Leveraging Novel Information for Coarse-Grained Prediction of Protein Motion

vorgelegt von
Dipl.-Inform.

Ines Putz

geb. in Deggendorf

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktorin der Naturwissenschaften
— Dr. rer. nat. —

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Marc Alexa
Gutachter:	Prof. Dr. Oliver Brock
Gutachterin:	Prof. Dr. Ivet Bahar
Gutachter:	Prof. Dr. Juri Rappsilber

Tag der wissenschaftlichen Aussprache: 06. November 2018

Berlin 2018

Declaration

Declarations according to §5, Sec. 1 of the Doctoral Regulations.

- re. (1): I hereby declare that I am acquainted with the current doctoral regulations of the TU Berlin as of 23 October 2006, last amended on 5 February 2014.
- re. (5): I hereby declare that all pre-publications of the dissertation or parts thereof and details of own contributions according to §2, subparagraph 4 doctoral regulations are listed in the attachment.
- re. (6): I hereby declare in lieu of an oath that I have independently completed the dissertation. All aids and sources have been listed and all details regarding own contributions according to (5) are correct.
- re. (7): I hereby declare that I have listed all applications (if any) for admission as a doctoral candidate or admission to doctoral procedure according in Sec. 4. of this form.

Berlin,

_____

Ines Putz

Thesis advisor: Professor Dr. Oliver Brock Ines Putz

Leveraging Novel Information for Coarse-Grained Prediction of Protein Motion

Abstract

Proteins are involved in almost all functions in our cells due to their ability to combine conformational motion with chemical specificity. Hence, information about the motions of a protein provides insights into its function. Proteins move on a rugged energy landscape with many local minima, which is imposed on their high-dimensional conformational space. Exhaustive sampling of this space exceeds the available computational resources for all but the smallest proteins. Computational approaches thus have to simplify the potential energy function and/or resolution of the model using information about what is relevant and what can be ignored. The accuracy of the approximation depends on the accuracy of the used information. Information that is specific to the problem domain, i.e. protein motion in our case, usually results in better models.

In this thesis, I propose a novel elastic network model of learned maintained contacts, *lmc*ENM. It expands the range of motions that can be captured by such simplified models by leveraging novel information about a protein's structure. This improves the general applicability of elastic network models.

Elastic network models (ENMs) are a highly popular coarse-grained method to study protein motions. They assume that protein motions are harmonic around an equilibrium conformation and largely governed by the protein's structural connectivity. This leads to the simplified representation of a protein as elastic mass-spring-network based on residue interactions. Despite their simplicity, ENMs predict intrinsic protein motions with surprising biological relevance. Accurate ENM predictions, however, require the initial contact topology to be maintained during a protein's motion. This is naturally fulfilled for highly collective motions resulting in successful predictions. But localized functional transitions involving substantial changes in the contact topology are often poorly explained. This limits the practical relevance of ENMs because the motion type of a protein is unknown a priori and thus it is unknown whether ENMs can capture it.

*lmc*ENM overcomes this limitation by leveraging information about the dynamic behavior of contacts, i.e. whether they break or are maintained when the protein moves. The maintained contacts remain after predicted breaking contacts have been removed from the initial network. In contrast to existing ENM variants, *lmc*ENM is able to accurately predict protein motions even for localized and uncorrelated functional transitions with changing contact topology.

In the first part of my thesis, I show that the absence of *observed* breaking contacts enables ENMs to accurately explain localized functional transitions. The resulting network of *observed* maintained contacts, *mc*ENM, can be built when start and end conformation of a functional transition are known. Of course, to apply this strategy in the standard case when only a single protein conformation is available, we need to be able to *predict* these breaking contacts.

In the second part of my thesis, I show how the breaking contacts can be *predicted*. To do so, I developed a machine-learning based classifier to differentiate breaking from maintained contacts based on a graph-based encoding of their structural context. The physicochemical characteristics of a contact's structural context capture how tightly different parts of the protein are bound to

each other, how this affects their movements, and ultimately their contact topology. To build *lmc*ENM the predicted breaking contacts are removed from the initial network. Using a large set of proteins covering different motion types I demonstrate the effectiveness of *lmc*ENM.

My thesis unlocks breaking contacts, or generally dynamic contact changes, as a novel source of information that has proven valuable in coarse-grained prediction of protein motion. Because they are defined on a simplified model of the structural connectivity of a protein, they are insensitive to structural details that would otherwise make their identification and prediction more difficult. The existence and usefulness of breaking contacts demonstrated in my thesis enables future research opportunities to study the conditions under which they occur and to examine the features that contributed the most to their accurate prediction. Our framework for predicting breaking contacts can be easily extended to further advance our understanding of protein motion.

Doktorvater: Professor Dr. Oliver Brock Ines Putz

Ausnutzung neuer Informationen für grobaufgelöste Vorhersage von Protein Bewegung

Zusammenfassung

Proteine sind an fast allen Funktionen in unseren Zellen beteiligt aufgrund ihrer Fähigkeit, Konformationsbewegungen mit chemischer Spezifität zu kombinieren. Informationen über die Bewegungen eines Proteins liefern somit Einblicke in seine Funktion. Proteine bewegen sich auf einer zerklüfteten Energielandschaft mit vielen lokalen Minima über ihrem hochdimensionalen Konformationsraum. Eine erschöpfende Abtastung dieses Raums übersteigt die verfügbaren Rechenressourcen für alle bis auf die kleinsten Proteine. Computergestützte Ansätze müssen daher die Energiefunktion und/oder die Auflösung des Modells vereinfachen aufgrund von Informationen darüber, was relevant ist und was ignoriert werden kann. Die Genauigkeit der Approximation hängt von der Genauigkeit der verwendeten Information ab. Informationen, die spezifisch für die Problemdomäne sind, d. h. Proteinbewegung in unserem Fall, führen normalerweise zu besseren Modellen.

In dieser Arbeit stelle ich ein neuartiges elastisches Netzwerkmodell von erlernten erhaltenen Kontakten, genannt *lmc*ENM, vor. Es erweitert die Bewegungsreichweite, die durch diese Netzwerke erfasst werden können, durch das Ausnutzen neuer Informationen über die Struktur eines Proteins. Dies verbessert die allgemeine Anwendbarkeit von elastischen Netzwerkmodellen.

Elastische Netzwerkmodelle (ENMs) sind eine sehr populäre grobkörnige Methode zur Untersuchung von Proteinbewegungen. Sie nehmen an, dass Proteinbewegungen harmonisch um eine Gleichgewichtskonformation verlaufen und weitgehend von der strukturellen Konnektivität des Proteins bestimmt werden. Dies führt zur vereinfachten Darstellung eines Proteins als elastisches Masse-Feder-Netzwerk auf der Basis von Residue-Interaktionen. Trotz ihrer Einfachheit sagen ENMs intrinsische Proteinbewegungen mit überraschender biologischer Relevanz voraus. Genaue ENM-Vorhersagen erfordern jedoch, dass die anfängliche Kontakttopologie während der Bewegung eines Proteins aufrechterhalten wird. Dies ist natürlicherweise für hoch kollektive Bewegungen erfüllt, was zu ihrer erfolgreichen Vorhersagen führt. Lokalisierte Funktionsbewegungen, die wesentliche Änderungen in der Kontakttopologie beinhalten, werden jedoch oft nur unzureichend erklärt. Dies begrenzt die praktische Relevanz von ENMs, da der Bewegungstyp eines Proteins a priori unbekannt ist und daher unbekannt ist, ob ENMs es erfassen können.

*lmc*ENM überwindet diese Einschränkung, indem Informationen über das dynamische Verhalten von Kontakten genutzt werden, d. h. ob sie brechen oder erhalten bleiben, wenn sich das Protein bewegt. Die erhaltenen Kontakte bleiben übrig, nachdem die brechenden Kontakte aus dem ursprünglichen Netzwerk entfernt wurden. Im Gegensatz zu existierenden ENM-Varianten ist *lmc*ENM in der Lage, Proteinbewegungen auch für lokalisierte und unkorrelierte Funktionstransitionen mit sich ändernder Kontakttopologie genau vorherzusagen.

Im ersten Teil meiner Arbeit zeige ich, dass die Abwesenheit von beobachteten brechenden Kontakten ENMs in die Lage versetzt, lokalisierte Funktionstransitionen genau zu erklären. Das resultierende Netzwerk von beobachteten bleibenden Kontakten, *mc*ENM, kann erstellt werden, wenn die Anfangs- und Endkonformation eines Funktionsübergangs bekannt ist. Um diese Strategie im Standardfall anzuwenden, wenn nur eine einzige Proteinkonformation zur Verfügung steht, müssen wir diese brechenden Kontakte natürlich vorhersagen können.

Im zweiten Teil meiner Arbeit zeige ich, wie die brechenden Kontakte vorhergesagt werden können. Um dies zu erreichen, entwickelte ich einen maschinell lernenden Klassifikator, der die brechenden von den bleibenden Kontakten unterscheidet auf Grundlage einer graph-basierten Kodierung ihres strukturellen Kontexts. Die physikalisch-chemischen Eigenschaften des strukturellen Kontexts eines Kontakts erfassen, wie stark verschiedene Teile des Proteins miteinander verbunden sind, wie sich dies auf ihre Bewegungen und letztendlich auf ihre Kontakttopologie auswirkt. Zum Erstellen von *lmc*ENM werden die vorhergesagten brechenden Kontakte aus dem ursprünglichen Netzwerk entfernt. Anhand eines großen Datensatzes von Proteinen, die verschiedene Bewegungstypen abdecken, demonstriere ich die Effektivität von *lmc*ENM.

Meine Dissertation erschließt brechende Kontakte oder allgemein dynamische Kontaktänderungen als eine neue Informationsquelle, die sich bei der grobkörnigen Vorhersage von Proteinbewegung als wertvoll erwiesen hat. Da diese dynamische Kontaktänderungen auf einem vereinfachten Modell der strukturellen Konnektivität eines Proteins definiert sind, sind sie unempfindlich gegenüber strukturellen Details, die ansonsten ihre Identifizierung und Vorhersage erschweren würden. Die Existenz und Nützlichkeit von brechenden Kontakten, die meine Dissertation zeigt, bietet die Grundlage dafür die Bedingungen für ihr Auftreten und die Eigenschaften, die am meisten zu ihrer Vorhersage beigetragen haben, weiter zu erforschen. Unser Framework für die Vorhersage von brechenden Kontakten kann leicht erweitert werden, um unser Verständnis der Proteinbewegung weiter voranzutreiben.

To Jari, Elin, and Kerstin.

Acknowledgments

I could never have done my PhD without the help of so many marvelleous people who supported me, motivated me and sometimes believed more in me than I did on myself on this exciting, challenging, and fun trip to my PhD.

First of all, I want to thank my family. Jari and Elin, you are the best kids I can ever imagine. Your curiosity, your courage, your empathy, your laughter motivates me every day. My deepest thanks goes to my wonderful wife Kerstin. I owe you more than I could express in words! I love you! Moogie, thank you for always being there for me without hesitation. Pap, thank you for teaching me about medicine and supporting me. Thanks, Sevi, Nick, Isa and Ina for being the best brothers and sisters.

I would also like to thank my friends for supporting me all these years: Conni, Burnd, Tom, Verena, Kuno, Caro, Steffi, Ruben, Vanessa, Dirk, Annie, Mandana, Abby, Anna, Milan, Christian, Kai, Gudrun, Seb, Carolina, Bine, Martina, Geli N., Theresa, Heiko, Simone, Dagmar, Hannes, Andrea, Björn, Mareike, Jan, Wanda, Frank, Andi, Julia, Björn B., and Eva. Thank you Conni, you motivated me to start this PhD and supported me all along that way. Burnd, I'm always happy when we meet and miss the parties on your balcony and POA. Tom and Carolina, thank you for churros.

Thank you RBO lab! It was a great pleasure to work with you, to discuss with you, to learn from you, to laugh with you, and to play table soccer. I will never forget our incredible X-mas parties. Thank you Alexander, Andreas, Angela, Arne, Armin, Can, Clemens, Dennis, Dov, Ely, Emily, Eveline, Florian, Freek, Gabriel L., Gabriel Z., Georg B., George, Henrietta, Ingmar, Ines, Janika, Jessica, John, José, Kolja, Mahmoud, Malte, Manuel, Marc, Marianne, Melinda, Michael, Nicolas, Oliver, Philip, Raphael, Rico, Robert, Roman, Sebastian H., Sebastian K., Serena, Stanio, Steffen, Tim, Thomas, Vincent, and Wolf.

A special thanks goes to Roberto and Rico, who graduated this year. You motivated me to follow your lead and also finish this year. Thank you Roberto for the amazing night at Monster's Karaoke Bar before you left for Stanford. I never imagined that karaoke could be so much fun! Thank you Rico for your pep talks during our bouldering sessions and for this feeling of incredible happiness and pride that I could see in your eyes after your defense. It gave me the final push to finish my thesis. Thank you Sebastian for always believing in me, for nice conversations, when we meet for lunch and for inviting me to great concerts. Thank you Janika for keeping the lab together and for our lovely lunches. Thank you Michael, Mahmoud, Tim, and Kolja for all the great discussions about proteins and high-dimensional spaces. It was so much fun to work with you and to learn from you. Thank you Nasir for being my first colleague at RBO and teaching me the first things about proteins. Thanks TJ for your support.

Thank you Kolja, Mahmoud, Arne, Sebastian, and Kerstin for proofreading my thesis and all your valuable comments.

I would also like to thank my team at SWP. Thank you Micha, Constanze, Andi, Patrick, Tom, Frank, and Danny for your support during the last phase of writing my thesis. Thank you Stefan for vacation days on short notice, for maximum flexibility, and for your incredible support.

Another big thanks goes to my committee. I thank Ivet Bahar and Juri Rappsilber for their interest in my work, their support, and the time they invested to give me valuable feedback. And last but not least, thank you Oliver for being my PhD advisor. You infected me with the fascination for proteins and all their complex interactions. You taught me how to become a researcher, how to question everything–even myself, how to learn from mistakes, how to make

great presentations, how to write clearly and concise, and how to become a critical thinker. Thank you for all your support even during the difficult times we had in between.

Finally, I would like to thank the institutions that made this dissertation possible, the Technische Universität Berlin and the Alexander-von-Humboldt Foundation.

Prepublication and Statement of Contribution

Part of this Thesis

Parts of this thesis have been previously published in the following peer-reviewed article:

A Putz I, Brock O (2017) *Elastic network model of learned maintained contacts to predict protein motion.* PLOS ONE 12(8): e0183889. https://doi.org/10.1371/journal.pone.0183889

Own contributions to [A]: I (IP) am the sole first author of this paper. I conceived the project idea together with the last author (OB). I conceived, designed, implemented and evaluated the algorithm and experiments presented in the paper and made the main contribution to paper writing. The last author (OB) gave scientific advice and contributed to paper writing.

Not Included in this Thesis

I also contributed to following peer-reviewed articles that are **not** part of this thesis:

B Mabrouk M, Werner T, Schneider M, Putz I, and Brock O (2016). *Analysis of Free Modeling Predictions by RBO Aleph in CASP11.* Proteins, 84 Suppl 1:87-104. https://doi.org/10.002/prot.24950

C Mabrouk M*, Putz I*, Werner T, Schneider M, Neeb M, Bartels P, and Brock O. (2015). *RBO Aleph: leveraging novel information sources for protein structure prediction.* Nucleic Acids Res., 43(W1):W343–W348.

* contributed equally

Own contributions to [B]: I am the fourth author of this paper. MM, TW, MS, I (IP), and OB conceived and designed experiments. MM, TW, MS, and I (IP) performed the experiments. MM, TW, MS, and I (IP) conceived and implemented analysis tools. MM, TW, MS, and I (IP) analyzed data. MM, TW, MS, I (IP), and OB contributed to paper writing. OB gave scientific advice.

Own contributions to [C]: I (IP) share the first authorship with MM. We contributed equally to design, implementation of the server, and to paper writing. MM, I (IP), TW, and MS conceived and designed the server. MM, I (IP), TW, and MS designed and implemented the backend of the server. My particular contribution here was the design and implementation of the domain prediction and splitting algorithm and its integration into the pipeline of the server. MM, I (IP), TW, PB, and MN designed and implemented the frontend of the web server. MM, I (IP), TW, and MS maintained the server during CASP11. MM, I (IP), TW, MS, and OB contributed to paper writing. OB gave scientific advice.

Appearance of Previous Publication in the Thesis

Chapter 1 provides the introduction for this thesis. Parts of it have been previously published in [A].

Chapter 2 reviews related work. Parts of this review have been previously published in [A].

Chapter 3 introduces the fundamentals of network analysis, machine learning, and elastic network models this thesis is based on. It is original to this thesis.

Chapter 4 introduces materials and methods relevant to the following two chapters 6 and 7. It contains parts that were previously published in [A].

Chapter 5 presents a novel elastic network model based on *observed* maintained contacts, *mc*ENM. It serves as proof for the assumption underlying our main contribution, which is presented in the following chapter. It contains parts that were previously published in [A]. The part on the relationship between the occurrence of observed breaking contacts and their effect on accuracy of *mc*ENM evaluated w.r.t. the function class of the proteins (5.4.5) is original to this thesis.

Chapter 6 presents the main contribution of this thesis, a novel elastic network based on *learned* maintained contacts, *lmc*ENM. It contains parts that were previously published in [A]. The part on the evaluation of binary classifiers and the third case study (6.4.7) are original to this thesis.

Chapter 7 concludes this thesis. It contains a discussion of potential applications of *lmc*ENM that was previously published in [A]. The remaining parts are original to this thesis.

Chapters (3-7) extend the previously published parts in [A] by providing additional background information and by relating the individual chapters to each other to present a concise story.

Table of Contents

List of Figures

List of Tables

Everything should be made as simple as possible, but not simpler.

Albert Einstein[1]

Proteins are not just good to eat, but they have specific shapes [...]. Although they consist of many thousands of atoms, they are governed by the same laws that govern the structure of bridges and houses. Everything is highly organized.

Michael Levitt[2]

1 Introduction

In 2013 Michael Levitt, Martin Karplus, and Arieh Warshel jointly won the Nobel Prize in Chemistry for their pioneering work on "developing multiscale models for complex chemical systems"[3]. It was the first time that the nobel price was awarded to computational research, thereby acknowledging the importance of combining experimental methods with computational approaches to further advance the field, now known as **computational structural biology**. The key quest behind the work of Levitt, Karplus, and Warshel is to find the *appropriate* degree of simplification that makes computational simulation of complex biochemical systems feasible but still yields biologically meaningful predictions.

To come up with a simplified model for a complex problem we need information that tells us what is important and what can be ignored. The accuracy of this information determines the quality of the approximation. Exploiting domain-specific information will usually result in better models because it is targeted to the actual problem. Nonetheless, every simplification introduces errors. Hence, the goal is to balance model complexity and (domain-specific) generalization error, which usually depends on the actual task, the availability of information, and the dedicated computational resources. The task defines the purpose of the simplified model, i.e. which questions should be answered, and in which context it is applied. For instance, if we are only interested in the flexibility of proteins the three-dimensional position of atoms can be ignored as demonstrated by rigidity analysis (Hermans et al., 2017).

This thesis aims to find such an *appropriate* simplification for coarse-grained prediction of protein motion using elastic network models (ENMs) (Tirion, 1996, Hinsen, 1998, Atilgan et al., 2001, Tama and Sanejouand, 2001). Based on strong assumptions, ENMs deliberately simplify

[1] This aphorism is attributed to Einstein although he may have never put it on paper. Instead he introduced the underlying idea in a lecture. The actual paraphrase may have been crafted by Roger Session when promoting Einstein's idea (https://quoteinvestigator.com/2011/05/13/einstein-simple/).

[2] https://news.stanford.edu/news/2013/october/levitt-nobel-chemistry-100913.html

[3] https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/advanced-chemistryprize2013.pdf

the model of a protein and yet successfully predict function-related protein motions. However, the gained computational efficiency comes at a certain cost. ENMs are limited to a particular motion type.

To overcome this limitation, we propose to leverage novel information about dynamic changes in the simplified network connectivity of ENMs. We present an approach to predict these dynamic changes using methods from graph theory and machine learning. By adjusting the network based on the predicted changes, we expand the range of motions that can be predicted by ENMs.

Before introducing our approach in more detail (section 1.3), I will give a basic intuition on

- *why* simplification is required to predict protein motion (section 1.1) and
- *how* simplification is facilitated by the use of information, focusing on the two most commonly used computational approaches to study protein motion (section 1.2).

1.1 Protein Motion

1.1.1 Functional Significance of Protein Motions

Proteins are essential building blocks in the cells of living organisms. Almost all cellular processes rely on their ability to combine chemical interaction with conformational motion (Alberts et al., 2002). They transport nutrients, transmit signals, catalyze enzymatic reactions, or regulate the metabolism. They stabilize the cells as structural scaffolds. They drive muscular contraction or vesicle transport in the cytoplasm as motor proteins. They support our immune system as antibodies by binding and neutralizing foreign microbes, and much more. All these functions require intensive interaction with binding partners, called ligands, which depends on the protein's ability to change its conformational shape.

