1cc(c(cc1Cl)Cl)S(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16450	<chem>c1cc(c(cc1F)Cl)S(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16451	<chem>c1cc(c(cc1Br)F)S(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16453	<chem>c1ccc2c(c1)nc(s2)S(=O)(=O)c3ccc(=O)[nH]n3</chem>
BindingDB_16455	<chem>c1ccc2c(c1)cc(s2)S(=O)(=O)c3ccc(=O)[nH]n3</chem>
BindingDB_16456	<chem>c1ccc2c(c1)cc(o2)S(=O)(=O)c3ccc(=O)[nH]n3</chem>
BindingDB_16457	<chem>c1cc2c(cc1Cl)cc(o2)S(=O)(=O)c3ccc(=O)[nH]n3</chem>
BindingDB_16458	<chem>c1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4cc(ccc4s3)C(F)(F)F</chem>
BindingDB_16459	<chem>Cc1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4cc(ccc4s3)C(F)(F)F</chem>
BindingDB_16460	<chem>c1cc2c(cc1C(F)(F)F)nc(s2)Cc3cn(c4c3cc(cc4)Cl)CC(=O)[O-]</chem>
BindingDB_16461	<chem>c1cc2c(cc1C(F)(F)F)nc(s2)Cc3cn(c4c3ccc(c4)Br)CC(=O)[O-]</chem>
BindingDB_16462	<chem>c1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4ccc(cc4s3)Cl</chem>
BindingDB_16463	<chem>c1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4c(cccc4s3)F</chem>
BindingDB_16464	<chem>c1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4cc(ccc4s3)F</chem>
BindingDB_16465	<chem>c1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4ccc(cc4s3)F</chem>
BindingDB_16466	<chem>c1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4cccc(c4s3)F</chem>
BindingDB_16467	<chem>c1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4cc(c(cc4s3)F)F</chem>
BindingDB_16468	<chem>c1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4c(cc(cc4s3)F)F</chem>
BindingDB_16469	<chem>c1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4c(c(cc(c4s3)F)F)F</chem>
BindingDB_16470	<chem>Cc1c(c2cccc2n1CC(=O)[O-])Cc3nc4c(c(cc(c4s3)F)F)F</chem>
BindingDB_16471	<chem>c1ccc(cc1)c2c(c3cccc3n2CC(=O)[O-])Cc4nc5c(c(cc(c5s4)F)F)F</chem>
BindingDB_16472	<chem>c1cc2c(c(c1)Cl)c(cn2CC(=O)[O-])Cc3nc4c(c(cc(c4s3)F)F)F</chem>
BindingDB_16473	<chem>c1cc2c(cc1Cl)c(cn2CC(=O)[O-])Cc3nc4c(c(cc(c4s3)F)F)F</chem>
BindingDB_16474	<chem>c1cc2c(cc1F)c(cn2CC(=O)[O-])Cc3nc4c(c(cc(c4s3)F)F)F</chem>
BindingDB_16475	<chem>c1cc2c(cc1Br)c(cn2CC(=O)[O-])Cc3nc4c(c(cc(c4s3)F)F)F</chem>
BindingDB_16476	<chem>Cc1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4c(c(cc(c4s3)F)F)F</chem>
BindingDB_16477	<chem>COc1ccc2c(c1)c(cn2CC(=O)[O-])Cc3nc4c(c(cc(c4s3)F)F)F</chem>
BindingDB_16478	<chem>c1ccc(cc1)COc2ccc3c(c2)c(cn3CC(=O)[O-])Cc4nc5c(c(cc(c5s4)F)F)F</chem>
BindingDB_16479	<chem>c1ccc(cc1)Oc2ccc3c(c2)c(cn3CC(=O)[O-])Cc4nc5c(c(cc(c5s4)F)F)F</chem>
BindingDB_16480	<chem>c1ccc(cc1)c2ccc3c(c2)c(cn3CC(=O)[O-])Cc4nc5c(c(cc(c5s4)F)F)F</chem>
BindingDB_16481	<chem>c1cc2c(cc1N3CCOCC3)c(cn2CC(=O)[O-])Cc4nc5c(c(cc(c5s4)F)F)F</chem>
BindingDB_16482	<chem>c1cc2c(cc1F)n(cc2Cc3nc4c(c(cc(c4s3)F)F)F)CC(=O)[O-]</chem>
BindingDB_16483	<chem>c1cc2c(cc1Cl)n(cc2Cc3nc4c(c(cc(c4s3)F)F)F)CC(=O)[O-]</chem>
BindingDB_16484	<chem>Cc1ccc2c(c1)n(cc2Cc3nc4c(c(cc(c4s3)F)F)F)CC(=O)[O-]</chem>
BindingDB_16485	<chem>COc1ccc2c(c1)n(cc2Cc3nc4c(c(cc(c4s3)F)F)F)CC(=O)[O-]</chem>
BindingDB_16486	<chem>c1ccc(cc1)c2ccc3c(c2)n(cc3Cc4nc5c(c(cc(c5s4)F)F)F)CC(=O)[O-]</chem>
BindingDB_16487	<chem>c1cc2c(cc1N3CCOCC3)n(cc2Cc4nc5c(c(cc(c5s4)F)F)F)CC(=O)[O-]</chem>
BindingDB_16488	<chem>c1cc2c(cn(c2c(c1)F)CC(=O)[O-])Cc3nc4c(c(cc(c4s3)F)F)F</chem>
BindingDB_16489	<chem>c1cc2c(cn(c2c(c1)Cl)CC(=O)[O-])Cc3nc4c(c(cc(c4s3)F)F)F</chem>

BindingDB_16490	<chem>c1cc2c(cnc2c(c1)Br)CC(=O)[O-]Cc3nc4c(c(cc4s3)F)F)F</chem>
BindingDB_16491	<chem>Cc1cccc2c1n(cc2Cc3nc4c(c(cc4s3)F)F)CC(=O)[O-]</chem>
BindingDB_16493	<chem>c1ccc2c(c1)c(cnc2CC(=O)[O-])CCc3nc4c(c(cc4s3)F)F)F</chem>
BindingDB_16494	<chem>c1ccc2c(c1)c(cnc2CCC(=O)[O-])Cc3nc4c(c(cc4s3)F)F)F</chem>
BindingDB_16495	<chem>CC(C(=O)[O-])n1cc(c2c1cccc2)Cc3nc4c(c(cc4s3)F)F)F</chem>
BindingDB_16496	<chem>c1cc2c(cc1Cl)n(c(=O)n(c2=O)Cc3ccc(cc3F)Br)CC(=O)[O-]</chem>
BindingDB_16512	<chem>c1cc2c(cc1F)C3(CC(O2)C(=O)N)C(=O)NC(=O)N3</chem>
BindingDB_16597	<chem>c1cc(ccc1S(=O)(=O)c2ccc(=O)[nH]n2)Br</chem>
BindingDB_16598	<chem>c1cc(cc(c1)Cl)S(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16599	<chem>c1cc(ccc1S(=O)(=O)c2ccc(=O)[nH]n2)Cl</chem>
BindingDB_16600	<chem>c1ccc(c(c1)F)S(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16601	<chem>c1cc(ccc1F)S(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16602	<chem>c1cc(c(c(c1)Cl)Cl)S(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16603	<chem>c1cc(c(cc1Cl)S(=O)(=O)c2ccc(=O)[nH]n2)Cl</chem>
BindingDB_16604	<chem>c1cc(c(c(c1)Cl)S(=O)(=O)c2ccc(=O)[nH]n2)Cl</chem>
BindingDB_16605	<chem>c1cc(c(c(c1)S(=O)(=O)c2ccc(=O)[nH]n2)F)F</chem>
BindingDB_16606	<chem>c1cc(c(cc1F)F)S(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16607	<chem>c1cc(cc(c1)S(=O)(=O)c2ccc(=O)[nH]n2)C(F)(F)F</chem>
BindingDB_16608	<chem>c1cc(ccc1C(F)(F)F)S(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16609	<chem>c1ccc2c(c1)cccc2Cc3cccc3S(=O)(=O)c4ccc(=O)[nH]n4</chem>
BindingDB_16610	<chem>c1ccc2cc(ccc2c1)Cc3cccc3S(=O)(=O)c4ccc(=O)[nH]n4</chem>
BindingDB_16611	<chem>c1ccc(cc1)CS(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16612	<chem>c1ccc(c(c1)CS(=O)(=O)c2ccc(=O)[nH]n2)Cl</chem>
BindingDB_16613	<chem>c1ccc(c(c1)CS(=O)(=O)c2ccc(=O)[nH]n2)F</chem>
BindingDB_16614	<chem>c1ccc(c(c1)CS(=O)(=O)c2ccc(=O)[nH]n2)C(F)(F)F</chem>
BindingDB_16615	<chem>c1cc(cc(c1)C(F)(F)F)CS(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16616	<chem>c1cc(c(c(c1)Cl)Cl)CS(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16617	<chem>c1cc(c(c(c1)Cl)CS(=O)(=O)c2ccc(=O)[nH]n2)Cl</chem>
BindingDB_16618	<chem>c1cc(c(c(c1)F)F)CS(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16619	<chem>c1cc(c(cc1Br)F)CS(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16620	<chem>c1cc(c(c(c1)Cl)CS(=O)(=O)c2ccc(=O)[nH]n2)F</chem>
BindingDB_16621	<chem>c1cc(c(c(c1)Cl)F)CS(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16622	<chem>c1cc(c(c(c1)C(F)(F)F)F)CS(=O)(=O)c2ccc(=O)[nH]n2</chem>
BindingDB_16624	<chem>c1cc(cc2c1cc([nH]2)S(=O)(=O)c3ccc(=O)[nH]n3)Cl</chem>
BindingDB_16625	<chem>c1cc2cc([nH]c2c(c1)Cl)S(=O)(=O)c3ccc(=O)[nH]n3</chem>
BindingDB_16628	<chem>c1cc(n[nH]c1=O)S(=O)(=O)c2cc3cc(cc3[nH]2)Cl)Cl</chem>
BindingDB_16629	<chem>Cc1c2cccc2oc1S(=O)(=O)c3ccc(=O)[nH]n3</chem>
BindingDB_16630	<chem>COc1ccc2c(c1)cc(o2)S(=O)(=O)c3ccc(=O)[nH]n3</chem>
BindingDB_16631	<chem>c1cc(n[nH]c1=O)S(=O)(=O)c2cc3cc(cc3o2)Cl)Cl</chem>
BindingDB_16632	<chem>Cc1c2ccc(cc2oc1S(=O)(=O)c3ccc(=O)[nH]n3)Cl</chem>
BindingDB_16633	<chem>Cc1c2cc(ccc2oc1S(=O)(=O)c3ccc(=O)[nH]n3)F</chem>
BindingDB_16634	<chem>Cc1c2cc(ccc2oc1S(=O)(=O)c3ccc(=O)[nH]n3)C(F)(F)F</chem>
BindingDB_16635	<chem>Cc1ccc2c(c1)c(c(o2)S(=O)(=O)c3ccc(=O)[nH]n3)C</chem>
BindingDB_16636	<chem>CCc1c2cc(ccc2oc1S(=O)(=O)c3ccc(=O)[nH]n3)Cl</chem>
BindingDB_16637	<chem>CC(C)c1c2cc(ccc2oc1S(=O)(=O)c3ccc(=O)[nH]n3)Cl</chem>
BindingDB_16638	<chem>c1ccc(cc1)c2c3cccc3oc2S(=O)(=O)c4ccc(=O)[nH]n4</chem>
BindingDB_16639	<chem>c1ccc2c(c1)c(c(o2)S(=O)(=O)c3ccc(=O)[nH]n3)c4ccc(cc4)F</chem>
BindingDB_16640	<chem>c1ccc(cc1)c2c3cc(ccc3oc2S(=O)(=O)c4ccc(=O)[nH]n4)Cl</chem>
BindingDB_16641	<chem>Cc1ccc2c(c1)cc(s2)S(=O)(=O)c3ccc(=O)[nH]n3</chem>
BindingDB_16642	<chem>Cc1c2cc(ccc2sc1S(=O)(=O)c3ccc(=O)[nH]n3)Cl</chem>
BindingDB_26193	<chem>c1ccc(c(c1)C(=O)[O-])O</chem>
BindingDB_50006353	<chem>CN1C(=O)Nc2cc(c(cc2C13C(=O)NC(=O)N3)[N+](=O)[O-])Cl</chem>
BindingDB_50006354	<chem>CN1C(=O)Nc2ccc(cc2C13C(=O)NC(=O)N3)[N+](=O)[O-]</chem>
BindingDB_50006355	<chem>CN1C(=O)Nc2cc(ccc2C13C(=O)NC(=O)N3)OC</chem>

BindingDB_50006356 Cc1cccc2c1C3(C(=O)NC(=O)N3)N(C(=O)N2)C
 BindingDB_50006357 CN1C(=O)Nc2c(cc(c(c2F)F)F)C13C(=O)NC(=O)N3
 BindingDB_50006358 CN1C(=O)Nc2c(cccc2F)C13C(=O)NC(=O)N3
 BindingDB_50006359 CN1C(=O)Nc2cccc2C13C(=O)NC(=S)N3
 BindingDB_50006361 CN1C(=O)Nc2ccc(cc2C13C(=O)NC(=O)N3)F
 BindingDB_50006362 Cc1cc2c(cc1Cl)C3(C(=O)NC(=O)N3)N(C(=O)N2)C
 BindingDB_50006363 CN1C(=O)Nc2cccc2C13C(=O)N(C(=O)N3)CC(=O)[O-]
 BindingDB_50006364 CN1C(=O)Nc2cccc2C13C(=O)NC(=O)N3
 BindingDB_50006365 c1ccc2c(c1)C3(C(=O)NC(=O)N3)N(C(=O)N2)O
 BindingDB_50006366 CC(C)CN1c2cccc2C3(C(=O)NC(=O)N3)N(C1=O)C
 BindingDB_50006367 CN1C(=O)Nc2cc(c(cc2C13C(=O)NC(=O)N3)[N+](=O)[O-])N4CCOCC4
 BindingDB_50006368 c1ccc(cc1)CN2C(=O)Nc3cccc3C24C(=O)NC(=O)N4
 BindingDB_50006369 CN1C(=O)Nc2ccc(cc2C13C(=O)NC(=O)N3)Cl
 BindingDB_50006370 CN1C(=O)Nc2cc(c(cc2C13C(=O)NC(=O)N3)Cl)Cl
 BindingDB_50006371 CN1C(=O)Nc2ccc(cc2C13C(=O)NC(=O)N3)N
 BindingDB_50006372 CN1C(=O)Nc2c(cc(cc2Cl)Cl)C13C(=O)NC(=O)N3
 BindingDB_50006373 CCCCN1C(=O)Nc2cccc2C13C(=O)NC(=O)N3
 BindingDB_50006374 CN1C(=O)Nc2ccc(c(c2C13C(=O)NC(=O)N3)Cl)[N+](=O)[O-]
 BindingDB_50006375 CN1C(=O)Nc2cc(c(cc2C13C(=O)NC(=O)N3)Cl)OC
 BindingDB_50006376 CN1C(=O)Nc2cccc2C13C(=O)NC(=O)N3
 BindingDB_50006377 CN1C(=O)Nc2c(cc(cc2N)Cl)C13C(=O)NC(=O)N3
 BindingDB_50006378 CN1C(=O)Nc2cccc2C13C(=O)NC(=O)N3C
 BindingDB_50006379 CC(C)N1C(=O)Nc2cccc2C13C(=O)NC(=O)N3
 BindingDB_50006380 CN1C(=O)Nc2cc(ccc2C13C(=O)NC(=O)N3)[N+](=O)[O-]
 BindingDB_50006382 CN1C(=O)Nc2ccc(cc2C13C(=O)NC(=O)N3)O
 BindingDB_50006384 CN1C(=O)Nc2cccc2C13C(=O)N=C(S3)N
 BindingDB_50006385 CN1C(=O)Nc2cccc2C13C(=O)NC(=[NH+])3N
 BindingDB_50006386 CN1c2cccc2C3(C(=O)NC(=O)N3)N(C1=O)C
 BindingDB_50006387 CN1C(=O)Nc2c(cc(cc2[N+](=O)[O-])[N+](=O)[O-])C13C(=O)NC(=O)N3
 BindingDB_50006388 c1ccc(cc1)N2C(=O)Nc3cccc3C24C(=O)NC(=O)N4
 BindingDB_50006389 CN1C(=O)Nc2cc(ccc2C13C(=O)NC(=O)N3)Cl
 BindingDB_50006390 Cc1ccc2c(c1)NC(=O)N(C23C(=O)NC(=O)N3)C
 BindingDB_50006391 CN1C(=O)Nc2ccc(cc2C13C(=O)NC(=O)N3)Cl
 BindingDB_50006392 CN1C(=O)Nc2c(cc(c(c2Cl)OC)Cl)C13C(=O)NC(=O)N3
 BindingDB_50006393 CCOC(=O)c1ccc2c(c1)C3(C(=O)NC(=O)N3)N(C(=O)N2)C
 BindingDB_50006394 Cc1ccc2c(c1)C3(C(=O)NC(=O)N3)N(C(=O)N2)C
 BindingDB_50006395 CCN1C(=O)Nc2cccc2C13C(=O)NC(=O)N3
 BindingDB_50006396 CN1C(=O)Nc2ccc(cc2C13C(=O)NC(=O)N3)F
 BindingDB_50006397 CN1C(=O)Nc2ccc(cc2C13C(=O)NC(=O)N3)Cl
 BindingDB_50006400 CN1C(=O)C2(c3cccc3NC(=O)N2C)NC1=O
 BindingDB_50006401 CN1C(=O)Nc2cccc(c2C13C(=O)NC(=O)N3)Cl
 BindingDB_50006402 CN1C(=O)C2(c3cccc3NC(=O)N2C)N(C1=O)C
 BindingDB_50006403 CN1C(=O)Nc2ccc(cc2C13C(=O)NC(=O)N3)Br
 BindingDB_50006404 CN1C(=O)Nc2c(cc(cc2[N+](=O)[O-])Cl)C13C(=O)NC(=O)N3
 BindingDB_50006458 c1cc2c(cc1C(F)(F)F)nc(s2)Cn3c4ccc(cc4c(n3)CC(=O)[O-])Cl
 BindingDB_50006459 c1c(cc(c2c1nc(s2)Cn3c(=O)c4c(c(n3)CC(=O)[O-])CCCC4)F)F
 BindingDB_50006461 c1cc(c(cc1Br)F)Cn2c(=O)c3c(c(n2)CC(=O)[O-])CCCC3
 BindingDB_50006462 c1ccc2cc3c(cc2c1)c(nn(c3=O)Cc4nc5cccc5s4)CC(=O)[O-]
 BindingDB_50006463 c1ccc2c(c1)c(nn(c2=O)Cc3nc4cccc4s3)CC(=O)[O-]
 BindingDB_50006464 Cc1cc(=O)n(nc1CC(=O)[O-])Cc2nc3cc(ccc3s2)C(F)(F)F
 BindingDB_50006465 c1ccc2c(c1)c(nn2Cc3nc4cc(cc4s3)F)F)CC(=O)[O-]
 BindingDB_50006466 Cc1c(c(=O)n(nc1CC(=O)[O-])Cc2ccc(cc2F)Br)C
 BindingDB_50006467 c1ccc2c(c1)c(nn2Cc3nc4cc(ccc4s3)C(F)(F)F)CC(=O)[O-]
 BindingDB_50006468 c1ccc2c(c1)nc(o2)Cn3c4ccc(cc4c(n3)CC(=O)[O-])Cl

BindingDB_50006469 c1cc2c(nn(c(=O)c2nc1)Cc3nc4cc(ccc4s3)C(F)(F)F)CC(=O)[O-]
BindingDB_50006470 c1cc2c(cc1F)nc(s2)Cn3c4ccc(cc4c(n3)CC(=O)[O-])Cl
BindingDB_50006471 c1cc2c(cc1C(F)(F)F)nc(s2)Cn3c(=O)c4c(c(n3)CC(=O)[O-])CCCC4
BindingDB_50006472 c1cc2c(cc1C(F)(F)F)nc(s2)Cn3c(=O)ccc(n3)CC(=O)[O-]
BindingDB_50006473 Cc1c(c(=O)n(nc1CC(=O)[O-])Cc2nc3cc(ccc3s2)C(F)(F)F)C
BindingDB_50006474 c1cc2c(c(nn(c2=O)Cc3nc4cc(ccc4s3)C(F)(F)F)CC(=O)[O-])nc1
BindingDB_50006475 Cc1c(c(=O)n(nc1CC(=O)[O-])Cc2nc(no2)c3cccc(c3F)F)C
BindingDB_50006476 c1cc2c(c(nn(c2=O)Cc3nc4cc(cc(c4s3)F)F)CC(=O)[O-])nc1
BindingDB_50006477 c1cc2c(cc1Cl)c(nn2Cc3nc4cc(cc(c4s3)F)F)CC(=O)[O-]
BindingDB_50006478 c1ccc2cc3c(cc2c1)c(nn(c3=O)Cc4ccc(cc4F)Br)CC(=O)[O-]
BindingDB_50006479 c1cc2c(nn(c(=O)c2nc1)Cc3ccc(cc3F)Br)CC(=O)[O-]
BindingDB_50006480 c1ccc2c(c1)c(nn2Cc3nc4cccc4s3)CC(=O)[O-]
BindingDB_50006481 Cc1c(c(=O)n(nc1CC(=O)[O-])Cc2nc3cc(cc(c3s2)F)F)C
BindingDB_50006482 c1cc2c(cc1Cl)c(nn2Cc3nc4cc(ccc4s3)Br)CC(=O)[O-]
BindingDB_50006483 c1ccc2c(c1)nc(s2)Cn3c4ccc(cc4c(n3)CC(=O)[O-])Cl
BindingDB_50006488 CCC1CC(c2cc(ccc2O1)F)(CC(=O)[O-])O
BindingDB_50006490 CC1CC(c2cc(c(cc2O1)Cl)F)(CC(=O)[O-])O
BindingDB_50006491 CCCC1CC(c2cc(ccc2O1)F)(CC(=O)[O-])O
BindingDB_50006493 CC1CC(c2ccc(cc2O1)C(=O)[O-])(CC(=O)[O-])O
BindingDB_50006496 CC1CC(c2cc(c(cc2O1)Br)Cl)(CC(=O)[O-])O
BindingDB_50006498 CC1CC(c2cc(c(cc2O1)F)F)(CC(=O)[O-])O
BindingDB_50006500 CC1CC(c2cc(c(cc2O1)OCc3cccc3)Cl)(CC(=O)[O-])O
BindingDB_50006503 CC(C)C1CC(c2cc(ccc2O1)F)(CC(=O)[O-])O
BindingDB_50006504 CC1CC(c2cc(c(cc2O1)Cl)[N+](=O)[O-])(CC(=O)[O-])O
BindingDB_50006506 CC1CC(c2cc(ccc2O1)F)(CC(=O)[O-])O
BindingDB_50006507 CC1CC(c2cc(c(cc2O1)Br)F)(CC(=O)[O-])O
BindingDB_50006508 CC(C)(C)C1CC(c2cc(ccc2O1)F)(CC(=O)[O-])O
BindingDB_50006509 CC1CC(c2cc(c(cc2O1)F)C#N)(CC(=O)[O-])O
BindingDB_50006510 CC1CC(c2ccc(c(cc2O1)Cl)Cl)(CC(=O)[O-])O
BindingDB_50006512 CC1CC(c2cc(ccc2O1)Cl)(CC(=O)[O-])O
BindingDB_50006513 CC1CC(c2cc(c(cc2O1)OC)Cl)(CC(=O)[O-])O
BindingDB_50006519 CC1CC(c2cc(c(cc2O1)F)Cl)(CC(=O)[O-])O
BindingDB_50006521 CC1CC(c2cc(cc(cc2O1)Cl)Cl)(CC(=O)[O-])O
BindingDB_50006523 c1cc2c(cc1F)C(CCO2)(CC(=O)[O-])O
BindingDB_50006524 Cc1cc2c(cc1Cl)C(CC(O2)C)(CC(=O)[O-])O
BindingDB_50006529 CCc1cc2c(cc1Cl)C(CC(O2)C)(CC(=O)[O-])O
BindingDB_50006530 CC1CC(c2cc(c(cc2O1)F)[N+](=O)[O-])(CC(=O)[O-])O
BindingDB_50006532 CC1CC(c2cc(ccc2O1)[N+](=O)[O-])(CC(=O)[O-])O
BindingDB_50006533 Cc1cc2c(cc1[N+](=O)[O-])C(CC(O2)C)(CC(=O)[O-])O
BindingDB_50006534 CC1(CC(c2cc(ccc2O1)[N+](=O)[O-])(CC(=O)[O-])O)C
BindingDB_50006535 CC1CC(c2cc(c3cccc3c2O1)Cl)(CC(=O)[O-])O
BindingDB_50006536 CC1CC(c2cc(ccc2O1)C(=O)C)(CC(=O)[O-])O
BindingDB_50006538 CC1C(Oc2ccc(cc2C1(CC(=O)[O-])O)F)C
BindingDB_50008435 c1ccc2c(c1)c(nn(c2=O)Cc3cc4cc(ccc4s3)F)CC(=O)[O-]
BindingDB_50008436 c1ccc2c(c1)c(nn(c2=O)C/C(=N/c3cc(ccc3Br)C(F)(F)F)/S)CC(=O)[O-]
BindingDB_50008438 c1ccc2c(c1)c(nn(c2=O)Cc3nc4cccc4o3)CC(=O)[O-]
BindingDB_50008440 c1ccc2c(c1)c(nn(c2=O)CC(=O)Nc3cc(cc(c3O)F)F)CC(=O)[O-]
BindingDB_50008441 Cc1cccc1c2nc(on2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
BindingDB_50008442 c1ccc2c(c1)c(nn(c2=O)Cc3nc4cc(cc(c4o3)Cl)Cl)CC(=O)[O-]
BindingDB_50008443 c1ccc2c(c1)c(nn(c2=O)Cc3cc4cc(ccc4s3)Br)CC(=O)[O-]
BindingDB_50008444 c1ccc2c(c1)c(nn(c2=O)C/C(=N/c3cccc(c3)C(F)(F)F)/S)CC(=O)[O-]
BindingDB_50008446 c1ccc2c(c1)c(nn(c2=O)Cc3nc4ccc(cc4o3)Br)CC(=O)[O-]
BindingDB_50008448 c1ccc2c(c1)c(nn(c2=O)Cc3nc4cc(ccc4o3)Cl)CC(=O)[O-]
BindingDB_50008449 c1ccc2c(c1)c(nn(c2=O)CC(=O)Nc3cccc(c3)F)CC(=O)[O-]