The function of a protein is tightly coupled to its structure-encoded motions (Bahar et al., 2010b, Haliloglu and Bahar, 2015, Orozco, 2014, Teilum et al., 2009, Rueda et al., 2007b, Karplus and Kuriyan, 2005). This became evident for the first time when Felix Haurowitz observed the conformational change of hemoglobin upon oxygen binding in 1938 (Haurowitz, 1938). Since then, it has been established that to carry out their function proteins need to interact with other molecules.

To enable the interaction with a binding partner proteins need to adapt their three-dimensional structure (conformational shape). Only if the shapes of protein and ligand match along their binding interface, also their chemical binding becomes strong enough to be effective (Alberts et al., 2002). To gain insights into the function of a protein therefore requires the ability to infer the motion abilities inherently encoded in its structure. Most importantly, it may advance therapeutic treatment and the design of novel drugs against severe diseases, such as Alzheimer's (Anand et al., 2014, Cope et al., 2018, Villemagne et al., 2018), HIV/AIDS (Ghosh et al., 2016, Chupradit et al., 2017, Wu et al., 2017), or influenza (Webster and Govorkova, 2014, Wang et al., 2015).

1.1.2 Complexity of Conformational Space and Dynamics

Proteins typically consist of tens of thousands of atoms. This number can grow up to hundreds of thousands when considering large biomolecular assemblies, such as the ribosomes, chaperons, or viruses (Voss, 2007). Hence, their conformational space is too large to be sampled exhaustively with current computational methods for all but the smallest proteins. Fig 1.1A shows such a case where the conformational space of the molecule is simple enough to be explored in full detail. Because the molecule has only two torsional angles along its backbone, its conformational space is sufficiently described by two dimensions. Each conformational change in this space also affects the free energy of the molecule, resulting in an energy landscape drawn along the third dimension.

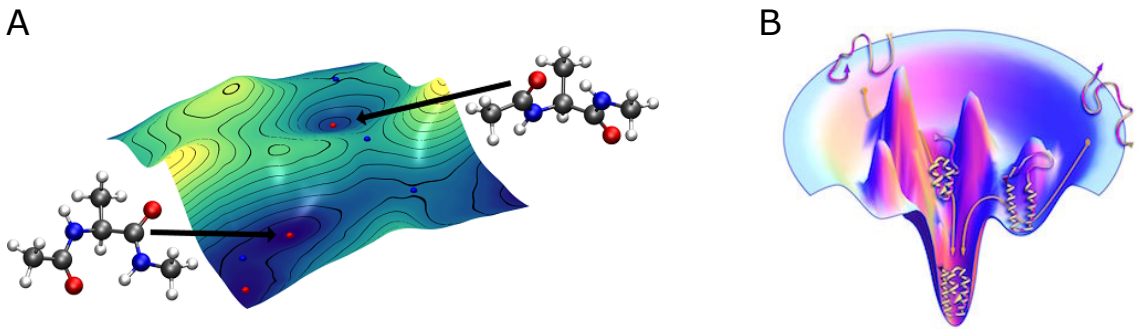

Figure 1.1: Illustration of free energy landscape and folding funnel. (A) Free energy surface of the Alanine Dipeptide imposed on their two-dimensional conformational space defined by two torsional angles. The two conformations represent low-energy states. Figure taken from[1]. (B) Folding funnel that leads over multiple partially folded states at intermediate energy levels towards few folded states with lowest energy. Figure source: Dill and MacCallum (2012). Reprinted with permission from AAAS.

The three-dimensional structure of a protein is determined by its sequence (Anfinsen, 1973), a chain of amino acids encoded in our DNA. To reach this folded state proteins must be guided efficiently because they cannot sample the space of possible conformations within realistic folding times of a few seconds (Levinthal, 1969). This has led to the view of the energy landscape as a rugged funnel-shaped surface with many local minima. At the bottom of the funnel is the global energy minimum containing a few low-energy folded states (Dill and Chan, 1997, Dill and MacCallum, 2012) as shown in Fig 1.1B.

Under physiological conditions in the cell, the native state is more accurately characterized as an ensemble of conformations (Frauenfelder et al., 1991, Henzler-Wildman and Kern, 2007, Orozco, 2014). Proteins continuously transition between stable, low-energy sub-states on the free energy landscape around their native state. Each transition involves overcoming energy barriers, which is more or less likely given the laws of thermodynamics and influence of the surrounding solvent. Consequently, some conformations are highly populated while others are rare. A major determinant of this distribution seems to be the nature's predisposition to optimize proteins for their biological function. There is growing evidence that evolution has shaped the protein's

[1] https://www.sfb716.uni-stuttgart.de/forschung/teilprojekte/projektbereich-c/teilprojekt-c6/beschreibung/index.en.html

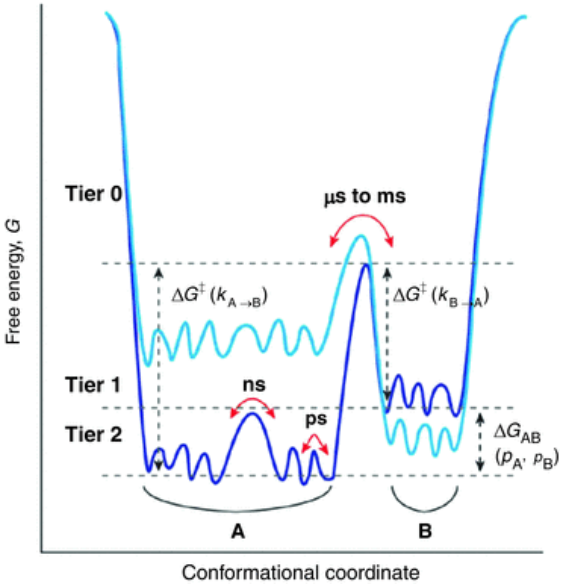

Figure 1.2: Hierarchy of motion amplitudes and timescales defined by the energy landscape. The landscape is organized in three tiers based on the energies, barrier height, and time scales of the conformational transitions. Transitions across high energy barriers have lower probability than over small barriers. Biomolecular processes at the micro- to milisecond scale, such as ligand binding or signal transmission, can change equilibrium of the energy landscape between states (from dark blue to light blue or vice versa). Figure source: Henzler-Wildman and Kern (2007). Reprinted with permission from Springer Nature.

energy landscape to favor the population of its functionally relevant sub-states (see Wei et al. (2016) and citations therein).

Binding a ligand may change this distribution. Two mechanisms have been proposed to explain how this might happen. The first mechanism, *induced-fit* (Koshland, 1958) occurs when the ligand actively induces a conformational change in the binding pocket to enable the binding process. The second mechanism *conformational selection mechanism* (Tsai et al., 1999, Goh et al., 2004) happens when the passive presence of a ligand shifts the population of states towards the reachable but only rarely visited bound conformation. In a large-scale study Stein et al. (2011) found that about half of the proteins showing conformational changes upon ligand binding follow the latter model, while the other half may be better explained by a combination of intrinsic and induced movements. Only a small portion of proteins showed purely induced motions. Hence, in most cases the conformational changes involved in ligand binding seem to be intrinsically encoded in a protein's structure.

To perform these conformational changes, a protein has to move along the energy landscape and cross energy barriers of different heights as shown in Fig 1.2. Consequently, these motions vary widely in their temporal and spatial scales (Henzler-Wildman and Kern, 2007). They range from fast, small-scale side chains fluctuations and loop motions up to slow, large-scale domain motions or even (partial) un- and refolding. Many functional processes in our cells, such as protein-ligand binding, enzyme catalysis, or signal transmission involve anharmonic transitions at the micro- to millisecond scale between different conformational states (see Fig 1.2). As we will see in the following, the temporal and spatial scales of protein motion impose a major challenge for studying them by experimental and computational means.

1.1.3 Limits of Experimental Protein Motion Determination

The wide range of temporal and spatial scales of protein motions make it difficult—if not impossible—to observe them directly with current experimental methods. Nonetheless, they are and will remain the gold standard for determining protein structure and dynamics. This is demonstrated by the amount of experimentally resolved structures that are deposited at the protein data bank (PDB) (Berman et al., 2000), which more than doubled over the past ten years. At the time of this writing the PDB contains a total of 118756 protein structures resolved by X-ray crystallography, 10753 by nuclear magnetic resonance tomography (NMR), and 1603 by cryo-electron microscopy (cryo-EM)[1].

X-ray crystallography (Drenth, 2007) provides the highest structural resolution and can be applied to a wide range of protein sizes. But its view on protein dynamics is limited to stable start or end conformations of functional transitions. To determine the three-dimensional positions of atoms x-rays rays are shot on a crystal of the protein, which consist of millions of protein instances arranged in a regular grid. The atom positions can be calculated from the resulting diffraction pattern. Crystallizing the protein at very low temperatures is necessary because at room temperature x-rays would rapidly destroy the protein. Besides being a time consuming process that is not guaranteed to work, it may introduce structural distortions (Dror et al., 2012, Wang et al., 2014a, Miller, 2014). Furthermore, there is no guarantee that these structural snapshots captured in the crystal match the most populated states in solution (Ma and Nussinov, 2016). Cryo-EM provides structural snapshots of large molecular complexes without the need of crystallization but cannot reach the atomic resolution of X-ray crystallography.

Dynamic methods, such as NMR (nuclear magnetic resonance) (Kovermann et al., 2016), FRET (fluorescence resonance energy transfer) (Lerner et al., 2018), AFM (atomic force microscopy, optical tweezers) (Pavliček and Gross, 2017), SAXS (Small-angle X-ray scattering) (Kikhney and Svergun, 2015) provide a time-resolved view on protein dynamics, but they are limited by the size of proteins or time-scales they can resolve (Dror et al., 2012, Wang et al., 2014a, Miller, 2014, Maximova et al., 2016). Recently, also temperature- and time-resolved x-ray-free-electron-laser (XFEL) crystallography (Keedy et al., 2015, Bostedt et al., 2016, Martin-Garcia et al., 2016, Johansson et al., 2017) has been developed, which overcomes the radiation damage potentially resulting from X-ray crystallography. However, there are only a few facilities world-wide to run these experiments, which has resulted in about 150 resolved protein structures deposited at the PDB so far (Johansson et al., 2017).

[1] Numbers are retrieved from `http://www.rcsb.org/stats/summary`, accessed on 2018-07-02.

1.2 Atomistic vs. Coarse-Grained Approaches to Study Protein Motions

Computational approaches to study protein motions aim to close the gap left by experimental methods. Nonetheless, due to the aforementioned complexity of protein motion they must find an appropriate level of simplification to balance computational cost and biological relevance. To do so, they replace the complex energy landscape of proteins by simpler energy potentials using knowledge about physics and reduce the resolution of the protein using knowledge about the highly organized structural shape of proteins that determines their motions.

Over the past decades numerous computational approaches to predict protein motions have been developed. They have been extensively reviewed in a series of recent publications (Maximova et al., 2016, Shehu and Plaku, 2016, Kmiecik et al., 2016, López-Blanco and Chacón, 2016, Orozco, 2014, Al-Bluwi et al., 2012). In this thesis we focus on the two widely used approaches to study protein motions, which span the range of used simplifications: most accurate molecular dynamics (MD) simulations and highly simplified elastic network models (ENMs).

Molecular dynamics approaches—on one end of the spectrum—simulate atomistic motions based on empirical physical force fields, which approximate the protein's energy landscape (McCammon et al., 1977, Karplus and Petsko, 1990, Karplus and McCammon, 2002, Karplus and Kuriyan, 2005). This results in what is believed to be a highly accurate understanding of protein motion. However, due to the computational requirements, only brief glimpses of protein motion can be obtained. In spite of increasing computational power, advances in parallelization (Buch et al., 2010, Stanley and De Fabritiis, 2015, Kutzner et al., 2015), and special-purpose supercomputers (Shaw et al., 2009, 2014, Ohmura et al., 2014), the practical usability of MD remains limited (Stanley and De Fabritiis, 2015).

On the other end of the spectrum, efficient computational approaches make drastic simplifications to the underlying physics—but at the same time maintain a surprising biological accuracy. They exploit the fact that much information about protein motion seems to be captured in the protein's contact topology, a simplified representation of the structural connectivity. These coarse-graining approaches, including the elastic network models (ENMs) (Tirion, 1996, Bahar et al., 1997, Hinsen, 1998, Haliloglu et al., 1997, Atilgan et al., 2001), deliberately decrease the resolution of the underlying model to gain computational power, yet predict intrinsic protein motions of biological relevance (Tama and Sanejouand, 2001, Krebs et al., 2002, Eyal et al., 2006, Ahmed et al., 2010, Bahar et al., 2010b).

Elastic network models, which will be the focus of this thesis, are one form of simplified model that has been very successful. They represent a protein as a network of masses connected by springs. Each mass corresponds to a residue of the protein. Two masses are connected by a virtual spring if the respective residues are within a certain distance in the protein structure (we will also say that the residues are in contact).

There is a cost associated with the reduction in model complexity realized by ENMs. The simplicity prevents them from capturing functional transitions if they are localized or uncorrelated

(low degree of collectivity) (Tama and Sanejouand, 2001, Ma, 2005, Cavasotto et al., 2005, Yang et al., 2007, Orellana et al., 2010). Making matters worse, it is difficult to know a priori whether ENMs can model a protein's motion accurately (Yang et al., 2007). As a result, ENMs currently are not only limited to a particular type of protein motion, it is also difficult to know if a given protein exhibits that motion type. These factors limit the practical relevance of ENMs.

1.3 This Thesis: Leveraging Novel Information for Coarse-Grained Protein Motion Prediction

In this thesis I propose a novel elastic network model that aims to improve the general applicability of ENMs by leveraging information to maintain the network's connectivity. The thesis consists of two main chapters: The first identifies the relevant information about dynamic contact changes that is required to improve the accuracy ENMs. The second presents our approach to predict these dynamic contact changes to be able to adjust the network of ENMs in the standard case, where only a single protein conformation is available.

- **Chapter 5 Elastic Network Model of Maintained Contacts (*mc*ENM)** investigates how the network connectivity of ENMs must be refined in order to capture also localized function-related protein motions. It is based on the insight that ENMs explain function-related transitions only if the initial network topology (the springs) is maintained during the protein's motion. Highly collective conformational changes naturally fulfill this requirement. Localized functional transitions, on the other hand, often lead to substantial changes in the contact topology and therefore in the corresponding network topology. I show that removing springs from the ENM for contacts that break during the motion enables ENMs to capture local and uncorrelated motions. This results in a novel elastic network model of maintained contacts (*mc*ENM) that can be applied when two conformations of a protein are available. Of course, to employ ENMs in situations when only a single conformation of the protein is known, we must also be able to *predict* these breaking contacts from that single conformation. My approach to predict these contacts is presented in the next chapter.

- **Chapter 6 Elastic Network Model of Learned Maintained Contacts (*lmc*ENM)** presents the core contribution of our approach, which is the ability to predict the dynamic behavior of contacts, i.e. whether they break or are maintained. To do so, I leverage information from the protein's structure. This information is captured in the physicochemical characteristics of local parts of the protein structure. While these parts largely maintain their structural shape when the protein moves, they move with respect to each other controlled by the strength of their physicochemical interactions. Consequently, the mobility and deformability of these parts also affect their underlying contact topology, causing some contacts to break during a functional transition. To predict these breaking contacts, I developed a machine-learning based classifier trained on a graph-

based representation of their structural context, which is based on the contact prediction framework for protein structure prediction introduced by Schneider and Brock (2014).

Based on the predicted contact changes, I build a novel elastic network model, called *lmc*ENM, which only consists of learned maintained contacts. These contacts form the connectivity of the ENM, after the predicted breaking contacts have been removed. The adjusted contact topology of *lmc*ENM more likely remains valid when the protein moves and thus helps capture localized conformational changes. Although *lmc*ENM encodes additional information about the dynamic behavior of contacts, it still preserves the simplicity of the original ENM approach. *lmc*ENM can be used to predict the motions of a protein based on a single conformation as input (standard case).

Before coming to these two main chapters, I will first review related work in the context of elastic network models (chapter 2), introduce the background required to understand the contributions of this thesis (chapter 3), and describe data set and methods relevant to both two main chapters (chapter 4). Each of the main chapters additionally introduces the methods that only apply in their own context. The final chapter concludes the thesis (chapter 7).

2 Related Work

In this chapter we review related work that aims to improve prediction accuracy and general applicability of elastic network models (ENMs). Before, we will give a brief introduction to the foundations of ENM that are required to understand the relation between previous work and our approach.

2.1 Elastic Network Models - Basics

Elastic network models (ENMs) approximate the structural connectivity of proteins to predict their structure-encoded, intrinsic motions. They describe proteins as mechanical networks of point masses (residues) that are linked by uniform elastic springs if their C_α atoms are within a predefined distance. Harmonic analysis of the resulting mechanical system then reveals the normal modes of the resulting mechanical system (Bahar et al., 2010b, López-Blanco et al., 2014). The most dominant, low-frequency modes are commonly associated with the protein's motion relevant for its function (Petrone and Pande, 2006).

Elastic network models (ENMs) derive information about protein motion based on two main assumptions: First, the intrinsic motions of a protein can be approximated by a simplified, harmonic potential (Tirion, 1996). Second, the coarse-grained structure of a protein largely encodes these motions (Bahar et al., 1997, Hinsen, 1998, Haliloglu et al., 1997, Atilgan et al., 2001).

Due to the harmonic approximation made by ENMs, the accuracy of motion predictions deteriorates with distance from the initial conformation. Nevertheless, often a few low-frequency modes suffice to accurately explain functional transitions of proteins that are large-scale and highly collective (Tama and Sanejouand, 2001, Krebs et al., 2002, Eyal et al., 2006, Ahmed et al., 2010). This ability to narrow down the relevant deformation space (spanned by the essential low-frequency modes) makes NMA-based approaches particularly suited to guide conformational exploration (Kirillova et al., 2008, Gur et al., 2013), docking simulations (Cavasotto

et al., 2005, Dobbins et al., 2008, Cavasotto, 2012), or refinement of experimentally resolved structures (Schröder et al., 2007, Gniewek et al., 2012).

ENMs often fail to capture localized or uncorrelated motions (Tama and Sanejouand, 2001, Ma, 2005, Cavasotto et al., 2005, Yang et al., 2007, Orellana et al., 2010, Dietzen et al., 2012, Globisch et al., 2013). In these cases, extraneous constraints, introduced by the simple construction of the model, stiffen the network, preventing the ENM from reflecting localized protein motion. To overcome this limitation, as we will see in this thesis, it is necessary to identify and remove these extraneous constraints from the network.

2.2 Elastic Network Model Variants

Refining elastic network models by exploiting additional information has a long tradition given their coarse-grained nature, see for example López-Blanco and Chacón (2016) for a recent review. However, one has to carefully balance how much and which additional information is actually relevant as computational cost increase with model complexity. We now briefly review related approaches that adjust network connectivity and/or stiffness, or interaction potential of ENMs. Based on the type of additional information we broadly categorize them into three groups: Methods that exploit (i) additional physcicochemical information, (ii) information about the protein's structure, or (iii) information about the protein's motion.

2.2.1 Exploiting Physicochemical Knowledge

ENMs rely on the fact that physical forces gradually decrease with distance, i.e. residues close in space are more likely to move together than more distant ones. Basic ENMs use an arbitrary fixed distance cut-off and constant spring stiffness, possibly oversimplifying matters. Alternative approaches connect all residues in the network and select spring stiffness as function of residue distance (Hinsen, 1998, Hinsen et al., 2000, Kovacs et al., 2004, Rueda et al., 2007a, Yang et al., 2009b). Apart from potentially over-constraining the network, a generic function for spring stiffness seems to be difficult to define (Lezon and Bahar, 2010).

Other approaches additionally consider the chemical type of the interaction. They vary spring stiffness between covalently bonded and non-bonded residue pairs (Hinsen et al., 2000, Kondrashov et al., 2006), or set them according to relative entropies between the interacting residues instead of relative energies (Sankar et al., 2018). Jeong et al. (2006) propose a chemical bond-cutoff ENM, where each CA-atom is connected to its four closest sequential neighbors and spring stiffness is varied with sequence distance. This implicitly guarantees network stability even for lower cutoffs that are usually not accessible for distance-cutoff based ENMs. Due to the sparser network they need to explicitly model chemical interactions, such as disulfide bridges, hydrogen bonds, or van-der-Waals forces. Recently, a mass-weighted variant has been proposed (Kim et al., 2013), which was further extended by symmetry constraints to better capture the packed state of protein crystals when their structure is determined experimentally (Kim et al., 2015). These models are particularly accurate in terms of B-factor prediction. However, B-factors themselves provide a

questionable source of information about protein motion due to the influence of crystal packing effects or errors introduced by molecular refinement (Fuglebakk et al., 2013).

2.2.2 Exploiting Structural Knowledge

Some approaches tailor the connectivity and/or potential of the ENM to knowledge of the protein's structure. The simplest way to achieve this is to consider interactions between more than two residues with a more complex potential (Stember and Wriggers, 2009, Lin and Song, 2010, Srivastava et al., 2012), or additionally incorporate side-chain connectivity and chemical type (Frappier and Najmanovich, 2014, Kaynak et al., 2017). It is also possible to consider additional backbone or side-chain atoms (Micheletti et al., 2004, Moritsugu and Smith, 2007), secondary structure (Tama et al., 2000), or information obtained from rigidity analysis (Ahmed and Gohlke, 2006, Ahmed et al., 2010, Hermans et al., 2017). While the former trade physical accuracy for computational cost, the latter may introduce errors due to the additional coarse-graining.

ENMs can also have mixed resolution, where functional relevant parts are modeled at the atomic level and other parts at the coarser residue level (Kurkcuoglu et al., 2009b). While this increases computational cost it also requires to know, where the functional relevant parts are in order to refine their resolution. To efficiently analyze large biomolecules Xia (2017) reduces resolution in a multi-scale virtual particle based ENM that accounts for mass distribution and distance relations of virtual particles (coarse-grained sub units of larger complexes). Xia et al. (2014) proposed an ENM derived from an alpha-shape based tesselation of the protein structure, which circumvents the definition of distance-cutoffs or distance-dependent functions to construct the network topology.

If aspects of the structure are known to remain constant during the protein's motion, it is possible to refine ENMs by adding additional constraints to maintain the overall structure, for example in the case of membrane proteins or larger protein complexes. Dony et al. (2013) augment ENMs by adding springs between buried residues as well as between hydrogen-bonded residues.

2.2.3 Exploiting Knowledge about Motion

The aforementioned approaches obtain the network topology of the ENM from a single, static protein conformation. Hence, there is no guarantee that the initial contact topology derived from this conformation remains valid when the protein moves. In some cases, however, we posses information about two or more conformations along the motion trajectory and use this information to improve the ENM.

One type of refinement is based on molecular dynamics (MD) simulations. Based on a single MD simulation, Hinsen et al. (2000) optimized a distance-dependent function to adjust spring stiffness. Orellana et al. (2010) optimized connectivity and stiffness of the ENM based on short MD trajectories. They propose a three-staged hybrid potential with strongly connected sequential neighbors, distance-weighted springs for residues close in space, and a protein-size dependent cutoff to ignore irrelevant, remote interactions (see 4.3 for details). Their approach outperforms simpler ENM variants, but it remains questionable whether MD trajectories in the nano-second regime

are able to cover the full space of motions accessible to proteins (Durrant and McCammon, 2011). Globisch et al. (2013) refine ENMs of protein complexes by analyzing short MD trajectories of their subunits. They reduce the network to bonds largely maintained throughout the simulations. The computational costs of the required MD simulations and the ability to only generate partial trajectories of the protein's motion limit the applicability of this approach.

Another source of information are ensembles obtained by Nuclear Magnetic Resonance (NMR) or X-ray. For instance, Lezon and Bahar (2010) derive optimal stiffness constants for secondary structure type and sequence distance between interacting residues using entropy maximation of NMR ensembles. Despite the good agreement between normal modes and PCA-modes from X-ray and NMR ensembles (Yang et al., 2009a), the structural diversity of the latter may be biased towards missing experimental data (Fuglebakk et al., 2013).

When two conformations of a protein are known (e.g. open and closed conformation), the structural differences between these conformation allow to infer aspects of the intermediate motion. Song and Jernigan (2006) and Yang et al. (2007) use this information to tailor ENMs to the observed collective motions by varying the spring stiffness within (stronger) and between (weaker) domains. The resulting ENMs are more accurate, but they can only be obtained when two different conformations are available.

2.3 Relation to Our Work

The main hypothesis of this thesis is that leveraging information about dynamic changes in the connectivity of elastic network models expands the range of motion types that they can capture. Hence, to advance the general applicability of ENMs we need to exploit additional information beyond the topological constraints imposed by the initial conformation. The aforementioned approaches suggest that additional information about the motions of a protein is encoded in a broad range of physicochemical, structural, and topological characteristics of their structure. While adjusting ENMs based on singular characteristics/properties or small subsets may improve their prediction accuracy in some cases, the most important aspects of function-related protein motions more likely result from the interplay of a broader set of properties (Jamroz et al., 2012)

Now the main question seems to be: how can we identify the combination of relevant characteristics to refine ENMs most effectively? We propose to learn these combinations from a large set of possible characteristics. In particular, we consider features that capture the influence of local and global structural topology on protein motion. Furthermore, we deliberately refine the network connectivity of ENMs without adjusting stiffness or interaction potential. This allows us to preserve the simplicity and computational efficiency of ENMs, while improving their general applicability. Still, the approach we present below can be used in conjunction with most of the previously mentioned methods of adjusting ENMs.

3 Background

In this chapter we introduce the theoretical foundations and concepts that lay out the basis for our approach. They are required to understand the contributions of this thesis.

In section 3.1 we introduce how networks/graphs can be used to analyze relations between data or objects. We use them in multiple ways in this thesis: (i) we model the structural connectivity of a protein as a network of inter-residue contacts (contact topology network), (ii) we augment this network with physicochemical and structural information to characterize the local environment of each contact, (iii) we analyze these local contact graphs and extract features from them to train a classifier to differentiate breaking from maintained contacts, and (iv) we obtain the elastic network models from the protein's inter-residue contact graph to determine its intrinsic motions. We introduce graph parameters, node and edge label statistics, spectral analysis of graphs, and centrality measures that we use to derive the graph features characterizing the local contact environment.

In section 3.2 we give an introduction into the basics of machine learning. In particular, we focus on support vector machines and how they can perform classification on graphs. We train them to predict the dynamic behavior of inter-residue contacts, i.e. if they are breaking or maintained when the protein moves, based on their local embedding in the contact topology network. We also introduce principal component analysis that we use to determine the most dominant movements in conformational ensembles of proteins to validate our approach.

Section 3.3 introduces coarse-grained prediction of protein motion using normal mode analysis and elastic network models. Here, we focus on the anisotropic network model because it forms the basis of our novel elastic network model of learned maintained contacts, *lmc*ENM (see chapter 6).

3.1 Network analysis

Many areas and activities in our daily live rely on the relationships between data, objects, or entities that are described, explored, and predicted by networks. For example, we interact and collaborate in social, political, or scientific networks (Scott, 2017, Ward et al., 2011, Maireder et al., 2017, Ebadi and Schiffauerova, 2015, Ding, 2011). We access search engines to find relevant information in the web (Brin and Page, 1998). We use optimized transportation systems (Guimerà et al., 2005, Arnold et al., 2004) or rely on the security of telecommunication networks (Gorman et al., 2004). We benefit from epidemics prevention (Luke and Harris, 2007), crisis management via social networks (Shi et al., 2017), or advances in network-based drug design (Csermely et al., 2013) and systems biology (Horvath, 2011, Albert, 2007, Böde et al., 2007). We improve our understanding of cognitive processes (Avena-Koenigsberger et al., 2018, Bassett et al., 2011) or how brain diseases, such as Alzheimer's, affect our brain's functional connectivity (Supekar et al., 2008).

Network analysis provides the tools and techniques to study and understand such complex systems of interactions and relationships in order to make future predictions. It reduces complexity by encoding relational data into networks of nodes that interact along edges. Nodes and edges can be attributed to capture additional properties of both, data and relations. Based on this simplified representation structural and topological properties of these networks can be analyzed.

In this thesis we aim to predict the dynamic behavior of inter-residue contacts, i.e. whether they break or are maintained, given their local embedding in the protein's contact network (see chapter 6 for more details on the contact prediction algorithm). To do so, we augment these local contact networks with physicochemical and structural properties, such as solvent accessibility, hydrogen bonding, associated secondary structure elements, closeness to a pocket, or being part of a symmetric arrangement. Based on the assumption that the local contact networks of breaking and maintained contacts differ in their properties and topology, we train a classifier to distinguish them based on their similarity. To compare the networks we derive features by analyzing their structural and topological properties (see the overview of features in 6.2.1 and the detailed list of features given in the appendix A.2). The similarity is then simply computed as the Euclidean distance between the feature vectors.

In the following we first introduce the basics of graphs and networks. Then, we focus on their analysis by introducing properties that characterize topology and spectrum of a graph, its node and edge label statistics, and centrality measures indicating the importance of individual nodes. These properties allow us to obtain the graph features required to classify contacts as breaking or maintained.

3.1.1 Graphs and Networks

Formally, a **graph** is defined as the ordered pair $G = (V, E)$ consisting of a set of vertices V (also called nodes) and edges E (links) (Cormen, 2009, Brandes, 2005). An edge $e = (u, v)$, where $u, v \in V$, connects exactly two vertices in graph G.

There are different types of graphs depending on the type of edges (see Fig. 3.1 for examples). The simplest form is an **undirected graph**. It consists of edges that have no direction, i.e. they link unordered vertex pairs with symmetric adjacency. Consequently, the edges (u, v) and (v, u) are equivalent. Simple, undirected graphs require the vertices of an edge to be distinct ($u \neq v$) and thus contain no self-loops. Edges can also be directed to represent asymmetric relationships between vertices in a so-called **directed graph**, or **digraph**. A directed edge $e = (u, v)$ starts at vertex u and ends at vertex v, i.e. the vertex pair (u, v) is ordered. In addition, undirected or directed edges can be weighted, for instance by their distance. The associated graphs are called **weighted graphs/digraphs**. However, we only employ undirected graphs/networks in this thesis.

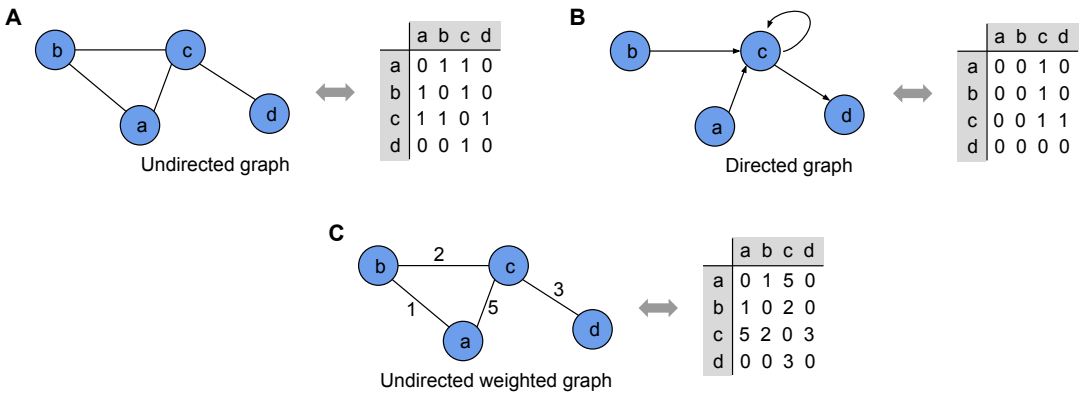

Figure 3.1: Examples of graph types and corresponding adjacency matrices. (A) Undirected graph with symmetric adjacency matrix. (B) Directed graph with asymmetric adjacency matrix. (C) Weighted undirected graph with weighted adjacency matrix.

Networks encode additional information in form of node and edge attributes, which goes beyond the adjacency relations typically modeled by graphs (Brandes, 2005). Nodes may be characterized by attributes such as label, size, or object category, while edges can be associated with properties such as capacity, time, similarity, or a function of other variables.

Fig. 3.2 shows a simplified network obtained from the contact topology of secondary structures of a protein that is augmented by structural and physicochemical properties. Two secondary structure elements (nodes) are in contact (linked) if at least one residue of element A is within a pre-defined distance of a residue of element B. The edges are attributed by the actual number of contacts, Euclidean distance between the centroids of the secondary structure elements, interaction energy, and the type of interacting secondary structure elements. Each secondary structure element is characterized by its type, its sequential length (number of amino acids), its 3d-length, and solvent accessibility.

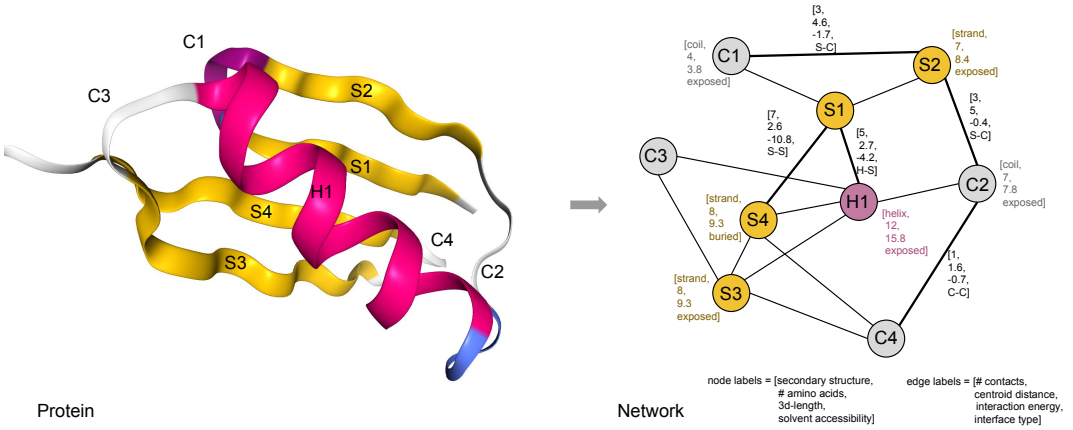

Figure 3.2: Example of a simplified network derived from the contact topology of the secondary structures of a protein structure. Nodes represent secondary structure elements. Edges link secondary structure elements that are in contact, i.e. at least one residue of element A is within a pre-defined distance with a residue in element B. Nodes and edges are characterized by structural and physicochemical properties.

3.1.2 Topological Analysis

Graph parameters (Brandes, 2005) capture topological properties of a graph/network. For instance, the **degree** of a vertex is defined by its number of edges to other vertices. The **distance** between two vertices in the graph is given by the length of the shortest path between them along edges from the graph. **Eccentricity** of a vertex denotes its largest distance to any other vertex in the graph. The minimum eccentricity in a graph defines the **radius** of the graph, while the maximum eccentricity denotes its **diameter**. A graph consisting of two disjoint vertex sets with edges only between but not within the sets is called a **bipartite graph**.

3.1.3 Spectral Analysis

Another way to represent graphs is to encode the relations between their nodes in matrices (Lovász, 2007, Brandes, 2005, Brouwer and Haemers, 2011). The most common matrices are the adjacency matrix and the Laplacian matrix. Spectral analysis based on the eigendecomposition of these matrices yields additional structural properties, such as spectrum or energy of a graph. Below we focus on the properties used in this thesis to further characterize the local embedding of an inter-residue contact in the contact graph. The derived features build the group of graph-spectrum features used for the breaking contact prediction (see 6.2.1 and Table A.5).

The **adjacency matrix** captures which nodes are adjacent to each other (see Fig. 3.1). For an undirected graph $G = (V, E)$ with n vertices $V = \{v_1, v_2, ..., v_n\}$, the adjacency matrix A has square shape $(n \times n)$. Its elements A_{ij} are defined as:

$$A_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

The diagonal of the matrix contains only zeros for simple graphs without self-loops. If the graph is weighted the non-zero entries are multiplied by the edge weight.

The adjacency matrix A of an undirected graph G is symmetric. Hence, its eigendecomposition $A = U\Lambda U^T$ results in a set of real eigenvalues and eigenvectors. The orthonormal eigenvectors build the columns of the matrix U. The set of eigenvalues λ_i along the diagonal of Λ is called the **spectrum** of graph G, where we assume that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. The first eigenvalue λ_1 relates to the average degree of G. Much more interesting is the so-called **eigenvalue or spectral gap** between the first and second largest eigenvalues because it is related to connectivity and expansion of a graph (Lovász, 2007, Brouwer and Haemers, 2011). A large gap indicates that the graph is highly connected and has large expansion, i.e. many edges must be removed in order to cut the graph into two parts. The **energy** (Li et al., 2012) of a simple graph is defined as the squared sum of the absolute eigenvalues of A. Dense graphs that are highly connected tend to have higher energy than sparse graphs of the same size, e.g. if edges have been removed (Shatto and Çetinkaya, 2017).

The **Laplacian** (sometimes called **Kirchhoff matrix**) of an undirected graph $G = (V, E)$ is built by subtracting the adjacency matrix A from the diagonal matrix D of the vertex degrees: $L = D - A$. The elements of the symmetric, $n \times n$ matrix L are given by:

$$L_{ij} = \begin{cases} -1 & \text{if } i \neq j \text{ and } (v_i, v_j) \in E \\ deg_G(v_i) & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

where $deg_G(v_i) = \sum_j A_{ij}$ denotes the **degree** of vertex v_i, $v_i \in V(G)$. For an undirected, simple graph, the spectrum of the Laplacian is called the **Laplacian spectrum** of the graph. We introduce the Laplacian here because it is used by the Gaussian Network Model to encode the inter-residue connectivity of a protein (see 3.3).

3.1.4 Distribution of Node and Edge Labels

Node/edge label statistics (Li et al., 2012) describe how node or edge labels are distributed in a network/graph $G = (V, E)$ with vertices V and edges E.

The **label entropy** E_G quantifies the probabilities of different node labels $l_1, ..., l_m$ given by

$$E_G = -\sum_{k=1}^{m} p(l_k) * \log p(l_k) \tag{3.3}$$

where $p(l_k) = |l_k|/|V|$ is the probability of observing the node label l_k in the graph.

The **neighborhood impurity** I_G of graph G measures the average distribution of labels in the neighborhood of all nodes. It is given by

$$I_G = \frac{\sum_{v \in V} |l(u) : u \in \text{nei}(v), l(v) \neq l(u)|}{|V|} \tag{3.4}$$

with nei(v) being the neighbor nodes of v. For instance, we calculate entropy and neighborhood impurity of secondary structure types or solvent accessibility among the residues in the local contact networks.

The **link impurity** L_G of graph G measures the impurity degree among all edges defined as

$$L_G = \frac{|(v,u) \in E : l(v) \neq l(u)|}{|E|} \tag{3.5}$$

For instance, we evaluate the link impurity considering chemical type, secondary structure, solvent accessibility, or symmetry.

3.1.5 Centrality Measures

Centrality measures indicate the relative importance of individual nodes or their influence on other nodes in the network (Brandes, 2005). For instance, in social networks, such as Twitter, information flow and content is often dominated by so-called *influencers*. They usually have many followers that retweet their posts and mention them on a regular basis (Cha et al., 2010). Although a large number of followers is not sufficient to be an influencer, it still increases the likelihood of being retweetet or mentioned if the followers find content and quality of the initial tweet worth sharing.

In general, there is no unique definition of node centrality as it depends on the context. Among other reasons nodes may be important because they are connected to many other nodes (degree centrality) or bridge between different parts of the network (betweenness centrality). We use these measures to characterize the importance of breaking and maintained contacts in different contexts. For instance, maintained contacts are important to stabilize the network when the protein moves, whereas breaking contacts can be neglected in this regard. In other words, one would expect that maintained contacts have higher degree centrality than breaking ones. In contrast, breaking contacts often occur between sparsely connected parts, such as flexible helices or loops or between two moving domains. Hence, breaking contacts should have higher betweenness centrality given that they reside in such bridge-like regions.

In the following we introduce three different centrality measures used in this thesis to characterize the importance of contacts within the whole network: degree, closeness, and betweenness centrality. While the degree centrality is agnostic to the topology of the network, closeness and betweenness centrality take the relative position of each node to the others into account. We define these measures for an undirected graph $G = (V, E)$ with vertex set V and edge set E as they are implemented in the Python library NetworkX (Hagberg et al., 2008). Fig. 3.3 exemplifies the different notions of importance captured by the depicted centrality measures.