BindingDB_50008450 c1ccc2c(c1)c(nnc2=O)Cc3c4cccc4ns3)CC(=O)[O-]
 BindingDB_50008452 c1ccc2c(c1)cc([nH]2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50008453 c1ccc2c(c1)c(nnc2=O)Cc3cc4c(s3)cccc4Cl)CC(=O)[O-]
 BindingDB_50008455 c1ccc2c(c1)c(nnc2=O)Cc3cc4ccccc4s3)CC(=O)[O-]
 BindingDB_50008456 c1ccc2c(c1)c(nnc2=O)CC(=O)Nc3cccc(c3)Cl)CC(=O)[O-]
 BindingDB_50008457 c1ccc2c(c1)c(nnc2=O)C/C(=N/c3cccc(c3)Cl)/S)CC(=O)[O-]
 BindingDB_50008460 c1ccc2c(c1)c(nnc2=O)Cc3csc(n3)c4cccc4F)CC(=O)[O-]
 BindingDB_50008461 c1ccc2c(c1)c(nnc2=O)Cc3nc(n3)c4cccc4C(F)(F)F)CC(=O)[O-]
 BindingDB_50008464 c1ccc(cc1)c2nc(on2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50008465 COc1cccc1c2nc(on2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50008468 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(ccc4o3)Br)CC(=O)[O-]
 BindingDB_50008469 c1ccc2c(c1)c(nnc2=O)CC(=O)Nc3cc(ccc3O)C(F)(F)F)CC(=O)[O-]
 BindingDB_50008470 c1ccc2c(c1)c(nnc2=O)Cc3nc(n3)c4cccc4)CC(=O)[O-]
 BindingDB_50008471 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(cc(c4o3)F)F)CC(=O)[O-]
 BindingDB_50008473 c1ccc(cc1)c2cnc(o2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50008475 c1ccc2c(c1)c(nnc2=O)CC(=O)Nc3cc(ccc3Br)C(F)(F)F)CC(=O)[O-]
 BindingDB_50008476 c1ccc2c(c1)c(nnc2=O)Cc3cc4cc(ccc4s3)Cl)CC(=O)[O-]
 BindingDB_50008477 c1ccc(cc1)c2cc(sn2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50008478 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(ccc4o3)C(F)(F)F)CC(=O)[O-]
 BindingDB_50008479 c1ccc2c(c1)c(nnc2=O)Cc3nc(n3)c4cccc(c4F)F)CC(=O)[O-]
 BindingDB_50008480 c1ccc2c(c1)c(nnc2=O)Cc3nc(n3)c4cccc4Br)CC(=O)[O-]
 BindingDB_50008481 c1ccc2c(c1)c(nnc2=O)CC(=O)Nc3ccc(cc3)Cl)CC(=O)[O-]
 BindingDB_50008482 c1ccc2c(c1)c(nnc2=O)Cc3nc(n3)c4cccc4Cl)CC(=O)[O-]
 BindingDB_50008483 c1ccc2c(c1)c(nnc2=O)Cc3nc(n3)c4c(cccc4Cl)F)CC(=O)[O-]
 BindingDB_50008484 c1ccc2c(c1)c(nnc2=O)CC(=O)Nc3cccc(c3)C(F)(F)F)CC(=O)[O-]
 BindingDB_50008486 c1ccc(cc1)c2csc(n2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50008489 c1ccc2c(c1)cc(s2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50008490 c1ccc2c(c1)c(nnc2=O)CC(=O)Nc3cccc3Cl)CC(=O)[O-]
 BindingDB_50008491 c1ccc2c(c1)c(nnc2=O)Cc3nc(n3)c4ccc(cc4F)F)CC(=O)[O-]
 BindingDB_50008492 c1ccc2c(c1)c(nnc2=O)Cc3cn4ccc(cc4n3)Cl)CC(=O)[O-]
 BindingDB_50008493 c1ccc2c(c1)c(nnc2=O)Cc3cc4cc(ccc4o3)Cl)CC(=O)[O-]
 BindingDB_50009748 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(c(cc4s3)F)Cl)CC(=O)[O-]
 BindingDB_50009749 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(cc(c4s3)Cl)F)CC(=O)[O-]
 BindingDB_50009750 c1ccc2c(c1)c(nnc2=O)Cc3nc4c(ccc(c4s3)F)F)CC(=O)[O-]
 BindingDB_50009751 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(cc(c4s3)Cl)C(F)(F)F)CC(=O)[O-]
 BindingDB_50009752 c1ccc2c(c1)nc(s2)Cn3c(=O)c4c(cccc4F)c(n3)CC(=O)[O-]
 BindingDB_50009753 c1ccc2c(c1)c(nnc2=O)CSc3nc4cccc4s3)CC(=O)[O-]
 BindingDB_50009754 c1ccc2c(c1)c(nnc2=O)Cc3nc4ccc(cc4s3)Cl)CC(=O)[O-]
 BindingDB_50009755 c1ccc2c(c1)c(nnc2=O)Cc3nc4cccc(c4s3)Br)CC(=O)[O-]
 BindingDB_50009756 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(ccc4s3)O)CC(=O)[O-]
 BindingDB_50009757 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(cc(c4s3)F)F)CC(=O)[O-]
 BindingDB_50009758 c1ccc2c(c1)c(nnc2=O)Cc3nc4c(cccc4s3)O)CC(=O)[O-]
 BindingDB_50009759 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(cc(c4s3)F)Cl)CC(=O)[O-]
 BindingDB_50009761 COc1ccc2c(c1)c(=O)n(nc2CC(=O)[O-])Cc3nc4cccc4s3
 BindingDB_50009762 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(cc(c4s3)C(F)(F)F)Cl)CC(=O)[O-]
 BindingDB_50009763 c1ccc2c(c1)c(nnc2=O)Cc3nc4c(cccc4s3)F)CC(=O)[O-]
 BindingDB_50009764 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(c(cc4s3)F)F)CC(=O)[O-]
 BindingDB_50009766 c1ccc2c(c1)c(nnc2=O)Cc3nc4c(s3)ccc(c4Cl)Cl)CC(=O)[O-]
 BindingDB_50009767 c1ccc2c(c1)c(nnc2=O)Cc3nc4ccc(cc4s3)C(F)(F)F)CC(=O)[O-]
 BindingDB_50009768 c1ccc2c(c1)ccc3c2nc(s3)Cn4c(=O)c5ccc(cc5c(n4)CC(=O)[O-])Cl
 BindingDB_50009769 c1ccc2c(c1)c(nnc2=O)Cc3nc4ccccc4s3)CC(=O)[O-]
 BindingDB_50009770 Cc1ccc2c(c1)sc(n2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50009771 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(cc(c4s3)O)O)CC(=O)[O-]
 BindingDB_50009772 COc1cc2c(c1)OC)sc(n2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]

BindingDB_50009773 c1ccc2c(c1)nc(s2)Cn3c(=O)c4ccc(cc4c(n3)CC(=O)[O-])C(F)(F)F
 BindingDB_50009774 Cc1cc(c2c(c1)nc(s2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-])C
 BindingDB_50009775 c1ccc2c(c1)nc(s2)Cn3c(=O)c4cc(ccc4c(n3)CC(=O)[O-])C(F)(F)F
 BindingDB_50009776 c1ccc2c(c1)nc(s2)Cn3c(=O)c4ccc(cc4c(n3)CC(=O)[O-])Br
 BindingDB_50009777 c1ccc2c(c1)c(nnc2=O)Cc3ccc(cc3F)Br)CC(=O)[O-]
 BindingDB_50009778 c1ccc2c(c1)c(nnc2=O)Cc3ccc(cc3)c4nc5cccc5s4)CC(=O)[O-]
 BindingDB_50009779 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(c(cc4s3)Cl)Cl)CC(=O)[O-]
 BindingDB_50009780 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(cc(c4s3)F)C(F)(F)F)CC(=O)[O-]
 BindingDB_50009781 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(ccc4s3)Br)CC(=O)[O-]
 BindingDB_50009782 CSc1ccc2c(c1)nc(s2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50009783 c1ccc2c(c1)nc(s2)Cn3c(=O)c4cc(c(cc4c(n3)CC(=O)[O-])Cl)Cl
 BindingDB_50009784 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(cc(c4s3)Cl)Cl)CC(=O)[O-]
 BindingDB_50009785 c1ccc2c(c1)nc(s2)Cn3c(=O)c4cc(ccc4c(n3)CC(=O)[O-])[N+](=O)[O-]
 BindingDB_50009786 c1ccc2c(c1)nc(s2)Cn3c(=O)c4cc(ccc4c(n3)CC(=O)[O-])Br
 BindingDB_50009787 COc1ccc2c(c1)nc(s2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50009788 c1ccc2c(c1)nc(s2)CN3CC4(c5c3ccc(c5)Cl)C(=O)NC(=O)N4
 BindingDB_50009790 c1cc(ccc1CN2CC3(c4c2ccc(c4)F)C(=O)NC(=O)N3)F
 BindingDB_50009791 COc1cccc2c1nc(s2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50009792 Cc1ccc2c(c1)nc(s2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50009793 c1ccc2c(c1)c(nnc2=O)Cc3nc4c(ccc4s3)C(F)(F)F)CC(=O)[O-]
 BindingDB_50009794 Cc1cccc2c1c(=O)n(nc2CC(=O)[O-])Cc3nc4c5cccc5ccc4s3
 BindingDB_50009795 Cc1ccc2c(c1F)sc(n2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50009796 c1ccc2c(c1)nc(s2)Cn3c(=O)c4cc(ccc4c(n3)CC(=O)[O-])Cl
 BindingDB_50009797 Cc1cccc2c1c(=O)n(nc2CC(=O)[O-])Cc3nc4cccc4s3
 BindingDB_50009798 Cc1ccc2c(c1)c(nnc2=O)Cc3nc4cccc4s3)CC(=O)[O-]
 BindingDB_50009800 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(c(cc4s3)O)C(F)(F)F)CC(=O)[O-]
 BindingDB_50009801 c1ccc2c(c1)c(nnc2=O)Cc3nc4c(ccc(c4s3)Cl)Cl)CC(=O)[O-]
 BindingDB_50009802 CC(c1nc2cc(ccc2s1)C(F)(F)F)n3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50009803 c1ccc2c(c1)nc(s2)Cn3c(=O)c4ccc(cc4c(n3)CC(=O)[O-])Cl
 BindingDB_50009804 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(ccc4s3)F)CC(=O)[O-]
 BindingDB_50009805 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(cc(c4s3)C(F)(F)F)CC(=O)[O-]
 BindingDB_50009806 c1ccc2c(c1)c(nnc2=O)Cc3nc4cccc(c4s3)Cl)CC(=O)[O-]
 BindingDB_50009810 c1ccc2c(c1)c(nnc2=O)Cc3nc4c(s3)ccc(c4F)F)CC(=O)[O-]
 BindingDB_50009811 c1ccc2c(c1)ccc3c2nc(s3)Cn4c(=O)c5cccc5c(n4)CC(=O)[O-]
 BindingDB_50009813 CC(c1nc2cc(ccc2s1)Cl)n3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50009814 c1ccc2c(c1)c(nnc2=O)Cc3nc4cccc(c4s3)C(F)(F)F)CC(=O)[O-]
 BindingDB_50009816 CC(c1nc2cccc2s1)n3c(=O)c4cccc4c(n3)CC(=O)[O-]
 BindingDB_50009817 c1ccc2c(c1)c(nnc2=O)Cc3nc4ccc(cc4s3)Br)CC(=O)[O-]
 BindingDB_50009818 CC(C)c1ccc2c(c1)c(=O)n(nc2CC(=O)[O-])Cc3nc4cccc4s3
 BindingDB_50009819 c1ccc2c(c1)c(nnc2=O)CCc3nc4cc(ccc4s3)C(F)(F)F)CC(=O)[O-]
 BindingDB_50009820 COc1cc(c2c(c1)nc(s2)Cn3c(=O)c4cccc4c(n3)CC(=O)[O-])C(F)(F)F
 BindingDB_50009821 CC(C)c1ccc2c(c1)c(nnc2=O)Cc3nc4cccc4s3)CC(=O)[O-]
 BindingDB_50009822 c1ccc2c(c1)c(nnc2=O)c3nc4cccc4s3)CC(=O)[O-]
 BindingDB_50009823 COc1ccc2c(c1)c(nnc2=O)Cc3nc4cccc4s3)CC(=O)[O-]
 BindingDB_50009824 c1ccc2c(c1)nc(s2)Cn3c(=O)c4ccc(cc4c(n3)CC(=O)[O-])[N+](=O)[O-]
 BindingDB_50009826 c1ccc2c(c1)c(nnc2=O)Cc3nc4ccc(cc4s3)O)CC(=O)[O-]
 BindingDB_50009827 c1ccc2c(c1)c(nnc2=O)Cc3nc4c(s3)cc(cc4Cl)Cl)CC(=O)[O-]
 BindingDB_50009828 c1ccc2c(c1)c(nnc2=O)Cc3nc4c(s3)cccc4Cl)CC(=O)[O-]
 BindingDB_50009829 c1ccc2c(c1)c(nnc2=O)Cc3nc4cc(ccc4s3)Cl)CC(=O)[O-]
 BindingDB_50009830 CCOC(=O)C1=C(C(=O)N(C1)Cc2cccn2)O
 BindingDB_50009831 CCOC(=O)C1=C(C(=O)N(C1)Cc2ccc(cc2)Br)O
 BindingDB_50009832 CCOC(=O)C1=C(C(=O)N(C1)Cc2cccc2)O
 BindingDB_50009833 CCOC(=O)C1=C(C(=O)N(C1)C2c3cccc3-c4c2cccc4)O
 BindingDB_50009835 CCOC(=O)C1=C(C(=O)N(C1)Cc2ccncc2)O

BindingDB_50009837 CCOC(=O)C1=C(C(=O)N(C1)Cc2cccs2)O
 BindingDB_50009841 C/C(=C\c1ccccc1)/C=C/2\C(=O)N(C(=S)S2)CC(=O)[O-]
 BindingDB_50009845 CCOC(=O)C1=C(C(=O)N(C1)Cc2ccc(c(c2)Cl)Cl)O
 BindingDB_50009848 CCOC(=O)C1=C(C(=O)N(C1)C(C)C2ccccc2)O
 BindingDB_50009849 CCOC(=O)C1=C(C(=O)N(C1)Cc2ccc(cc2)Cl)O
 BindingDB_50009851 CCOC(=O)C1=C(C(=O)N(C1)Cc2ccc(cc2)OC)O
 BindingDB_50010279 c1cc2c(cc1F)C3(CCO2)C(=O)NC(=O)N3
 BindingDB_50010283 c1cc-2c(cc1F)C3(c4c2ccc(c4)F)C(=O)NC(=O)N3
 BindingDB_50012759 c1c(cnc2c1C3(CCO2)C(=O)NC(=O)N3)F
 BindingDB_50012761 CC1CC2(c3cc(cnc3O1)Br)C(=O)NC(=O)N2
 BindingDB_50012762 CC1CC2(c3cc(c[n+](c3O1)[O-])Cl)C(=O)NC(=O)N2
 BindingDB_50012764 CC1CC2(c3cc(cnc3O1)Cl)C(=O)NC(=O)N2
 BindingDB_50012767 CC1CC2(c3cc(cnc3O1)Cl)C(=O)NC(=O)N2
 BindingDB_50016628 CS(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)F
 BindingDB_50016629 c1ccc(cc1)CCCNS(=O)(=O)c2ccc(cc2C3C(=O)NC(=O)N3)Cl
 BindingDB_50016630 CCCCCS(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016631 CS(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016632 c1cc(ccc1NS(=O)(=O)c2ccc(cc2C3C(=O)NC(=O)N3)Cl)F
 BindingDB_50016633 CCCCCNS(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016634 CS(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016635 c1cc(ccc1CS(=O)c2ccc(cc2C3C(=O)NC(=O)N3)Cl)Cl
 BindingDB_50016637 Cc1cc(cc(c1S(=O)(=O)C)C2C(=O)NC(=O)N2)Cl
 BindingDB_50016638 CCCS(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016640 CNS(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016641 CS(=O)(=O)c1c(cc(cc1C(F)(F)F)Cl)C2C(=O)NC(=O)N2
 BindingDB_50016642 c1cc(c(cc1Cl)C2C(=O)NC(=O)N2)S(=O)(=O)N
 BindingDB_50016644 CCS(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016645 c1cc(ccc1CS(=O)(=O)c2ccc(cc2C3C(=O)NC(=O)N3)Cl)Cl
 BindingDB_50016646 CCCCCS(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016647 CCCNS(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016649 CSc1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016652 c1cc(ccc1CCS(=O)(=O)c2ccc(cc2C3C(=O)NC(=O)N3)Cl)Cl
 BindingDB_50016653 c1ccc(cc1)CNS(=O)(=O)c2ccc(cc2C3C(=O)NC(=O)N3)Cl
 BindingDB_50016654 CCCCNS(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016655 CC(C)S(=O)(=O)c1ccc(cc1C2C(=O)NC(=O)N2)Cl
 BindingDB_50016656 CS(=O)(=O)c1c(cc(cc1Cl)Cl)C2C(=O)NC(=O)N2
 BindingDB_50022260 c1ccc2c(c1)CCc3c(ccc4c3OC(=O)C4)C2=O
 BindingDB_50022263 c1ccc2c(c1)ccc3cc(cc(c3c2=O)O)CC(=O)[O-]
 BindingDB_50022264 c1cc2c(cc1CC(=O)[O-])CCc3ccc(cc3C2=O)O
 BindingDB_50022265 c1ccc2c(c1)COc3cc(ccc3C2=O)CC(=O)[O-]
 BindingDB_50022268 c1ccc2c(c1)ccc3cc(ccc3C2=O)CC(=O)[O-]
 BindingDB_50022273 CC(c1cc2c(c(c1)O)C(=O)c3ccccc3CC2)C(=O)[O-]
 BindingDB_50022274 c1ccc2c(c1)CSc3cc(ccc3C2=O)CC(=O)[O-]
 BindingDB_50022391 c1cc2c(cc1Br)C3(CCS2)C(=O)NC(=O)N3
 BindingDB_50022398 c1cc2c(cc1Cl)C3(CCO2)C(=O)NC(=O)O3
 BindingDB_50022401 c1ccc(cc1)c2cccc3c2OCCC34C(=O)NC(=O)N4
 BindingDB_50022407 c1cc2c(cc1F)C3(CCS2(=O)=O)C(=O)NC(=O)N3
 BindingDB_50022419 CN1C(=O)C2(CCOc3c2cc(cc3)F)NC1=O
 BindingDB_50022434 c1cc2c(ccc3c2OCCC34C(=O)NC(=O)N4)nc1
 BindingDB_50022435 c1cc2c(cc1F)C3(CCS2)C(=O)NC(=O)N3
 BindingDB_50022436 c1cc2ccc(=O)n3c2c(c1)C4(CC3)C(=O)NC(=O)N4
 BindingDB_50022458 c1cc2c(cc1Cl)C3(CCS2)C(=O)NC(=O)N3
 BindingDB_50022459 c1cc2c(cc1F)C3(CCS2=O)C(=O)NC(=O)N3
 BindingDB_50022473 c1cc2c(cc1Br)C3(CCO2)C(=O)NC(=O)N3

BindingDB_50022481 CN1C(=O)C2(CCOc3c2cc(cc3)F)N(C1=O)C
 BindingDB_50038839 c1ccc2c(c1)C(=O)N(C(=O)C23CC(=O)NC3=O)Cc4ccc(cc4F)Br
 BindingDB_50038843 c1cc2c(cc1F)C3(CC(=O)NC3=O)C(=O)N(C2=O)Cc4ccc(cc4F)Br
 BindingDB_50038846 c1cc2c(cc1C(F)(F)F)nc(s2)CN3C(=O)c4ccc(cc4C5(C3=O)CC(=O)NC5=O)F
 BindingDB_50038847 c1cc2c(cc1F)C3(CC(=O)NC3=O)C(=O)N(C2=O)Cc4ccc(cc4F)Br
 BindingDB_50067397 c1cc2n(c1)C3(CC(=O)NC3=O)C(=O)N(C2=O)Cc4ccc(cc4F)Br
 BindingDB_50067407 c1cc2n(c1)C3(CC(=O)NC3=O)C(=O)N(C2=O)Cc4ccc(cc4F)Br
 BindingDB_50086037 CN1C(=O)c2ccc(cc2C3(C1=O)CC(=O)NC3=O)F
 BindingDB_50086038 CN1C(=O)c2ccc(cc2C3(C1=O)CC(=O)NC3=O)Cl
 BindingDB_50086039 c1cc2c(cc1F)C(C(=O)N(C2=O)Cc3ccc(cc3F)Br)CC(=O)[O-]
 BindingDB_50089423 c1cc2c(cc1F)C3(CC(O2)C(=O)N)C(=O)NC(=O)N3
 BindingDB_50222609 c1cc2c(cc1C(F)(F)F)nc(s2)CNC(=O)c3ccc(cc3OCC(=O)[O-])Cl
 BindingDB_50222610 c1ccc(c(c1)C(=O)NCc2ccc(cc2F)Br)OCC(=O)[O-]
 BindingDB_50222611 c1cc(c(cc1F)OCC(=O)[O-])C(=O)NCc2nc3c(c(cc(c3s2)F)F)F
 BindingDB_50222612 Cc1cccc1CC(=O)Nc2cc(c(c(c2)C)S(=O)(=O)NCC(=O)[O-])C
 BindingDB_50222613 c1cc(cc(c1)[N+](=O)[O-])CNC(=O)c2cc(c(cc2OCC(=O)[O-])F)C#N
 BindingDB_50222614 c1cc2c(cc1C(F)(F)F)nc(s2)CNC(=O)c3ccc(cc3OCC(=O)[O-])F
 BindingDB_50240584 c1cc(=O)n(nc1CC(=O)[O-])Cc2nc3cc(cc(c3s2)F)F
 BindingDB_50240587 c1cc2c(c(nn(c2=O)Cc3nc4cc(ccc4s3)F)CC(=O)[O-])nc1
 BindingDB_50240837 c1ccc2c(c1)c(nn(c2=O)Cc3nc(no3)c4cccc4F)CC(=O)[O-]
 BindingDB_50270640 c1cc(ccc1/C=C\2/C(=O)N(C(=O)S2)CC(=O)[O-])O

13.11 Publication VI

Hot Spots and Transient Pockets: Predicting the Determinants of Small-Molecule Binding to a Protein-Protein Interface

§Metz, A., §Pfleger, C., Kopitz, H., Pfeiffer-Marek, S., Baringhaus, K.-H., Gohlke, H.

J. Chem. Inf. Model. (2012), 52, 120-133

Author contribution to the publications:

My contribution to this publication was (I) running constrained geometrical FRODA simulations, (II) developing the PPIAnalyzer method to investigate structural features of protein-protein interfaces, and (III) running docking experiment of known IL-2 ligands into representative protein conformations. This results in a contribution of **35%** to this publication.


§ Both authors contributed equally to this work.

Hot Spots and Transient Pockets: Predicting the Determinants of Small-Molecule Binding to a Protein–Protein Interface

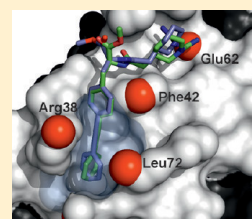
Alexander Metz,^{†,§} Christopher Pflieger,^{†,§} Hannes Kopitz,[†] Stefania Pfeiffer-Marek,[‡] Karl-Heinz Baringhaus,[‡] and Holger Gohlke^{*,†}

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

[‡]Sanofi-Aventis Deutschland GmbH, LGCR Drug Design, Frankfurt am Main, Germany

 Supporting Information

ABSTRACT: Protein–protein interfaces are considered difficult targets for small-molecule protein–protein interaction modulators (PPIMs). Here, we present for the first time a computational strategy that simultaneously considers aspects of energetics and plasticity in the context of PPIM binding to a protein interface. The strategy aims at identifying the determinants of small-molecule binding, hot spots, and transient pockets, in a protein–protein interface in order to make use of this knowledge for predicting binding modes of and ranking PPIMs with respect to their affinity. When applied to interleukin-2 (IL-2), the computationally inexpensive constrained geometric simulation method FRODA outperforms molecular dynamics simulations in sampling hydrophobic transient pockets. We introduce the PPIAnalyzer approach for identifying transient pockets on the basis of geometrical criteria only. A sequence of docking to identified transient pockets, starting structure selection based on hot spot information, RMSD clustering and intermolecular docking energies, and MM-PBSA calculations allows one to enrich IL-2 PPIMs from a set of decoys and to discriminate between subgroups of IL-2 PPIMs with low and high affinity. Our strategy will be applicable in a prospective manner where nothing else than a protein–protein complex structure is known; hence, it can well be the first step in a structure-based endeavor to identify PPIMs.



INTRODUCTION

Protein–protein interactions (PPIs) are involved in nearly all biological processes. Due to their universal occurrence, protein–protein interfaces provide an important, yet neglected, new class of drug targets.¹ At present, the design of small-molecule protein–protein interaction modulators (PPIMs) encounters at least two challenges. First, in contrast to enzymes, protein–protein interfaces are rather flat and usually lack a distinct binding pocket.² Second, due to the often large size of protein–protein interfaces (~ 1200 to ~ 4660 Å²)^{3,4} interactions that are favorable for binding can be widely distributed over the interface.

Experimental evidence suggests that these challenges can be overcome.^{5–8} Most strikingly, residues participating in important interactions have been shown to be spatially clustered in protein–protein interfaces, forming so-called “hot spot” regions.^{4,9–11} Mimicking localized interactions at these hot spots provides a possibility for PPIM development.^{1,4,12–14} Furthermore, an opening of so-called transient pockets was observed in protein–protein interfaces.⁵ In fact, binding of several PPIMs to transient pockets in protein–protein interfaces has been reported.⁷ The most prominent example is given by small-molecule inhibitors binding to interleukin-2 (IL-2). These PPIMs inhibit the interaction with the IL-2 α -receptor (IL-2R α).^{5,15,16} Notably, the small molecule-bound IL-2 exhibits pockets in the interface region that are present neither in the unbound nor in the IL-2R α -bound crystal structure.^{5,17}

Computational methods can aid in finding and in the design of PPIMs if they are able to provide an accurate description of the energetics and dynamics of small-molecule binding to protein–protein interfaces.^{7,13,18–21} While many computational studies have been reported that deal with the energetics^{13,22–33} and dynamics^{7,13,19,34–39} of protein–protein interfaces *per se*, only a few have focused on the aspect of small-molecule binding to protein–protein interface regions,^{7,14,20,40–46} and none has considered aspects of energetics and interface plasticity simultaneously in this context so far. Thus, in the present study, we set out to evaluate the capability of state-of-the-art computational methods to predict small-molecule binding to protein–protein interfaces. In particular, we address five major questions that consider aspects of structure, dynamics, and energetics important for binding to protein–protein interfaces: (i) Can one identify hot spots in a protein–protein interface based on a protein–protein complex structure and make use of these hot spots for predicting the binding mode of small-molecule ligands? (ii) Can one sample the opening of transient pockets comparable to those observed in the protein–protein interface of a bound protein, as suggested by the conformational selection model,⁴⁷ starting from an unbound protein conformation? (iii) Is it possible to identify protein conformations with transient pockets by energetic or

Received: July 13, 2011

Published: November 17, 2011

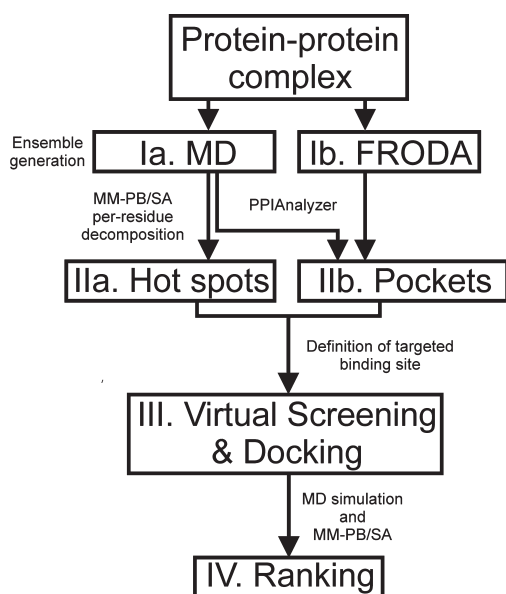


Figure 1. Outline of the strategy for hot spot and transient pocket identification, docking, and ranking of PPIMs using only the structure of a protein–protein complex as a starting point. A similarity-based virtual screening was not performed in this study. See text for details.

geometrical criteria? (iv) By docking to transient pockets, can one reproduce binding modes of known PPIMs and enrich PPIMs by virtual screening? (v) Can one rank known PPIMs with respect to their affinity?

To answer these questions, we chose IL-2 as a model system because of the wealth of information available for this system in terms of crystal structures of unbound IL-2, IL-2 bound to IL-2R α , and IL-2 bound to five PPIMs as well as experimental binding and inhibition data for the wild type and mutant protein.^{5,51,48} As to the methodological strategy (Figure 1), we apply and compare molecular dynamics (MD) and constrained geometric (FRODA) simulations for generating structural ensembles, introduce the PPIAnalyzer approach for investigating structural properties of protein–protein interfaces within these ensembles, and apply the MM-PB(GB)SA (molecular mechanics Poisson–Boltzmann (generalized Born) surface area) approach to identify hot spots and rank PPIMs. We note that when pursuing this strategy, we paid particular attention to mimicking a “real-life” scenario in structure-based ligand design. Thus, our strategy will be applicable also in a prospective study where nothing else than a protein–protein complex structure is known at the beginning.

MATERIALS AND METHODS

Overall Strategy. As to the methodological strategy (Figure 1), we pursued the following steps:

- Ia/b. Starting from a given protein–protein complex structure, conformational ensembles are generated on the basis of MD and FRODA simulations.
- Ila. Hot spot residues of a protein–protein complex structure are identified by MM-PBSA free energy decomposition on the basis of the structural ensemble generated by MD simulation.

- Ilb. Transient pockets in the protein–protein interface are identified by energetic or geometrical criteria in conformational ensembles generated either by MD or by FRODA.
- III. PPIM binding poses are predicted by molecular docking using the hot spot and transient pocket information for guidance. This docking setup is also used to enrich PPIMs from a large set of decoys, thus performing virtual screening.
- IV. PPIMs are ranked by their MM-PBSA binding effective energies ($\Delta G_{\text{eff}} = \text{gas phase energy} + \text{solvation free energy}$ according to a continuum solvent model) calculated for conformational ensembles from MD simulations that were started from the docked binding poses.

These steps will now be described in more detail. Detailed information about structure preparation and protocols for molecular dynamic simulations and docking experiments is provided in the Supporting Information.

FRODA Simulations. FRODA is a geometrical simulation method that explores the internal mobility of biomolecular systems. Details of the algorithm can be found in ref 49. The FRODA simulation was performed with the FIRST 6.2 suite of programs. For the rigid cluster decomposition with FIRST, a hydrogen bond energy cutoff of $-1.0 \text{ kcal mol}^{-1}$ was applied together with the “H 1” function for hydrophobic interactions. For each system, 10 000 000 conformations were sampled, of which every 10 000th conformation was stored. The random displacement distance of the mobile atoms was set to 0.1 Å, and the continuous motion (CM) method was applied.

MM-PB(GB)SA Calculations. MM-PBSA^{34,50} calculations were carried out according to the “multiple trajectory method” and the “single trajectory method”. For the multiple trajectory method, snapshots of all of the IL-2 complexes, the unbound IL-2 structures, and the IL-2R α subunit as well as the small-molecule ligands were extracted from independent MD trajectories. Alternatively, for the single trajectory method, snapshots of the binding partners were extracted from MD trajectories of the complexes only. All counterions and water molecules were stripped from the snapshots. Snapshots were extracted every 10 ps. Autocorrelation analysis of the effective energy of snapshots revealed that this time interval is sufficient for generating statistically independent snapshots. The gas phase energy was calculated on the basis of the ff99SB force field⁵¹ without applying any non-bonded cutoff. The polar part of the solvation free energy was determined by solving the linearized Poisson–Boltzmann (PB) equation⁵² or by applying the “OBC” generalized Born (GB) method ($igb = 5$) using mbondi2 radii.⁵³ A dielectric constant of 1 and 80 for the interior and exterior of the solute was applied, respectively. The polar contributions were computed at 100 mM ionic strength. Nonpolar solvation energies were calculated by a solvent-accessible surface area (SASA) dependent term, using a surface tension proportionality constant of $\gamma = 0.0072 \text{ kcal mol}^{-1} \text{ Å}^{-2}$. Contributions from vibrational entropy were neglected,⁵⁴ which can be justified with the small-molecule ligands being very similar. For calculating per residue contributions, the decomposition scheme²² implemented in the SANDER and MM-PBSA code of AMBER 10 was extended to also consider the PB reaction field energy. This is done on the basis of the concept of induced surface charges on the dielectric boundary.⁵⁵ The contribution of a residue k to the reaction field energy E_{rf} is then calculated without additional computational costs as the sum of Coulomb interactions of all of

its atomic charges q_i with all induced surface charges q_j , with r being the distance between the two charges (eq 1):

$$E_{\text{eff}}(k) = \sum_{i \in k} \sum_j \frac{q_i q_j}{4\pi\epsilon_0 r} \quad (1)$$

Identification of Residues in the Interface of IL-2/IL-2R α

Residues that are ≤ 5 Å apart from IL-2R α in the IL-2/IL-2R α complex structure (PDB code: 1z92) were chosen as interface residues of IL-2. Residues pointing toward the interior of IL-2 and thus not contributing to the binding of IL-2R α or small ligands were excluded upon visual inspection. This resulted in a set of 31 IL-2 interface residues: Tyr31, Asn33, Pro34, Lys35, Thr37, Arg38, Met39, Thr41, Phe42, Lys43, Phe44, Tyr45, Glu60, Glu61, Glu62, Lys64, Pro65, Leu66, Glu67, Glu68, Val69, Asn71, Leu72, Met104, Cys105, Glu106, Tyr107, Ala108, Asp109, Glu110, and Thr111.

Structural Analysis of Protein–Protein Interfaces. For investigating structural properties of protein–protein interface regions from conformational ensembles, we developed the PPIAnalyzer method (Figure S1). The method comprises three steps:

- I. Structural changes of the interface are determined in terms of root mean-square deviations (RMSD) of backbone and side chain atoms.
- IIa. The steric quality of the generated conformations is assessed.
- IIb. Distinct interface conformations are selected on the basis of a clustering with respect to the RMSD of heavy atoms of interface residues.
- III. Transient pockets are identified in those conformations.

In more detail, heavy atom RMSD values of interface residues are calculated with respect to the small molecule-bound (PDB codes: 1m48, 1m49, 1pw6, 1py2, and 1qvn) and unbound crystal structures (PDB code: 1m47) after structurally aligning the interface regions. We note that this RMSD calculation was only done to retrospectively assess the sampling of bound interface conformations; at no time of the study was knowledge of bound conformations applied for the identification of protein conformations that were subsequently used in docking experiments. The stereochemical quality of each snapshot was assessed using PROCHECK.⁵⁶ Snapshots not satisfying all of the following stereochemical criteria were excluded from further investigation:

- I. At most, two bad contacts are present.
- II. Less than 5% of the amino acids are in disallowed regions of the Ramachandran plot.
- III. At most two unfavorable main chain or side chain parameters are present. Of the remaining snapshots, 100 representative structures were selected using a k -medoids clustering algorithm⁵⁷ with respect to the RMSD of the interface residues.

Finally, potential binding pockets were detected using the PocketAnalyzer program⁵⁸ that implements a pocket identification strategy similar to the one proposed by Hendlich et al.⁵⁹ Details of the algorithm can be found in ref 58. Here, the following parameters were used: minimal degree of buriedness: 9; minimal number of neighbors: 9; minimal cluster size: 50; grid spacing: 1.0 Å. Snapshots for the subsequent docking experiments were chosen with respect to their identified pocket volume.

Data Set of Useful Decoys. Following the procedure described to generate the directory of useful decoys (DUD),⁶⁰ we selected compounds from the “purchasable subset” of the ZINC

database⁶¹ (as of May 19, 2010) that are similar with respect to physicochemical properties to the five IL-2 ligands in complex structures (Table 1). Descriptors for the ZINC compounds were downloaded from the ZINC Web site or were calculated for the IL-2 ligands using Molinspiration.⁶² The number of functional groups was calculated using the OpenEye FILTER program.⁶³ The pairwise dissimilarity between compounds was calculated as the *weighted root mean-cubed difference* (RMCD; eq 2) of the differences of normalized descriptors X_i (weights w_i in parentheses) logP (8); molecular weight (4); number of hydrogen bond donors (4) and acceptors (4); number of rotatable bonds (4); number of amide (1), amino (1), and carboxylic acid (1) groups; and the sum of the numbers of amidino and guanidino groups (1).

$$\text{RMCD} = \sqrt[3]{\frac{\sum_i w_i |X_i|^3}{\sum_i w_i}} \quad (2)$$

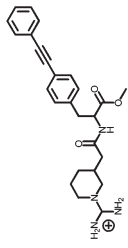
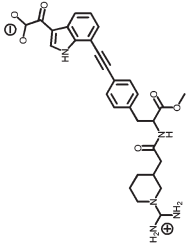
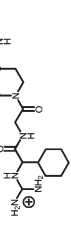
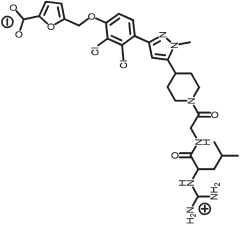
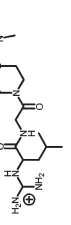
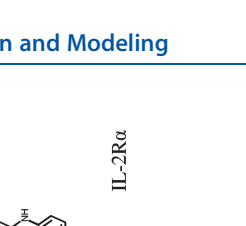
The ZINC compounds were then sorted by their pairwise RMCDs to all five IL-2 ligands in complex structures. The 10 000 most similar ZINC compounds were clustered into 1000 clusters on the basis of the pairwise RMCD using hierarchical clustering according to Ward's method as implemented in the hclust module of R.⁶⁴ Out of each cluster, the compound with the smallest RMSD of pairwise RMCD to the reference ligands was selected. Four compounds were not retrievable from the ZINC. For the remaining 996 unique decoy structures, a total of 1297 protonation and tautomerization states, as stored in the ZINC database, were considered.

Ranking of Docked Structures. For ranking docked structures by MM-PBSA, appropriate starting structures for the MD-based snapshot generation must be chosen initially. For this, consider that 100 docking runs were performed for each of the 10 structures with the largest interface pocket volumes obtained by a FRODA simulation of the unbound IL-2 structure and for each ligand. The starting structure selection was based on hot spot information, RMSD clustering, and intermolecular docking energy. First, it was required that a ligand's guanidinium group be within 5 Å of the side chain heavy atoms of the hot spot amino acid Glu62. Second, all of the remaining docking poses were clustered with respect to heavy atom RMSD of the PPIMs in the docked complexes after aligning only the proteins. Hierarchical complete linkage clustering was performed with R⁶⁴ with a cluster distance of 5 Å. Compared to clustering docking results above, a larger clustering distance of 5 Å was chosen to account for the fact that the clustering is performed over complex structures with different receptor conformations. Finally, the structure with the lowest intermolecular docking energy from the largest cluster was chosen as a starting structure for MD simulation and subsequent MM-PB(GB)SA binding effective energy calculation. Equilibrations, production runs, and MM-PB(GB)SA calculations were performed as described in the Supporting Information for the crystal structures. To allow the complex structures to relax after geometrical FRODA simulations and subsequent docking, we performed 20 ns of unrestrained MD simulation, with only the last 10 ns being used for snapshot extraction.

RESULTS AND DISCUSSION

Structures from MD Simulations. In order to investigate the opening of transient pockets in the IL-2 interface and to calculate effective energies by MM-PB(GB)SA, conformational ensembles of

Table 1. Structures of IL-2, IL-2 Complexes, and IL-2 Binding Partners with Respective Structural, Experimental, and Computed Data

PDB code	1m47 ^a	1m4c ^a	1m48 ^b	1m49 ^b	1pw6 ^b	1py2 ^b	1qvn ^b	1z92 ^c
Ligand	—	—						
Ligand abbreviation	—	—	FRG	CMM	FRB	FRH	FRI	IL-2Ra
Resolution ^d	1.99	2.40	1.95	2.00	2.60	2.80	2.70	2.80
Experimental affinity ^e	—	—	-6.90 ^j	-7.18 ^k	-7.02 ^l	-9.60 ^m	-8.03 ⁿ	-10.97 ^o
Calculated affinity	—	—	-44.50	-47.84	-49.72	-56.32	-49.13	-77.19
(SEM) ^f	—	—	(±0.14)	(±0.16)	(±0.13)	(±0.16)	(±0.17)	(±0.37)
MD length ^g	11.21	12.47	9.11	13.79	6.97	6.17	9.15	10.00
Drift ^h	—	—	0.20	0.19	-0.27	-0.63	0.74	0.49
IL-2 energy	-3181.50	-3182.45	-3203.77	-3211.09	-3212.32	-3210.79	-3188.06	-3190.96
(SEM) ^j	(± 1.06)	(± 1.02)	(± 1.25)	(± 1.19)	(± 1.36)	(± 1.47)	(± 1.36)	(± 1.33)

^a Unbound IL-2. ^b IL-2 in complex with a small molecule. ^c IL-2 in complex with the extracellular IL-2 binding domain of IL-2Ra. ^d In Ångströms. ^e In kilocalories per mole. ^f Binding effective energy calculated by the MM-PBSA single trajectory method, in kilocalories per mole. Standard error of the mean (SEM) in parentheses. ^g In nanoseconds. ^h Average drift of the binding effective energy in the MD ensemble calculated by linear regression in kilocalories per mole per nanosecond. ⁱ Calculated average absolute effective energy of IL-2 in the unbound state or the bound state as extracted from the complex trajectories, in kilocalories per mole. Standard error of the mean (SEM) in parentheses. ^j Derived by van't Hoff analysis of surface plasmon resonance data. ^k Derived from the average of $K_d = 7.5 \mu\text{M}$ and $K_d = 4.2 \mu\text{M}$ as determined by equilibrium analytical ultracentrifugation. ^l Calculated from $K_d = 7 \mu\text{M}$ as determined by surface plasmon resonance. ^m Calculated from $K_d = 100 \text{ nM}$ as determined by surface plasmon resonance. ⁿ Calculated from $K_d = 1.4 \mu\text{M}$, which was calculated by $K_{d,AB} = K_{d,BA}$ from the reported $K_d = 7 \mu\text{M}$ of FRB and the reported IC_{50} values of FRB ($\text{IC}_{50} = 10 \mu\text{M}$)⁶ and FRI ($\text{IC}_{50} = 2 \mu\text{M}$)¹⁵ the latter two being measured by ELISA under identical conditions. ^o As denoted by Thanos et al.¹⁵ and in agreement with the calculation from $K_d = 10 \text{ nM}$ as denoted by Rickert et al.⁶⁶

unbound IL-2 and IL-2 bound to either IL-2R α or five PPIMs were generated by MD simulations of at least 6 ns in length (Table 1).

For all systems, the RMSD of heavy atoms with respect to structures at the end of the equilibration procedure (see Supporting Information) was determined (Table S1, Supporting Information). Over all trajectories, IL-2 showed mean RMSD values of 2.55–3.46 Å. These values are in good agreement with those generally found during other MD simulations.⁶⁵ The interface regions of IL-2 showed generally lower RMSD values of 1.94–2.88 Å, with the interface region of the two unbound IL-2 structures showing the largest structural deviations (2.44 and 2.88 Å). This agrees well with observations from X-ray crystallography that show an opening of transient pockets in this region (see below). The small-molecule ligands bound to IL-2 showed RMSD values between 0.89 and 2.20 Å and, thus, stayed close to the initial binding region. Overall, after an initial rise during the first 2–4 ns, the RMSD values remain constant for the remainder of the MD simulations (Table S1).

To investigate the mobility of IL-2 and its interface region, we calculated root-mean-square fluctuations (RMSFs) of heavy atoms of protein residues. Starting structures of MD simulations with RMSF values mapped in a color-coded fashion are shown in Figure S2 (Supporting Information). Unsurprisingly, the largest fluctuations up to 9.58 Å were found for flexible loop regions and the termini. Many of these mobile regions have not been resolved in several of the crystallographic structures,^{5,66} which already provides a hint as to their mobility. In contrast, all of the interface residues of IL-2 show RMSF values <2.50 Å. Interestingly, the mobility of Phe42 of IL-2, whose conformational transition is crucial for the opening of a transient pocket (see below), was found to be significantly higher in the unbound structure (RMSF = 1.86 Å (1.60 Å) for 1m47 (1m4c)) than in the bound structures (RMSF between 0.57 and 1.15 Å).

MM-PB(GB)SA calculations, which make use of a continuum electrostatic model for evaluating (de)solvation effects, may fail if structural waters are present in or close to the binding interface.^{67,68} To identify such water molecules, we calculated the RMSFs of all water molecules and, subsequently, investigated the residence times of waters with low RMSFs by visual inspection. First, the analysis did not reveal any long-lasting (residence time >1 ns) water molecule on the outer surface of the IL-2 interface except in the case of the IL-2/IL-2R α complex. However, none of these water molecules formed strong interactions with the protein for the complete simulation time. Second, in the interior of IL-2 adjacent to the binding interface, long-lasting (residence time >1 ns) water molecules were found at three distinct sites (I, in the interior of IL-2 close to Glu62; II, inside of a loop region enclosed by Tyr45, Ala108, Asp109, and Glu110; III, at the N-terminal end of helix D enclosed by Met39, Phe42, and Leu114). However, none of these waters is in direct contact with any of the IL-2 ligands. Furthermore, these waters are conserved in almost all simulations of unbound and bound IL-2 so that potential effects on MM-PB(GB)SA results should cancel. Overall, these findings lead us to expect only minor influences due to structural waters on MM-PB(GB)SA results for hot spot prediction and ligand ranking.

Identification of Hot Spots by MM-PBSA Free Energy Decomposition. Mimicking localized interactions in hot spot regions of protein–protein interfaces has proven valuable for PPIM development.⁴ We thus set out to computationally identify

hot spots in the protein–protein interface of IL-2/IL2R α and IL-2/small-molecule complexes (Figure 2). For this, we implemented and applied the MM-PBSA per residue effective energy decomposition,^{25,69} which complements the MM-GBSA effective energy decomposition introduced by us.²² Here, we applied the MM-PBSA single trajectory method. While this method neglects energetic contributions due to conformational changes of the binding partners, it leads to a drastic reduction in the statistical uncertainty of the free energy components.^{34,70,71}

For validation, computed effective energy components were compared to experimentally determined changes in the binding free energy of IL-2/IL-2 R α and IL-2/FRH complexes upon mutations of IL-2 interface residues to alanine.^{6,15} The experiments showed that for both IL-2R α and FRH binding was strongly disrupted ($EC_{50,Ala}/EC_{50,WT} \geq 100$, equivalent to $\Delta G \geq 2.8$ kcal mol⁻¹ at 37 °C) when Phe42, Tyr45, or Glu62 were mutated to alanine. Encouragingly, the effective energy decomposition also identified Phe42 and Glu62 as hot spot residues ($\Delta G_{eff} = -2.84$ to -4.45 kcal mol⁻¹) and only slightly underestimated the contribution of Tyr45 ($\Delta G_{eff} \approx -1.3$ kcal mol⁻¹). Thus, experimental and computational predictions of hot spots are in good agreement. Moderate changes in the binding affinity ($EC_{50,Ala}/EC_{50,WT} \geq 10$, equivalent to $\Delta G \geq 1.4$ kcal mol⁻¹ at 37 °C) were observed for IL-2/IL-2R α when IL-2 residues Thr41, Lys43, or Phe44 were mutated to alanine.¹⁵ Computed ΔG_{eff} values are in the range of -0.4 to -0.8 kcal mol⁻¹ in these cases, demonstrating that smaller effects on the binding affinity could be well identified by the MM-PBSA effective energy decomposition, too. Finally, residues Lys35, Arg38, and Leu72 were also identified as hot spots by the MM-PBSA effective energy decomposition. However, except for Leu72, which moderately disrupted the IL-2/FRH complex ($EC_{50,Ala}/EC_{50,WT} \geq 10$) when mutated, all others did only show a weak disrupting effect on IL-2R α and FRH binding ($EC_{50,Ala}/EC_{50,WT} = 3-5$).

The seemingly prominent interactions of Arg83 ($\Delta G_{eff} \approx -2.3$ kcal mol⁻¹) are dominated by an intermittent salt bridge to Asp56 (IL-2R α residues are highlighted in italics, whereas IL-2 residues are depicted in “normal” font), for which the proteins needed to undergo structural changes during the MD simulation. As Arg83 is far apart from the localized cluster formed by the other hot spots, it was neglected for the guidance of the subsequent docking. Finally, our calculations predict Glu68 to contribute unfavorably ($\Delta G_{eff} = +1.72$ kcal mol⁻¹) to the binding of IL-2R α ; Glu68 thus is a “cold spot”. In summary, the identified hot spots cluster together and form a functional epitope localized on helices A' and B' with an approximate area of 500 Å² corresponding to 20% of the total protein–protein interface area.

Mimicry of Localized Interactions in Hot Spot Regions by PPIMs. Next, we investigated to what extent the PPIMs mimic IL2R α as an interaction partner. If such mimicry existed, hot spots identified in a protein–protein complex could be used for guiding PPIM identification and development. Figure 2 reveals that the energetic fingerprint of IL-2/IL-2R α is indeed highly similar to the energetic fingerprints of the IL-2/small-molecule complexes. Three amino acids stand out in that respect: I. Phe42 is the center of a hydrophobic core forming contacts to other IL-2 residues (Met39, Val69, Leu72) as well as Met25, Asn27, Leu42, and His120, which explains its hot spot character in the IL-2/IL-2R α case. Moreover, Phe42 forms favorable interactions with the piperidine (FRI, FRH), pyrazole (FRB), or central phenyl

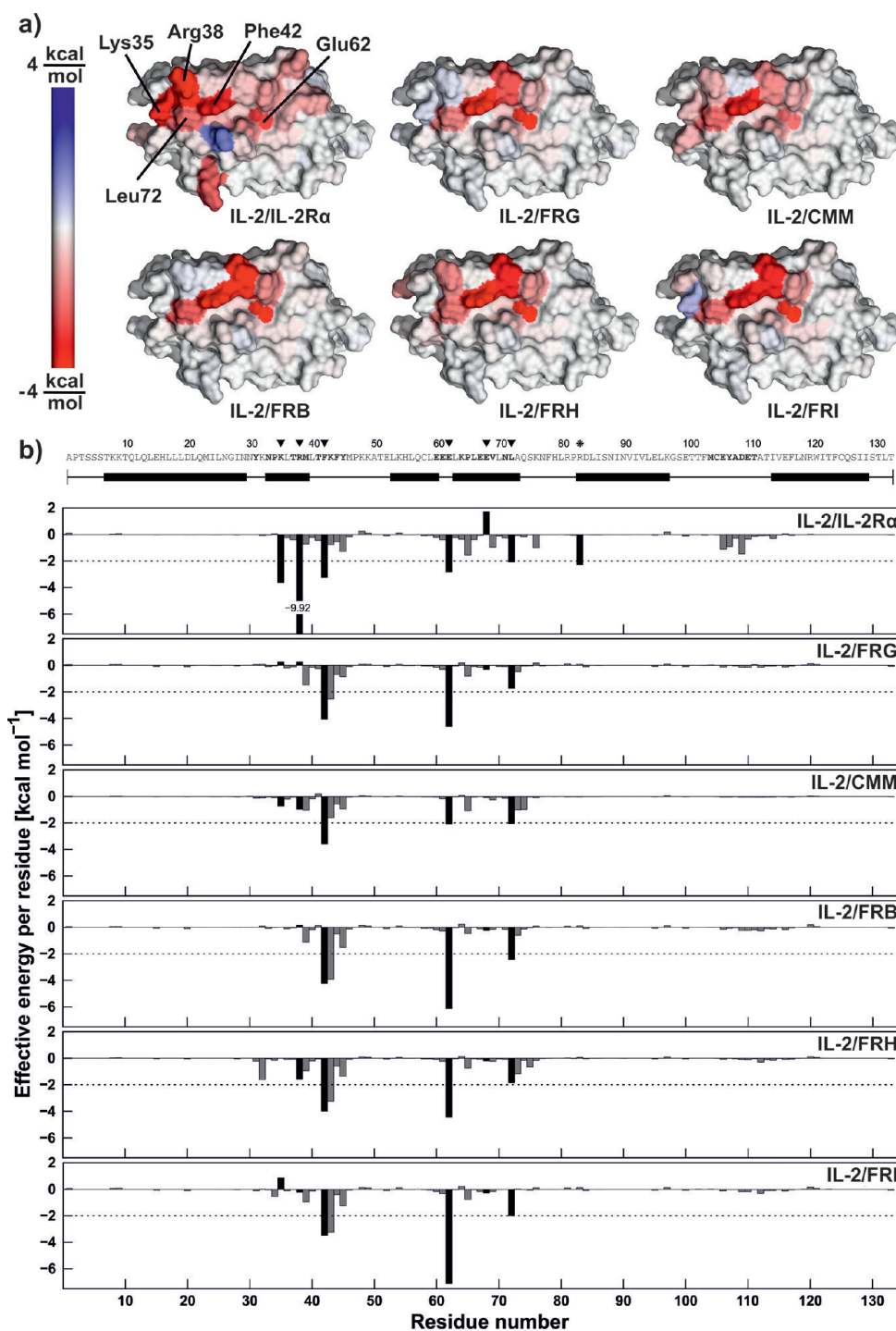


Figure 2. Per-residue contributions to the binding effective energy as calculated by MM-PBSA decomposition. The per-residue contributions were calculated by applying the single trajectory MM-PBSA method to the MD trajectories of IL-2 in complex with IL-2R α (PDB code: 1z92) and the five PPIMs FRG, CMM, FRB, FRH, and FRI (PDB codes: 1m48, 1m49, 1pw6, 1py2, and 1qvn). (a) The per-residue contribution is mapped onto the crystal structure of IL-2 bound to IL-2R α using a color-code with a linear scale for IL-2/IL-2R α , IL-2/FRG, IL-2/CMM, IL-2/FRB, IL-2/FRH, and IL-2/FRI. In (b) per-residue contributions of the six complexes are depicted as bar plots. At the top, the IL-2 sequence is depicted in single letter code with all interface residues highlighted (bold). α -helices are marked as horizontal boxes in the line below. Hot spots (highlighted as black bars in the energy plot and marked by black triangles (▼) above the sequence) were selected based on the per-residue decomposition of the IL-2/IL-2R α trajectory by applying an energy cutoff of 2 kcal mol $^{-1}$ (dashed line). See text for details on Arg83, which is marked by an asterisk (*) above the sequence.

ring of the tolane moiety (FRG, CMM) of the PPIMs and their adjacent amide moieties, which makes Phe42 also a hot spot for the binding of these PPIMs. II. Glu62 forms a stable salt bridge with Arg36 in the protein–protein complex. Glu62 is also the strongest anchor for the binding of the five PPIMs by salt-bridge formation with the guanidinium groups present in all ligands. III. The hydrophobic interaction of Leu72 with Met25 and Leu2 is replaced by the phenyl and 1,2-dichlorophenyl moieties of FRI, the 1,2-dichlorophenyl moieties of FRH or FRB, the indolyl moiety of CMM, or the terminal phenyl ring of the tolane moiety of FRG. In contrast, interactions involving Lys35 and Arg38, which are important for IL-2/IL-2R α affinity, are not or only weakly mimicked by the PPIM.