The **degree centrality** c_D (Brandes, 2005) of vertex v is given by its degree $deg_G(v)$, i.e. the number of its neighbors. We use the **normalized degree centrality** c_{D^n} as implemented in NetworkX (Hagberg et al., 2008), which is defined as

$$c_{D^n}(v) = \frac{deg_G(v)}{(|V| - 1)} \tag{3.6}$$

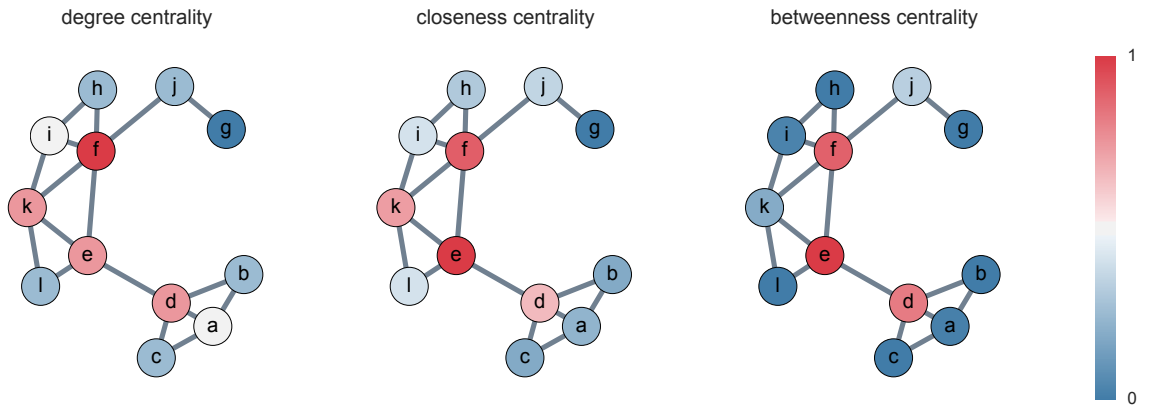

Figure 3.3: Centrality measures for a graph. The three different color codings illustrate how degree, closeness, and betweenness centralities capture different meanings of importance for the same graph. All measures are normalized. Hence, red indicates a high value, blue a low value. Node f has the highest degree centrality because it has the largest number of neighbors compared to all other nodes. Node e is the one closest to all other nodes in the graph. It can be seen as a "hub" that distributes messages to the other nodes. Node e has also the highest betweenness centrality because it "bridges" between the two parts of the graph.

where $|V|$ is the number of vertices in the graph. Hence, the degree of each vertex gets normalized by the maximum possible degree of the graph.

The **normalized closeness centrality** c_C^n (Freeman, 1978, Brandes, 2005) of vertex v is defined as the inverse of the average shortest path distance to u over all $n-1$ reachable nodes in the graph

$$c_{C^n}(v) = \frac{n-1}{\sum_{u \in V, u \neq v} d(v,u)} \tag{3.7}$$

where $d(v,u)$ is the shortest-path distance between vertex v and u, and n denotes the number of nodes with a path to v. The multiplication with $n-1$ normalizes the measure. Intuitively, the node with highest closeness centrality has minimum effort to communicate to all other nodes in the network. This refers to the notion of a "hub".

The **betweenness centrality** c_B (Brandes, 2008, 2005) of vertex v is defined as

$$c_B(v) = \sum_{s,t \neq v \in V} \frac{\rho_{st}(v)}{\rho_{st}} \tag{3.8}$$

where ρ_{st} is the number of shortest paths between all pairs of vertices $(s,t) \in V$ and $\rho_{st}(v)$ refers to the fraction of such shortest-paths that pass along vertex v. The betweenness centrality can be interpreted as how much a node controls the communication between all other node pairs, which have this node on their shortest-path.

3.2 Machine Learning

In this section we will give a brief introduction into basic concepts of machine learning with a particular focus on the algorithms used in this thesis: (i) classification using support vector machines and graphs, and (ii) dimensionality reduction using principal component analysis. More details can be found in a series of comprehensive books about this topic, such as Mitchell (1997), Bishop (2006), Murphy (2012), Goodfellow et al. (2016), Michalski et al. (2013), Géron (2017), James et al. (2013). They cover the depths of machine learning and its various subfields in theory and practice and serve as basis for this introduction.

Machine learning aims to empower computers to automatically learn from data (Samuel, 1959) or more specifically (Mitchell, 1997):

> A computer program is said to **learn** from experience E with respect to some task T and some performance measure P, if its performance on T as measured by P, improves with experience E.

Given a set of experiences, also called training data or samples, a machine learning algorithm searches for the hypothesis that best explains this data in order to make predictions for new, unseen data. For example, our task is to predict, whether a tumor is benign or malignant (task) based on its size (Kourou et al., 2015). As input, we get histological data of the tumor (training samples) obtained from microarray analysis of a patient's tissue. We measure the performance of the algorithm based on the match between predicted label and actual label, which was assigned by a medical expert. Based on this feedback the algorithm trains to improve its performance. Using the trained model it can then predict the label of unseen histological data.

Depending on the type of training data machine learning problems are broadly categorized into supervised, semi-supervised, unsupervised, and reinforcement learning. Machine learning approaches are also labeled based on the learning methodology into online methods, which learn while making predictions, and batch learning, which trains once given all available data in order to make future predictions afterwards.

Supervised Learning has access to labeled training data. Given a set of labeled training data $D = \{\mathbf{x}_i, y_i\}_{i=1}^n$, it aims to learn a function $f : X \to Y$ that maps the input $\mathbf{x}_i \in X$ (training samples) to the output $y_i \in Y$ (labels). If the labels are discrete, i.e. belong to two (*binary*) or more classes (*multi-class*), we call it a *classification* problem. If the labels are real-valued we call it a *regression* problem. The example above falls in the category of supervised learning because a human expert (teacher) labeled the input data used to train the algorithm.

Unsupervised Learning, in contrast, has no access to the labels of the training data. Its objective is to structure the given input data $D = \{\mathbf{x}_i\}_{i=1}^n$ on its own by identifying patterns in the training samples. If the data is categorized into discrete groups, where similar samples belong to the same group, we call it *clustering*. If the data is projected onto a lower dimensional continuous representation we call it *dimensionality reduction*. An example for unsupervised learning is the segmentation of tumors given medical images, e.g. from MRI-scans. The algorithm

aims to detect boundaries in these images by grouping similar pixels together, i.e. to segment the tumor from its environment.

Semi-Supervised Learning gets training data as input that is only partially labeled, while most of it is unlabeled. Using unsupervised learning it categorizes the unlabeled data based on similarity. Knowing the label for one sample of a category is then enough to label the whole category accordingly. The previous example can be extended to semi-supervised learning by providing the algorithm with a few images, where the segmented tumors were labeled by an medical expert as benign or malign.

Reinforcement Learning (RL) learns to sequentially optimize a decision policy based on a reward signal. In RL the learning system is called an agent that sequentially decides about the next action to take given its current state in order to maximize its total reward. The received feedback signal in form of positive (good choice) or negative (bad choice) rewards guides the agent to optimize its decision policy, i.e. which action to take next based on the current state. An example for reinforcement learning is the control of drug dose in cancer treatment. Based on the biomarkers of a patient the RL agent decides to increase or decrease the drug dose. It then gets feedback, whether the biomarker of the patient improved or not, which helps the algorithm to optimize the drug dose over time.

We will now introduce support vector machines in more detail because we use them to differentiate breaking from maintained contacts in this thesis (see chapter 6).

3.2.1 Support Vector Machines (SVMs)

Support vector machines (Boser et al., 1992, Cortes and Vapnik, 1995) are popular supervised learning models that are particularly good at solving complex classification tasks. In the biomedical field they have many applications (Ben-Hur et al., 2008), among them the prediction of cancer recurrence, susceptibility or survival (Kourou et al., 2015), the identification of drug targets (Wang et al., 2017), or computer-aided diagnosis of Alzheimer's (Khedher et al., 2017). SVMs can also be used for regression tasks. However, here we focus on SVM-based classification because we use it to differentiate breaking from maintained contacts in order to improve elastic network models (see chapter 6). This introduction is based on the following publications (Boser et al., 1992, Cortes and Vapnik, 1995, Burges, 1998, Vert et al., 2004, Noble, 2006, Ben-Hur et al., 2008, James et al., 2013, Géron, 2017), which provide additional details for interested readers.

In their simplest form, SVMs aim to find the hyperplane that separates data points into two classes while maximizing the perpendicular distance, called margin, to the closest samples in each class. This is called binary classification, which discriminates unseen data depending on which side of the hyperplane it falls. The maximum margin criterion is motivated by the assumption that the distance of unseen data to the decision boundary is approximately the same as of the training data. Hence, the maximal margin hyperplane reduces the risk of misclassification compared to any other separating hyperplane, which improves the generalization performance of SVMs for previously unseen data (Fig 3.4A).

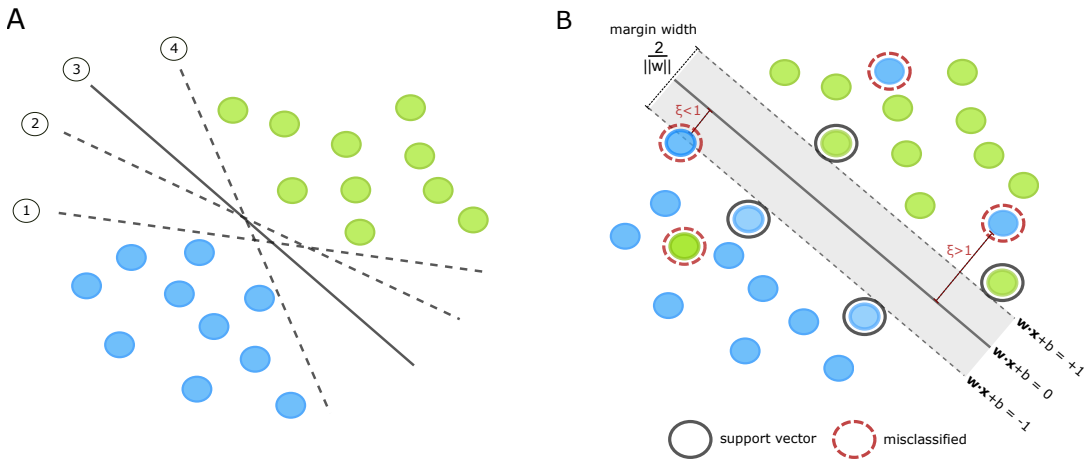

Figure 3.4: Examples of separating hyperplanes and soft-margin support vector machine. (A) Examples of hyperplanes separating the training data in two classes with different margins. All separate the samples well. But hyperplane no. 3 with the largest margin to the closest samples in each class has the best generalization performance, assuming that training data and unseen data are similarly distributed. (B) Illustration of a soft-margin support vector machine. It maximizes the margin, while tolerating margin violations measured by the slack-variables ξ to some extent, regulated by the cost parameter C. The support vectors define the class hyperplanes (dashed lines) parallel to the maximal margin hyperplane (solid line).

The training samples lying on the hyperplanes defining the class boundaries are called **support vectors**. Removing one of these support vectors from the training set most likely changes the boundary, while the absence of other training samples has no effect on the decision boundary.

In many real-world problems, a decision boundary that separates all data points without error does not exist. Because larger margins improve generalization performance, we would thus accept some misclassifications, as long as their number remains small enough. The **soft margin hyperplane** fulfills this purpose by maximizing the margin while minimizing the number of errors (Cortes and Vapnik, 1995). It extends the original hard-margin SVM by adding **slack variables** ξ_i and a user-defined cost parameter C. The former relaxes the margin by tolerating errors to some degree, while the latter regulates the trade-off between margin size and error acceptance (Fig 3.4B). To discourage misclassification, errors must be penalized by a larger cost.

Given a set of labeled training data $D = \{\mathbf{x}_i, y_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^d$ being the i-th vector in D and corresponding labels $y_i \in \{-1, 1\}$, the optimization objective of an SVM can be formalized as follows

$$\underset{\mathbf{w}, \xi, b}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{3.9a}$$

$$\text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \tag{3.9b}$$

$$\xi_i > 0, \quad \text{for} \quad i = 1, \ldots, n \tag{3.9c}$$

where $\mathbf{w}_0 \cdot \mathbf{x} + b_0 = 0$ is the equation of the maximum margin hyperplane. It can be shown that the margin between the two parallel hyperplanes, defining the class boundary, is $\frac{2}{\|\mathbf{w}\|^2}$ (Burges, 1998). Thus, we aim to minimize the norm (length) $\|\mathbf{w}\|^2$ in order to maximize the distance between the two hyperplanes.

Finding a solution to this primal minimization problem is equivalent to solving its dual maximization problem. Because the latter is a quadratic programming problem it can be solved efficiently. For $\alpha_i \in \mathbb{R}^n$ the dual problem using Lagrange multipliers is defined as

$$\underset{\alpha_i \leq 0}{\text{maximize}} \quad L(\alpha) \equiv \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \, \mathbf{x}_i \cdot \mathbf{x}_j \tag{3.10a}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C \quad \forall i, \tag{3.10b}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{3.10c}$$

The dual Lagrangian has the advantage that the α_i are bounded only by the regulation parameter C, whereas the slack variables ξ and their Lagrange multipliers do not appear. Now the weight vector of a large margin hyperplane can be formulated by linearly combining the solutions α_i of the optimization problem above (eqn. 3.10a) and the input samples $\mathbf{x}_i$:

$$\mathbf{w} = \sum_{i=1}^{n_s} \alpha_i y_i \mathbf{x}_i \tag{3.11}$$

where n_s denotes the number of support vectors.

Linear SVM-classifiers measure the similarity between input samples by their inner product (also called dot product). However, many real-world problems are not linearly separable. To overcome this problem, Boser et al. (1992) introduced the **kernel-trick**. It uses the transformation $\phi : \mathbb{R}^d \to \mathcal{H}$ to map the data from the d-dimensional input space to a higher (infinite) dimensional feature space $\mathcal{H}$, also called Hilbert space (see Fig 3.5 for an illustration). Using eqn. 3.11 the discriminant function becomes

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w} \cdot \phi(\mathbf{x}) + b \\ &= \sum_{i=1}^{n} \alpha_i y_i \, \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b \\ &= \sum_{i=1}^{n} \alpha_i y_i \, k(\mathbf{x}_i, \mathbf{x}) + b \end{aligned}$$

Using a kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ all dot products of the algorithm in the input space can be replaced by dot products in the feature space. Instead of doing the linear separation in the input space, which is not possible for these kinds of problems, it can be done in the infinite dimensional feature space. Hence, using the kernel-trick enables support vector machines to solve non-linear classification problems.

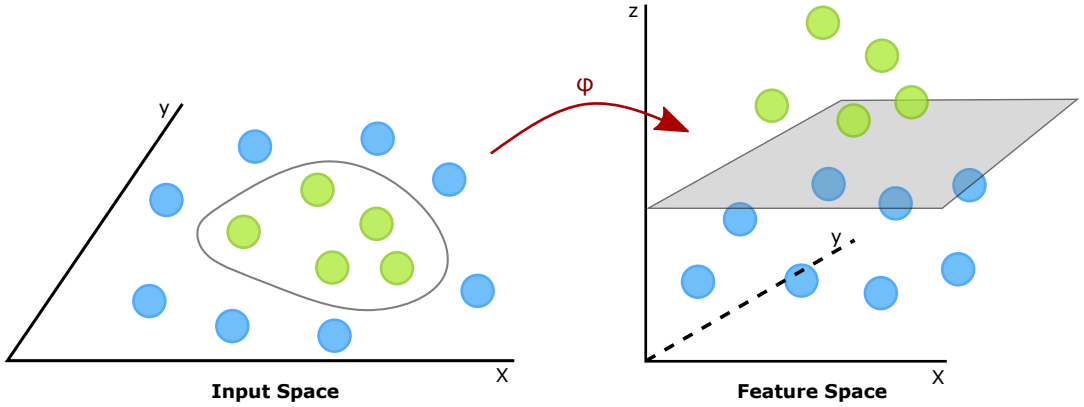

Figure 3.5: Illustration of the Kernel-Trick. The dataset is not linearly separable in the two-dimensional input space. By transforming it with the transformation $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ it becomes linearly separable by a hyperplane in the higher-dimensional feature space, shown at the right. The gray curve in the left panel refers to the decision boundary from feature space back-projected to the input space. Figure adapted from MIT OpenCourseWare[1].

We use two of the most popular kernels in this thesis (Vert et al., 2004):

- **Linear Kernel**

$$k_L(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j = \sum_{i,j=1}^{n} x_i x_j \tag{3.12}$$

- **Gaussian radial basis function (RBF) Kernel**

$$k_G(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2) \tag{3.13}$$

  where $\gamma > 0$ is a user-defined parameter, which specifies the width of the Gaussians. Large values of γ may lead to overfitting and poor generalization performance. Hence, this parameter has to be tuned in advance.

Besides these two kernels many others have been proposed and implemented[2], even for non-vectorial data. Examples are string-kernels (Lodhi et al., 2002, Saunders et al., 2003, Leslie and Kuang, 2004) or graph-kernels (Borgwardt et al., 2005, Vishwanathan et al., 2010, Yanardag and Vishwanathan, 2015), which measure the similarity between strings or graphs, respectively.

A binary SVM classifier usually outputs a binary value indicating the class a test sample belongs to. To predict probability values instead Platt's scaling method (Wu et al., 2004) can be used. It performs logistic regression on the binary output values of the SVM using additional cross-validation.

[1] https://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/lecture-notes/MIT15_097S12_lec13.pdf

[2] see for instance https://github.com/gmum/pykernels.

3.2.2 Graph Classification

As introduced before (see chapter 1 and section 3.1), we aim to predict the dynamic behavior of inter-residue contacts given their structural context in the protein. We claim that the properties of this contact environment determine whether contacts break or are maintained when the protein moves. Graphs/networks are a well-suited data structures to encode this contact environment and its properties due to their ability to capture interactions (edges) between objects (nodes).

Classifying contacts as breaking or maintained given the graph-based representation of their structural embedding is a particular instance of methods that apply machine learning on graphs. Learning on graph-structured data has many applications, for instance to detect anomalies in social networks (Kang et al., 2014), to differentiate proteins by their roles in protein-protein interaction networks (Hamilton et al., 2017), or to predict whether therapeutic protein drugs are potentially usable in the treatment of other diseases (Duvenaud et al., 2015).

Being able to apply machine learning methods on graphs requires a way to compute their similarity. In general, the problem of finding an exact mapping between two graphs or subgraphs ((sub)graph isomorphism) is NP-complete, i.e. there exists no algorithm that is able to compute it in polynomial time unless $P = NP$ (Vishwanathan et al., 2010). Common approaches tackle this problem by using a simplified, but more practical estimation of similarity, among them are kernel-based (Vishwanathan et al., 2010) and feature-based methods (Li et al., 2012).

Graph kernels use subgraphs, subtrees, shortest paths, cycles, or random walks to decompose a graph into parts (Vishwanathan et al., 2010, Li et al., 2012). By mapping the derived patterns into the lower-dimensional feature space their similarity can be efficiently calculated using the dot-product (see 3.2.1 for an explanation of the "kernel-trick"). An example of a graph kernel is the shortest-path-kernel (Borgwardt and Kriegel, 2005). Given two graphs G_1 and G_2 it computes the shortest paths between all pairs of nodes in each graph. The kernel function now compares the similarity between the two graphs by comparing the lengths of all these shortest-paths based on their inner product, for instance using a linear kernel:

$$\begin{aligned} k(G_1, G_2) &= \sum_{s,t \in V_1} \sum_{k,l \in V_2} k(d(s,t), d(k,l)) \\ &= d(s,t) \cdot d(k,l) \end{aligned}$$

Graph features encode topological metrics and label statistics of graphs in order to compare them (Li et al., 2012). For instance, average degree or average path lengths (closeness centrality) attribute the global topology of a graph, whereas neighborhood impurity or label entropy capture the distribution of edge or node labels (see 3.1.1 for details on calculating graph attributes). The similarity between two graphs can now easily be calculated as the Euclidean distance between their fixed-length feature vectors.

In this thesis we use the feature-based approach to extract information encoding the dynamic behavior of contacts from their structural context. This is mainly motivated by the fact that incorporating domain-specific information is much easier in the feature-based approach because

they can simply be added to the feature vector. In addition, feature-based approaches are computationally more efficient and scale better to larger graphs than kernel-based methods (Li et al., 2012).

With the recent advent of deep learning an alternative approach became quite popular, which is called **representation learning on graphs** (Hamilton et al., 2018). Instead of human-engineered features or kernels for graph comparison, these methods automatically learn a mapping of graphs into a lower-dimensional space, while optimizing the match between original graph and learned mapping considering geometric relations. Thus, they can be incorporated directly into the machine learning algorithm itself. However, to be effective such approaches typically require large amounts of data.

3.2.3 Principal Component Analysis (PCA)

Principal component analysis belongs to the category of unsupervised learning problems, which aim to find structure in the input data without prior knowledge, e.g. known labels. It is a highly popular statistical technique to identify essential patterns, called principal components, in data, which represent the orthogonal axes of largest variance (David and Jacobs, 2014, Smith, 2002). The principal components are linear combinations of the original data and can be used to project the data into a lower-dimensional space with minimum loss. Hence, PCA is typically used for **dimensionality reduction**, for instance to compress high-throughput gene-expression data in bioinformatics (Ma and Dai, 2011) or to filter MRI-images used to predict brain healthiness, e.g. to support early Alzheimer's diagnosis (Gewers et al., 2018).

In its basic form, dimensionality reduction with PCA relies on five steps:

1. normalize the vectorial input data by subtracting the mean
2. build the **covariance matrix** (or correlation matrix)
3. decompose it into eigenvectors (principal components) and eigenvalues (amplitudes)[1]
4. project the input data into the lower dimensional space using a subset of main principal components

In this thesis, we apply PCA to identify the dominant directions of structural displacements captured by protein conformational ensembles (e.g. experimentally determined or snap-shots from molecular dynamics simulations) (David and Jacobs, 2014). These so-called essential dynamics (ED) are often used to validate the intrinsic motions of proteins predicted by elastic network models (Sankar et al., 2018, Yang et al., 2009a). In the context of proteins, normalization of the conformational ensemble is equivalent to finding their optimal structural superposition, which can be calculated by a least-squares fit (Kabsch, 1978) to a pre-defined reference structure (e.g. the unbound conformation or an average structure for MD/NMR ensembles). This removes global rotational and translational movements of the conformers, which yields the internal motions in the ensemble.

[1]For square matrices, which can be diagonalized, eigen-decomposition is used, otherwise the more general singular value decomposition (SVD).

From this superimposed ensemble we can obtain the fluctuations of the C_α-atoms[1] from their average position. The covariances between atom i and j define the elements Q_{ij} of the covariance matrix $\mathbf{Q}$ as

$$Q_{ij} = \left\langle (q_i - \langle q_i \rangle)(q_j - \langle q_j \rangle)^T \right\rangle \tag{3.14}$$

where $q_1, ..., q_{3N}$ represent the mass-weighted, three-dimensional coordinates of the C_α atoms, N refers to the number of residues of the protein, and $\langle \cdot \rangle$ denotes the average over the ensemble. The variances on the diagonal of $\mathbf{Q}$ capture the average motion amplitude along one coordinate. The covariances (cross-correlations) in the off-diagonal elements reveal the relationship between the motions.

Diagonalization of this matrix yields $3N - 6$ eigenvectors with non-zero eigenvalues, where the first six trivial modes representing global rotation and translation are ignored. The eigenvectors, ranked by decreasing variance, capture the collective motions of the ensemble, where the corresponding eigenvalues indicate their mean square fluctuations.

In contrast to normal mode analysis, introduced below, PCA can explain also non-harmonic conformational changes, which show displacements beyond the harmonic approximation from the equilibrium conformation. Nonetheless, the most dominant ED-modes most often agree well with the lowest-frequency normal modes. For a more detailed derivation of PCA in the context of molecular ensembles analysis please refer to David and Jacobs (2014), Yang et al. (2009a).

3.3 Coarse-Grained Normal Mode Analysis with Elastic Network Models

The main contribution of this thesis is a novel elastic network model of learned maintained contacts (*lmc*ENM in chapter 6). Elastic network models (ENMs) represent proteins as mass-spring-networks to examine their structure-encoded motions at a coarse-grained scale using normal mode analysis (NMA). In the following, we will introduce the basic theoretical foundations of normal mode analysis and elastic network models, which are required to understand the technical details and contributions of our approach.

3.3.1 Normal Mode Analysis (NMA)

Normal mode analysis (NMA) is a widely used method to determine the motions of a protein intrinsically accessible to its structure. It is based on the assumption that pairs of interacting atoms behave as coupled harmonic oscillators and that the motions of a protein can thus be approximated by the sum of these pairwise vibrations around a given equilibrium conformation (Bahar et al., 2010a, Bastolla, 2014, López-Blanco et al., 2014). Fig 3.6A illustrates a simple network of three point masses coupled by harmonic oscillators.

[1] PCA is usually performed using C_α-atoms representing the residues of the protein. However, any atom-subset, e.g. all backbone or heavy atoms, can be used.

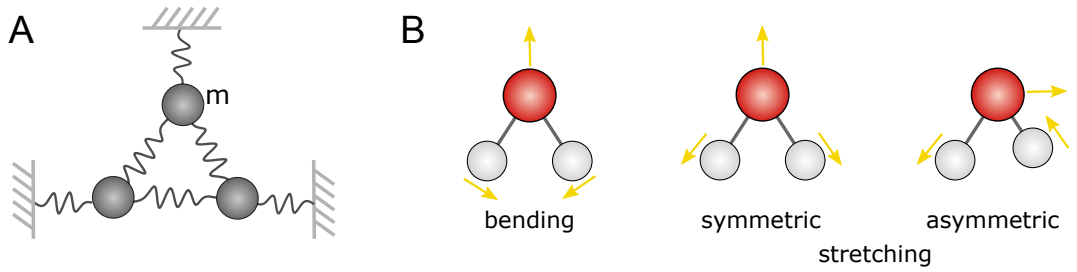

Figure 3.6: Coupled harmonic oscillators and normal modes of water molecule. (A) Network of point masses (m) coupled by simple harmonic oscillators (springs). (B) Normal modes predicted by NMA for the water molecule: a bending mode and two stretching modes (symmetric and asymmetric). The mode directions of each atom are indicated by the yellow arrows.

Commonly used in physics, NMA can be applied to any particle network to analyze its spectrum of vibrations, for instance to determine the vibrations of crystals (Rousseau et al., 1981) or to calculate the elastic deformations of smart hybrid materials, such as magnetic gels, to infer their capabilities as vibration absorbers or soft actuators (Pessot et al., 2016).

Researchers started to use NMA to study protein motions more than 30 years ago (Brooks and Karplus, 1983, Go et al., 1983, Levitt et al., 1985). Since then NMA-based prediction of protein motions became highly popular because the predicted low-frequency motions were found to agree well with functional protein motions even for coarse-grained normal mode analysis based on elastic network models (ENMs 3.3.2) (Sankar et al., 2018, Bahar et al., 2015, Kurkcuoglu et al., 2012, Meireles et al., 2011, Bahar et al., 2010a).

NMA analytically solves the equations of motions based on the assumption that the potential energy landscape of a protein–despite its many local minima–is approximately parabolic (called the **harmonic hypothesis/approximation**, see Fig 3.8 on page 33). As such, NMA complements MD simulations, which need to numerically integrate the equations of motions to sample motion trajectories along this rugged energy landscape. In principle, any differentiable force field, e.g. semi-empirical force fields from MD, can be used to study the collective motions of proteins. But due to their lower computational costs, simpler force fields requiring no energy minimization of the initial conformation, such as the ones used by ENMs (see 3.3.3), are nowadays routinely used.

In the following we will give a brief introduction into the theoretical foundations of NMA. A full derivation is presented in a series of publications (Bahar et al., 2010a, Bastolla, 2014, López-Blanco et al., 2014), which served as basis for this introduction.

Given a protein with N atoms, we can represent a particular conformation $\mathbf{q}$ as a $3N$-dimensional vector of its atom coordinates in Cartesian space given by

$$\mathbf{q} = (x_1, y_1, z_1, ..., x_N, y_N, z_N)^T \tag{3.15}$$

With V_{NMA} being the potential energy of the protein defined by any force field, we can express V_{NMA} around the equilibrium conformation $\mathbf{q}^0$ as an expanding Taylor series given by

$$V_{\mathrm{NMA}}(\mathbf{q}) \cong V_{\mathrm{NMA}}(\mathbf{q}^0) + \sum_{i}^{3N} \frac{\delta V_{\mathrm{NMA}}}{\delta q_i}\bigg|_{\mathbf{q}^0} (q_i - q_i^0) + \frac{1}{2}\sum_{i,j}^{3N} \frac{\delta^2 V_{\mathrm{NMA}}}{\delta q_i \delta q_j}\bigg|_{\mathbf{q}^0} (q_i - q_i^0)(q_j - q_j^0) + ... \quad (3.16)$$

where $V_{\mathrm{NMA}}(\mathbf{q}^0)$ refers to the potential at the energy minimum and is zero per definition. Also the second term, the first derivative, vanishes to zero because the equilibrium conformation q^0 is defined to be in the energy minimum of the potential function. Due to the harmonic assumption, small displacements around the equilibrium are sufficiently approximated by the second-order term and therefore higher-order terms can be ignored. Thus, the approximate potential from eqn. 3.16 can be reduced to

$$V_{\mathrm{NMA}}(\mathbf{q}) \cong \frac{1}{2}\sum_{i,j}^{3N} \frac{\delta^2 V_{\mathrm{NMA}}}{\delta q_i \delta q_j}\bigg|_{\mathbf{q}^0} (q_i - q_i^0)(q_j - q_j^0) \quad (3.17)$$

Let $\mathbf{H}$ denote the **Hessian matrix** of second partial derivatives of the potential function w.r.t. the positions of the network nodes ($\Delta\mathbf{q}^{\mathbf{T}}$). The elements of $\mathbf{H}$ are given by

$$H_{ij} = \frac{\delta^2 V_{\mathrm{NMA}}}{\delta q_i \delta q_j}\bigg|_{\mathbf{q}^0} \quad (3.18)$$

Then we can rewrite eqn. 3.17 as

$$V_{\mathrm{NMA}}(\mathbf{q}) \cong \frac{1}{2}\sum_{i,j}^{3N} (q_i - q_i^0) H_{ij} (q_j - q_j^0) = \frac{1}{2}\Delta\mathbf{q}^{\mathbf{T}}\mathbf{H}\Delta\mathbf{q} \quad (3.19)$$

$\mathbf{H}$ is organized as NxN-matrix consisting of $3x3$-submatrices. Each of this submatrices captures the effect of two interacting atoms onto the energy of the system.

Normal mode analysis decomposes the Hessian matrix with respect to the mass-weighted positions of the network nodes into a set of orthogonal eigenvectors (normal modes) and eigenvalues (frequencies) based on the following generalized eigenvalue problem:

$$\mathbf{HU} = \mathbf{\Lambda TU} \quad (3.20)$$

where $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N)$ is the matrix of eigenvectors, which diagonalizes the Hessian matrix $\mathbf{H}$, $\mathbf{\Lambda} = (\lambda_1, \lambda_2, ..., \lambda_N)$ is the diagonal matrix of corresponding eigenvalues, and $\mathbf{T}$ is the kinetic energy matrix.

$\mathbf{H}$ is symmetric and positive-semidefinite because it resides at the local minimum of an harmonic energy potential. Thus, it can be diagonalized without negative eigenvalues, i.e. the local curvature of the potential can only be zero or positive. The first six normal modes with zero-frequency are typically ignored because they correspond to the external rigid body motions (three rotations and three translations in 3D) of the protein.

Hence, the complete deformation space of a protein is spanned by the remaining $3N-6$ normal modes, i.e. any motion in this space can be described by the linear combination of these modes. The normal modes are ranked by increasing frequency according to their corresponding eigenvalues, which specify the energetic cost of a deformation along the mode direction. Low-frequency modes are easily accessible for the protein due to their low energetic costs. Hence, they represent the most dominant, collective motions of the system and are also called **global** or **soft modes**. High-frequency modes, in contrast, are energetically less favorable and encode local movements with low degree of collectivity. Fig 3.6B on page 28 shows the normal modes of a water molecule.

The harmonic assumption greatly simplifies the prediction of protein motions using NMA. But at the same time it is the source of its main limitations:

1. To avoid energetic instability the energy of the initial conformation of a protein must be minimized before analyzing its normal modes. Besides being a time-consuming step, this energy minimization may introduce structural distortions, which in turn may reduce the accuracy of the predicted motions.

2. Due to the harmonic approximation structural displacements along the mode directions are only valid close to the equilibrium conformation.

3. NMA ignores molecular constraints, such as fixed bond lengths of covalent bonds or fixed dihedral angles. Hence, additional efforts are required to preserve structural integrity beyond infinitesimal small displacements from the equilibrium along the normal modes, e.g. by iterating between moving along normal mode directions and re-evaluating of normal modes based on the energy minimized intermediate conformations.

4. The eigenvalue decomposition of a matrix has in general cubic complexity, although highly efficient solvers exist. Therefore, all-atom NMA based on detailed force fields, as used for instance in atomistic MD simulations, is practically limited to small and mid-sized molecules.

To reduce computational costs and mitigate these limitations coarse-grained NMA based on elastic networks deliberately simplify potential function and/or model resolution. We will introduce them in the following.

3.3.2 Elastic Network Models (ENMs)

More than 20 years ago, Monique Tirion proposed the first elastic network model (ENM) based on a highly simplified harmonic potential instead of the complex force fields commonly used by NMA (Tirion, 1996). She modeled a protein as network of atoms, where spatially close neighbors are linked by uniform elastic springs (based on a distance cutoff). Despite its simplicity, the elastic network closely resembled the low-frequency modes and fluctuations predicted by standard NMA.

Several advantages arise from this single-parameter model: First, there is no need to minimize the energy of the initial conformation because it is defined to be the minimum energy state of the harmonic potential. Besides saving computational time, this also avoids the risk of structural distortions. Second, it reduces the computational costs for diagonalizing the Hessian of the network potential to infer normal modes (eigenvectors) and associated frequencies (eigenvalues). And most importantly, studying conformational transitions by the means of predicted normal modes became computationally feasible.

Apart from simplifying the potential function also the resolution of the underlying model can be reduced, which inspired the development of several ENM variants shortly after Tirion's model was published. The **Gaussian Network Model (GNM)** (Haliloglu et al., 1997, Bahar et al., 1997) is an elastic network model defined on residue- instead of atomic-level. Because the GNM only considers topological constraints between residues it predicts collective dynamics in form of residue fluctuations and their cross-correlations but cannot provide information about their directionality. The **Anisotropic Network Model (ANM)** (Atilgan et al., 2001, Tama and Sanejouand, 2001) extends the GNM with directional information by considering the three-dimensional coordinates of the C_α-atoms representing the residues. Both ENMs presented in this thesis (chapters 5 and 6) are based on the ANM. Thus, we will introduce its theoretical foundations in more detail below (see 3.3.3).

Coarse-grained NMA based on ENMs became a quite popular tool for studying protein motions due to its ability to combine computational efficiency with surprising biological accuracy. Numerous studies have demonstrated that the most dominant normal modes capture collective protein motions with functional relevance (Tama and Sanejouand, 2001, Krebs et al., 2002, Eyal et al., 2006, Rueda et al., 2007a, Ahmed et al., 2010, Bahar et al., 2010a, Kurkcuoglu et al., 2012, Meireles et al., 2011, Mahajan and Sanejouand, 2015, Bahar et al., 2015, Sankar et al., 2018). These motions are robustly encoded by the overall geometric shape of the protein without being affected by small variations in structural details or potential function. Consequently, ENMs allow the analysis of large biomolecular assemblies that exceed the range of atomistic MD simulations.

Today, ENMs have reached a broad range of applications (Fig 3.7) (López-Blanco and Chacón, 2016). For instance, they are employed to study the intrinsic motions of large molecular complexes, such as viral capsids (Lee et al., 2018, Hsieh et al., 2016) or the ribosome (Wang et al., 2004, Kurkcuoglu et al., 2009a). They provide insights into the mechanism of allosteric effects (Guzel and Kurkcuoglu, 2017) or reveal structural dynamics of membrane proteins (Bahar et al., 2010a, Di Luca and Kaila, 2018). They are be used to produce pools of candidate structures for protein docking (Cavasotto et al., 2005, Dobbins et al., 2008, Meireles et al., 2011, Cavasotto, 2012, Kurkcuoglu and Doruker, 2016). They guide fine-grained conformational sampling, such as MD simulations or geometric exploration along the most dominant motion directions of a protein (Kirillova et al., 2008, Gur et al., 2013). They assist the refinement of experimentally determined models through normal-mode guided flexible fitting of candidate structures into low-resolution density maps (Hinsen et al., 2010, Schröder et al., 2007, Gniewek et al., 2012). They have been applied to study the dynamics of RNA (Mailhot et al., 2017, Zimmermann and Jernigan, 2014, Pinamonti et al., 2015), to predict the effect of sequence mutations on the

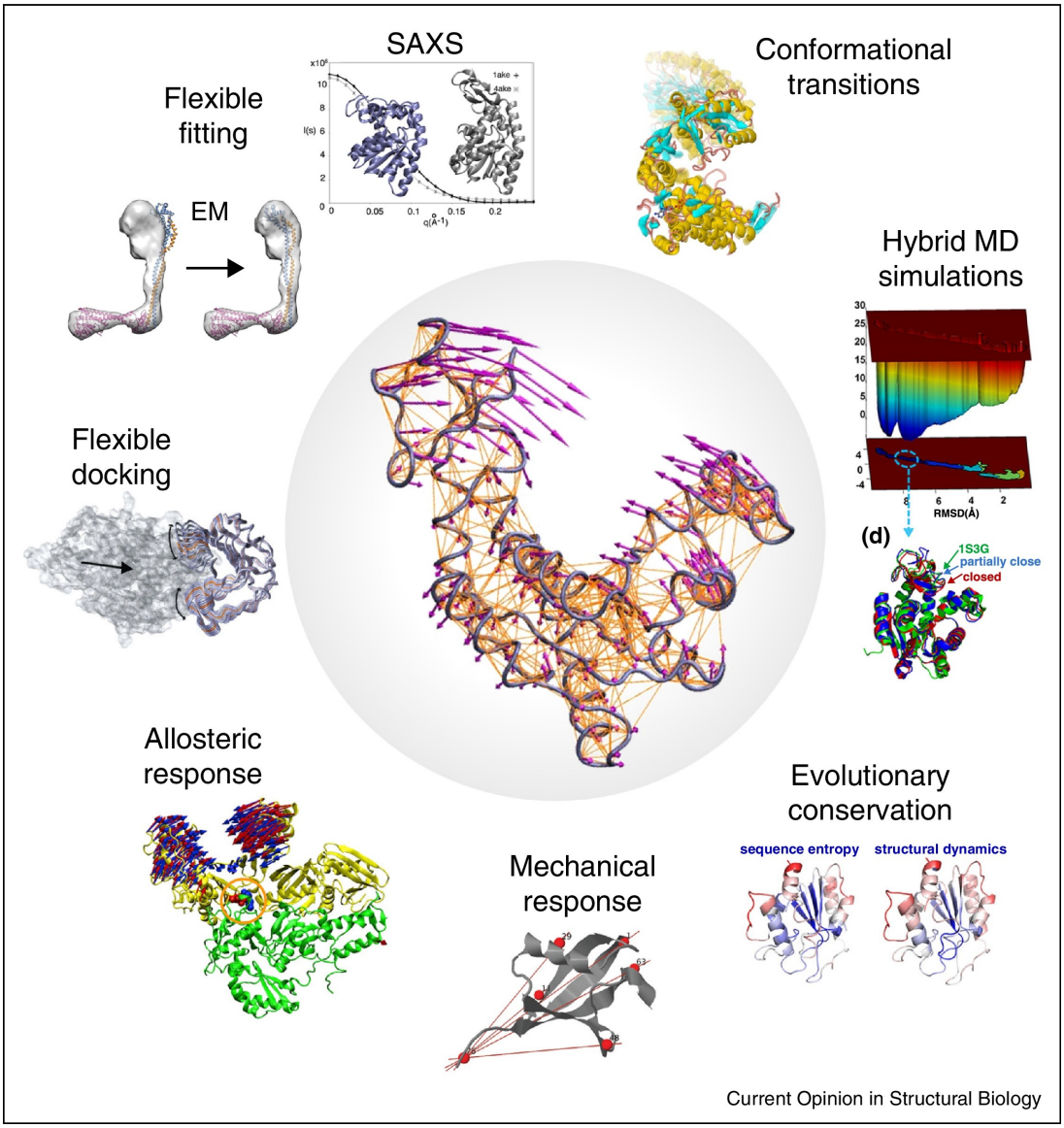

Figure 3.7: Representative applications of ENMs. Application scenarios range from analysis and prediction of biologically relevant motions to generation of conformational ensembles used in various structural biology scenarios. The central image shows the elastic network model of adenylate kinase with arrows indicating the motion direction corresponding to the lowest energy normal mode. Figure source: Reprinted from López-Blanco and Chacón (2016), Copyright (2018), with permission from Elsevier and additional permission from Refs. (Lopéz-Blanco et al., 2011, Oot et al., 2012, Miyashita et al., 2011, Wang et al., 2014b, May and Zacharias, 2008, Liu and Bahar, 2012, Bahar et al., 2015) to reprint the images around the central image.

structure-encoded protein motions (Frappier and Najmanovich, 2014), or to examine the link between structure-encoded motions of a protein and evolution of structure and sequence (Haliloglu and Bahar, 2015, Liu and Bahar, 2012).

3.3.3 Anisotropic Network Model (ANM)

The elastic network models, *mc*ENM and *lmc*ENM (see chapters 5-6), presented in this thesis base on the widely used anisotropic network model (ANM) (Atilgan et al., 2001, Tama and Sanejouand, 2001). Essentially, the ANM is a low-resolution variant of Tirion's ENM (see 3.3.2) using the same simplified harmonic potential. But instead of considering atomic interactions, the ANM captures interactions between the C_α-atoms of spatially close residues ("in contact") by connecting them with uniform elastic springs.