In summary, these findings demonstrate that three out of five computationally identified hot spots of the IL-2/IL-2R α complex are equally important for small-molecule binding to IL-2. Furthermore, all five computed hot spots cluster in a subregion of the structural epitope of IL-2/IL-2R α . Together, this strongly suggests that hot spots computed from protein–protein complexes can be used for guiding the identification and optimization of PPIM.

Opening of a Transient Pocket during Simulations Started from the Unbound State. PPIMs have been found to be particularly effective when they bind to well-defined clefts or grooves in the protein–protein interface.^{5,9,72,73} Here, we investigate whether an opening of transient pockets in the rather flat protein–protein interface of IL-2 can be observed when sampling the conformational space of unbound IL-2, following the “conformational selection” model.⁴⁷ Two conformational sampling techniques were used. First, as a state-of-the-art method, we applied all-atom MD simulations in explicit solvent of 10 ns in length. This setup is similar to a study by Helms and Eyrich.⁷ Second, as a computationally cheaper alternative, we applied the geometrical simulation method FRODA.⁴⁹ FRODA has already been successfully applied for identifying spontaneous and relevant *apo*-to-*holo* conformational transitions of HIV-1 TAR RNA.⁷⁴ FRODA relies upon a decomposition of a biomacromolecule into rigid and flexible regions.⁷⁵ In the unbound structure of IL-2, helices A, A', B, B', C, and D form discrete rigid clusters that are interconnected by flexible hinges (Figure S3, Supporting Information). With respect to the interface region, only 25% of the atoms belong to rigid clusters. This ensures that the majority of the interface atoms can move freely during the FRODA simulation.

The interface heavy atom RMSD of all snapshots of MD and FRODA simulations were calculated with respect to the PPIM-bound and unbound IL-2 structures (Table S2, Supporting Information). The overall structural changes between bound and unbound IL-2 structures are small: in the best case, an interface conformation coming as close as 1.42 Å RMSD to a bound state was found, starting from a structural deviation between bound and unbound structures of 1.50 Å. This can be explained in that structural deviations between bound and unbound structures are uniformly distributed over the interface rather than caused by a large-scale collective movement. Interestingly, when comparing the performance of FRODA and MD simulations, FRODA snapshots were generally found to be more similar to four out of five bound IL-2 structures than were MD snapshots. We attribute this fact to an appropriate coarse-graining of the unbound IL-2 structure prior to the FRODA simulation. Apparently, residues were correctly identified to be part of a rigid cluster that is not involved in pocket opening, leading to a focusing of movements to that region where a pocket opens up (Figure S4,

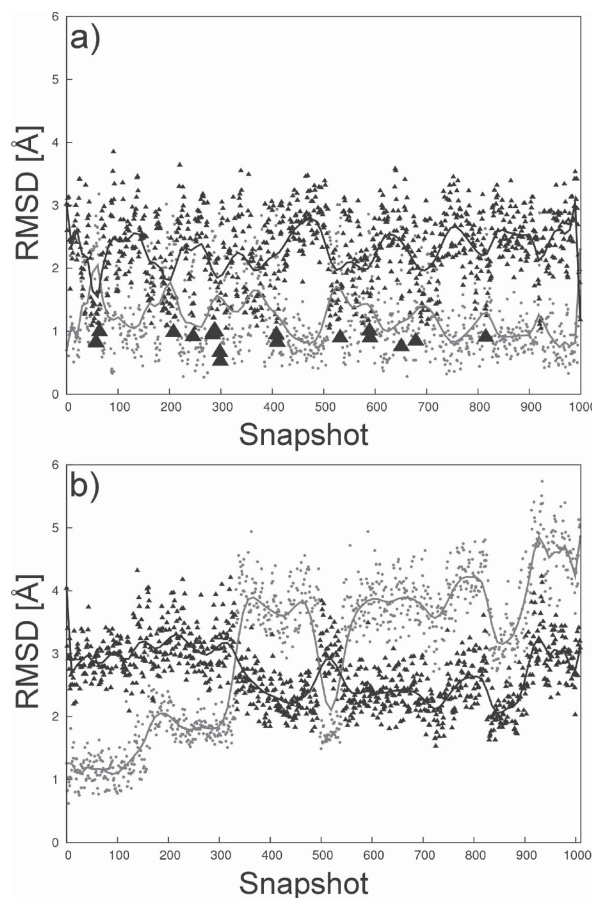


Figure 3. Conformational analysis of the interface residue Phe42. Consecutive snapshots generated by (a) FRODA and (b) MD simulation of unbound IL-2 are compared to the bound (black, PDB code: 1m48) and unbound IL-2 structures (gray, PDB code: 1m47) regarding the RMSD of Phe42 based on a structural alignment of heavy atoms of the interface. For the FRODA simulation, the RMSD shows a clear antidromic character that indicates the flipping of the aromatic ring. In contrast, for the MD simulation, the antidromic character of the RMSD curve is much less pronounced. Phe42 conformations that come closer to the bound IL-2 state than 1 Å RMSD are marked by black triangles.

Supporting Information). In contrast, in the MD simulations, all residues are allowed to move freely, leading to larger overall structural deviations in the protein–protein interface that do not necessarily lead to a structure close to a bound conformation.

Of the residues that line the PPIM binding pocket, Phe42 has been described as functioning as a gate keeper by flipping its phenyl ring.^{5,15} In Figure 3, RMSD time series calculated from all heavy atoms of Phe42 with respect to both unbound and bound IL-2 structures are shown for FRODA and MD simulations. The FRODA simulation shows an antidromic behavior of both time series, and the simulated conformations repeatedly approach the bound structure and depart from it. This leads to Phe42 approaching the small molecule bound conformation to <1.0 Å RMSD in 18 cases during the simulation. In contrast, the RMSD time series of the MD simulation showed a much less pronounced antidromic behavior, and a conformation of Phe42 with an RMSD <1.0 Å with respect to bound IL-2 was never detected during the simulation.

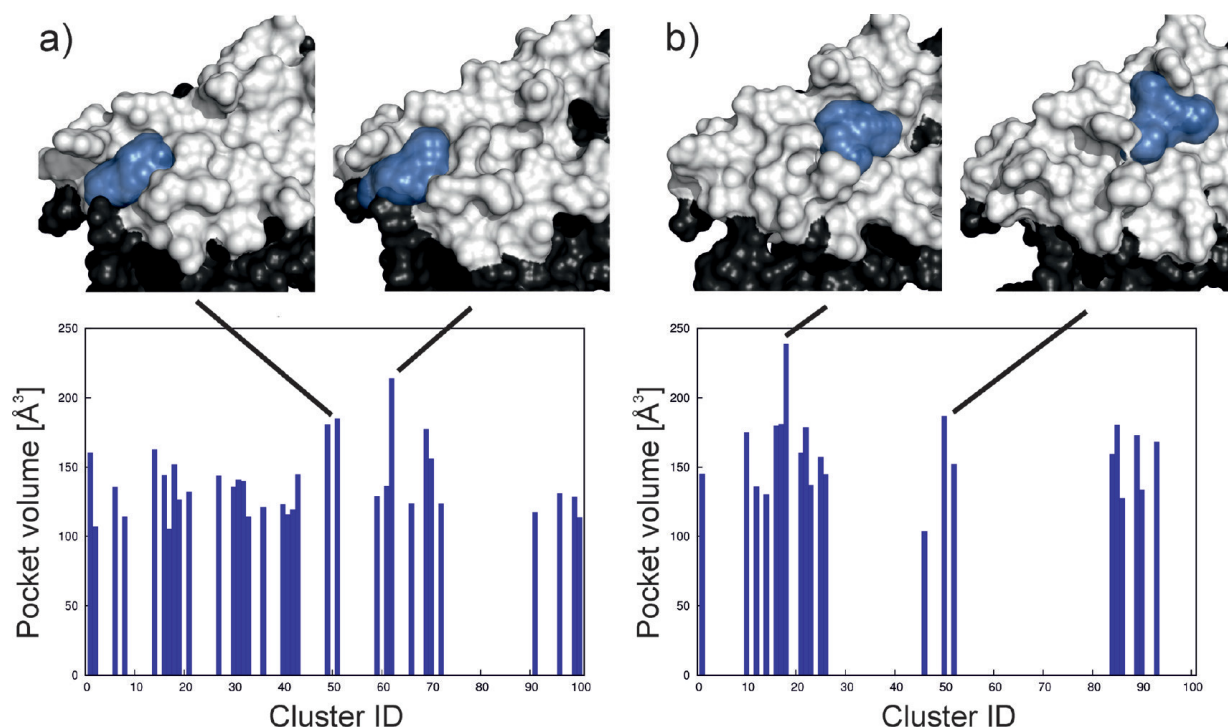


Figure 4. Detection of interface pockets in the cluster representatives of IL-2 structures generated by (a) FRODA and (b) MD simulation. The box plots depict pocket volumes computed by PocketAnalyzer. In addition, the two largest pockets found in IL-2 structures generated by either simulation method are shown. Notably, the locations of these pockets differ for both methods.

Identification of Transient Pockets in Structural Ensembles. Next, the question needs to be addressed as to how one can identify binding-competent conformations from the generated ensembles without already knowing the bound conformation from experimental results. An evaluation based on energetic criteria appears to be difficult^{34,76} (see the Supporting Information on “Statistical significance of MM-PBSA results” and Figures S5 and S8) because the expected change in free energy accompanying the opening of a transient pocket is on the order of $k_B T$, which is several orders of magnitude smaller than the total conformational energy of the protein.⁷⁷ Instead, we resorted to a sequential scheme that involved checking the stereochemical quality of the simulated conformations, clustering of similar conformations, and identification of transient interface pockets based on volume and the degree of “buriedness”. Thus, we only considered geometrical parameters for the identification of transient pockets.

Conformations sampled by either MD or FRODA simulations generally showed a high degree of stereochemical quality, as determined by PROCHECK.⁵⁶ In fact, none of the FRODA conformations had to be excluded from further investigations, whereas only 45 MD conformations were discarded because they had more than two unfavorable main chain parameters. In the next step, 100 structurally varying interface conformations were selected as representatives from each simulation by *k*-medoids clustering. The interface RMSD of the selected representatives ranges from 0.85 to 3.43 Å for MD-generated conformations and from 0.78 to 2.14 Å for FRODA-generated conformations. Finally, the PocketAnalyzer program⁸⁸ was applied for pocket detection. Pockets embraced by at least 70% of the interface residues were identified as interface pockets. As for the crystal

structures, all small-molecule bound structures displayed interface pockets with volumes ranging from 107 to 234 Å³, with all of these pockets being located between residues Lys35, Arg38, and Phe42. No pocket was present in the unbound IL-2 structure as well as the IL-2/IL-2R α complex.

In 33% of the selected FRODA conformations, an interface pocket was detected (Figure 4). The average volume of these pockets is 138 Å³, with minimal (maximal) values of 104 Å³ (215 Å³). Similar to the crystal structures, all interface pockets from FRODA-generated conformations were located between residues Lys35, Arg38, and Phe42. As for MD-generated conformations, an interface pocket was identified in 22% of the conformations, with an average volume of 159 Å³ (min., 103 Å³; max., 240 Å³). Notably, 2/3 of these interface pockets are located between Lys43, Tyr45, and Phe42 and, thus, deviate in position from the pockets found in the bound crystal structures.

Docking into Transient Pockets. We then investigated whether simulated IL-2 conformations with transient pockets in the protein–protein interface can be used as receptor structures for docking. Therefore, we selected those 10 representative structures from each simulation that showed the largest interface pocket volume (Table S3, Supporting Information). Each of the five known IL-2 ligands (Table 1) was then docked into this set of conformations. To exclude any bias due to the knowledge of the experimentally determined complex structures, the placement of the potential grids for docking (Figure 5) was solely based on (I) all hot spots identified by MM-PBSA except Arg83 (see above, Figure 2) and (II) all amino acids lining the identified interface pocket (Table S4, Supporting Information). Docking was considered successful when the ligand pose with the lowest

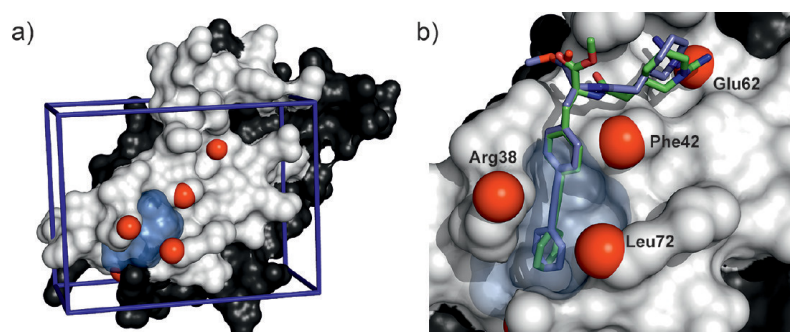


Figure 5. Definition of the potential grid and exemplary docking result for IL-2. The interface region of IL-2/IL-2R α is colored white. (a) The hot spots identified by MM-PBSA decomposition (red spheres) and the transient pocket (blue surface) were used to define the location and size of the potential grid (see Materials and Methods). The grid dimensions are represented by a blue box. (b) Predicted binding pose of the ligand FRG (blue sticks) docked into a FRODA snapshot containing an identified transient pocket. The RMSD between the predicted and crystallographic binding pose (PDB code: 1m48, green sticks) is 1.28 Å.

Table 2. Results of Redocking and *apo*-Docking

redocking					<i>apo</i> -docking				
ligand	PDB code ^a	RMSD ^b	score ^c	cluster size	ligand	PDB code ^a	RMSD ^b	score ^c	cluster size
FRG	1m48	1.50	−13.84	84	FRG	1m47	2.58	−10.93	19
CMM	1m48	1.37	−15.69	82	CMM	1m47	3.08	−12.53	40
FRB	1pw6	0.59	−13.01	25	FRB	1m47	4.63	−10.82	51
FRH	1py2	0.87	−15.38	53	FRH	1m47	3.80	−14.10	9
FRI	1qvn	1.14	−14.65	27	FRI	1m47	8.10	−13.30	9

^a PDB code of the IL-2 complex structure. ^b RMSD of the ligand pose with the lowest energy in the largest cluster with respect to the native structure, in Ångströms. ^c In kilocalories per mole.

intermolecular docking energy in the largest cluster had an RMSD < 2.0 Å to the native pose.

For comparison, we first redocked the five IL-2 ligands into the corresponding IL-2 complex structure. Likewise, we also performed docking of these ligands to an *apo* structure of IL-2 (PDB code: 1m47). Since no interface pocket could be detected in this *apo* structure, the same potential grid definition as for the redocking approach was used for the *apo*-docking. The redocking was successful in all cases, whereas *apo*-docking failed (Table 2). The latter is not unexpected due to the absence of any pronounced indentation in the protein–protein interface. Additionally, the success and convergence of the redocking is demonstrated by the occurrence of ≥ 25 poses in the largest cluster for IL-2/FRB, IL-2/FRI, and IL-2/FRH and >80 poses in the largest cluster for IL-2/FRG and IL-2/CMM.

The results of docking into transient pockets of simulated conformations are summarized in Table 3. Notably, docking into MD-generated conformations was not successful because in no case was a docked pose with RMSD < 2 Å obtained, and in only 2 out of 50 dockings was a pose with RMSD < 2.5 Å identified. This failure is a result of the interface pockets being located at a different position than the pockets found in the bound crystal structures. A more detailed inspection of the MD-generated conformations showed that correctly localized transient pockets do exist (data not shown). However, these pockets are much less pronounced than those occurring in FRODA-generated snapshots and, hence, are not among the 10 largest pockets chosen for the docking experiments. This is because a hydrophobic channel

Table 3. Number of Successful Attempts of Docking into Snapshots with Identified Transient Pockets

ligand	RMSD ^a							
	MD				FRODA			
	<2.0	<2.5	<3.5	≥ 3.5	<2.0	<2.5	<3.5	≥ 3.5
FRG	0	2	1	7	2	6	1	1
CMM	0	0	4	6	2	4	4	0
FRB	0	0	3	7	0	0	7	2
FRH	0	0	0	10	1	2	4	3
FRI	0	0	0	10	1	1	0	8

^a RMSD of the ligand pose with the lowest energy in the largest cluster with respect to the native structure, in Ångströms.

embraced by Lys35, Arg38, and Phe42 must open for the transient pockets to be correctly localized.⁵ Such an opening tends to be less pronounced in MD-generated ensembles, as demonstrated by the lower hydrophobicity of these pockets when compared to pockets found in crystal structures.⁷

In contrast, we were able to identify the hydrophobic channel in all of the selected FRODA-generated snapshots, possibly due to the absence of solvent in the simulation process. As a consequence, docking to at least one transient pocket was successful for all IL-2 ligands but FRB (Table 3). This is exemplarily shown for the ligand FRG in Figure 5b.

These results are encouraging in that a drop in docking accuracy compared to redocking is often found to be mirrored by the degree to which a protein moves upon ligand binding.^{78,79} Thus, docking to an *apo* form usually shows the largest deterioration.⁸⁰ Being able to start from the *apo* IL-2 structure and identify transient pockets in trajectories of computationally inexpensive FRODA simulations that are adequate for ligand docking thus is a valuable achievement.

Docking Enrichment in a Large Set of Decoys. In order to demonstrate that the identified hot spots and transient pockets

could also be used for structure-based virtual screening (VS), we performed a retrospective VS for IL-2 PPIMs. As known binders, the five IL-2 ligands in complex structures (Table 1) and 52 structures with similar scaffolds and known IC₅₀ values were used.^{6,17,48} Decoys were selected from the “purchasable subset” of the ZINC database⁶¹ following the procedure described for creating the directory of useful decoys (DUD).⁶⁰ The DUD procedure aims at selecting decoy structures that are physicochemically similar to known binders in order to avoid any bias in enrichment calculations. This led to 996 unique decoy structures with a total of 1297 protonation and tautomerization states. We note that during docking, the docking scores were normalized by the square root of the molecular weight of a ligand in order to correct for any size related bias, too.⁸¹

For the 57 (S) known IL-2 ligands (in complex structures) we found good enrichments for the individual transient binding pockets (Table 4, Figure 6, Figure S6, Supporting Information) with EF_{max} = 16.2–23.8 (EF_{max} = 57.8–260.4) and EF₁ = 13.6–23.8 (EF₁ = 37.2–92.2), and area under the curve values of receiver operator curves of AUC ≥ 0.89 (AUC ≥ 0.93).

We note that these enrichments may be too optimistic compared to a real-life scenario and, hence, should be interpreted cautiously because the VS has likely benefitted from the fact that the known IL-2 ligands were structurally optimized for binding to IL-2 and partially violate Lipinski's rules.⁸² Hence, even though the decoys were selected following the DUD procedure, in some cases, a perfect match of the property distribution curves between binders and decoys could not be achieved (Figure S7, Supporting Information). This is particularly true for the properties “molecular weight”, “no. of hydrogen bond donors”, and “no. of amidino and guanidino groups”. Still, with respect to the aim of this study, our results demonstrate that known IL-2 ligands could successfully be screened from a set of decoys using only information about hot spots and transient pockets on the protein side.

Ranking of IL-2 Ligands. Binding effective energies calculated by the MM-PBSA single trajectory method appear to be converged and remain stable throughout simulation lengths of 6–14 ns (Table S5 and Figure S8, Supporting Information).

Table 4. Docking Enrichment of Known IL-2 Ligands^a

FRODA snapshot ^b	EF _{max} ^{c,d}	EF ₂₀ ^{c,d}	EF ₃ ^{c,d}	EF ₁ ^{c,d}	AUC ^{d,e}
6	23.7 (260.2)	4.6 (5.0)	17.4 (26.0)	22.0 (74.3)	0.95 (0.99)
117	23.8 (173.6)	4.4 (5.0)	18.0 (32.5)	22.1 (74.4)	0.95 (0.99)
169	23.7 (260.0)	4.8 (5.0)	22.0 (32.5)	23.7 (92.9)	0.98 (1.00)
301	23.8 (260.4)	4.9 (5.0)	19.7 (26.0)	18.7 (55.8)	0.99 (0.99)
418	23.8 (260.4)	4.8 (5.0)	10.4 (26.0)	11.9 (37.2)	0.94 (0.98)
514	19.0 (97.6)	5.0 (5.0)	16.2 (32.5)	17.0 (55.8)	0.98 (0.99)
534	23.8 (86.8)	4.4 (5.0)	11.0 (26.0)	13.6 (37.2)	0.92 (0.98)
657	16.2 (57.8)	4.0 (4.0)	10.4 (19.5)	15.3 (55.8)	0.89 (0.93)
698	23.8 (260.4)	4.1 (5.0)	16.2 (32.5)	23.8 (55.8)	0.94 (0.99)
729	23.8 (86.8)	4.6 (5.0)	12.2 (26.0)	13.6 (55.8)	0.95 (0.99)

^a The set of known IL-2 ligands consists of five IL-2 ligands with available complex crystal structures as well as 52 structures with similar scaffolds and known IC₅₀.^{6,17,48} The set of decoys consists of 996 unique structures with a total of 1297 protonation and tautomerization states.

^b Consecutive number of the snapshot from a total of 1000 snapshots uniformly extracted from the 10 000 000 FRODA-generated structures starting from the unbound IL-2 structure (PDB code: 1m47). The 10 snapshots with the largest pocket volumes were used. ^c EF₁, EF₃, EF₂₀, and EF_{max} correspond to the enrichment factors at 1%, 3%, and 20% of the ranked database and the maximal enrichment factors over the whole data set. ^d Values correspond to all 57 known IL-2 ligands, while values in parentheses correspond to the five IL-2 ligands with available complex crystal structures only. ^e Area under the receiver operator curve (ROC).

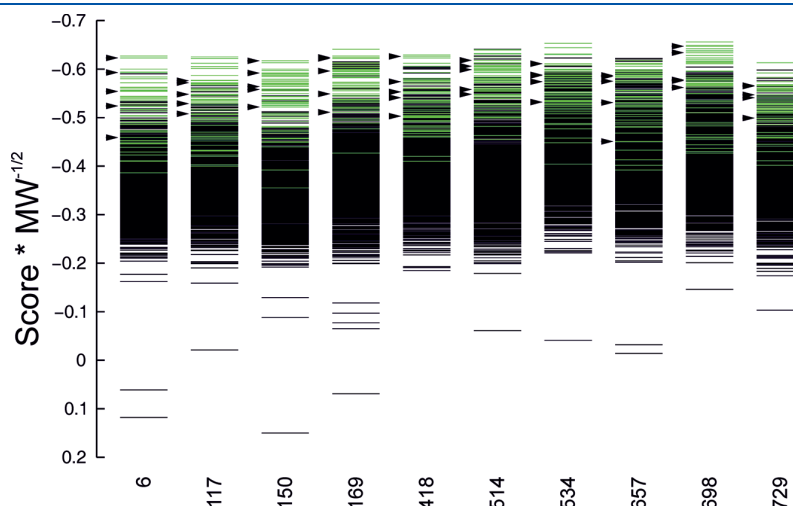


Figure 6. Ranking of docked structures. The best poses (as defined in Materials and Methods) of docked binders and decoy structures were ranked by intermolecular energy divided by the square root of the molecular weight. The 10 FRODA structures with the largest transient pocket are indicated by their snapshot number at the bottom. The 57 known IL-2 ligands and the decoys are depicted with green and black lines, respectively. In addition, the five IL-2 ligands with available complex crystal structures are highlighted by arrows.

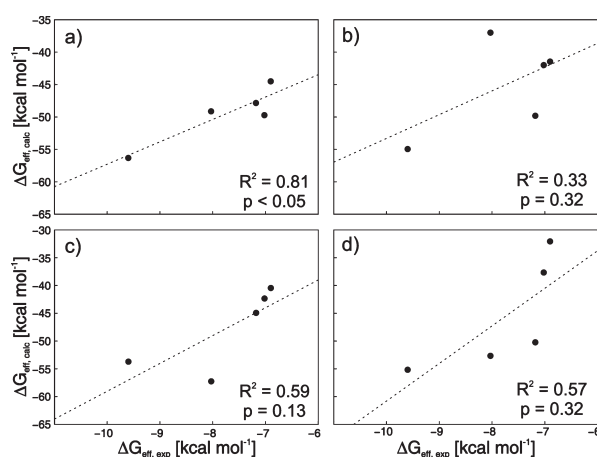


Figure 7. Correlation of computed binding effective energies ($\Delta G_{\text{eff,calc}}$) with respect to experimental free enthalpies of binding ($\Delta G_{\text{eff,exp}}$). Binding effective energies were calculated using the Poisson–Boltzmann continuum solvation model (a and c) and the generalized Born continuum solvation model (b and d) within the MM-PB(GB)SA single trajectory method. MM-PB(GB)SA calculations were based on either crystallographic (a and b) or docked (c and d) starting structures.

Absolute binding effective energies computed for the five IL-2 ligands starting from *crystal structures* of the complexes are about 45 kcal mol^{-1} more negative than the experimentally determined free enthalpies of binding (Figure 7, Table 1). Two reasons account for this. First, unfavorable energetic contributions due to conformational strain of the binding partners are not taken into account in the single trajectory method. These contributions can be as high as 36 kcal mol^{-1} .⁸³ Second, we neglect any changes in the configurational entropy of the binding partners, which accounted for contributions to the free energy of $20\text{--}30 \text{ kcal mol}^{-1}$ at 300 K in related studies.^{22,84} However, these two points do not have a major impact on relative binding effective energies as demonstrated by a good ($R^2 = 0.81$) and significant ($p < 0.05$) linear correlation of computed and experimentally determined binding energies, allowing for a successful ranking of four out of the five ligands (Figure 7a). This result is all the more remarkable in that the range of experimentally determined binding affinities is only 3 kcal mol^{-1} . To our knowledge, this is one of only a few reports so far of successfully applying MM-PBSA to rank PPIMs.^{69,85} When using the GB model, no significant correlation is obtained. In particular, the binding effective energies of ligands FRI and CMM show strong deviations from the correlation line (Figure 7b). A per residue decomposition of the binding effective energies using the GB model did not allow one to assign the origin of these deviations to contributions by a particular set of residues (data not shown).

For ranking *docked* IL-2/small-molecule complex structures (Figure 7), first, reasonable poses obtained by docking into FRODA-generated structures with transient pockets were selected without making use of any knowledge of the bound crystal structures (Table S6, Supporting Information). Then, MD simulations and MM-PB(GB)SA calculations were applied as in the case of the crystal structures. The linear correlation of computed MM-PBSA binding effective energies with respect to experimentally determined binding free energies is fair ($R^2 = 0.59$) and weakly significant ($p = 0.13$; Figure 7c). Again, no significant correlation was obtained in the case of MM-GBSA. The largest deviations from the correlation line are observed for ligands

FRH and FRI in the case of MM-PBSA, which also show the largest structural deviations from the native pose in the starting structures (Table S6). These results demonstrate that for MM-PBSA calculation to be successful in ranking PPIMs, (at least) good starting structures (RMSD $< 2.5 \text{ \AA}$) are required. Nevertheless, it is encouraging to note that the quality of the generated docking poses was still sufficient to successfully discriminate between the subgroups of high and low affinity ligands.

CONCLUSION

We have presented for the first time a computational strategy that simultaneously considers aspects of energetics and plasticity in the context of PPIM binding to a protein interface. In particular, our strategy aims at identifying the determinants of small-molecule binding, hot spots and transient pockets, in a protein–protein interface in order to make use of this knowledge for predicting binding modes of and ranking PPIMs with respect to their affinity. Although performed in a retrospective manner on the well-investigated system of IL-2, we note that at no point in the study did we utilize information from the experiment about the binding mode and affinity of PPIMs. Thus, our strategy will be applicable also in a prospective manner where nothing other than a protein–protein complex structure is known; hence, it can well be the first step in a structure-based endeavor to identify PPIMs.