Fig 3.8 illustrates the coarse-grained quadratic approximation of a protein's detailed energy landscape. The native conformation of the protein defining the energy minimum of the harmonic potential is modeled as ANM. By deforming the initial conformation along the predicted global modes two substates S1 and S2 can be reached, which have been revealed at an intermediate resolution of the energy profile.

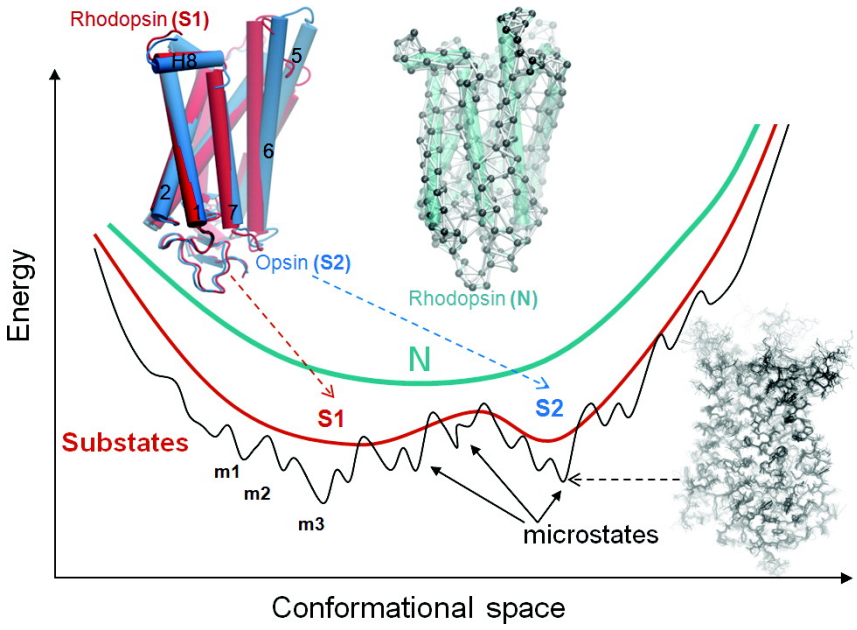

Figure 3.8: Coarse-grained approximation of the energy landscape (2D-profile) around the native conformation of a protein. N refers to the native conformation modeled as a coarse-grained elastic network model, which resides at the single energy minimum of the quadratic approximation (green curve) of the energy profile (black curve). In between the detailed energy profile with several microstates (m1, m2, m3, ...) and its harmonic approximation is an energy profile at intermediate resolution, which reveals two and more substates (S1, S2, S3, ...). The structural models of substates S1 and S2 result from sampling along global modes around the native conformation. Small fluctuations around S2 produce a conformational ensemble at higher structural resolution (bottom right). Figure source: Reprinted with permission from Bahar et al. (2010a). Copyright (2018) American Chemical Society.

In general, the spring stiffness can be varied as done by several ENM variants (Fuglebakk et al., 2015) (see also chapter 2). In this thesis, we rely on the ANM in its classical form using

an uniform spring stiffness to purely focus on the effect of dynamic contact changes on ENM accuracy (see chapters 5 and 6).

Formally, two residues i and j are in contact if their distance is shorter than a pre-determined cutoff r_c. The binary contact matrix $\mathbf{C}$ captures this network in its elements C_{ij}:

$$C_{ij} = \begin{cases} 1, & \text{if } d_{ij} \leq r_c \\ 0, & \text{otherwise} \end{cases} \tag{3.21}$$

where d_{ij} denotes the Euclidean distance between the three-dimensional coordinates of the C_α atoms, which represent residues i and j. The cutoff distance r_c depends on the type of ENM used and is often tailored protein- or problem-wise.

The generalized form of the entire network potential of the ANM is defined as

$$V_{\text{ANM}} = \sum_{i,j}^{N} \frac{K_{ij}}{2}(d_{ij} - d_{ij}^0)^2 \tag{3.22}$$

where d_{ij} and d_{ij}^0 denote the instantaneous and equilibrium distance of residues i and j measured between their C_α atoms, N is the number of residues of the protein, and K_{ij} refers to the elements of the stiffness matrix defined below.

The stiffness matrix $\mathbf{K}$ specifies the spring stiffness between residues i and j as

$$K_{ij} = \gamma \cdot C_{ij} \tag{3.23}$$

where γ is a uniform stiffness constant and $C_{ij} \in \{0, 1\}$ refers to the entry of residues i and j in the contact topology matrix of the initial conformation as defined in eq. 3.21.

For a protein with N residues, the Hessian is a $3N \times 3N$ matrix that is constructed from 3×3 submatrices. Using eqn. 3.22 in eqn. 3.18, the off-diagonal elements H_{ij} are given by

$$H_{ij} = -\frac{K_{ij}}{d_{ij}^2} \begin{bmatrix} (x_j - x_i)^2 & (x_j - x_i)(y_j - y_i) & (x_j - x_i)(z_j - z_i) \\ (y_j - y_i)(x_j - x_i) & (y_j - y_i)^2 & (y_j - y_i)(z_j - z_i) \\ (z_j - z_i)(x_j - x_i) & (z_j - z_i)(y_j - y_i) & (z_j - z_i)^2 \end{bmatrix} \tag{3.24}$$

and the diagonal elements H_{ii} as

$$H_{ii} = -\sum_{i,j} H_{ij} \tag{3.25}$$

Diagonalization of the Hessian matrix yields the normal modes (eigenvectors) and frequencies (eigenvalues) of the ANM, as described in detail above (see 3.3.1).

4

Materials and Methods

In this chapter we present materials and methods that are relevant to both proposed elastic network models in this thesis, namely the **elastic network model of maintained contacts (*mc*ENM)** in chapter 5 and the **elastic network model of *learned* maintained contacts (*lmc*ENM)** in chapter 6. First, we introduce the set of proteins that is used for development and evaluation of *mc*ENM and *lmc*ENM. Next, we introduce the measures that we use to evaluate the ENMs in terms of biological accuracy, captured deformation space, as well as alignment of predicted low-frequency modes to essential dynamics of conformational ensembles. Last, we introduce the reference ENMs that we use to validate *mc*ENM and *lmc*ENM including their parametrization.

4.1 Protein Data Set

To train and test our classifier as well as to evaluate the performance of *lmc*ENM, we chose a set of proteins with known motion type. The *Protein Structural Change DataBase (PSCDB)* (Amemiya et al., 2011, 2012) provides motion classified protein pairs, each representing the functional transition of one protein family in the SCOP (Structural Classification of Proteins) database. A pair consists of two conformations, marking start and end of the functional transition, where only the latter is bound to a ligand, the former is unbound. The PSCDB classifies each of these functional transitions into six motion types (see below). In particular, it distinguishes highly collective, domain motions from localized, uncorrelated transitions. This allows us to explicitly assess the ability of our approach to explain localized, functional transitions that are elusive for classical ENMs.

We applied several filters to extract a meaningful and consistent data set from the PSCDB and excluded proteins

(a) without significant motion (root mean squared distance (RMSD) $\leq$ 1.0Å),

(b) with less than 70 residues alignment length,

(c) with resolution higher than 2.5Å,

(d) including chain breaks (defined as more than 4.2Å Euclidean distance between two consecutive C_α atoms along the sequence (Li et al., 2011)),

(e) including a peptide with more than six non-hydrogen atoms in the unbound conformation (Brylinski and Skolnick, 2008),

(f) with largely extended or disordered structures.

Furthermore, we limited ourselves to single-chain proteins to enable faster development and testing. Filters (a) and (d) exclude proteins encoding little to no information about protein motions, whereas (b), (c), (e), and (f) exclude proteins for which this information is distorted due to low structural quality, highly specialized structural topology, or interaction with other chains.

Our final data set of 90 protein pairs is distributed across the following motion classes (see Dataset A in the supplementary file S2 of our paper (Putz and Brock, 2017)): coupled domain motion (short: CDM, 21 protein pairs), independent domain motion (IDM, 14), coupled local motion (CLM, 27), independent local motion (ILM, 18), buried ligand motion (BLM, 4), and other types of motion (OTM, 6). Both domain and local motions can be associated with ligand binding (coupled) or without (independent). Proteins that are bound to a ligand in the end conformation, but lack considerable movement between start and end, are categorized as buried ligand motions. Although, these proteins move to bind the ligand, the structural differences between the two conformations are small because the ligand-free conformation seems to imitate the shape of the ligand using occluded water molecules (Amemiya et al., 2011). All remaining proteins fall into the category other types of motions that are larger (RMSD $>$ 1.0Å) but do not match the criteria for local or domain motions.

The length of proteins in our data set ranges from 70 to 712 amino acids. The RMSDs (root mean squared distances) between the unbound and bound conformation range between 1.1Å and 9.6Å.

4.2 Evaluation of Elastic Network Models

Coarse-grained ENMs often guide more detailed exploration of protein motions (see Elastic Network Models (ENMs)). The value of this guidance largely depends on two factors: First, how much can the guidance be trusted, i.e. how accurate is the prediction of the essential deformation space that has to be searched. Second, how much can it reduce computational cost by narrowing down the search space for conformational exploration.

In addition, we evaluate how well the predicted low-frequency ENM-modes match to the dominant structural deformations captured by conformational ensembles.

4.2.1 Assessing the Biological Accuracy

A common measure to evaluate the accuracy of ENMs is the **mode overlap** O_j (Marques and Sanejouand, 1995, Tama and Sanejouand, 2001). It specifies the amount of conformational change captured by a single mode j based on the angle between conformational displacement vector and mode direction vector M_j, as defined in:

$$O_j = \frac{\left| \sum^{3N} M_j \Delta r_i \right|}{\left[\sum^{3N} M_j^2 \cdot \sum^{3N} \Delta r_i^2 \right]^{1/2}} \tag{4.1}$$

where $\Delta r_i = (r_i^S - r_i^E)$ denotes the displacement vector from start (r_i^S) to end conformation (r_i^E) at residue i; N is the number of residues of the protein. The measure ranges between 0 and 1 (perfect match).

By summing up the individual mode overlaps of the first k low-frequency modes, we now can specify their **cumulative mode overlap** $CO(k)$ (Yang et al., 2007). It indicates how accurate the deformation space spanned by these modes captures the functional transition, given by:

$$CO(k) = \left[\sum_{j=1}^{k} O_j^2 \right]^{1/2}. \tag{4.2}$$

In principle, the number of low-frequency modes required to span the *essential* deformation space is unknown. This is due to its strong coupling to the collectivity of the functional transition (see 3.3.2). However, usually less than ten modes suffice to accurately capture function-related movements that are highly collective. In the results section, we thus assess the cumulative mode overlap of the first ten low-frequency modes $CO(10)$ unless stated otherwise. We use $CO(10)$ as main measure for benchmarking the different ENM variants. To avoid over-fitting to a single measure we evaluate a set of other commonly used metrics described below.

The **Pearson correlation coefficient** is used to measure the similarity between predicted residue fluctuations and observed displacements, as well as between predicted fluctuations and experimental B-factors from the unbound conformation. Predicted fluctuations were scaled to observed displacements and B-factors, respectively. The correlation coeffient ranges between -1 (total negative correlation), 0 (no correlation) and +1 (total positive correlation).

The **fraction of variance** of a mode measures how much of the structural variance it explains. It is defined by the variance of mode j divided by the trace of the covariance matrix of the model. The **cumulative fraction of variance** (CFV(k)) sums up the individual contributions of the first k low-frequency modes.

The **degree of collectivity** κ_i (Brüschweiler, 1995) of a protein motion quantifies the number of involved residues. It is given by:

$$\kappa_j = \frac{1}{N} \exp(-\sum_i^N u_{j,i}^2 \log u_{j,i}^2) \tag{4.3}$$

where N denotes the number of residues of the protein and $u_{j,i}^2$ is defined as $u_{j,i}^2 = \alpha \frac{1}{m_j}(M_{j,X}^2 + M_{j,Y}^2 + M_{j,Z}^2)$ with M_j being the j-th mode vector and m_j its mass; α is a normalization factor to ensure that $\sum_i N u_{j,i}^2 = 1$. The measure varies between $1/N$ (only one residue affected) and 1 (maximally collective).

4.2.2 Assessing the Dimensionality of Deformation Space

The dimensionality of the essential deformation space depends on the desired accuracy. Therefore, we assess the number of modes required to capture 70%, 80%, and 90% of the functional transition (measured in percent cumulative mode overlap). Lower dimensionality effectively reduces computational cost for subsequent exploration of this space.

In addition, we report the **maximum overlap** Max$O(j)$ among the first j modes, together with the rank of the corresponding mode (rank 0 refers to the first mode), its collectivity, and fraction of variance.

4.2.3 Comparing against Essential Dynamics of Conformational Ensembles

Conformational ensembles obtained from structural databases provide an additional source to characterize protein flexibility (Best et al., 2006, Burra et al., 2009, Monzon et al., 2016). For a subset of proteins we obtained such an ensemble and analyzed its Essential Dynamics (ED) using Principal Component Analysis (PCA) as implemented in ProDy (Bakan et al., 2011). We use the following measures to analyze the similarity between ENM deformation space and principal components space.

The **Pearson correlation coefficient** CC is used to determine the similarity between the mean square fluctuations captured by ED and the squared fluctuations of the ENM. It varies between -1 (total negative correlation), 0 (no correlation) and +1 (total positive correlation).

The **root mean square inner product (RMSIP)** (Amadei et al., 1999) measures the similarity of two vectorial spaces by the overlap of their k-dimensional subspaces:

$$\mathrm{RMSIP}(k) = \left[\frac{\sum_{i,j=1}^{k} (U_i \cdot V_j)^2}{k} \right]^{1/2} \tag{4.4}$$

where U_i and V_j are the eigenvectors/principal components of the compared covariance matrices; k is the dimensionality of the subset of low-frequency modes/principal components. Commonly, k is set to an arbitrary value of 10. RMSIP ranges between 0 and 1 (perfect match).

A related measure of vector space similarity is the **root weighted square inner product (RWSIP)** (Carnevale et al., 2007). In contrast to the RMSIP it considers the relative contribution of each eigenvector (direction) weighted by its corresponding eigenvalue (magnitude). Further, it takes into account the full spaces to be compared instead of a small subspace. The RWSIP is given by:

$$\text{RWSIP} = \left[\frac{\sum_{i,j=1}^{N} u_i v_j (U_i \cdot V_j)^2}{\sum_{i=1}^{N} u_i v_i} \right]^{1/2} \tag{4.5}$$

where U_i and V_j are the eigenvectors/principal components of the compared covariance matrices; u_i and v_j are the eigenvalues; N is the number of non-trivial eigenvectors in each mode set. ENM eigenvalues have been inverted to be proportional to the relative amplitudes captured by PCA eigenvalues. RWSIP ranges between 0 and 1 (perfect overlap).

4.3 Reference Elastic Network Models Used for Evaluation

To validate our approach, we compare it against the **baseline ENM** defined by the classical ANM as well as three other reference ENM variants:

- **HCA-method** - a cutoff-free elastic network model with distance-dependent force constants (Hinsen et al., 2000),
- **OFC-ENM** - a model with optimized force constants based on structural properties (Lezon and Bahar, 2010),
- **edENM** - a hybrid elastic network model combining a bond-cutoff strategy for close sequential neighbors and distance-dependent force constants for remote interactions (Orellana et al., 2010).

All variants use the general form of network potential, V_{ANM}, as defined in eq. 3.22 on page 34. In the following we introduce the technical details and parametrization of each reference ENM.

4.3.1 Baseline ENM

All ENM variants proposed in this thesis rely on the Anisotropic Network Model (ANM) invented by Bahar et al. (Bahar et al., 1997) (see 3.3.3). As such it defines the natural *baseline* for evaluating our approach.

In their simplest form ANMs assign a uniform stiffness constant to all springs in the network and only parametrize the distance-cutoff to determine residues in contact, which influences the overall density of the network. A common strategy is to adjust the cutoff distance of ANMs within a range of 8-15Å based on the given protein or problem (Atilgan et al., 2001, Eyal et al., 2006, Kondrashov et al., 2007, Leioatts et al., 2012, Fuglebakk et al., 2013, 2015). Because sparser networks at smaller cutoffs tend to become unstable, cutoff values of 12Å and larger are typically chosen (Jeong et al., 2006, Eyal et al., 2006, Fuglebakk et al., 2013). While these artificial constraints usually do not alter collective motions of the network they often reduce the actual mobility of local parts of the network. As a consequence, ANMs at these cutoffs are less suitable to predict localized functional transitions, which is the main goal of this thesis. We therefore tried to lower the cutoff distance for our data set as much as possible without making the network unstable.

To do so, we evaluated the performance of ANM at cutoff values ranging from 8-18Å by measuring their cumulative mode overlap of the first ten low-frequency modes. At cutoffs lower than 11Å, some networks became unstable yielding more than the trivial six zero eigenvalues: 20 cases for ANM8, 9 cases for ANM9, and 3 cases for ANM10 from our full set of 90 proteins. We therefore tried to stabilize the ANMs to work at cutoffs lower than 11Å .

To be stable elastic network models must fulfill two requirements (Jeong et al., 2006): (i) each node must be connected to at least four other nodes, (ii) the network must have at least $3N - 6$ edges, where N refers to the number of nodes. As first step towards stabilizing the ANM at lower cutoffs we therefore enforce that each C_α atom is constrained by at least four neighbors (node degree >= 4). Under-constrained C_α atom get connected to their closest–not yet connected–neighbors along the sequence irrespective of their distance. Please note that this also adds artificial constraints but only locally, which is unlikely to alter the intrinsic motions encoded by the ANM in contrast to an overall larger cutoff as discussed above.

Table 4.1 shows the performance of this network, called $\text{ANM}_{minDeg4}$, at different cutoffs. The 4-neighbor-connectedness criterion largely reduces the number of networks yielding more than six zero eigenvalues, yet some networks remain unstable at cutoffs 8 and 9Å. However, $\text{ANM10}_{minDeg4}$ now fulfills the criterion of six zero eigenvalues for all proteins and yields largest agreement between predicted and actual motion directions. Hence, we chose it as baseline for our approach.

This is in line with Kondrashov et al. (2007) who found that a distance-cutoff of 10Å yields largest agreement in overlap of motion directions, but less accurate prediction of motion magnitudes. In turn, the best match of motion amplitudes (fluctuation profiles) requires cutoffs larger than 15Å, thereby reducing the overlap in motion directions by increasing structural stiffness and collectivity

Table 4.1: Performance of ANM$_{minDeg4}$ at cutoff value r_c ranging between 8 and 18Å measured by the cumulative mode overlaps of the first ten low-frequency modes evaluated on the LMC_all data set (90 proteins). Cutoffs 9Å and 10Å yield best median overlap. However, only ANM10$_{minDeg4}$ fulfills the criterion of six zero eigenvalues making it the better choice as baseline for our approach. Table source: Putz and Brock (2017).

r_c (Å)	8[a]	9[b]	10	11	12	13	14	15	16	17	18
median	0.670	0.685	0.685	0.677	0.673	0.675	0.665	0.660	0.647	0.624	0.615
mean	0.661	0.665	0.665	0.667	0.664	0.660	0.651	0.646	0.639	0.631	0.622

[a] 10 cases with more than the six trivial zero eigenvalues
[b] 3 cases with more than the six trivial zero eigenvalues

of motion. This counteracts our goal to accurately model localized functional transitions with low degree of collectivity.

We also find that the simple 4-neighbor-connectedness criterion above does not guarantee network stability of distance-cutoff based ANMs (see ANM8$_{minDeg4}$ and ANM9$_{minDeg4}$ in Table S13. This is interesting because also the second requirement for stabilized networks is maintained as it is for the bond-cutoff based ENM proposed by Jeong et al. (Jeong et al., 2006). They connect each CA-atom with its closest four sequential neighbors and scale spring stiffness based on sequence distance. But, due to the sparser network they need to explicitly model chemical interactions, such as disulfide bridges, hydrogen bonds, of van-der-Waals forces, which makes the network construction more complex than for the distance-cutoff based ENM. In contrast, we only ensure that each CA-atoms is constrained by four contacts that are not necessarily the closest four sequential neighbors and model springs with uniform stiffness. This allows us to keep the network construction simple but also to solely focus on the effect of changes in the network topology.

In the rest of this thesis, we refer to this ANM10$_{minDeg4}$ simply as ENM or ***baseline* ENM**.

4.3.2 HCA

The HCA method (Hinsen et al., 2000) defines the spring stiffness between all residue pairs in the network using a fast decaying distance-dependent function:

$$K_{ij} = \begin{cases} a \cdot d_{ij} - b, & \text{if } d_{ij} < r_c \\ c \cdot (d_{ij})^{-d}, & \text{otherwise} \end{cases} \tag{4.6}$$

where d_{ij} is the Cartesian distance between residues i and j. We used the parametrization of the original publication ($a = 205.5$ kcal mol^{-1}Å^{-3}, $b = 571.2$ kcal mol^{-1}Å^{-2}, $c = 3.059 * 10^5$ kcal mol^{-1}Å^4, $d = 6$, and $r_c = 4.0$Å).

4.3.3 OFC-ENM

OFC-ENM (Lezon and Bahar, 2010) scales spring stiffness based on secondary structure type and sequential distance between interacting residues. The optimal stiffness constants are obtained by analyzing NMR-ensembles using entropy maximization. We use OFC-ENM with the distance cutoff 10Å and the default parameter set as implemented in ProDy (Bakan et al., 2011).

4.3.4 EDENM

edENM (Orellana et al., 2010) is a hybrid elastic network model that distinguishes three types of interactions. Residues close in sequence (up to three sequence positions apart) build a fully connected network where spring stiffness depends on sequence distance. Interactions between residues within a protein-size-dependent cutoff, r_c, are modeled with distance-dependent springs. Irrelevant, remote interactions above the cutoff are excluded from the network. This leads to the following definition of spring stiffness between residues i and j:

$$K_{ij} = \begin{cases} a/(s_{ij})^b, & \text{if } s_{ij} \leq 3 \\ (c/d_{ij})^d, & \text{if } s_{ij} > 3 \text{ and } d_{ij} \leq r_c \\ 0, & \text{otherwise} \end{cases} \tag{4.7}$$

where d_{ij} (s_{ij}) is the Cartesian (sequential) distance between residues i and j, respectively; $r_c = 2.9 * \ln(N) - 2.9$ is a size-dependent distance cutoff with N being the number of residues of the protein. We used the default parametrization of the original publication, which was optimized based on MD simulations ($a = 60$ kcal mol^{-1}Å^{-2} and $b = 2$; $c = 6$ kcal mol^{-1}Å^{-2} and $d = 6$).

5

Elastic Network Model of Maintained Contacts (*mc*ENM)

5.1 Introduction

This chapter addresses a major shortcoming of elastic network models (ENMs): While they reliably predict collective protein motions, they poorly capture localized functional transitions in most cases. In chapter 1 we argued that this limitation questions the practical relevance of ENMs because the motion type of a protein is usually unknown a priori. Hence, we have no guarantee that the motions predicted by ENMs match the actual motions of the protein.

The goal of this chapter is to show that this limitation can be overcome by accounting for dynamic changes in the network connectivity of ENMs. This is based on the key insight that localized, function-related transitions often involve substantial changes in the contact topology of a protein. To reliably predict its motions, we therefore need to adjust its initial contact topology to reflect these dynamic changes. In case we know at least two conformations of a protein we will see that this task simply becomes extracting the differences between their contact topologies, i.e. which contacts break, form, or are maintained. We call these differences *observed* contact changes. Of course, to see effective changes in the contact topologies, the structures of the conformations must differ enough. Good candidates for such a pair of conformations are start and end of a functional transition.

By investigating the effect of these dynamic contact changes on ENM accuracy we identify breaking contacts as the relevant contact change. Removing the springs associated with observed breaking contacts releases constraints on local parts of the elastic network, which were imposed by the initial contact topology of the protein. We call the resulting network the **elastic network of maintained contacts** (***mc*ENM**, see Fig 5.1).

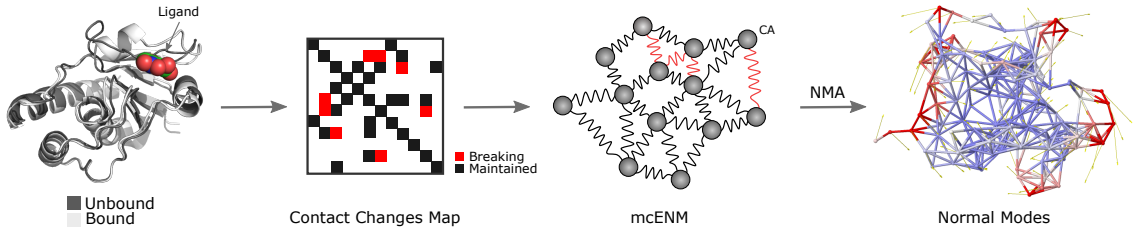

Figure 5.1: Flowchart overview of *mc*ENM construction and analysis. First, we identify observed breaking contacts between start (unbound) and end (bound) conformation of a functional transition. Next, we remove the springs associated with the observed breaking contacts from the initial elastic network model of the start conformation, resulting in the network of maintained contacts (*mc*ENM). Last, we analyze *mc*ENM using normal mode analysis (NMA) to predict the intrinsic deformations (normal modes) of the protein (image generated with ANM 2.0 web server (Eyal et al., 2015)).

We evaluate the performance of *mc*ENM on a set of 90 pairs of protein conformations covering different motion types and compare it to the classical, distance-cutoff based ENM. Our results show that *mc*ENM is indeed capable of predicting localized functional transitions, thereby expanding the range of motions that can be captured by ENMs. Furthermore, *mc*ENM requires fewer modes to capture such localized function-related movements, which alleviates another problem of ENMs not to know how many and which modes to consider. Finally, we investigate the relationship between the occurrence of observed breaking contacts and their effect on ENM accuracy by considering motion type, structural fold, and function class of the proteins in our data set.

This chapter provides the foundation for the core contribution of this dissertation–presented in the following chapter 6–in two regards: First, *mc*ENM marks the theoretical upper bound on ENM precision that can be achieved with a "perfect" network, which is useful when evaluating the performance of different ENM variants. Second, the observed breaking contacts contain valuable information that enables their prediction, as we will see in the following chapter.

5.1.1 Contributions

In this chapter, we make the following contributions:

Conceptual Contributions

- We propose that accounting for dynamic changes in the contact topology of proteins expands the range of motion types that can be explained by elastic network models (ENMs).
- We identify *observed* breaking contacts in the initial contact topology as a major obstacle for ENMs to capture localized function-related protein motions. These breaking contacts can be obtained if at least two conformations of a protein, preferably start and end of a functional transition, are known.

Technical Contributions

- We present a novel elastic network model of maintained contacts (*mc*ENM) that accounts for a particular type of observed dynamic changes in the underlying contact topology of proteins. It ignores springs that are associated with contacts that have been observed to break when the protein moves. *mc*ENM can be applied to proteins, when more than one conformation is known.

Empirical Contributions

- We show that the absence of *observed* breaking contacts enables *mc*ENM to capture function-related protein motions not only when they are collective but also when they are local and uncorrelated. It also substantially reduces the dimensionality of the essential deformation space required to explain localized functional transitions. If more than one conformation of a protein is known, *mc*ENM provides a more accurate prediction of the motions required to transition between these two conformational states than the widely used, distance-cutoff based ENMs.

- We show that *mc*ENM improves prediction accuracy not only for local movers but also for proteins in other motion categories. We also show that certain structural folds and functional classes of proteins promote dynamic changes in the contact topology more than others and that breaking contacts differ in their impact on ENM accuracy.

5.1.2 Outline

The rest of this chapter is organized as follows:

- **Section 5.2 Methods** introduces the concept of dynamic contacts, i.e. observed contact changes, their definition, and the strategies to identify, which contact changes are relevant to improve ENM accuracy.

- **Section 5.3 Implementation** describes the parametrization of *mc*ENM and its baseline, the distance-cutoff based ENM.

- **Section 5.4 Results and Discussion** describes the experimental setup and analyzes and discusses the experimental results, which contain an analysis of breaking contact occurrence in relation to their effect on ENM accuracy w.r.t. motion type, structural fold, and functional class of proteins.

- **Section 5.5 Conclusion** summarizes the findings of this chapter, discusses its limitations, and establishes observed breaking contacts as a novel source of information to improve prediction accuracy and general applicability of ENMs.

5.2 Methods

In this section we introduce our definition of contact changes and the resulting types of dynamic contacts. Based on these changes we then adjust the initial contact topology of proteins to study the effect of contact changes on ENM accuracy.

As introduced in 3.3.2 elastic network models build upon the contact topology of proteins, which simply models them as network of interacting residues. Two residues interact, i.e. are *in contact*, if the distance between their C_α atoms is within a pre-defined distance cutoff (see eqn. 3.21). The contact topology of a protein is derived from a single conformation. Therefore it encodes *static* structural connectivity. But proteins move between different conformations, for instance to accommodate their shape to fit a binding partner. Depending on the type of motion these changes may also affect the coarse-grained contact topology of a protein.

5.2.1 Definition of Contact Changes and Contact Types

To identify function-related changes in connectivity, we thus compare the contact matrices between start (S) and end (E) conformation of a functional transition. Taking the endpoints of a conformational transition increases the chance that associated changes in the contact topology are substantial and relevant for the protein to function.

We observe three different types of contact changes yielding the following contact types, which we call *dynamic contacts*:

- **maintained** contacts preserve their original distance within a pre-defined threshold,
- **breaking** contacts exceed or shorten their original distance above this threshold, and
- **forming** contacts establish between residues that are in contact distance after the movement, i.e. in the end conformation.

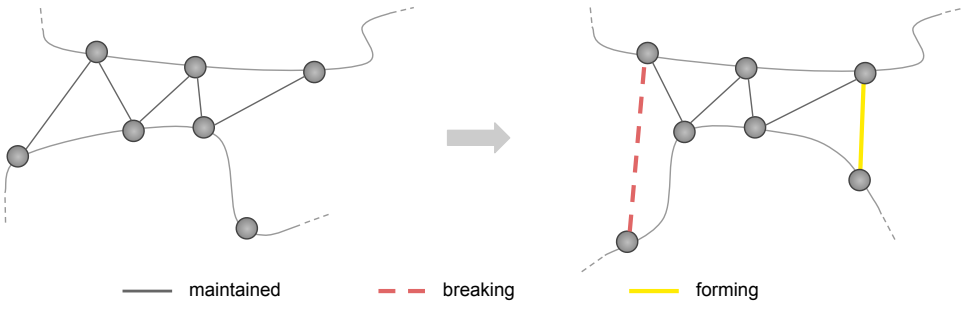

Figure 5.2: Simplified illustration of types of contact changes. Deformation of the contact network causes a contact to break, another one to form, while most contacts maintain their distance.

Based on the contact matrix C as defined in eqn. 3.21 we formalize the dynamic contacts in the transition matrix T, whose elements, T_{ij}, encode the three different types of contact transitions, defined as follows:

$$T_{ij} = \begin{cases} \text{maintained contact,} & \text{if} \quad C_{ij}^S = 1 \quad \text{and} \quad e_{ij} \triangleq |\frac{\Delta d_{ij}}{d_{ij}^S}| \leq e_c \\ \text{breaking contact,} & \text{if} \quad C_{ij}^S = 1 \quad \text{and} \quad e_{ij} \triangleq |\frac{\Delta d_{ij}}{d_{ij}^S}| > e_c \\ \text{forming contact,} & \text{if} \quad C_{ij}^S = 0 \quad \text{and} \quad C_{ij}^E = 1 \\ \text{no contact,} & \text{otherwise} \end{cases} \tag{5.1}$$

where C_{ij}^S (C_{ij}^E) refers to the entry for residues i and j in the contact matrix of the start and end conformation, respectively; e_{ij} denotes the distance change between residues i and j relative to their initial distance in the start conformation, where $\Delta d_{ij} \triangleq d_{ij}^S - d_{ij}^E$. Intuitively, e_{ij} can be interpreted as strain measuring how much the distance between two particles in a body elongates ("stretch") or shortens ("compression") relative to their original distance. We limit the distance change by an upper bound e_c to distinguish breaking from maintained contacts.

Fig. 5.3C shows an example of a contact transition matrix that is derived from the conformational transition of Arsenate reductase (ArsC) upon binding a ligand.

5.2.2 Identification of Relevant Contact Changes

Based on the observed contact changes we consider two of the three possible strategies to adjust the initial contact topology of ENMs:

(I) removing breaking contacts, and

(II) removing breaking contacts **and** adding forming contacts.

We explicitly exclude the third possible strategy from our analysis, which is to add forming contacts without removing the breaking ones. This is because it immediately contradicts our main hypothesis that erroneous constraints prevent parts of the initial network from performing localized motions. Every forming contact imposes another constraint on the network rendering it less flexible.

Strategy I - Removing Breaking Contacts

What does it actually mean if we observe contacts to break when the protein moves? Two parts of the protein, initially in contact distance, are driven apart throughout the function-related movements of the protein. Thus, elastic network models should model these contacts much weaker than contacts that maintain their distance during the motion. Otherwise they would locally render the network stiffer than it actually is, thereby preventing localized functional transitions.

However, the widely used, distance-cutoff based ENMs treat every contact the same due to the uniform spring stiffness. This leaves us with two options: Either we adjust the spring stiffness accordingly, for instance based on the observed relative distance change of contacts, or we completely ignore the springs associated with observed breaking contacts in the network.

We decided to do the latter for two reasons: First, only the network topology is adjusted without making the simplified model more complex. This is motivated by a widely accepted principle in

science called *Occam's Razor*, which states that simpler hypotheses making fewer assumptions should be preferred over more complex ones because they can easier be tested[1]. Second, weakening the springs associated with observed breaking contacts may still underestimate the mobility of functionally relevant loops that are highly flexible. In particular, if partial unfolding and refolding is involved complete removal of these erroneous constraints may be necessary to explain the flexibility of these parts. The case study 6.4.7 in the following chapter presents an example for partial unfolding and refolding of functionally relevant loops in an outer membrane transporter. Its functional transitions are more accurately captured by ENMs if the observed breaking contacts are ignored.

We call the resulting model without the springs associated with the observed breaking contacts elastic network of maintained contacts (*mc*ENM). The steps to build *mc*ENM are illustrated in Fig 5.3. *mc*ENM allows us to solely examine the effect of a refined network topology on ENM accuracy. Optimizing spring stiffness can be considered a follow up step to further tune ENM performance. In general, *mc*ENM can be combined with any other ENM formulation that optimizes spring stiffness.

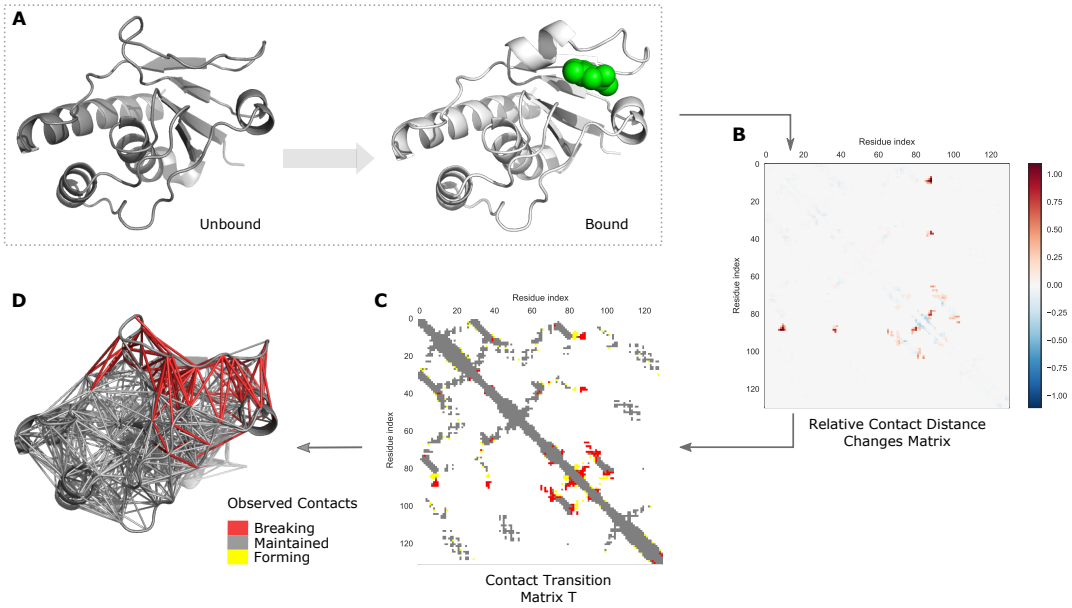

Figure 5.3: Illustration of *mc*ENM construction steps. (A) Conformational transition from unbound to bound conformation of Arsenate reductase (ArsC). (B) Distance changes of contacts relative to their initial distance in the unbound conformation. (C) Contact transition matrix based on predefined extension threshold to distinguish observed breaking from maintained contacts (see eqn. 5.1 for details). Color coding of observed contact changes applies to subfigures C and D. (D) Removing the observed breaking contacts (highlighted in red) results in the elastic network model of maintained contacts (*mc*ENM). Observed forming contacts are only shown in the contact transition matrix (subfigure C).

[1] https://plato.stanford.edu/entries/simplicity/

Strategy II - Removing Breaking Contacts and Adding Forming Contacts

Above we argued that adding observed forming contacts will most likely counteract increased mobility in local parts of the network gained by removing breaking contacts. However, there may be cases, where adding forming contacts to the elastic network in combination with removing breaking ones more accurately captures the intrinsic motions of proteins. To test this hypothesis, we build the elastic network of maintained and forming contacts (*mfc*ENM).

5.2.3 Protein Data Set

To identify contact changes related to function-related protein motions we obtained a set of 90 conformational pairs from the Protein Structural Change DataBase (PSCDB) (Amemiya et al., 2011, 2012). Each pair captures start and end of a functional transition, which is classified by motion type, such as collective domain motions or localized motions involved in ligand binding. We also utilize these conformational pairs to evaluate the performance of the ENM variants w.r.t. to different types of function-related protein motions. To do so, we predict intrinsic motions based on the start conformation and validate their match with the observed conformational displacement. Detailed information on the data set and the motion classification can be found in section 4.1.

5.2.4 Evaluation of Elastic Network Models

We assess the performance of elastic network models by evaluating their biological accuracy and the dimensionality of their essential deformation space. The employed measures are introduced in detail in section 4.2. For convenience, we briefly summarize them here.

The biological accuracy of ENMs indicates how much we can trust the predicted ENM-motions to use them as guidance for subsequent applications, such as conformational exploration or ensemble generation for protein docking. We use **mode overlap** (Marques and Sanejouand, 1995, Tama and Sanejouand, 2001) and **cumulative mode overlap** (Yang et al., 2007) of the first ten low-frequency modes to measure the alignment between predicted motion directions and actual conformational displacement. We calculate the **Pearson correlation coefficient** to assess the similarity between predicted and actual residue fluctuation profiles. We measure how much structural variance can be explained by individual or a subset of low-frequency modes by the **(cumulative) fraction of variance**. Finally, we quantify the amount of residues involved in the protein's motion by the **degree of collectivity** (Brüschweiler, 1995) of a mode.

The dimensionality of the essential deformation space of ENMs influences the computational costs required to search the space spanned by a subset of the most dominant low-frequency modes. Hence, we analyze the number of modes that are required to capture 70%, 80%, and 90% of the actual conformational displacement. We also report rank, collectivity, and fraction of variance of the mode with maximum overlap among the first ten low-frequency modes together with its overlap.

5.3 Implementation

In the following we describe the implementation details of the two ENM variants proposed and evaluated in this chapter, *mc*ENM and *mfc*ENM. This includes their parametrization and information about the software, which we used to implement them. The two ENM variants, *mc*ENM and *mfc*ENM, are based on the widely used distance-cutoff based anisotropic network model introduced in 3.3.3. Thus, it naturally marks the lower bound on ENM performance and we will refer to it simply as ENM or *baseline* ENM for the rest of this thesis.

A detailed description of the parametrization of the baseline ENM is presented in 4.3.1. We find that a cutoff distance of 10Å yields the best performance of the baseline ENM in terms of capturing the functional transitions in our data set while maintaining network stability by enforcing the four-neighbor-connectedness criterion.

5.3.1 Parametrization of *mc*ENM and *mfc*ENM

Before adjusting the contact networks of *mc*ENM and *mfc*ENM they are essentially the same as the baseline ENM at cutoff 10Å including the stability modifications. To distinguish breaking from maintained contacts, we use an empirically defined extension threshold of 9% of the initial contact distance that maximizes the median accuracy improvement of *mc*ENM for our dataset (Table 5.1).

Table 5.1: Performance of *mc*ENM at different extension thresholds e_c used to distinguish breaking from maintained contacts. Contacts that extend their distance by less than e_c percent of their initial distance are considered maintained, otherwise they are labeled as breaking. *mc*ENM is based on the best performing $\text{ANM}_{minDeg4}$ at cutoff value 10Å (see Table S10). The performance is measured by the cumulative mode overlaps of the first ten low-frequency modes evaluated on the whole data set (90 proteins). Extension thresholds between 5% and 9% reach similar performance, with slightly better median at threshold 9%. Removing too many breaking contacts as for threshold 5% may lead to instable networks. Hence, we chose extension threshold 9% to build *mc*ENM in this study.

e_c (%)	5[a]	7	9	11	13	15	17	19	21	23	25
median	0.819	0.819	0.820	0.815	0.807	0.801	0.795	0.790	0.777	0.778	0.779
mean	0.804	0.800	0.799	0.793	0.787	0.781	0.774	0.769	0.767	0.763	0.762

[a] 1 case with more than the six trivial zero eigenvalues

However, *mc*ENM and *mfc*ENM became instable in several cases after removing observed breaking contacts based on the above defined extension threshold. This was due to the removal of contacts with less than four amino acids sequence separation. Hence, we had to tighten the stability criterion of the baseline ENM that required at least four neighbors in contact for each residue, which must not necessarily be the closest four along the sequence. To maintain network stability for *mc*ENM and *mfc*ENM, we remove breaking contacts only if their residues are at least four sequence positions apart. The criterion for six zero eigenvalues is maintained after the

removal of breaking contacts in both, *mc*ENM and *mfc*ENM, for all proteins (Table F in the supplementary file S2 of our paper (Putz and Brock, 2017)).