Perhaps the most surprising result from a methodological point of view is that the computationally much cheaper constrained geometric simulation method FRODA outperforms state-of-the-art MD simulations in *sampling* transient pockets in the IL-2 interface. Apparently, the neglect of solvent in FRODA not only leads to a reduced computational burden but also facilitates the opening of a hydrophobic channel. Although applied to only one protein–protein interface in the present study, we note that the good performance of FRODA in sampling relevant conformational transitions is in line with results obtained by this⁷⁴ and a related method⁸⁶ on other systems.

It is encouraging that geometrical parameters summarized in the PPIAnalyzer method sufficed to successfully *identify* transient pockets. On the one hand, this finding alleviates the need to resort to conformational free/effective energies for identifying such pockets. Using energetic criteria is hampered by the demand for very precise computations due to the fact that small differences in conformational energies must be calculated from large absolute values. On the other hand, this finding reconfirms the FRODA results, as it demonstrates that the most pronounced pockets only opened up where expected. As FRODA strongly depends on a preceding flexibility analysis of the protein, it is thus tempting to speculate that regions that are prone to open transient pockets could be identified by such a flexibility analysis. This knowledge could then be used to focus other locally enhanced sampling schemes on that particular region.⁸⁷

Finally, we consider it a valuable achievement that the sequence of, first, docking to identified transient pockets; second, starting structure selection based on hot spot information, RMSD clustering, and intermolecular docking energies; and third, MM-PBSA calculations allowed one to discriminate between subgroups of IL-2 PPIMs with low and high affinity. Also, we obtained good enrichments for the individual transient binding pockets in a docking-based, retrospective virtual screening for IL-2 PPIMs. Together with the fact that the known PPIMs of IL-2 were identified to mimic many of the interactions also found in the

IL-2/IL-2R α region, this suggests that current computational methods can assist the knowledge-driven process of PPIM identification when starting from a given protein–protein complex.

■ ASSOCIATED CONTENT

S Supporting Information. Tables with the heavy atom RMSD of IL-2 and its complexes during MD simulations (Table S1), heavy atom RMSD of the IL-2 interface region of experimentally determined bound IL-2 conformations (Table S2), the largest pocket volumes for selected FRODA and MD snapshots (Table S3), selected pocket residues of IL-2 for the definition of the potential energy grids (Table S4), the drift of the effective energies during MD simulations (Table S5), and the selection of the best docking poses in the FRODA snapshots with the largest pocket volume (Table S6) as well as graphical representations of the workflow of the PPIAnalyzer method (Figure S1); the RMSF values of IL-2 during MD simulations of the unbound and bound states (Figure S2); the rigid cluster decomposition of unbound IL-2 obtained by FIRST (Figure S3); the overlay of the protein–protein interface region of IL-2 in unbound, bound, and FRODA sampled conformation (Figure S4); the calculated mean absolute effective energies G_{eff} of IL-2 in its unbound and bound conformations (Figure S5); the enrichment of known IL-2 ligands (Figure S6); the property distributions of known IL-2 ligands and decoys (Figure S7); the time series of the effective energies (Figure S8); and the multiple sequence alignment of sequences of IL-2 crystal structures (Figure S9). This information is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

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

Author Contributions

^SBoth authors contributed equally to this work.

■ ACKNOWLEDGMENT

We are grateful to Teresa Jiménez Vaquero for providing an initial version of the grid-based pocket identification algorithm and to Dr. Simone Fulle for critically reading the manuscript. This work was supported by Sanofi-Aventis Deutschland GmbH (LGCR Drug Design). We are grateful to the “Zentrum fuer Informations- und Medientechnologie” (ZIM) at the Heinrich-Heine-University, Düsseldorf for computational support. We are grateful to OpenEye Scientific Software for granting a no-cost academic license to us. Figures were generated by gnuplot and PyMOL.

■ ABBREVIATIONS:

PPI, protein–protein interaction; PPIM, small-molecule protein–protein interaction modulator; RMSD, root mean-square deviation; RMCD, root mean-cubed difference; MD, molecular dynamics; IL-2, interleukin-2; IL-2R α , α -subunit of the interleukin-2 receptor; GB, generalized Born; PB, Poisson–Boltzmann; MM-PBSA, molecular mechanics Poisson–Boltzmann surface area; MM-GBSA, molecular mechanics generalized Born; DUD, directory of useful

decoys; ROC, receiver operator curve; AUC, area under the curve; EF_X, enrichment factor at X% of the data set.

■ REFERENCES

- (1) Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **2007**, *450* (7172), 1001–1009.
- (2) Jones, S.; Thornton, J. M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93* (1), 13–20.
- (3) Lo Conte, L.; Chothia, C.; Janin, J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **1999**, *285* (5), 2177–2198.
- (4) Clackson, T.; Wells, J. A. A Hot-Spot of Binding-Energy in a Hormone-Receptor Interface. *Science* **1995**, *267* (5196), 383–386.
- (5) Arkin, M. R.; Randal, M.; DeLano, W. L.; Hyde, J.; Luong, T. N.; Oslob, J. D.; Raphael, D. R.; Taylor, L.; Wang, J.; McDowell, R. S.; Wells, J. A.; Braisted, A. C. Binding of small molecules to an adaptive protein-protein interface. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (4), 1603–1608.
- (6) Raimundo, B. C.; Oslob, J. D.; Braisted, A. C.; Hyde, J.; McDowell, R. S.; Randal, M.; Waal, N. D.; Wilkinson, J.; Yu, C. H.; Arkin, M. R. Integrating fragment assembly and biophysical methods in the chemical advancement of small-molecule antagonists of IL-2: An approach for inhibiting protein-protein interactions. *J. Med. Chem.* **2004**, *47* (12), 3111–3130.
- (7) Eyrisch, S.; Helms, V. What induces pocket openings on protein surface patches involved in protein-protein interactions? *J. Comput.-Aided Mol. Des.* **2009**, *23* (2), 73–86.
- (8) Oltersdorf, T.; Elmore, S. W.; Shoemaker, A. R.; Armstrong, R. C.; Augeri, D. J.; Belli, B. A.; Bruncko, M.; Deckwerth, T. L.; Dinges, J.; Hajduk, P. J.; Joseph, M. K.; Kitada, S.; Korsmeyer, S. J.; Kunzer, A. R.; Letai, A.; Li, C.; Mitten, M. J.; Nettesheim, D. G.; Ng, S.; Nimmer, P. M.; O'Connor, J. M.; Oleksijew, A.; Petros, A. M.; Reed, J. C.; Shen, W.; Tahir, S. K.; Thompson, C. B.; Tomaselli, K. J.; Wang, B.; Wendt, M. D.; Zhang, H.; Fesik, S. W.; Rosenberg, S. H. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* **2005**, *435* (7042), 677–681.
- (9) Keskin, Z.; Gursoy, A.; Ma, B.; Nussinov, R. Principles of protein-protein interactions: What are the preferred ways for proteins to interact? *Chem. Rev.* **2008**, *108* (4), 1225–1244.
- (10) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Hot spots-A review of the protein-protein interface determinant amino-acid residues. *Proteins* **2007**, *68* (4), 803–812.
- (11) Ozbabacan, S. E.; Gursoy, A.; Keskin, O.; Nussinov, R. Conformational ensembles, signal transduction and residue hot spots: application to drug discovery. *Curr. Opin. Drug Discovery Dev.* **2010**, *13* (5), 527–537.
- (12) Berg, T. Small-molecule inhibitors of protein-protein interactions. *Curr. Opin. Drug Discovery Dev.* **2008**, *11* (5), 666–674.
- (13) Gonzalez-Ruiz, D.; Gohlke, H. Targeting protein-protein interactions with small molecules: Challenges and perspectives for computational binding epitope detection and ligand finding. *Curr. Med. Chem.* **2006**, *13* (22), 2607–2625.
- (14) Zhong, S.; Macias, A. T.; MacKerell, A. D., Jr. Computational identification of inhibitors of protein-protein interactions. *Curr. Top. Med. Chem.* **2007**, *7* (1), 63–82.
- (15) Thanos, C. D.; DeLano, W. L.; Wells, J. A. Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (42), 15422–15427.
- (16) Wilson, C. G.; Arkin, M. R. Small-molecule inhibitors of IL-2/IL-2R: lessons learned and applied. *Curr. Top. Microbiol. Immunol.* **2011**, *348*, 25–59.
- (17) Braisted, A. C.; Oslob, J. D.; DeLano, W. L.; Hyde, J.; McDowell, R. S.; Waal, N.; Yu, C.; Arkin, M. R.; Raimundo, B. C. Discovery of a potent small molecule IL-2 inhibitor through fragment assembly. *J. Am. Chem. Soc.* **2003**, *125* (13), 3714–3715.
- (18) Fuller, J. C.; Burgoyne, N. J.; Jackson, R. M. Predicting druggable binding sites at the protein-protein interface. *Drug Discovery Today* **2009**, *14* (3–4), 155–161.

- (19) Brown, S. P.; Hajduk, P. J. Effects of conformational dynamics on predicted protein druggability. *ChemMedChem* **2006**, *1* (1), 70–72.
- (20) Czarna, A.; Beck, B.; Srivastava, S.; Popowicz, G. M.; Wolf, S.; Huang, Y.; Bista, M.; Holak, T. A.; Domling, A. Robust generation of lead compounds for protein-protein interactions by computational and MCR chemistry: p53/Hdm2 antagonists. *Angew. Chem., Int. Ed. Engl.* **2010**, *49* (31), 5352–5356.
- (21) Corradi, V.; Mancini, M.; Manetti, F.; Petta, S.; Santucci, M. A.; Botta, M. Identification of the first non-peptidic small molecule inhibitor of the c-Abl/14–3-3 protein-protein interactions able to drive sensitive and Imatinib-resistant leukemia cells to apoptosis. *Bioorg. Med. Chem. Lett.* **2010**, *20* (20), 6133–6137.
- (22) Gohlke, H.; Kiel, C.; Case, D. A. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RaGDS complexes. *J. Mol. Biol.* **2003**, *330* (4), 891–913.
- (23) Gohlke, H.; Kuhn, L. A.; Case, D. A. Change in protein flexibility upon complex formation: Analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* **2004**, *56* (2), 322–337.
- (24) Ababou, A.; van der Vaart, A.; Gogonea, V.; Merz, K. M., Jr. Interaction energy decomposition in protein-protein association: a quantum mechanical study of barnase-barstar complex. *Biophys. Chem.* **2007**, *125* (1), 221–236.
- (25) Lafont, V.; Schaefer, M.; Stote, R. H.; Altschuh, D.; Dejaegere, A. Protein-protein recognition and interaction hot spots in an antigen-antibody complex: free energy decomposition identifies "efficient amino acids". *Proteins* **2007**, *67* (2), 418–434.
- (26) Wichmann, C.; Becker, Y.; Chen-Wichmann, L.; Vogel, V.; Vojtkova, A.; Herglotz, J.; Moore, S.; Koch, J.; Lausen, J.; Mantele, W.; Gohlke, H.; Grez, M. Dimer-tetramer transition controls RUNX1/ETO leukemogenic activity. *Blood* **2010**, *116* (4), 603–613.
- (27) Zoete, V.; Michielin, O. Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein-protein association using MM-GBSA: application to the TCR-pMHC complex. *Proteins* **2007**, *67* (4), 1026–1047.
- (28) Krueger, D. M.; Gohlke, H. DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. *Nucleic Acids Res.* **2010**, *38* (Suppl), W480–W486.
- (29) Cole, D. J.; Skylaris, C. K.; Rajendra, E.; Venkitaraman, A. R.; Payne, M. C. Protein-protein interactions from linear-scaling first-principles quantum-mechanical calculations. *Europhys. Lett.* **2010**, *91* (3), 37004.
- (30) Dastidar, S. G.; Madhumalar, A.; Fuentes, G.; Lane, D. P.; Verma, C. S. Forces mediating protein-protein interactions: a computational study of p53 "approaching" MDM2. *Theor. Chem. Acc.* **2010**, *125* (3–6), 621–635.
- (31) Wong, S.; Amaro, R. E.; McCammon, J. A. MM-PBSA Captures Key Role of Intercalating Water Molecules at a Protein-Protein Interface. *J. Chem. Theory Comput.* **2009**, *5* (2), 422–429.
- (32) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Protein-protein recognition: a computational mutagenesis study of the MDM2-P53 complex. *Theor. Chem. Acc.* **2008**, *120* (4–6), 533–542.
- (33) Cui, Q. Z.; Sulea, T.; Schrag, J. D.; Munger, C.; Hung, M. N.; Naim, M.; Cygler, M.; Purisima, E. O. Molecular dynamics-solvated interaction energy studies of protein-protein interactions: The MP1-p14 scaffolding complex. *J. Mol. Biol.* **2008**, *379* (4), 787–802.
- (34) Gohlke, H.; Case, D. A. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* **2004**, *25* (2), 238–250.
- (35) Joce, C.; Stahl, J. A.; Shridhar, M.; Hutchinson, M. R.; Watkins, L. R.; Fedichev, P. O.; Yin, H. Application of a novel in silico high-throughput screen to identify selective inhibitors for protein-protein interactions. *Bioorg. Med. Chem. Lett.* **2010**, *20* (18), 5411–5413.
- (36) Neumann, J.; Gottschalk, K. E. The Effect of Different Force Applications on the Protein-Protein Complex Barnase-Barstar. *Biophys. J.* **2009**, *97* (6), 1687–1699.
- (37) Andrusier, N.; Mashiach, E.; Nussinov, R.; Wolfson, H. J. Principles of flexible protein-protein docking. *Proteins* **2008**, *73* (2), 271–289.
- (38) Chaudhury, S.; Gray, J. J. Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. *J. Mol. Biol.* **2008**, *381* (4), 1068–1087.
- (39) Chang, C. E. A.; McLaughlin, W. A.; Baron, R.; Wang, W.; McCammon, J. A. Entropic contributions and the influence of the hydrophobic environment in promiscuous protein-protein association. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (21), 7456–7461.
- (40) Enyedy, I. J.; Ling, Y.; Nacro, K.; Tomita, Y.; Wu, X. H.; Cao, Y. Y.; Guo, R. B.; Li, B. H.; Zhu, X. F.; Huang, Y.; Long, Y. Q.; Roller, P. P.; Yang, D. J.; Wang, S. M. Discovery of small-molecule inhibitors of bcl-2 through structure-based computer screening. *J. Med. Chem.* **2001**, *44* (25), 4313–4324.
- (41) Lugovskoy, A. A.; Degterev, A. I.; Fahmy, A. F.; Zhou, P.; Gross, J. D.; Yuan, J. Y.; Wagner, G. A novel approach for characterizing protein ligand complexes: Molecular basis for specificity of small-molecule Bcl-2 inhibitors. *J. Am. Chem. Soc.* **2002**, *124* (7), 1234–1240.
- (42) Koehler, N. K. U.; Yang, C. Y.; Varady, J.; Lu, Y. P.; Wu, X. W.; Liu, M.; Yin, D. X.; Bartels, M.; Xu, B. Y.; Roller, P. P.; Long, Y. Q.; Li, P.; Kattah, M.; Cohn, M. L.; Moran, K.; Tilley, E.; Richert, J. R.; Wang, S. M. Structure-based discovery of nonpeptidic small organic compounds to block the T cell response to myelin basic protein. *J. Med. Chem.* **2004**, *47* (21), 4989–4997.
- (43) Nikolovska-Coleska, Z.; Xu, L.; Hu, Z. J.; Tomita, Y.; Li, P.; Roller, P. P.; Wang, R. X.; Fang, X. L.; Guo, R. B.; Zhang, M. C.; Lippman, M. E.; Yang, D. J.; Wang, S. M. Discovery of embelin as a cell-permeable, small-molecular weight inhibitor of XLAP through structure-based computational screening of a traditional herbal medicine three-dimensional structure database. *J. Med. Chem.* **2004**, *47* (10), 2430–2440.
- (44) Gao, Y.; Dickerson, J. B.; Guo, F.; Zheng, J.; Zheng, Y. Rational design and characterization of a Rac GTPase-specific small molecule inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (20), 7618–7623.
- (45) Fujii, N.; Haresco, J. J.; Novak, K. A. P.; Stokoe, D.; Kuntz, I. D.; Guy, R. K. A selective irreversible inhibitor targeting a PDZ protein interaction domain. *J. Am. Chem. Soc.* **2003**, *125* (40), 12074–12075.
- (46) Debnath, A. K.; Radigan, L.; Jiang, S. B. Structure-based identification of small molecule antiviral compounds targeted to the gp41 core structure of the human immunodeficiency virus type 1. *J. Med. Chem.* **1999**, *42* (17), 3203–3209.
- (47) Boehr, D. D.; Nussinov, R.; Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009**, *5* (11), 789–796.
- (48) Arkin, M.; Lear, J. D. A new data analysis method to determine binding constants of small molecules to proteins using equilibrium analytical ultracentrifugation with absorption optics. *Anal. Biochem.* **2001**, *299* (1), 98–107.
- (49) Wells, S.; Menor, S.; Hespeneide, B.; Thorpe, M. F. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* **2005**, *2* (4), S127–S136.
- (50) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33* (12), 889–897.
- (51) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65* (3), 712–725.
- (52) Lu, Q.; Luo, R. A Poisson-Boltzmann dynamics method with nonperiodic boundary condition. *J. Chem. Phys.* **2003**, *119* (21), 11035–11047.
- (53) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **2004**, *55* (2), 383–394.
- (54) Zoete, V.; Irving, M. B.; Michielin, O. MM-GBSA binding free energy decomposition and T cell receptor engineering. *J. Mol. Recognit.* **2010**, *23* (2), 142–152.

- (55) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *J. Comput. Chem.* **2002**, *23* (1), 128–137.
- (56) Laskowski, R. A.; Macarthur, M. W.; Moss, D. S.; Thornton, J. M. Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.
- (57) de Hoon, M. J. L.; Imoto, S.; Nolan, J.; Miyano, S. Open source clustering software. *Bioinformatics* **2004**, *20* (9), 1453–1454.
- (58) Craig, I. R.; Pfleger, C.; Gohlke, H.; Essex, J. W.; Spiegel, K. Pocket-Space Maps To Identify Novel Binding-Site Conformations in Proteins. *J. Chem. Inf. Model.* **2011**, *51* (10), 2666–2679.
- (59) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Modell.* **1997**, *15* (6), 359–363.
- (60) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.
- (61) Irwin, J. J.; Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182.
- (62) Molinspiration, version 1; Molinspiration Cheminformatics: Slovensky Grob, Slovak Republic, 2008. www.molinspiration.com (accessed Nov. 2011).
- (63) Filter, version 2.0.2; OEChem, version 1.4.2; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2010. www.eyesopen.com (accessed Nov. 2011).
- (64) R: A Language and Environment for Statistical Computing, version 2.6.2; R Development Core Team: Vienna, Austria, 2008.
- (65) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330* (6002), 341–346.
- (66) Rickert, M.; Wang, X. Q.; Boulanger, M. J.; Goriatcheva, N.; Garcia, K. C. The structure of interleukin-2 complexed with its alpha receptor. *Science* **2005**, *308* (5727), 1477–1480.
- (67) Jiang, L.; Kuhlman, B.; Kortemme, T. A.; Baker, D. A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins* **2005**, *58* (4), 893–904.
- (68) Mobley, D. L.; Dill, K. A. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure* **2009**, *17* (4), 489–498.
- (69) Hu, G. D.; Wang, D. Y.; Liu, X. G.; Zhang, Q. G. A computational analysis of the binding model of MDM2 with inhibitors. *J. Comput.-Aided Mol. Des.* **2010**, *24* (8), 687–697.
- (70) Bradshaw, R. T.; Patel, B. H.; Tate, E. W.; Leatherbarrow, R. J.; Gould, I. R. Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction. *Protein Eng., Des. Sel.* **2011**, *24* (1–2), 197–207.
- (71) Hou, T. J.; Yu, R. Molecular dynamics and free energy studies on the wild-type and double mutant HIV-1 protease complexed with amprenavir and two amprenavir-related inhibitors: Mechanism for binding and drug resistance. *J. Med. Chem.* **2007**, *50* (6), 1177–1188.
- (72) Chene, P. Drugs targeting protein-protein interactions. *ChemMedChem* **2006**, *1* (4), 400–411.
- (73) Li, Y.; Huang, Y.; Swaminathan, C. P.; Smith-Gill, S. J.; Mariuzza, R. A. Magnitude of the hydrophobic effect at central versus peripheral sites in protein-protein interfaces. *Structure* **2005**, *13* (2), 297–307.
- (74) Fulle, S.; Christ, N. A.; Kestner, E.; Gohlke, H. HIV-1 TAR RNA spontaneously undergoes relevant apo-to-holo conformational transitions in molecular dynamics and constrained geometrical simulations. *J. Chem. Inf. Model.* **2010**, *50* (8), 1489–1501.
- (75) Gohlke, H.; Thorpe, M. F. A natural coarse graining for simulating large biomolecular motion. *Biophys. J.* **2006**, *91* (6), 2115–2120.
- (76) Zoete, V.; Meuwly, M.; Karplus, M. Study of the insulin dimerization: Binding free energy calculations and per-residue free energy decomposition. *Proteins* **2005**, *61* (1), 79–93.
- (77) Ballet, T.; Boulange, L.; Brechet, Y.; Bruckert, F.; Weidenhaupt, M. Protein conformational changes induced by adsorption onto material surfaces: an important issue for biomedical applications of material science. *Bull. Pol. Acad. Sci.: Tech. Sci.* **2010**, *58* (2), 303–315.
- (78) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47* (12), 3032–3047.
- (79) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-Ligand Docking against Non-Native Protein Conformers. *J. Chem. Inf. Model.* **2008**, *48* (11), 2214–2225.
- (80) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47* (1), 45–55.
- (81) Jacobsson, M.; Karlen, A. Ligand bias of scoring functions in structure-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46* (3), 1334–1343.
- (82) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23* (1–3), 3–25.
- (83) Ahmed, A.; Kazemi, S.; Gohlke, H. Protein flexibility and mobility in structure-based drug design. *Front. Drug Des. Discovery* **2007**, *3*, 455–476.
- (84) Kongsted, J.; Ryde, U. An improved method to predict the entropy term with the MM/PBSA approach. *J. Comput.-Aided Mol. Des.* **2009**, *23* (2), 63–71.
- (85) Ahmed, S.; Metpally, R. P. R.; Sangadala, S.; Reddy, B. V. B. Virtual screening and selection of drug-like compounds to block noggin interaction with bone morphogenetic proteins. *J. Mol. Graphics Modell.* **2010**, *28* (7), 670–682.
- (86) Gohlke, H.; Ahmed, A.; Rippmann, F.; Barnickel, G. A Normal Mode-Based Geometric Simulation Approach for Exploring Biologically Relevant Conformational Transitions in Proteins. *J. Chem. Inf. Model.* **2011**, *51* (7), 1604–1622.
- (87) Simmerling, C.; Elber, R. Hydrophobic Collapse in a Cyclic Hexapeptide - Computer-Simulations of Chdlfc and Caaaac in Water. *J. Am. Chem. Soc.* **1994**, *116* (6), 2534–2547.

13.12 Publication VI – Supporting Information

Hot Spots and Transient Pockets: Predicting the Determinants of Small-Molecule Binding to a Protein-Protein Interface

§Metz, A., §Pfleger, C., Kopitz, H., Pfeiffer-Marek, S., Baringhaus, K.-H., Gohlke, H.

J. Chem. Inf. Model. (2012), 52, 120-133

§ Both authors contributed equally to this work.

Supporting Information

Hot spots and transient pockets: Predicting the determinants of small-molecule binding to a protein-protein interface

Alexander Metz,^{1§} Christopher Pflieger,^{1§} Hannes Kopitz,¹ Stefania Pfeiffer-Marek,²
Karl-Heinz Baringhaus,² Holger Gohlke^{1*}

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

²Sanofi-Aventis Deutschland GmbH, CAS Drug Design, Frankfurt am Main, Germany

[§]Both authors contributed equally to this work.

^{*}Universitätsstr. 1, 40225 Düsseldorf, Germany. Phone: (+49) 211 81-13662.

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

Structure preparation

Starting structures for the simulations of human IL-2 and its complexes were taken from the Protein Data Bank¹ (PDB codes: 1m47, 1m4c, 1m48, 1m49, 1pw6, 1py2, 1qvn, and 1z92). These structures were modified to achieve consistency with respect to the sequence and number of amino acids. Solvent and buffer molecules were removed except for crystal waters bound to protein chains, which were considered in the MD simulations. Histidine protonation and rotation states were assigned manually such that all IL-2 chains have the same constitution and that histidines can form optimal local interactions. In the case of multiple identical chains, the one with the lowest number of unresolved residues was chosen. Missing residues (Figure S9) were modeled with MODELLER 7v7² using other IL-2 structures as templates, as was done for the Ala69Val mutation in the structure with PDB code 1QVN. The flexible loop between *Ser64* and *Leu100* (all IL-2R α residues are highlighted in italics, whereas all IL-2 residues are depicted in “normal” font) of IL-2R α was not resolved in the crystal structure (PDB code: 1z92).³ As the loop does not contact the binding interface,³ it was not considered any further. This should not influence the structural integrity of IL-2R α during MD simulations because either end of the loop is bound to the residual IL-2R α structure by a disulfide bond. Ligand structures were extracted from the complexes. For docking, the ligands were converted to MOL2 files using the PRODRG2⁴ server. Atom types were corrected manually if necessary. Flexible torsions were determined by AutoTors from the AutoDock suite of programs.⁵

Molecular dynamics simulations

MD simulations were performed with the AMBER 9 package of molecular simulation programs⁶ using the Cornell *et al.* force field⁷ with modifications introduced by Hornak *et al.* (ff99SB)⁸ and the general amber force field (GAFF)⁹ for proteins and small molecules, respectively. Partial charges of small molecules were generated according to the RESP procedure.⁹⁻¹⁰ The structures were solvated in a truncated octahedron of TIP3P water¹¹ such that the distance between the edges of the box and the closest solute atom was at least 11 Å. Periodic boundary conditions were applied using the particle mesh Ewald (PME) method¹² to treat long-range electrostatic interactions. Bond lengths involving bonds to hydrogen atoms were constrained by SHAKE.¹³⁻¹⁴ The time step for all MD simulations was 2 fs, and a direct-space non-bonded cutoff of 8 Å was applied. After minimization the system was heated from 100 K to 300 K using canonical ensemble (NVT) MD. Then, the solvent density was adjusted

using isothermal-isobaric ensemble (NPT) MD. Positional restraints applied during equilibration were reduced in a stepwise manner over 50 ps followed by 50 ps of unrestrained canonical ensemble (NVT) MD at 300 K with a time constant of 2 ps for heat bath coupling. Snapshots were extracted every 10 ps from production runs for further analysis (Table 1).

Docking

All docking runs were performed with AutoDock 3.05⁵ using DrugScore pair potentials¹⁵ as a scoring function.¹⁶⁻¹⁷ The docking protocol for flexible ligand docking comprised 100 independent runs per ligand using an initial population size of 100 individuals, 5.0×10^3 generations, a maximum number of 10.0×10^6 energy evaluations, a mutation rate of 0.02, a crossover rate of 0.8, and an elitism value of 1. For the enrichment evaluation the maximum number of energy evaluations and the population size were reduced to 3.0×10^6 and 50, respectively. Before calculating the DrugScore potential grids, all structures were aligned to the *x/y*-plane of the Cartesian coordinate system such that the rms distance between the interface amino acids and the plane is minimal. By doing so, the potential grids are optimally positioned for the mainly flat interface region of IL-2. The dimensions of the grids were chosen such that the grids extend beyond all hot spots as well as amino acids lining the identified interface pockets by at least 2.5 Å. In the case of *apo*-docking where no transient pocket is available, the same potential grid definition was chosen as for the re-docking approach. We note that this way no information about the known binding modes of the PPIM was considered for setting up the docking. The grid spacing was set to 0.375 Å. Similar docking poses (RMSD < 1 Å) were clustered, and the intermolecular docking energy was calculated. As the final docking result, the ligand pose with the lowest intermolecular docking energy from the largest cluster was chosen. A docking experiment was considered successful when this ligand pose had an RMSD < 2.0 Å to the native pose.

Statistical significance of MM-PB/SA results

To investigate the energetics of IL-2/IL-2R α and IL-2/small-molecule complex formation, the MM-PB/SA method was applied to compute effective energies as the sum of gas-phase energies and solvation free energies. Entropic terms resulting from translational, rotational, and vibrational contributions of the solutes were omitted. The gas-phase and solvation free energy values were averaged over 617 – 1379 snapshots (Table 1) taken at 10 ps intervals from the trajectories of the MD simulations. The correlation time for relaxation of effective energy fluctuations was computed to < 10 ps (data not shown), in agreement with related studies.¹⁸ Hence, the extracted snapshots should be uncorrelated, and mean values of binding effective energies computed by the single trajectory MM-PB/SA method can be estimated to within a standard error of the mean (SEM) between 0.13 – 0.37 kcal mol⁻¹ (Table 1).

Time-series of effective energies computed using the MM-PB/SA method are displayed in Figure S8 for snapshots of the unbound solutes and the IL-2/IL-2R α and IL-2/small-molecule complexes. In all cases, significant drifts and fluctuations in the absolute effective energies were found, which demonstrates the sensitivity of these values to conformational details and reflects structural variations throughout the MD trajectories. The observed energy drift (Table S5) depends on the size and conformational complexity of the solutes (Table 1) with IL-2/IL-2R α showing the largest drift (-11.02 kcal mol⁻¹ ns⁻¹), unbound IL-2 and the IL-2/small-molecule complexes showing drifts of -0.56 – -6.03 kcal mol⁻¹ ns⁻¹, and the small molecules showing negligible drifts of -0.30 – 0.20 kcal mol⁻¹ ns⁻¹.

These analyses indicate as to why MM-PB/SA binding effective energies computed by the multiple trajectory method for IL-2/small-molecule complexes do not correlate with experimental results ($R^2 < 0.1$, data not shown). In contrast, in the case of the single trajectory method, binding effective energies show a much smaller drift (-0.63 – 0.74 kcal mol⁻¹ ns⁻¹, Table S5) due to a cancellation of internal energies.¹⁸ These results also provide an explanation as to why differentiating between conformational states of IL-2 based on absolute effective energies is not successful (Figure S8), in addition to the error introduced by neglecting changes in the solute's configurational entropy. As the energy drifts are mainly caused by conformational transitions of the solute that occur, in particular, in modeled regions, loops, and termini, much longer simulation times would be required to obtain mean absolute effective energies that are stable over time. However, even when simulating for up to 10 ns in related studies,¹⁸⁻²³ this problem could not be alleviated, and comparable drifts were observed.

Tables

Table S1: Heavy atom RMSD during MD simulation

PDB code	RMSD ^a				
	Complex	IL-2 ^b	IL-2 interface ^c	Bound ligand ^d	Unbound ligand ^d
1m47	—	3.17 (3.63)	2.88 (3.70)	—	—
1m4c	—	3.16 (3.83)	2.44 (3.13)	—	—
1m48	2.57 (3.08)	2.55 (3.03)	2.42 (3.00)	1.46 (2.02)	2.66 (4.21)
1m49	2.66 (3.06)	2.68 (3.09)	2.13 (2.60)	1.06 (1.89)	2.64 (4.29)
1pw6	2.70 (3.25)	2.72 (3.27)	2.20 (2.86)	0.89 (1.46)	2.01 (4.76)
1py2	2.59 (3.09)	2.61 (3.12)	1.94 (2.36)	1.55 (2.03)	2.92 (4.94)
1qvn	2.88 (3.21)	2.81 (3.17)	2.27 (3.03)	2.20 (3.94)	2.81 (5.52)
1z92	3.50 (4.52)	3.46 (4.32)	2.49 (3.18)	2.98 (4.07)	3.77 (5.73)

^a Mean heavy atom RMSD with respect to the equilibrated structure; in Å. Five N-terminal amino acids of IL-2 were omitted. Maximum RMSD in parentheses.

^b Unbound IL-2 or IL-2 extracted from the MD trajectory of the complex.

^c IL-2 residues Tyr31, Asn33-Lys35, Thr37-Met39, Thr41-Tyr45, Glu60-Glu62, Lys64-Val69, Asn71, Leu72, and Met104-Thr111.

^d Aligned with respect to the ligand.

Table S2: RMSD of the IL-2 interface region of experimentally determined bound IL-2 conformations^a

IL-2 structure	IL-2/FRG	IL-2/CMM	IL-2/FRB	IL-2/FRH	IL-2/FRI
Unbound structure ^b	1.77	1.69	3.08	1.50	1.69
MD ^c	1.83	1.74	2.99	1.74	1.78
FRODA ^c	1.66	1.58	2.98	1.42	1.51

^a All heavy atoms of the interface region (IL-2 residues Tyr31, Asn33-Lys35, Thr37-Met39, Thr41-Tyr45, Glu60-Glu62, Lys64-Val69, Asn71, Leu72, Met104-Thr111) are considered; in Å.

^b RMSD from the unbound IL-2 conformation (PDB code: 1m47).

^c Minimal RMSD obtained from snapshots generated by either MD or FRODA simulation starting from the unbound IL-2 conformation (PDB code: 1m47).

Table S3: Ten largest pocket volumes of selected FRODA and MD snapshots

FRODA snapshot ^a	Volume ^b	MD snapshot ^c	Volume ^b
6	159	123	175
117	157	199	182
169	148	215	180
301	145	234	240
418	148	269	181
514	185	498	184
534	188	782	163
657	215	794	180
698	176	843	173
729	156	940	167

^a Consecutive number of the snapshot from a total of 1,000 snapshots uniformly extracted from the 10,000,000 FRODA-generated snapshots starting from the unbound IL-2 structure (PDB code: 1m47).

^b In Å³.

^c Consecutive number of the snapshot from a total of 1,021 snapshots 10 ps apart that were generated by MD starting from the unbound IL-2 structure (PDB code: 1m47).

Table S4: Pocket residues of IL-2 selected for the definition of the potential energy grids

Protein structure	Residues
Crystal structures ^a	Ile28, Tyr31-Tyr45, Cys58, Glu61-Pro65, Glu68-Lys76, Tyr107, Ile114
MD	Tyr31, Pro34-Lys35, Arg38-Met39, Thr41-Tyr45, Glu60-Glu62, Lys64-Val69, Asn71-Leu72, Cys105, Tyr107, Thr111
FRODA	Tyr31, Pro34-Lys35, Thr37-Met39, Thr41-Tyr45, Glu61-Glu62, Lys64-Val69, Asn71-Leu72

^a Pocket residues identified in PDB codes 1m48 (chain A,B), 1m49 (chain A,B), 1pw6 (chain A), 1py2 (chain A,B,C,D), and 1qvn (chain B,C,D).

Table S5: Drift of the effective energy

PDB code	Drift of effective energy ^a					
	Multiple trajectory method ^b			Single trajectory method ^c		Binding effective energy from single trajectory method ^c
	IL-2	ligand	complex	IL-2	ligand	
1m47	-1.45	—	—	—	—	—
1m4c	-0.56	—	—	—	—	—
1m48	—	0.03	-2.38	-2.33	0.13	0.20
1m49	—	-0.09	-3.08	-3.22	-0.05	0.19
1pw6	—	0.20	-4.55	-4.05	-0.23	-0.27
1py2	—	-0.30	-5.33	-4.96	0.16	-0.63
1qvn	—	-0.04	-6.03	-6.78	0.00	0.74
1z92	—	0.16	-11.02	-8.24	-3.27	0.49

^a In kcal mol⁻¹ ns⁻¹.^b Structures of IL-2, ligand, and complex were generated by separate MD simulations.^c Structures of IL-2 and ligand were extracted from the MD trajectory of the respective complexes.

Table S6: Selection of poses from docking into FRODA snapshots

Ligand	PDB code ^a	RMSD ^b	Score ^c	Clustered poses ^d	Cluster size ^e
FRG	1m48	2.08	-13.20	213	204
CMM	1m48	2.59	-14.81	335	327
FRB	1pw6	2.48	-12.78	105	94
FRH	1py2	4.30	-15.77	272	78
FRI	1qvn	3.22	-14.92	133	46

^a PDB code of the corresponding IL-2 complex structure.^b RMSD of the ligand pose with the lowest intermolecular docking energy in the largest cluster with respect to the native pose; in Å.^c In kcal mol⁻¹.^d The number of poses out of 1,000 docked poses (10 FRODA simulated structures with largest pocket volume times 100 docking runs) where the ligand's guanidinium group is within 5 Å of the side chain heavy atoms of Glu62 that were subjected to hierarchical complete linkage clustering with R²⁴ with a cluster distance of 5 Å.^e Number of ligand poses in the largest cluster.

Figures

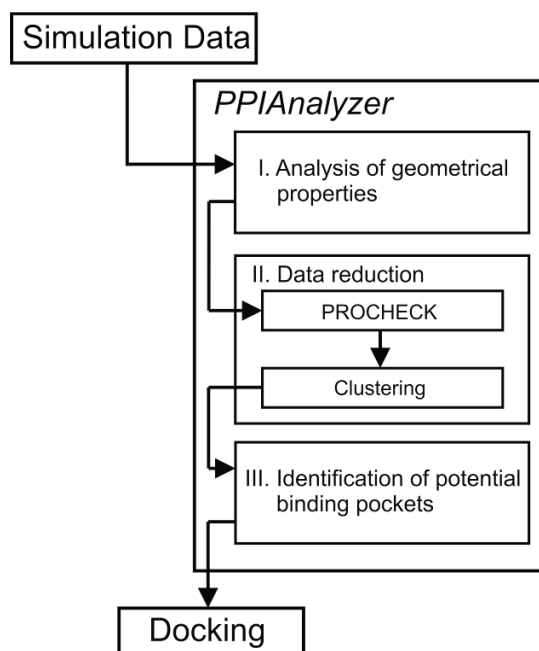


Figure S1: Workflow of the PPIAnalyzer method. The method contains three main steps: I. Analysis of geometrical properties in terms of root mean-square deviations (RMSD) and rotamer analysis. II. Reduction of the dataset by assessing the steric quality of the generated conformations and clustering with respect to the RMSD of heavy atoms of interface residues. III. Identification of transient pockets in the remaining conformations. Representative structures that show the largest interface pocket volume are then selected for the subsequent docking experiments.

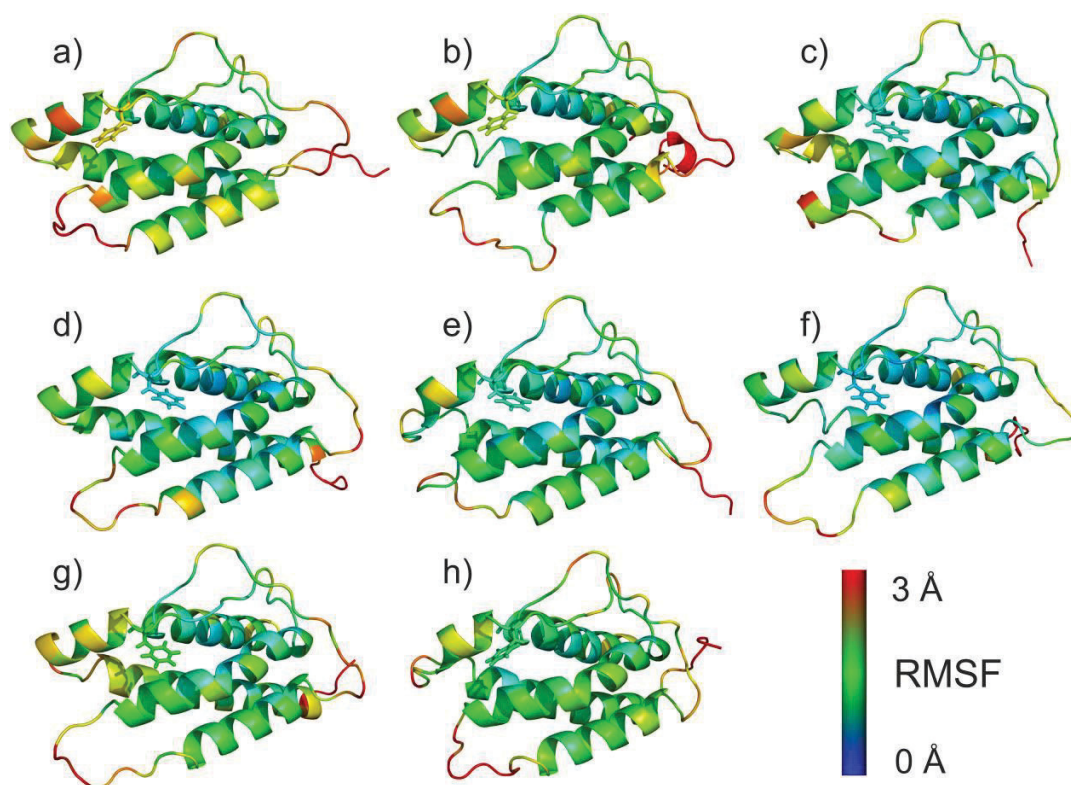


Figure S2: RMSF values of IL-2 residues obtained by MD simulations of the unbound and bound states. The RMSF value of each residue is calculated as the average over all atoms. RMSF values are color-coded onto the respective starting structure of the MD simulations: a) 1m47 and b) 1m4c for unbound IL-2; c) 1m48, d) 1m49, e) 1pw6, f) 1py2, and (g) 1qvn for IL-2 bound to PPIM; h) 1z92 for IL-2 bound to IL-2R α . The RMSF values were calculated for snapshots 10 ps apart. Prior to the RMSF calculations, all snapshots were structurally aligned to the starting structure of the MD simulation considering all heavy protein atoms. The highly mobile five N-terminal residues of IL-2 were neglected in the structural alignment. The protein is depicted in cartoon representation. Phe42 is depicted in stick representation to indicate the location of the small molecule binding pocket. Figures were generated by PyMOL.²⁵

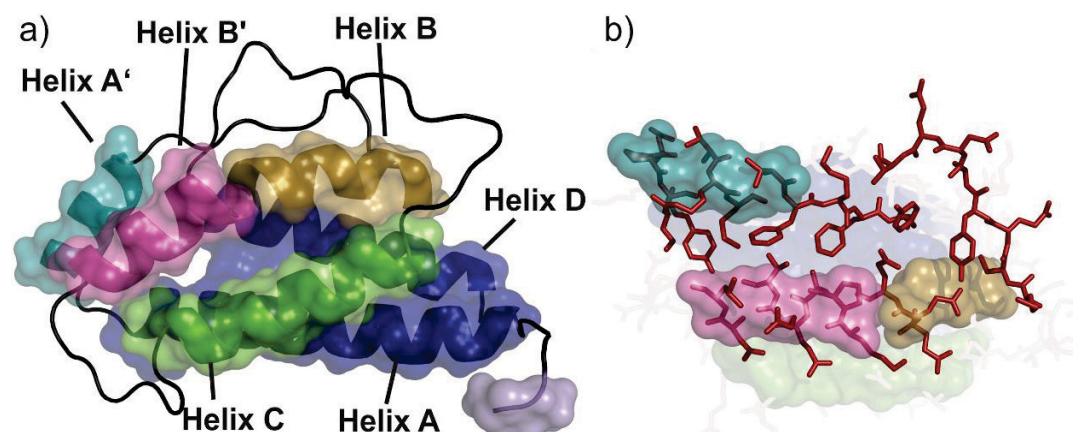


Figure S3: Rigid cluster decomposition obtained by FIRST. (a) The rigid clusters (transparent surfaces) are denominated RC1-6 in the order of decreasing size. RC1 (blue) covers helices A and D, RC2 (green) covers helix C, RC3 (magenta) covers parts of helix B', RC4 (turquoise) covers helix A', RC5 (gold) covers parts of helix B, and RC6 (light blue) is located at the N-terminus of IL-2. (b) 25.5% of all interface atoms are part of the rigid clusters RC3, RC4, and RC5. All flexible atoms (red) can move freely in FRODA simulations. Figures were generated by PyMOL.²⁵

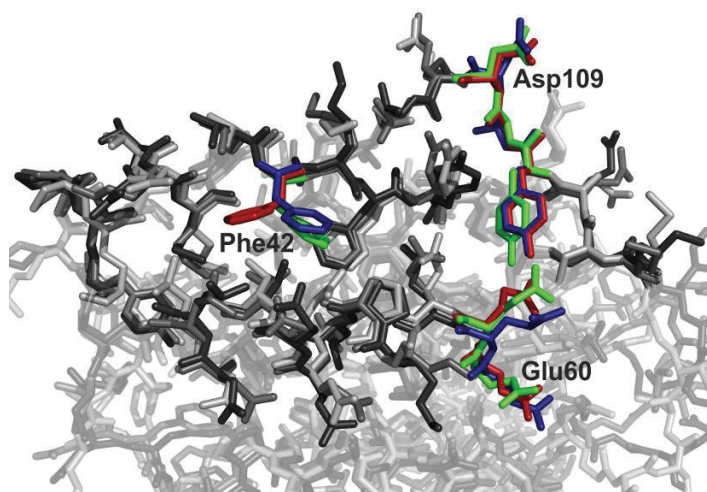


Figure S4: Overlay of the protein-protein interface region of IL-2 in unbound (red) and bound (green) conformation. Exemplarily, one snapshot from a FRODA simulation started from the unbound state is shown (blue), demonstrating that the movement of Phe42 can even be observed in the absence of the ligand, leading to a transient pocket opening. Regions for which no movements were observed by experiment (around Glu60 and Asp109) also remain immobile during the simulation. Figure was generated by PyMOL.²⁵

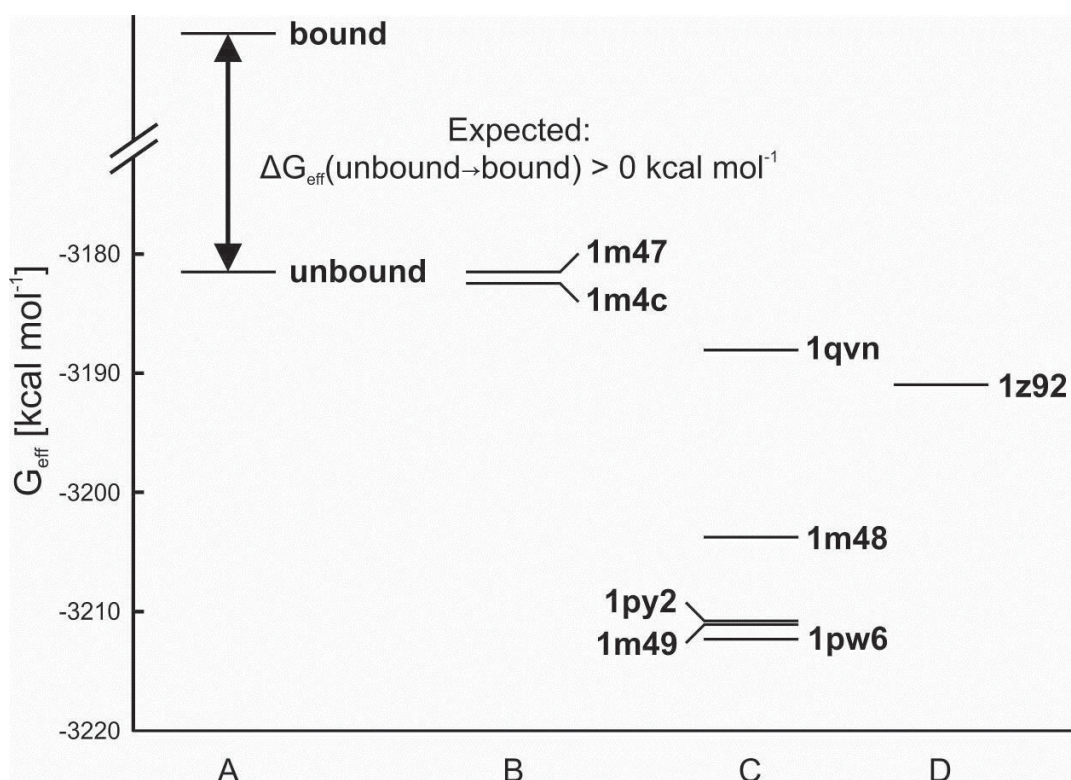


Figure S5: Mean absolute effective energies G_{eff} of IL-2 in its unbound and bound conformations. Conformational stress and changes in solvation and configurational entropy are expected to increase the free energy of a bound conformation over an unbound one (lane A). In contrast, computed G_{eff} of IL-2 extracted from MD trajectories of IL-2/small molecule complexes (lane C) or from the IL-2/IL-2R α complex (lane D) are lower than G_{eff} of unbound IL-2 (lane B). We attribute this observation to neglecting changes in configurational entropy upon the conformational transitions and the occurrence of significant drifts of G_{eff} over time (see Table S5 and Figure S8). Figure was generated by gnuplot.²⁶

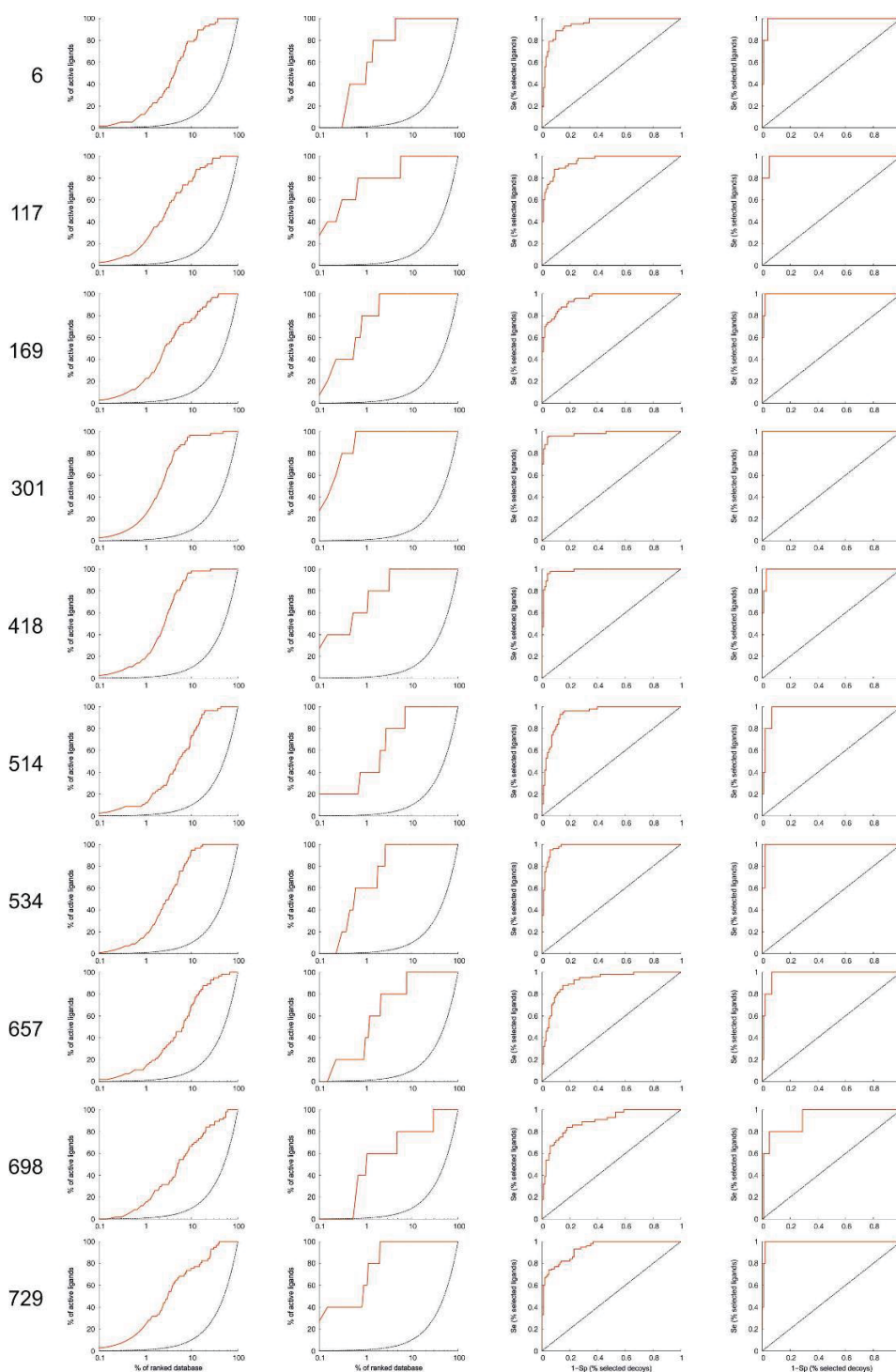


Figure S6: Docking enrichment of known IL-2 ligands. The number of the FRODA snapshot with a transient pocket used for docking is indicated in the left row. Enrichment plots for all 57 IL-2 ligands (1st vertical lane) and the five IL-2 ligands with available complex crystal structure (2nd lane) as well as ROC curves for all 57 IL-2 ligands (3rd lane) and the five IL-2 ligands with available complex crystal structure (4th lane) are given.

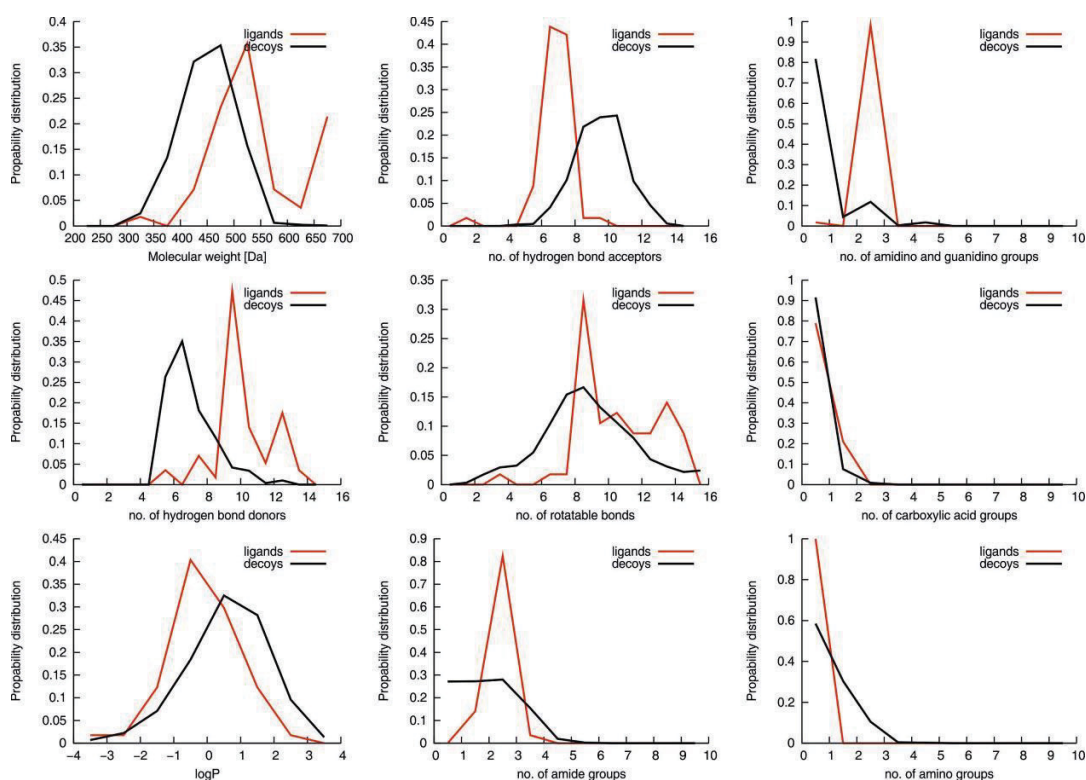


Figure S7: Property distribution of the known IL-2 ligands and decoys. The red line represents all 57 IL-2 ligands. The black line represents the decoy set generated with the aim of similar physicochemical properties to the five IL-2 ligands with available complex crystal structures following the DUD procedure. Figures were generated by gnuplot.²⁶

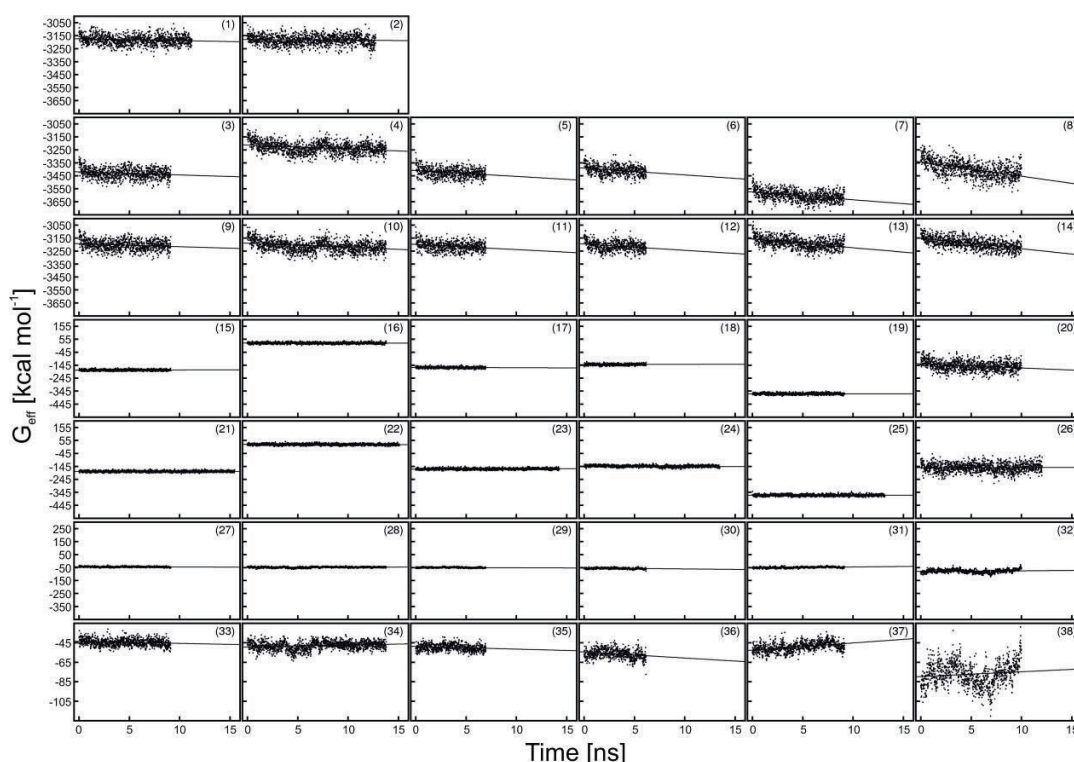


Figure S8: Time series of effective energies. The effective energies were calculated by applying the MM-PB/SA method to snapshots extracted every 10 ps from MD trajectories for: (1) unbound IL-2 [PDB-code: 1m47]; (2) unbound IL-2 [1m4c]; IL-2 in complex with (3) FRG [1m48], (4) CMM [1m49], (5) FRB [1pw6], (6) FRH [1py2], (7) FRI [1qvn], and (8) IL-2R α [1z92]; IL-2 extracted from the trajectories of the complexes of IL-2 with (9) FRG [1m48], (10) CMM [1m49], (11) FRB [1pw6], (12) FRH [1py2], (13) FRI [1qvn], and (14) IL-2R α [1z92]; IL-2 ligands extracted from the trajectories of the complexes of IL-2 with (15) FRG [1m48], (16) CMM [1m49], (17) FRB [1pw6], (18) FRH [1py2], (19) FRI [1qvn], and (20) IL-2R α [1z92]; unbound ligands of IL-2 (21) FRG [1m48], (22) CMM [1m49], (23) FRB [1pw6], (24) FRH [1py2], (25) FRI [1qvn], and (26) IL-2R α [1z92]. In addition, MM-PB/SA single trajectory binding effective energies are depicted for the complexes of IL-2 with (27) FRG [1m48], (28) CMM [1m49], (29) FRB [1pw6], (30) FRH [1py2], (31) FRI [1qvn], and (32) IL-2R α [1z92]. The range of the ordinate values is identical in all plots (1) – (32). For reasons of clarity, MM-PB/SA single trajectory binding effective energies are depicted again with a magnified ordinate scale for the complexes of IL-2 with (33) FRG [1m48], (34) CMM [1m49], (35) FRB [1pw6], (36) FRH [1py2], (37) FRI [1qvn], and (38) IL-2R α [1z92]. Figures were generated by gnuplot.²⁶

Determinants of small-molecule binding to a protein-protein interface – Metz, Pfeleger, Kopitz, Pfeiffer-Marek, Baringhaus, Gohlke

19

```

wt      APTSSSTKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEEELKPLE
1m47    -----STKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEEELKPLE
1m4C    -----STKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEEELKPLE
1m48    ---SSSTKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEEELKPLE
1m49    ---SSSTKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEEELKPLE
1pw6    ----SSTKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEEELKPLE
1py2    -----STKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEEELKPLE
1qvn    ---SSSTKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEEELKPLE
1z92    -----STKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEEELKPLE
          *****
wt      EVLNLAQSKNFHLRPRDLISNINVIVLELKGSETTFMCEYADETATIVEFLNRWITFCQSIISTLT
1m47    EVLNLAQ--NFHLRPRDLISNINVIVLELKG----FMCEYADETATIVEFLNRWITFCQSIISTLT
1m4C    EVLNLA-----RDLISNINVIVLELKG---FMCEYADETATIVEFLNRWITFCQSIISTLT
1m48    EVLNLAQSK---NFRDLISNINVIVLELKGSETTFMCEYADETATIVEFLNRWITFCQSIISTLT
1m49    EVLNLAQ-----RPRDLISNINVIVLELKGSETTFMCEYADETATIVEFLNRWITFCQSIISTLT
1pw6    EVLNLAQSKNFHLRPRDLISNINVIVLELKGSETTFMCEYADETATIVEFLNRWITFCQSIISTLT
1py2    EVLNLAQ-----RPRDLISNINVIVLELKG-ETTFMCEYADETATIVEFLNRWITFCQSIISTLT
1qvn    EALNLAQ-----RPRDLISNINVIVLELKGSETTFMCEYADETATIVEFLNRWITFCQSIISTLT
1z92    EVLNLA-----RPRDLISNINVIVLELKGSETTFMCEYADETATIVEFLNRWITFCQSIISTLT
          * . ****          *****          *****

```

Figure S9: Multiple sequence alignment of sequences of IL-2 crystal structures (PDB codes: 1m47, 1m4c, 1m48, 1m49, 1pw6, 1py2, 1qvn, and 1z92). Residues that have not been resolved are indicated by a dash (–) and were modeled using MODELLER 7v7² to match the full length wild-type sequence (wt). Ala69 of one of the crystal structures (PDB code: 1qvn) was mutated to alanine using MODELLER 7v7 to match the wt sequence. The multiple sequence alignment was created using CLUSTAL-W.²⁷

References

- (1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235-242.
- (2) Sali, A.; Blundell, T. L., Comparative Protein Modeling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234* (3), 779-815.
- (3) Rickert, M.; Wang, X. Q.; Boulanger, M. J.; Goriatcheva, N.; Garcia, K. C., The structure of interleukin-2 complexed with its alpha receptor. *Science* **2005**, *308* (5727), 1477-1480.
- (4) Schuettelkopf, A. W.; van Aalten, D. M. F., PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 1355-1363.
- (5) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19* (14), 1639-1662.
- (6) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A., AMBER 9, University of California, San Francisco. **2006**.
- (7) Wang, J. M.; Cieplak, P.; Kollman, P. A., How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21* (12), 1049-1074.
- (8) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65* (3), 712-725.
- (9) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25* (9), 1157-1174.
- (10) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A., Application of the Multimolecule and Multiconformational Resp Methodology to Biopolymers - Charge Derivation for DNA, Rna, and Proteins. *J. Comput. Chem.* **1995**, *16* (11), 1357-1377.
- (11) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926-935.
- (12) Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089-10092.
- (13) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327-341.
- (14) Miyamoto, S.; Kollman, P. A., Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. *J. Comput. Chem.* **1992**, *13* (8), 952-962.
- (15) Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295* (2), 337-356.
- (16) Sotriffer, C. A.; Gohlke, H.; Klebe, G., Docking into knowledge-based potential fields: A comparative evaluation of DrugScore. *J. Med. Chem.* **2002**, *45* (10), 1967-1970.
- (17) Pfeiffer, P.; Gohlke, H., DrugScore(RNA) - Knowledge-based scoring function to predict RNA-ligand interactions. *J. Chem. Inf. Model.* **2007**, *47* (5), 1868-1876.
- (18) Gohlke, H.; Kiel, C.; Case, D. A., Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RaIGDS complexes. *J. Mol. Biol.* **2003**, *330* (4), 891-913.

- (19) Gohlke, H.; Case, D. A., Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* **2004**, *25* (2), 238-250.
- (20) Cui, Q. Z.; Sulea, T.; Schrag, J. D.; Munger, C.; Hung, M. N.; Naim, M.; Cygler, M.; Purisima, E. O., Molecular dynamics-solvated interaction energy studies of protein-protein interactions: The MP1-p14 scaffolding complex. *J. Mol. Biol.* **2008**, *379* (4), 787-802.
- (21) Deng, N. J.; Cieplak, P., Insights into affinity and specificity in the complexes of alpha-lytic protease and its inhibitor proteins: binding free energy from molecular dynamics simulation. *Phys. Chem. Chem. Phys.* **2009**, *11* (25), 4968-4981.
- (22) Zoete, V.; Michielin, O., Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein-protein association using MM-GBSA: application to the TCR-p-MHC complex. *Proteins* **2007**, *67* (4), 1026-47.
- (23) Tuncel, A.; Kavakli, I. H.; Keskin, O., Insights into subunit interactions in the heterotetrameric structure of potato ADP-glucose pyrophosphorylase. *Biophys. J.* **2008**, *95* (8), 3628-3639.
- (24) R Development Core Team, *A Language and Environment for Statistical Computing*, 2.6.2; Vienna, Austria. **2008**.
- (25) DeLano, W. L. *PyMOL Molecular Graphics System*, 0.99rc6; DeLano Scientific LLC: 2006.
- (26) Williams T.; Kelley C.; Broeker H.B.; E.A., M., *gnuplot - An Interactive Plotting Program*. **2007**.
- (27) Thompson, J. D.; Higgins, D. G.; Gibson, T. J., Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Res.* **1994**, *22* (22), 4673-4680.

14 CURRICULUM VITAE

Personal Information

Name: Christopher Pfleger
Address: Ickerswarder Str. 11, 40589 Düsseldorf
Date and place of birth: 24th of September, 1977 in Frankfurt am Main

Education and Professional Training

- 2009 – to date **PhD Student**, Heinrich-Heine-University, Düsseldorf
Advisor: Prof. Dr. Holger Gohlke
- 2008 – 2009 **PhD Student**, Christian-Albrechts-University, Kiel
Advisor: Prof. Dr. Holger Gohlke
- 2006 – 2007 **Diploma in Bioinformatics (FH)**, University of Applied Science, Bingen;
work done at Goethe-University, Frankfurt
Topic : *“Protein-protein interfaces investigated by computer simulations”*
Advisor: Prof. Dr. Holger Gohlke
- 2005 – 2006 **Practical semester**, department „Bio- and Chemoinformatics”,
Merck KGaA, Darmstadt
Topic: *“Evaluation of a database for X-ray crystal structures.
Development of a Perl/CGI program to prepare X-ray data in an
automated way”*
Advisor: Dr. Ulrich Grädler
- 2002 – 2007 **Studies in Bioinformatics**, University of Applied Science, Bingen
- 1998 – 2001 **Training as laboratory assistant**, Boehringer Ingelheim Pharma KG,
Ingelheim am Rhein

Work experience

- 2007 – 2008 **Freelancer**, Goethe-University, Frankfurt
- 2003 – 2005 **Student assistant**, department „Drug Delivery”,
Boehringer Ingelheim Pharma KG, Ingelheim am Rhein
- 2001 – 2002 **Technical assistant**, department „Drug Delivery”,
Boehringer Ingelheim Pharma KG, Ingelheim am Rhein

Publications

Pfleger C, Minges ARM, Gohlke H, *Allosteric Coupling deduced from Altered Rigidity Percolation in Biomacromolecules*. Submitted manuscript (2014).

Pfleger C, Gohlke H, *Efficient and Robust Analysis of Biomacromolecular Flexibility Using Ensembles of Network Topologies Based on Fuzzy Noncovalent Constraints*. Structure **2013**, 21, 1725-1734.

Krüger DM, Rath PC, Pfleger C, Gohlke H, *CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo)stability, and function*. Nucleic Acids Res. **2013**, 41, W340-348.

Pfleger C[§], Rath PC[§], Klein DL, Radestock S, Gohlke H, *Constraint Network Analysis (CNA): A Python Software Package for Efficiently Linking Biomolecular Structure, Flexibility, (Thermo-)Stability, and Function*. J. Chem. Inf. Model. **2013**, 53, 1007-1015.

Pfleger C, Radestock S, Schmidt E, Gohlke H, *Global and Local Indices for Characterizing Biomolecular Flexibility and Rigidity*. J. Comput. Chem. **2013**, 34, 220-233.

Metz A[§], Pfleger C[§], Kopitz H, Pfeiffer-Marek S, Baringhaus K-H, Gohlke H, *Hot Spots and Transient Pockets: Predicting the Determinants of Small-Molecule Binding to a Protein-Protein Interface*. J. Chem. Inf. Model. **2012**, 52, 120-133.

Craig IR, Pfleger C, Gohlke H, Essex JW, Spiegel K, *Pocket-Space Maps to Identify Novel Binding-Site Conformations in Proteins*. J. Chem. Inf. Model. **2011**, 51, 2666-2679.

Rath PC, Pfleger C, Fulle S, Klein DL, Gohlke H: *Modeling of Molecular Properties. Statics of Biomacromolecules*, ed. Comba P (Wiley-VCH, Weinheim), **2011**, pp 281-299.

[§] Both authors contributed equally to this work.

Oral and poster presentations

Poster presentation at the 3rd Indo-German conference on Modelling Chemical and Biological (Re) activity (MCBR3), Chandigarh, India, **2013**. *Allosteric regulation based on changes in biomolecular flexibility and rigidity*. (Awarded the "best poster award".)

Oral presentation at the 25th Molecular Modeling Workshop 2011, Erlangen, Germany, **2011**. *Robust and efficient analysis of biomacromolecular stability using ensembles of random network topologies*.

Poster presentation at the 23th Molecular Modeling Workshop 2009, Erlangen, Germany, **2009**. *Improved consistency of protein flexibility analysis by fluctuating hydrogen bond networks*.

Poster presentation at the 21st "Darmstadt" Molecular Modeling Workshop, Erlangen, Germany, **2007**. *Plasticity of protein-protein interfaces investigated by MD and constrained geometric simulations*.

15 REFERENCES

1. Hardy JA, Wells JA (2004) Searching for new allosteric sites in enzymes. *Curr. Opin. Struct. Biol.* 14(6):706-715.
2. Pardee AB (2006) Regulatory molecular biology. *Cell Cycle* 5(8):846-852.
3. Monod J, Wyman J, Changeux JP (1965) On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* 12:88-118.
4. Tsai CJ, del Sol A, Nussinov R (2008) Allostery: absence of a change in shape does not imply that allostery is not at play. *J. Mol. Biol.* 378(1):1-11.
5. Lee GM, Craik CS (2009) Trapping moving targets with small molecules. *Science* 324(5924):213-215.
6. Holst B, Brandt E, Bach A, Heding A, Schwartz TW (2005) Nonpeptide and peptide growth hormone secretagogues act both as ghrelin receptor agonist and as positive or negative allosteric modulators of ghrelin signaling. *Mol. Endocrinol.* 19(9):2400-2411.
7. Jeffrey PD, *et al.* (1995) Mechanism of CDK activation revealed by the structure of a cyclin-CDK2 complex. *Nature* 376(6538):313-320.
8. Leger D, Herve G (1988) Allostery and pKa changes in aspartate transcarbamoylase from *Escherichia-Coli* - analysis of the pH-dependence in the isolated catalytic subunits. *Biochemistry* 27(12):4293-4298.
9. McMahon B, Frauenfelder H, Austin R, Chu K, Groves JT (2001) The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Biophys. J.* 80(1):286a-286a.
10. Fan YX, McPhie P, Miles EW (2000) Regulation of tryptophan synthase by temperature, monovalent cations, and an allosteric ligand. Evidence from arrhenius plots, absorption spectra, and primary kinetic isotope effects. *Biochemistry* 39(16):4692-4703.
11. Ozkan E, Yu HT, Deisenhofer J (2005) Mechanistic insight into the allosteric activation of a ubiquitin-conjugating enzyme by RING-type ubiquitin ligases. *Proc. Natl. Acad. Sci. U. S. A.* 102(52):18890-18895.
12. Luther KB, Haltiwanger RS (2009) Role of unusual O-glycans in intercellular signaling. *Int. J. Biochem. Cell Biol.* 41(5):1011-1024.
13. Volkman BF, Lipson D, Wemmer DE, Kern D (2001) Two-state allosteric behavior in a single-domain signaling protein. *Science* 291(5512):2429-2433.
14. Cho HS, *et al.* (2000) NMR structure of activated CheY. *J. Mol. Biol.* 297(3):543-551.
15. Kemp RG, Foe LG (1983) Allosteric regulatory properties of muscle phosphofructokinase. *Mol. Cell. Biochem.* 57(2):147-154.
16. Goodey NM, Benkovic SJ (2008) Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.* 4(8):474-482.
17. Bohr C, Hasselbalch K, Krogh A (1904) Ueber einen in biologischer Beziehung wichtigen Einfluss, den die Kohlensäurespannung des Blutes auf dessen Sauerstoffbindung übt1. *Skand. Arch. Physiol.* 16(2):402-412.
18. Hill AV (1910) The possible effects of the aggregation of the molecules of hæmoglobin on its dissociation curves. *J. Physiol.* 40:i--vii.
19. Kendrew JC, *et al.* (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181(4610):662-666.
20. Perutz MF, *et al.* (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis. *Nature* 185(4711):416-422.

21. Changeux JP (1961) The feedback control mechanisms of biosynthetic L-threonine deaminase by L-isoleucine. *Cold Spring Harbor Symp. Quant. Biol.* 26:313-318.
22. Changeux JP (2011) 50th anniversary of the word "allosteric". *Protein Sci.* 20(7):1119-1124.
23. Koshland DE, Jr., Nemethy G, Filmer D (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* 5(1):365-385.
24. Cooper A, Dryden DT (1984) Allostery without conformational change. A plausible model. *Eur. Biophys. J.* 11(2):103-109.
25. Williamson MP, Havel TF, Wuthrich K (1985) Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *J. Mol. Biol.* 182(2):295-315.
26. Popovych N, Sun SJ, Ebright RH, Kalodimos CG (2006) Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.* 13(9):831-838.
27. Nelson DL, Cox MM, Lehninger AL, Begnen K (2005) *Lehninger Biochemie* (Springer).
28. Kern D, Zuiderweg ER (2003) The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.* 13(6):748-757.
29. Conway A, Koshland DE, Jr. (1968) Negative cooperativity in enzyme action. The binding of diphosphopyridine nucleotide to glyceraldehyde 3-phosphate dehydrogenase. *Biochemistry* 7(11):4011-4023.
30. Levitzki A, Koshland DE, Jr. (1969) Negative cooperativity in regulatory enzymes. *Proc. Natl. Acad. Sci. U. S. A.* 62(4):1121-1128.
31. Changeux JP (2013) The concept of allosteric modulation: an overview. *Drug Discov. Today* 10(2):e223-228.
32. Koshland DE, Hamadani K (2002) Proteomics and models for enzyme cooperativity. *J. Biol. Chem.* 277(49):46841-46844.
33. Perutz MF (1970) Stereochemistry of cooperative effects in haemoglobin. *Nature* 228(5273):726-&.
34. Perutz MF, Wilkinson AJ, Paoli M, Dodson GG (1998) The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *Annu. Rev. Biophys. Biomol. Struct.* 27:1-34.
35. Gunasekaran K, Ma BY, Nussinov R (2004) Is allostery an intrinsic property of all dynamic proteins? *Proteins* 57(3):433-443.
36. Swain JF, Gierasch LM (2006) The changing landscape of protein allostery. *Curr. Opin. Struct. Biol.* 16(1):102-108.
37. Frauenfelder H, Parak F, Young RD (1988) Conformational substates in proteins. *Annu. Rev. Biophys. Biophys. Chem.* 17:451-479.
38. Cui Q, Karplus M (2008) Allostery and cooperativity revisited. *Protein Sci.* 17(8):1295-1307.
39. Weber G (1972) Ligand binding and internal equilibiums in proteins. *Biochemistry* 11(5):864-878.
40. Pan H, Lee JC, Hilser VJ (2000) Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble. *Proc. Natl. Acad. Sci. U. S. A.* 97(22):12020-12025.
41. Daily MD, Gray JJ (2007) Local motions in a benchmark of allosteric proteins. *Proteins* 67(2):385-399.
42. Ho BK, Agard DA (2010) Conserved tertiary couplings stabilize elements in the PDZ fold, leading to characteristic patterns of domain conformational flexibility. *Protein Sci.* 19(3):398-411.
43. Petit CM, Zhang J, Sapienza PJ, Fuentes EJ, Lee AL (2009) Hidden dynamic allostery in a PDZ domain. *Proc. Natl. Acad. Sci. U. S. A.* 106(43):18249-18254.

44. Fuentes EJ, Der CJ, Lee AL (2004) Ligand-dependent dynamics and intramolecular signaling in a PDZ domain. *J. Mol. Biol.* 335(4):1105-1115.
45. Hilser VJ, Wrabl JO, Motlagh HN (2012) Structural and energetic basis of allostery. *Annu. Rev. Biophys.* 41:585-609.
46. Clarkson MW, Gilmore SA, Edgell MH, Lee AL (2006) Dynamic coupling and allosteric behavior in a nonallosteric protein. *Biochemistry* 45(25):7693-7699.
47. Clarkson MW, Lee AL (2004) Long-range dynamic effects of point mutations propagate through side chains in the serine protease inhibitor eglin c. *Biochemistry* 43(39):12448-12458.
48. Williams DC, Benjamin DC, Poljak RJ, Rule GS (1996) Global changes in amide hydrogen exchange rates for a protein antigen in complex with three different antibodies. *J. Mol. Biol.* 257(4):866-876.
49. Freire E (1999) The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme. *Proc. Natl. Acad. Sci. U. S. A.* 96(18):10118-10122.
50. Amaro RE, Sethi A, Myers RS, Davisson VJ, Luthey-Schulten ZA (2007) A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. *Biochemistry* 46(8):2156-2173.
51. Masterson LR, Mascioni A, Traaseth NJ, Taylor SS, Veglia G (2008) Allosteric cooperativity in protein kinase A. *Proc. Natl. Acad. Sci. U. S. A.* 105(2):506-511.
52. Rakauskaitė R, Dinman JD (2008) rRNA mutants in the yeast peptidyltransferase center reveal allosteric information networks and mechanisms of drug resistance. *Nucleic Acids Res.* 36(5):1497-1507.
53. Torres M, Fernandez-Fuentes N, Fiser A, Casadevall A (2007) Exchanging murine and human immunoglobulin constant chains affects the kinetics and thermodynamics of antigen binding and chimeric antibody autoreactivity. *PLoS One* 2(12).
54. Zhuravleva A, Gierasch LM (2011) Allosteric signal transmission in the nucleotide-binding domain of 70-kDa heat shock protein (Hsp70) molecular chaperones. *Proc. Natl. Acad. Sci. U. S. A.* 108(17):6987-6992.
55. Rivalta I, *et al.* (2012) Allosteric pathways in imidazole glycerol phosphate synthase. *Proc. Natl. Acad. Sci. U. S. A.* 109(22):E1428-E1436.
56. Malmendal A, Evenas J, Forsen S, Akke M (1999) Structural dynamics in the C-terminal domain of calmodulin at low calcium levels. *J. Mol. Biol.* 293(4):883-899.
57. Berger C, *et al.* (1999) Antigen recognition by conformational selection. *FEBS Lett.* 450(1-2):149-153.
58. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286(5438):295-299.
59. Tsai CJ, del Sol A, Nussinov R (2009) Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. *Mol. Biosyst.* 5(3):207-216.
60. del Sol A, Tsai CJ, Ma BY, Nussinov R (2009) The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure* 17(8):1042-1050.
61. Fenton AW, Reinhart GD (2003) Mechanism of substrate inhibition in Escherichia coli phosphofructokinase. *Biochemistry* 42(43):12676-12681.
62. Fenton AW, Paricharttanakul NM, Reinhart GD (2004) Disentangling the web of allosteric communication in a homotetramer: Heterotropic activation in phosphofructokinase from Escherichia coli. *Biochemistry* 43(44):14104-14110.
63. Fenton AW, Reinhart GD (2002) Isolation of a single activating allosteric interaction in phosphofructokinase from Escherichia coli. *Biochemistry* 41(45):13410-13416.

-
64. Ortigosa AD, Kimmel JL, Reinhart GD (2004) Disentangling the web of allosteric communication in a homotetramer: heterotropic inhibition of phosphofructokinase from *Bacillus stearothermophilus*. *Biochemistry* 43(2):577-586.
 65. Kimmel JL, Reinhart GD (2001) Isolation of an individual allosteric interaction in tetrameric phosphofructokinase from *Bacillus stearothermophilus*. *Biochemistry* 40(38):11623-11629.
 66. Nelson SW, Honzatko RB, Fromm HJ (2002) Hybrid tetramers of porcine liver fructose-1,6-bisphosphatase reveal multiple pathways of allosteric inhibition. *J. Biol. Chem.* 277(18):15539-15545.
 67. Grant GA, Xu XL, Hu ZQ (2004) Quantitative relationships of site to site interaction in *Escherichia coli* D-3-phosphoglycerate dehydrogenase revealed by asymmetric hybrid tetramers. *J. Biol. Chem.* 279(14):13452-13460.
 68. Faga LA, Sorensen BR, VanScyoc WS, Shea MA (2003) Basic interdomain boundary residues in calmodulin decrease calcium affinity of sites I and II by stabilizing helix-helix interactions. *Proteins* 50(3):381-391.
 69. Jaren OR, Kranz JK, Sorensen BR, Wand AJ, Shea MA (2002) Calcium-induced conformational switching of *Paramecium* calmodulin provides evidence for domain coupling. *Biochemistry* 41(48):14158-14166.
 70. VanScyoc WS, *et al.* (2002) Calcium binding to calmodulin mutants monitored by domain-specific intrinsic phenylalanine and tyrosine fluorescence. *Biophys. J.* 83(5):2767-2780.
 71. Sorensen BR, Faga LA, Hultman R, Shea MA (2002) An interdomain linker increases the thermostability and decreases the calcium affinity of the calmodulin N-domain. *Biochemistry* 41(1):15-20.
 72. Sun H, Yin D, Coffeen LA, Shea MA, Squier TC (2001) Mutation of Tyr(138) disrupts the structural coupling between the opposing domains in vertebrate calmodulin. *Biochemistry* 40(32):9605-9617.
 73. van Westen GJ, Gaulton A, Overington JP (2014) Chemical, target, and bioactive properties of allosteric modulation. *PLoS Comput. Biol.* 10(4):e1003559.
 74. Salvesen GS, Riedl SJ (2007) Caspase inhibition, specifically. *Structure* 15(5):513-514.
 75. Schweizer A, *et al.* (2007) Inhibition of caspase-2 by a designed ankyrin repeat protein: Specificity, structure, and inhibition mechanism. *Structure* 15(5):625-636.
 76. Christopoulos A, Kenakin T (2002) G protein-coupled receptor allostery and complexing. *Pharmacol. Rev.* 54(2):323-374.
 77. May LT, Leach K, Sexton PM, Christopoulos A (2007) Allosteric modulation of G protein-coupled receptors. *Annu. Rev. Pharmacol. Toxicol.* 47:1-51.
 78. Noble MEM, Endicott JA, Johnson LN (2004) Protein kinase inhibitors: insights into drug design from structure. *Science* 303(5665):1800-1805.
 79. Vintonyak VV, Waldmann H, Rauh D (2011) Using small molecules to target protein phosphatases. *Bioorg. Med. Chem.* 19(7):2145-2155.
 80. Wiesmann C, *et al.* (2004) Allosteric inhibition of protein tyrosine phosphatase 1B. *Nat. Struct. Mol. Biol.* 11(8):730-737.
 81. Zhang S, Zhang ZY (2007) PTP1B as a drug target: recent developments in PTP1B inhibitor discovery. *Drug Discov. Today* 12(9-10):373-381.
 82. Montalibet J, Kennedy BP (2005) Therapeutic strategies for targeting PTP1B in diabetes. *Drug Discov. Today Ther. Strateg.* 2(2):129-135.
 83. Galzi JL, Changeux JP (1994) Neurotransmitter-gated ion channels as unconventional allosteric proteins. *Curr. Opin. Struct. Biol.* 4(4):554-565.
 84. Richter L, *et al.* (2012) Diazepam-bound GABAA receptor models identify new benzodiazepine binding-site ligands. *Nat. Chem. Biol.* 8(5):455-464.

85. Sawyer GW, Chiara DC, Olsen RW, Cohen JB (2002) Identification of the bovine gamma-aminobutyric acid type A receptor alpha subunit residues photolabeled by the imidazobenzodiazepine [3H]Ro15-4513. *J. Biol. Chem.* 277(51):50036-50045.
86. Arkin MR, Wells JA (2004) Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat. Rev. Drug Discov.* 3(4):301-317.
87. Jubb H, Higuero AP, Winter A, Blundell TL (2012) Structural biology and drug discovery for protein-protein interactions. *Trends Pharmacol. Sci.* 33(5):241-248.
88. Metz A, Ciglia E, Gohlke H (2012) Modulating protein-protein interactions: from structural determinants of binding to druggability prediction to application. *Curr. Pharm. Des.* 18(30):4630-4647.
89. Jones S, Thornton JM (1996) Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* 93(1):13-20.
90. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* 285(5):2177-2198.
91. Liu G (2001) Small molecule antagonists of the LFA-1/ICAM-1 interaction as potential therapeutic agents. *Expert Opin. Ther. Pat.* 11(9):1383-1393.
92. Gadek TR, *et al.* (2002) Generation of an LFA-1 antagonist by the transfer of the ICAM-1 immunoregulatory epitope to a small molecule. *Science* 295(5557):1086-1089.
93. Crump MP, *et al.* (2004) Structure of an allosteric inhibitor of LFA-1 bound to the I-domain studied by crystallography, NMR, and calorimetry. *Biochemistry* 43(9):2394-2404.
94. Guckian KM, *et al.* (2008) Design and synthesis of a series of meta aniline-based LFA-1 ICAM inhibitors. *Bioorg. Med. Chem. Lett.* 18(19):5249-5251.
95. Weitz-Schmidt G, Chreng S, Riek S (2009) Allosteric LFA-1 inhibitors modulate natural killer cell function. *Mol. Pharmacol.* 75(2):355-362.
96. Nussinov R, Tsai CJ, Ma BY (2013) The underappreciated role of allostery in the cellular network. *Annu. Rev. Biophys.* 42:169-189.
97. Canals M, *et al.* (2012) A Monod-Wyman-Changeux mechanism can explain G protein-coupled receptor (GPCR) allosteric modulation. *J. Biol. Chem.* 287(1):650-659.
98. Liu W, *et al.* (2012) Structural basis for allosteric regulation of GPCRs by sodium ions. *Science* 337(6091):232-236.
99. Huang ZM, *et al.* (2011) ASD: a comprehensive database of allosteric proteins and modulators. *Nucleic Acids Res.* 39:D663-D669.
100. Koon N, Squire CJ, Baker EN (2004) Crystal structure of LeuA from Mycobacterium tuberculosis, a key enzyme in leucine biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* 101(22):8295-8300.
101. Zhang P, *et al.* (2009) Molecular basis of the inhibitor selectivity and insights into the feedback inhibition mechanism of citramalate synthase from Leptospira interrogans. *Biochem. J.* 421:133-143.
102. Cross PJ, Dobson RC, Patchett ML, Parker EJ (2011) Tyrosine latching of a regulatory gate affords allosteric control of aromatic amino acid biosynthesis. *J. Biol. Chem.* 286(12):10216-10224.
103. Pedreno S, Pisco JP, Larrouy-Maumus G, Kelly G, de Carvalho LPS (2012) Mechanism of feedback allosteric inhibition of ATP phosphoribosyltransferase. *Biochemistry* 51(40):8027-8038.
104. Frantom PA, Zhang HM, Emmett MR, Marshall AG, Blanchard JS (2009) Mapping of the allosteric network in the regulation of alpha-isopropylmalate synthase from Mycobacterium tuberculosis by the feedback inhibitor L-leucine: solution-phase H/D exchange monitored by FT-ICR mass spectrometry. *Biochemistry* 48(31):7457-7464.

105. Wooll JO, *et al.* (2001) Structural and functional linkages between subunit interfaces in mammalian pyruvate kinase. *J. Mol. Biol.* 312(3):525-540.
106. Lee JC (2008) Modulation of allostery of pyruvate kinase by shifting of an ensemble of microstates. *Acta Biochim. Biophys. Sin.* 40(7):663-669.
107. Yan YW, *et al.* (2004) Inhibition of a mitotic motor protein: Where, how, and conformational consequences. *J. Mol. Biol.* 335(2):547-554.
108. Kleckner IR, Foster MP (2011) An introduction to NMR-based approaches for measuring protein dynamics. *Biochim. Biophys. Acta, Proteins Proteomics* 1814(8):942-968.
109. Bhattacharya A, *et al.* (2009) Allostery in Hsp70 chaperones is transduced by subdomain rotations. *J. Mol. Biol.* 388(3):475-490.
110. Tzeng SR, Kalodimos CG (2009) Dynamic activation of an allosteric regulatory protein. *Nature* 462(7271):368-372.
111. Busenlehner LS, Armstrong RN (2005) Insights into enzyme structure and dynamics elucidated by amide H/D exchange mass spectrometry. *Arch. Biochem. Biophys.* 433(1):34-46.
112. Manley G, Loria JP (2012) NMR insights into protein allostery. *Arch. Biochem. Biophys.* 519(2):223-231.
113. Selvaratnam R, Chowdhury S, VanSchouwen B, Melacini G (2011) Mapping allostery through the covariance analysis of NMR chemical shifts. *Proc. Natl. Acad. Sci. U. S. A.* 108(15):6133-6138.
114. Dawson JE, Farber PJ, Forman-Kay JD (2013) Allosteric coupling between the intracellular coupling helix 4 and regulatory sites of the first nucleotide-binding domain of CFTR. *PLoS One* 8(9).
115. Forster T (1948) Zwischenmolekulare Energiewanderung Und Fluoreszenz. *Ann. Phys. (Berlin)* 2(1-2):55-75.
116. Ramanoudjame G, Du M, Mankiewicz KA, Jayaraman V (2006) Allosteric mechanism in AMPA receptors: A FRET-based investigation of conformational changes. *Proc. Natl. Acad. Sci. U. S. A.* 103(27):10473-10478.
117. Johnson CK (2006) Calmodulin, conformational states, and calcium signaling. A single-molecule perspective. *Biochemistry* 45(48):14233-14246.
118. Taraska JW, Puljung MC, Olivier NB, Flynn GE, Zagotta WN (2009) Mapping the structure and conformational movements of proteins with transition metal ion FRET. *Nat. Meth.* 6(7):532-U594.
119. Salsbury FR (2010) Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Curr. Opin. Pharmacol.* 10(6):738-744.
120. Durrant JD, McCammon JA (2011) Molecular dynamics simulations and drug discovery. *BMC Biol.* 9.
121. Karplus M, Kuriyan J (2005) Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U. S. A.* 102(19):6679-6685.
122. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. *Nature* 267(5612):585-590.
123. Mike WR, W. (2013) AMBER 12 NVIDIA GPU ACCELERATION SUPPORT.
124. Martino AJ, Ferrone FA (1989) Rate of allosteric change in hemoglobin measured by modulated excitation using fluorescence detection. *Biophys. J.* 56(4):781-794.
125. Mouawad L, Perahia D, Robert CH, Guilbert C (2002) New insights into the allosteric mechanism of human hemoglobin from molecular dynamics simulations. *Biophys. J.* 82(6):3224-3245.

126. Ota N, Agard DA (2005) Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *J. Mol. Biol.* 351(2):345-354.
127. Sharp K, Skinner JJ (2006) Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling. *Proteins* 65(2):347-361.
128. Kong Y, Karplus M (2009) Signaling pathways of PDZ2 domain: a molecular dynamics interaction correlation analysis. *Proteins* 74(1):145-154.
129. Li LW, Uversky VN, Dunker AK, Meroueh SO (2007) A computational investigation of allostery in the catabolite activator protein. *J. Am. Chem. Soc.* 129(50):15668-15676.
130. Sessions RB, Dauberosguthorpe P, Osguthorpe DJ (1989) Filtering molecular-dynamics trajectories to reveal low-frequency collective motions - phospholipase-A2. *J. Mol. Biol.* 210(3):617-633.
131. McClendon CL, Friedland G, Mobley DL, Amirkhani H, Jacobson MP (2009) Quantifying correlations between allosteric sites in thermodynamic ensembles. *J. Chem. Theory Comput.* 5(9):2486-2502.
132. Vesper MD, de Groot BL (2013) Collective dynamics underlying allosteric transitions in hemoglobin. *PLoS Comput. Biol.* 9(9):e1003232.
133. Lange OF, Grubmuller H, de Groot BL (2005) Molecular dynamics simulations of protein G challenge NMR-derived correlated backbone motions. *Angew. Chem., Int. Ed. Engl.* 44(22):3394-3399.
134. Hinsen K (1998) Analysis of domain motions by approximate normal mode calculations. *Proteins: Struct., Funct., Genet.* 33(3):417-429.
135. Hinsen K, Thomas A, Field MJ (1999) Analysis of domain motions in large proteins. *Proteins: Struct., Funct., Genet.* 34(3):369-382.
136. Atilgan AR, *et al.* (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80(1):505-515.
137. Tama F, Sanejouand YH (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng.* 14(1):1-6.
138. Orozco M, *et al.* (2011) Coarse-grained representation of protein flexibility. Foundations, successes, and shortcomings. *Adv. Protein Chem. Struct. Biol.* 85:183-215.
139. Zheng WJ, Brooks BR, Doniach S, Thirumalai D (2005) Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved. *Structure* 13(4):565-577.
140. Zheng WJ, Brooks BR, Thirumalai D (2006) Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl. Acad. Sci. U. S. A.* 103(20):7664-7669.
141. Ming D, Wall ME (2005) Allostery in a coarse-grained model of protein dynamics. *Phys. Rev. Lett.* 95(19):198103.
142. Ming D, Wall ME (2005) Quantifying allosteric effects in proteins. *Proteins* 59(4):697-707.
143. Zheng W, Brooks BR, Thirumalai D (2007) Allosteric transitions in the chaperonin GroEL are captured by a dominant normal mode that is most robust to sequence variations. *Biophys. J.* 93(7):2289-2299.
144. Gerek ZN, Ozkan SB (2011) Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. *PLoS Comput. Biol.* 7(10).
145. Atilgan C, Atilgan AR (2009) Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput. Biol.* 5(10).
146. Atilgan C, Gerek ZN, Ozkan SB, Atilgan AR (2010) Manipulation of conformational change in proteins by single-residue perturbations. *Biophys. J.* 99(3):933-943.

147. Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* 10(1):59-69.
148. Hatley ME, Lockless SW, Gibson SK, Gilman AG, Ranganathan R (2003) Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl. Acad. Sci. U. S. A.* 100(24):14445-14450.
149. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R (2004) Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* 116(3):417-429.
150. Ferguson AD, *et al.* (2007) Signal transduction pathway of TonB-dependent transporters. *Proc. Natl. Acad. Sci. U. S. A.* 104(2):513-518.
151. Reynolds KA, McLaughlin RN, Ranganathan R (2011) Hot spots for allosteric regulation on protein surfaces. *Cell* 147(7):1564-1575.
152. Chi CN, *et al.* (2008) Reassessing a sparse energetic network within a single protein domain. *Proc. Natl. Acad. Sci. U. S. A.* 105(12):4679-4684.
153. Liu ZX, Chen J, Thirumalai D (2009) On the accuracy of inferring energetic coupling between distant sites in protein families from evolutionary imprints: illustrations using lattice model. *Proteins* 77(4):823-831.
154. Aftabuddin M, Kundu S (2007) Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophys. J.* 93(1):225-231.
155. Atilgan AR, Akan P, Baysal C (2004) Small-world communication of residues and significance for protein dynamics. *Biophys. J.* 86(1):85-91.
156. Bagler G, Sinha S (2005) Network properties of protein structures. *Physica A* 346(1-2):27-33.
157. Böde C, *et al.* (2007) Network analysis of protein dynamics. *FEBS Lett.* 581(15):2776-2782.
158. Brinda KV, Vishveshwara S (2005) A network representation of protein structures: implications for protein stability. *Biophys. J.* 89(6):4159-4170.
159. Dokholyan NV, Li L, Ding F, Shakhnovich EI (2002) Topological determinants of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* 99(13):8637-8641.
160. Greene LH, Higman VA (2003) Uncovering network systems within protein structures. *J. Mol. Biol.* 334(4):781-791.
161. Heringa J, Argos P, Egmond MR, Devlieg J (1995) Increasing thermal stability of subtilisin from mutations suggested by strongly interacting side-chain clusters. *Protein Eng.* 8(1):21-30.
162. Kanna N, Vishveshwara S (1999) Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.* 292(2):441-464.
163. Krishnan A, Zbilut JP, Tomita M, Giuliani A (2008) Proteins as networks: usefulness of graph theory in protein science. *Curr. Protein Pept. Sci.* 9(1):28-38.
164. Kundu S (2005) Amino acid network within protein. *Physica A* 346(1-2):104-109.
165. Vishveshwara S, Ghosh A, Hansia P (2009) Intra and inter-molecular communications through protein structure network. *Curr. Protein Pept. Sci.* 10(2):146-160.
166. Angelova K, *et al.* (2011) Conserved amino acids participate in the structure networks deputed to intramolecular communication in the lutropin receptor. *Cell. Mol. Life Sci.* 68(7):1227-1239.
167. Atilgan AR, Turgut D, Atilgan C (2007) Screened nonbonded interactions in native proteins manipulate optimal paths for robust residue communication. *Biophys. J.* 92(9):3052-3062.
168. Daily MD, Upadhyaya TJ, Gray JJ (2008) Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins* 71(1):455-466.
169. del Sol A, Fujihashi H, Amoros D, Nussinov R (2006) Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol.* 2:2006 0019.

-
170. Fanelli F, Felling A (2011) Dimerization and ligand binding affect the structure network of A(2A) adenosine receptor. *Biochim. Biophys. Acta* 1808(5):1256-1266.
 171. Ghosh A, Sakaguchi R, Liu C, Vishveshwara S, Hou YM (2011) Allosteric communication in cysteinyl tRNA synthetase: a network of direct and indirect readout. *The Journal of biological chemistry* 286(43):37721-37731.
 172. Ghosh A, Vishveshwara S (2007) A study of communication pathways in methionyl- tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proc. Natl. Acad. Sci. U. S. A.* 104(40):15711-15716.
 173. Pandini A, Fornili A, Fraternali F, Kleinjung J (2012) Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J.* 26(2):868-881.
 174. Sethi A, Eargle J, Black AA, Luthey-Schulten Z (2009) Dynamical networks in tRNA:protein complexes. *Proc. Natl. Acad. Sci. U. S. A.* 106(16):6620-6625.
 175. Tang S, *et al.* (2007) Predicting allosteric communication in myosin via a pathway of conserved residues. *J. Mol. Biol.* 373(5):1361-1373.
 176. Daily MD, Gray JJ (2009) Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput. Biol.* 5(2):e1000293.
 177. Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435(7043):814-818.
 178. Hendrickson B (1992) Conditions for unique graph realizations. *Siam. J. Comput.* 21(1):65-84.
 179. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) Protein flexibility predictions using graph theory. *Proteins* 44(2):150-165.
 180. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. U. S. A.* 99(6):3540-3545.
 181. Gohlke H, Thorpe MF (2006) A natural coarse graining for simulating large biomolecular motion. *Biophys. J.* 91(6):2115-2120.
 182. Gohlke H, Kuhn LA, Case DA (2004) Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* 56(2):322-337.
 183. Jacobs DJ, Thorpe MF (1995) Generic rigidity percolation: The pebble game. *Phys. Rev. Lett.* 75(22):4051-4054.
 184. Jacobs DJ, Hendrickson B (1997) An algorithm for two-dimensional rigidity percolation: the pebble game. *J. Comput. Phys.* 137(2):346-365.
 185. Jacobs DJ (1998) Generic rigidity in three-dimensional bond-bending networks. *J. Phys. A: Math. Gen.* 31(31):6653-6668.
 186. Katoh N, Tanigawa S (2011) A proof of the molecular conjecture. *Discrete Comput. Geom.* 45(4):647-700.
 187. Jacobs DJ, Dallakyan S, Wood GG, Heckathorne A (2003) Network rigidity at finite temperature: relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* 68(6).
 188. Lei M, Zavodszky MI, Kuhn LA, Thorpe MF (2004) Sampling protein conformations and pathways. *J. Comput. Chem.* 25(9):1133-1148.
 189. Wells S, Menor S, Hespenheide BM, Thorpe MF (2005) Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* 2(4):S127-S136.
 190. Ahmed A, Gohlke H (2006) Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. *Proteins* 63(4):1038-1051.

191. Ahmed A, Villinger S, Gohlke H (2010) Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses. *Proteins* 78(16):3341-3352.
192. Rathi PC, Pfeleger, C., Fulle, S., Klein, D.L., Gohlke, H. (2011) Modeling of molecular properties. *Statics of biomacromolecules*, ed Comba P (Wiley-VCH, Weinheim), pp 281-299.
193. Maxwell JC (1864) On the calculation of the equilibrium and stiffness of frames. *Philos. Mag.* 27(182):294-299.
194. Laman G (1970) On graphs and rigidity of plane skeletal structures. *J. Eng. Math.* 4(4):331-340.
195. Tay TS, Whiteley W (1984) Recent advances in the generic rigidity of structures. *Struct. Topol.* 9:31-38.
196. Katoh N, Tanigawa S (2009) A proof of the molecular conjecture. *Proceedings of the 25th annual symposium on Computational geometry*, (ACM), pp 296-305.
197. Welzenbach K, Hommel U, Weitz-Schmidt G (2002) Small molecule inhibitors induce conformational changes in the I domain and the I-like domain of lymphocyte function-associated antigen-1 - Molecular insights into integrin inhibition. *J. Biol. Chem.* 277(12):10590-10598.
198. Hespenheide BM, Jacobs DJ, Thorpe MF (2004) Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J. Phys.: Condens. Matter* 16(44):S5055-S5064.
199. Whiteley W (2005) Counting out to the flexibility of molecules. *Phys. Biol.* 2(4):S116-126.
200. Dahiyat BI, Gordon DB, Mayo SL (1997) Automated design of the surface positions of protein helices. *Protein Sci.* 6(6):1333-1337.
201. Radestock S, Gohlke H (2008) Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.* 8(5):507-522.
202. Hespenheide BM, Rader AJ, Thorpe MF, Kuhn LA (2002) Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graphics Modell.* 21(3):195-207.
203. Rader AJ, Bahar I (2004) Folding core predictions from network models of proteins. *Polymer* 45(2):659-668.
204. Rathi PC, Radestock S, Gohlke H (2012) Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* 159(3):135-144.
205. Makhatadze GI, Privalov PL (1995) Energetics of protein structure. *Adv. Protein Chem.* 47:307-425.
206. Zaccai G (2000) Biochemistry - How soft is a protein? A protein dynamics force constant measured by neutron scattering. *Science* 288(5471):1604-1607.
207. Livesay DR, Dallakyan S, Wood GG, Jacobs DJ (2004) A flexible approach for understanding protein stability. *FEBS Lett.* 576(3):468-476.
208. Livesay DR, Jacobs DJ (2006) Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins* 62(1):130-143.
209. Jacobs DJ, Livesay DR, Hules J, Tasayco ML (2006) Elucidating quantitative stability/flexibility relationships within thioredoxin and its fragments using a distance constraint model. *J. Mol. Biol.* 358(3):882-904.
210. Livesay DR, Huynh DH, Dallakyan S, Jacobs DJ (2008) Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family. *Chem. Cent. J.* 2(17):1-20.
211. Mottonen JM, Xu M, Jacobs DJ, Livesay DR (2009) Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family. *Proteins* 75(3):610-627.

-
212. Mottonen JM, Jacobs DJ, Livesay DR (2010) Allosteric response is both conserved and variable across three CheY orthologs. *Biophys. J.* 99(7):2245-2254.
213. Verma D, Jacobs DJ, Livesay DR (2012) Changes in lysozyme flexibility upon mutation are frequent, large and long-ranged. *PLoS Comput. Biol.* 8(3):e1002409.
214. Taverna DM, Goldstein RA (2002) Why are proteins marginally stable? *Proteins* 46(1):105-109.
215. Fulle S, Gohlke H (2009) Statics of the ribosomal exit tunnel: implications for cotranslational peptide folding, elongation regulation, and antibiotics binding. *J. Mol. Biol.* 387(2):502-517.
216. Mamonova T, Hespenheide B, Straub R, Thorpe MF, Kurnikova M (2005) Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.* 2(4):S137-S147.
217. Pfleger C, Gohlke H (2013) Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. *Structure* 21(10):1725-1734.
218. Seidelt B, *et al.* (2009) Structural insight into nascent polypeptide chain-mediated translational stalling. *Science* 326(5958):1412-1415.
219. Vázquez-Laslop N, Ramu H, Klepacki D, Kannan K, Mankin A (2010) The key function of a conserved and modified rRNA residue in the ribosomal response to the nascent peptide. *EMBO J.* 29(18):3108-3117.
220. Rader AJ, Brown SM (2011) Correlating allostery with rigidity. *Mol. Biosyst.* 7(2):464-471.
221. Radestock S, Gohlke H (2011) Protein rigidity and thermophilic adaptation. *Proteins* 79(4):1089-1108.
222. Radestock S (2010) Entwicklung eines rechnerischen Verfahrens zur Simulation der thermischen Entfaltung von Proteinen und zur Untersuchung ihrer Thermostabilität. Erlangung des Dotoktorgrades (Goethe-Universität Frankfurt, Frankfurt am Main).
223. Pfleger C, Radestock S, Schmidt E, Gohlke H (2013) Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.* 34(3):220-233.
224. Rader AJ (2009) Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys. Biol.* 7(1):16002.
225. Andraud C, Beghdadi A, Lafait J (1994) Entropic analysis of random morphologies. *Physica A* 207(1-3):208-212.
226. Crivelli S, *et al.* (2002) A physical approach to protein structure prediction. *Biophysical Journal* 82(1):36-49.
227. Forli S, Olson AJ (2012) A Force field with discrete displaceable waters and desolvation entropy for hydrated ligand docking. *J. Med. Chem.* 55(2):623-638.
228. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry* 28(6):1145-1152.
229. Pfleger C, Rathi PC, Klein DL, Radestock S, Gohlke H (2013) Constraint Network Analysis (CNA): a Python software package for efficiently linking biomolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model.*
230. Beazley DM (2003) Automated scientific software scripting with SWIG. *Future Gener Comp Sy* 19(5):599-609.
231. Gohlke H, Case DA (2004) Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* 25(2):238-250.
232. Haliloglu T, Bahar I (1999) Structure-based analysis of protein dynamics: comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. *Proteins: Struct., Funct., Genet.* 37(4):654-667.
233. Radford SE, Buck M, Topping KD, Dobson CM, Evans PA (1992) Hydrogen exchange in native and denatured states of hen egg-white lysozyme. *Proteins* 14(2):237-248.

-
234. McCammon JA, Gelin BR, Karplus M, Wolynes PG (1976) The hinge-bending mode in lysozyme. *Nature* 262(5566):325-326.
 235. Smith LJ, Sutcliffe MJ, Redfield C, Dobson CM (1993) Structure of hen lysozyme in solution. *J. Mol. Biol.* 229(4):930-944.
 236. Craig IR, Pfleger C, Gohlke H, Essex JW, Spiegel K (2011) Pocket-space maps to identify novel binding-site conformations in proteins. *J. Chem. Inf. Model.* 51(10):2666-2679.
 237. Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* 15(6):359-+.
 238. Halgren TA (2009) Identifying and characterizing binding sites and assessing druggability. *Journal of chemical information and modeling* 49(2):377-389.
 239. Amaro RE, *et al.* (2007) Remarkable loop flexibility in avian influenza N1 and its implications for antiviral drug design. *J. Am. Chem. Soc.* 129(25):7764-+.
 240. Landon MR, *et al.* (2008) Novel druggable hot spots in avian influenza neuraminidase H5N1 revealed by computational solvent mapping of a reduced and representative receptor ensemble. *Chem. Biol. Drug Des.* 71(2):106-116.
 241. Bowman GR, Geissler PL (2012) Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc. Natl. Acad. Sci. U. S. A.* 109(29):11681-11686.
 242. Kunze J, *et al.* (2014) Targeting dynamic pockets of HIV-1 protease by structure-based computational screening for allosteric inhibitors. *J. Chem. Inf. Model.* 54(3):987-991.
 243. Metz A, *et al.* (2011) Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein-protein interface. *J. Chem. Inf. Model.* 52(1):120–133.
 244. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* 295(2):337-356.
 245. Sotriffer CA, Gohlke H, Klebe G (2002) Docking into knowledge-based potential fields: A comparative evaluation of DrugScore. *J. Med. Chem.* 45(10):1967-1970.