5.3.2 Used Software

All evaluated ENM variants in this thesis have been implemented and analyzed using the open-source Python framework ProDy (Bakan et al., 2011) in version 1.8.2, which provides various tools and methods to analyze protein structural dynamics. To produce the figures, tables, and plots presented in this chapter, we used ProDy (Bakan et al., 2011), Matplotlib (Hunter, 2007), Seaborn (mwa), Pandas (McKinney et al., 2010), IPython (Pérez and Granger, 2007), Jupyter (Kluyver et al., 2016), and Pymol (Schrödinger, LLC, 2015).

5.4 Results and Discussion

This section provides the biological grounding of our approach. By assuming "perfect" knowledge we examine whether ENMs indeed are capable of explaining *localized*, functional transitions of proteins. We propose that in order to do so ENMs have to account for dynamic changes in the initial contact topology of a protein, which are required to capture these localized movements.

We start by explaining the experimental setup to *observe* these dynamic contact changes and how our refined ENM variants are evaluated. We structure the remainder of this section into three parts: (i) identification of breaking contacts as relevant contact change, (ii) the effect of the refined network on ENM accuracy, in particular w.r.t. different motion types, and (iii) the relation between breaking contact occurrence and accuracy improvement considering motion type, structural fold, and functional class of the studied proteins.

First, we analyze if our refined ENM variants based on *observed* contact changes improve the match between predicted and actual motions at all. We will see that only the absence of observed breaking contacts yields a better match of *mc*ENM-predicted motions to actual motions than the baseline ENM, whereas also adding forming contacts results in a drastically reduced prediction accuracy that is far below the baseline ENM.

After identifying breaking contacts as the relevant contact change we continue with an extensive evaluation of *mc*ENM compared to the baseline ENM on the data set of 90 proteins performing different motion types. We start by examining if removing observed breaking contacts helps to capture localized functional transitions. Next, we analyze and discuss how this affects the dimensionality of the essential deformation space needed to explain these motions. This has implications for other methods and applications that use ENM predictions as guidance, such as molecular dynamics, or protein docking.

Finally, we analyze the occurrence of observed breaking contacts depending on motion type, structural fold, and function class of the proteins in our data set and how this relates to the achieved accuracy improvement by *mc*ENM.

5.4.1 Experimental Setup

We obtain observed contact changes by examining the differences in the contact maps of start and end conformation capturing a functional transition of a protein. We chose to use conformational pairs determined by high-resolution X-ray crystallography as basis for our approach. This is in contrast to the growing work favoring MD simulations to optimize and benchmark ENMs (see recent reviews (Fuglebakk et al., 2015, López-Blanco and Chacón, 2016) and citations therein). Clearly, two conformations capture only part of the structural variability of conformational ensembles. However, the chosen conformations represent the end points of a functional transition having largest structural difference among known conformations of a protein family (Amemiya et al., 2011). This increases the chances that the associated function-related structural changes are not only relevant but also appear in their coarse-grained contact topology, which may be more difficult to identify in structural ensembles obtained by MD simulations or Nuclear Magnetic Resonance. Both have a limited view on actual structural variance due to inaccuracies in sampling or measurement and still have restrictions on the protein's size. Further, the captured structural differences may be too small to effectively change the simplified contact topology. Also, energy barriers may prevent MD simulations from accessing certain conformational states. Such an effect is commonly associated with the induced-fit mechanism, where the presence of the binding partner triggers the required conformational change for successful binding (Stein et al., 2011).

For all tested ENM variants we only analyze ENMs based on unbound (start) protein conformations, because they generally capture more of the functional transition than the more compact bound (end) conformation (Tama and Sanejouand, 2001, Yang et al., 2007, Frappier and Najmanovich, 2014). However, preliminary results indicate that *mc*ENM is also well suited to explain the backwards movements starting from the more compact bound conformation. Further analysis is left out for future research.

Section 4.2 introduces the measures used in this thesis to assess biological accuracy, dimensionality of the essential deformation space, and agreement with essential dynamics of conformational ensembles.

5.4.2 Observed Breaking Contacts Matter

We observe three different types of dynamic contacts, namely breaking, forming, and maintained ones. Hence, the first step is to identify which of these contact changes is actually relevant to improve prediction accuracy of ENMs.

Breaking contacts seem to be weaker than maintained ones because they loose contact during a functional transition. Modeling them as strong as other contacts (uniform springs) thus inhibits actually accessible movements. The simplest approach to release these artificial/erroneous constraints is to remove breaking contacts from the initial contact network, yielding *mc*ENM, the elastic network model of maintained contacts.

Forming contacts, in contrast, establish towards the end of a conformational change due to the more compact fold of the bound conformation. Hence, they further constrain the initial contact network. Even if we incorporate them into the less constrained network of maintained contacts to

build *mfc*ENM, they rather inhibit required movements than enable them. Therefore, we expect improvement in prediction accuracy for *mc*ENM, but not for *mfc*ENM.

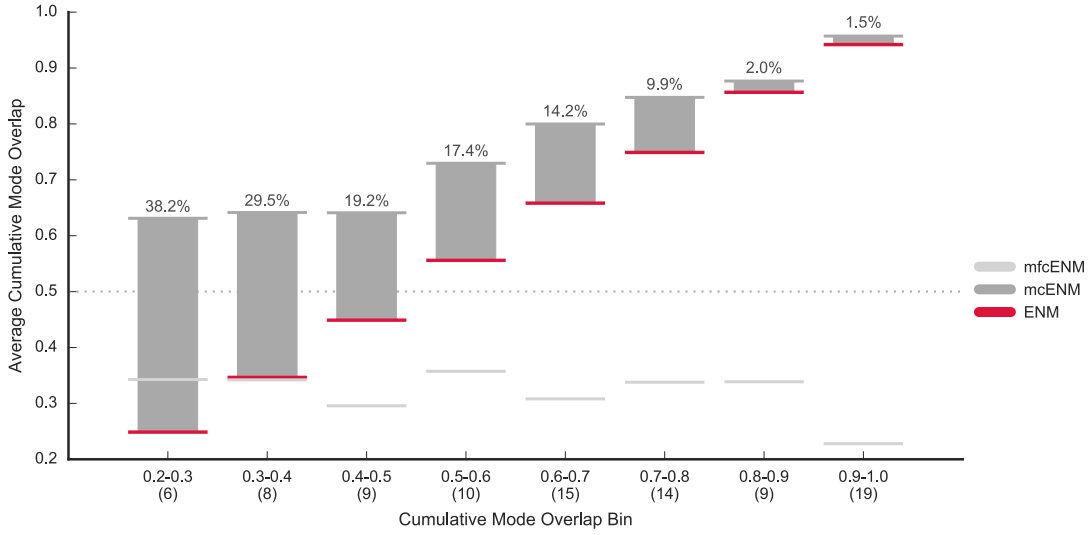

Figure 5.4: Accuracy of *mc*ENM and *mfc*ENM compared to ENM on full data set (90 proteins). Accuracy is measured by the cumulative mode overlap of the first ten low-frequency normal modes ($CO(10)$). Proteins are binned based on the cumulative mode overlap reached by ENM (#proteins per bin is given in brackets). The horizontal lines mark the average accuracy per bin (absolute improvement of *mc*ENM over ENM given by numbers above each bin). *mc*ENM consistently improves over ENM being particularly effective for proteins poorly captured by ENM (indicated by the gray dotted line). In contrast, *mfc*ENM performs much worse than ENM. Figure source: Putz and Brock (2017).

In Fig 5.4, we compare the accuracy of *mc*ENM and *mfc*ENM with the baseline ENM in terms of their cumulative mode overlap. Detailed results for every protein are given in Table C in supplementary file S2 of our paper (Putz and Brock, 2017). *mc*ENM consistently improves over ENM, whereas *mfc*ENM drops far below the baseline in almost all cases. This proofs our initial hypothesis that added forming contacts artificially stiffen the network, thereby preventing the ENM from capturing the functional transitions. Thus, we identify removing observed breaking contacts as the relevant contact change that will improve ENM accuracy and exclude *mfc*ENM from the rest of the evaluation.

*mc*ENM is particularly effective for proteins that are most difficult to capture with ENM (indicated by a cumulative mode overlap smaller than 0.6). For proteins in these four leftmost bins, *mc*ENM gains between 7.0% up to 58.7% improvement in accuracy. As expected, *mc*ENM improves less in accuracy for proteins, whose functional transitions are well captured by ENM. Furthermore, *mc*ENM substantially increases the number of proteins reaching 60% coverage of the functional transition with only ten lowest-frequency normal modes (*mc*ENM: 92% of proteins, ENM: 63%).

Table 5.2 shows that the improvement of *mc*ENM over ENM is consistent over all evaluated metrics. For detailed results see Table C in supplementary file S2 of our paper (Putz and Brock, 2017). *mc*ENM more accurately captures the functional transition not only in terms of

motion directions (overlap, structural variance), but also w.r.t. motion amplitudes (correlations between fluctuation profiles and temperature factors, where *mc*ENM is on par with ENM). Further, *mc*ENM reaches higher overlap and better agreement in structural variance for the best-overlapping mode, which is shifted towards the lower frequency spectrum of modes (rank). Hence, this mode becomes more dominant, which is desired for the most relevant mode. *mc*ENM also largely reduces the amount of modes required to explain a certain percentage of cumulative mode overlap. Only when considering the degree of collectivity, i.e. how many residues are involved in the movement, *mc*ENM reaches lower values than ENM. We will investigate this further when analyzing the effect of the refined *mc*ENM-network on the dimensionality of the essential deformation space below (see 5.4.4).

Table 5.2: Evaluated similarity measures for ENM and *mc*ENM. Table source: Putz and Brock (2017).

	ENM (median/mean)	*mc*ENM (median/mean)
Cumul. Mode Overlap (10)	0.69/0.66	0.82/0.80
Cumul. Fraction of Variance (10)	0.35/0.38	0.57/0.59
CorrCoeff Fluctuations - Displacements (10)	0.52/0.50	0.81/0.78
CorrCoeff Temperature Factors - Betas (10)	0.40/0.40	0.41/0.40
Max Overlap	0.47/0.50	0.60/0.62
Rank (Max Overlap Mode)	1.00/11.08	0.00/1.93
Degree of Collectivity (Max Overlap Mode)	0.38/0.39	0.27/0.31
Fraction of Variance (Max Overlap Mode)	0.05/0.08	0.12/0.20
#Modes Cumul. Mode Overlap (70%)	11.00/34.51	3.00/6.92
#Modes Cumul. Mode Overlap (80%)	35.00/79.07	7.50/19.41
#Modes Cumul. Mode Overlap (90%)	164.50/200.60	39.00/75.56

For several measures we consider only the subset of the first ten low-frequency modes indicated by (10) after the measure's name. Except for Rank and Collectivity of the best-overlapping mode higher values are better. A lower rank of the best overlapping mode with the observed displacement vector indicates that the most relevant motion captured by the elastic network is also more dominant. In terms of Degree of Collectivity, we find that lower values indicate that less collective, localized functional transitions are better captured (see next paragraph for more details).

Our results indicate that observed breaking contacts actually matter in contrast to forming ones. Their absence improves ENM accuracy, and is most effective in capturing otherwise poorly explained function-related movements.

5.4.3 *mc*ENM Accurately Captures Localized Functional Transitions

Now we need to show that our strategy works in particular for proteins with localized, functional transitions. To validate this assumption we analyzed the performance of *mc*ENM w.r.t. the motion type of the proteins (see 4.1 for details on motion classification of our data set).

Fig 5.5 shows the distribution of cumulative mode overlap of *mc*ENM and ENM for proteins classified as local vs. domain movers. Both categories are further subdivided into ligand-coupled or independent motions. *mc*ENM consistently improves over ENM for the shown motion types. However, proteins with localized functional transitions benefit by far the most. Here, *mc*ENM captures both coupled (independent) transitions on average 21% (15%) more accurate than ENM. For the domain motions already well captured by ENM, *mc*ENM still improves between 4% and 7% on average.

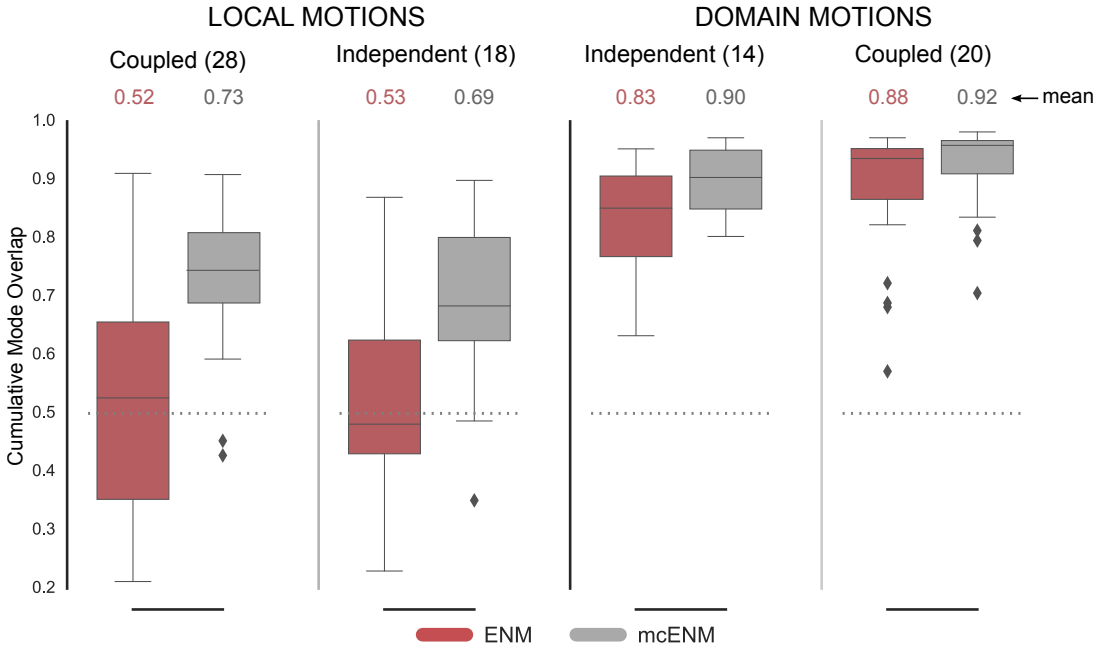

Figure 5.5: Accuracy of *mc*ENM compared to ENM measured by cumulative mode overlap, subset of local and domain motions (80 proteins). The distribution of cumulative mode overlap is evaluated for the first ten low-frequency normal modes ($CO(10)$). *mc*ENM consistently improves over ENM in each motion category. *mc*ENM is particularly effective for proteins with localized functional transitions yielding an improvement between 15% and 21% for independent and coupled local motions. Figure adapted from: Putz and Brock (2017).

*mc*ENM substantially improves over ENM also in terms of other metrics, such as the structural variance captured by the lowest frequency modes and the similarity between predicted and observed fluctuation profiles (Fig 5.6 (A,B)). Again, local movers benefit the most. Correlating predicted and experimentally observed temperature factors yields comparable performance of *mc*ENM and ENM (Fig 5.6 (C)). To better capture experimental B-factors ENMs require larger distance cutoffs (>16 Å) thereby increasing structural stiffness and collectivity of motion (Kondrashov

et al., 2007). This counteracts our goal to accurately model localized functional transitions with low degree of collectivity. Hence, this metric has little relevance in our context.

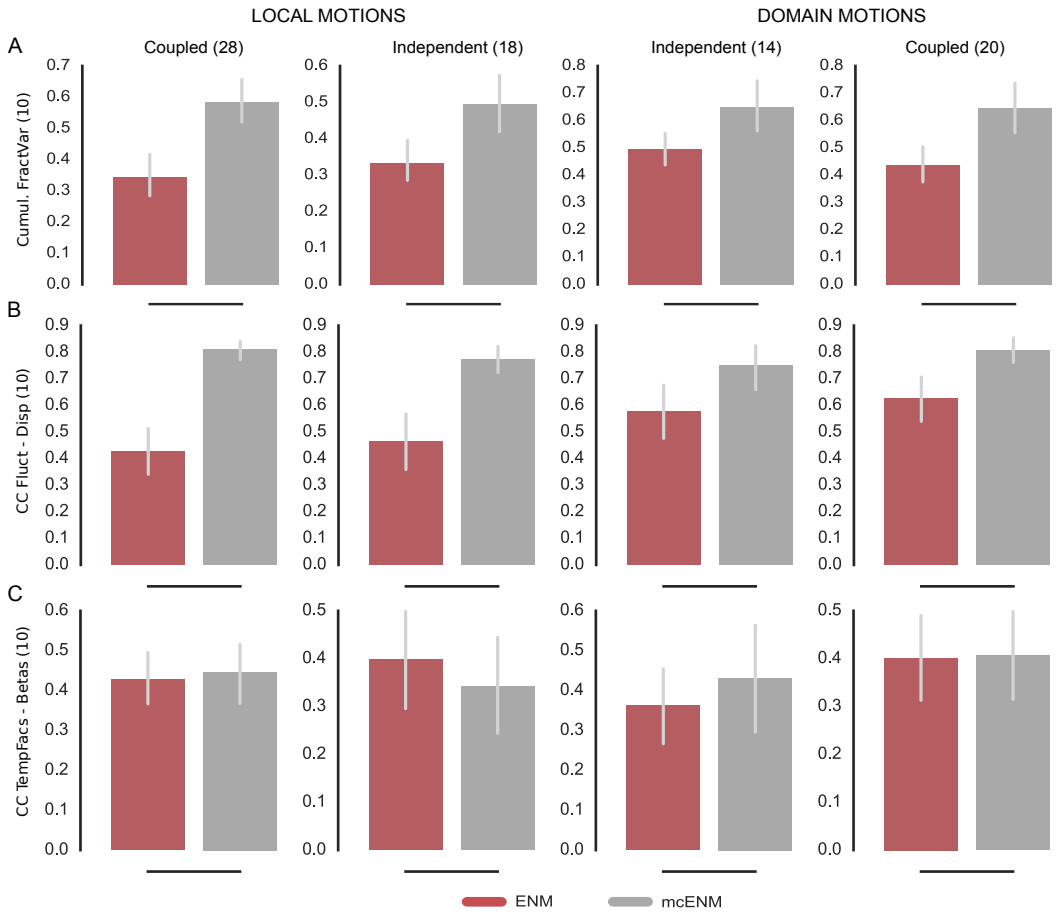

Figure 5.6: Accuracy of *mc*ENM compared to ENM using additional metrics on proteins grouped by motion type, subset of local and domain motions (80 proteins). (A) Cumulative Fraction of Variance (10 modes). (B) Correlation coefficient between predicted residue fluctuations and observed displacement magnitudes (10 modes). (C) Correlation coefficient between predicted Temperature factors and experimental Beta factors (10 modes). *mc*ENM consistently outperforms ENM considering the first two measures. The improvement is largest for proteins with localized functional transitions. Considering the similarity of temperature factor profiles *mc*ENM and ENM perform roughly the same. Figure source: Putz and Brock (2017).

Our results show that *mc*ENM, in fact, is able to capture localized, functional transitions while largely outperforming the distance-cutoff based ENM. Apart from comparing the agreement between motion directions and magnitude, we will now continue our analysis by evaluating the complexity of the resulting essential deformation space.

5.4.4 *mc*ENM Reduces Dimensionality of Essential Deformation Space

ENMs often guide more fine-grained exploration by narrowing down the search space (essential deformation space). The computational costs of searching this space increase with dimensionality (number of spanning modes). Hence, lower dimensional search spaces are desirable as long as they are accurate enough. As mentioned above, the common strategy to consider between 10-20 lowest-frequency modes works well in capturing highly collective functional transitions, but fails for localized functional transitions with low degree of collectivity. Here, the relevant modes (usually less than 10) are often spread among higher frequencies (Cavasotto et al., 2005). Consequently, a much larger number of ENM modes would need to be considered to capture them, which in turn yields a higher dimensional search space. In the following, we analyze how the absence of breaking contacts affects the relationship between desired accuracy and number of required modes. Again, we focus on local and domain motions.

Fig 5.7 depicts the median number of modes required to achieve a cumulative overlap of 70%, 80%, and 90%. *mc*ENM needs much less modes to be as accurate as ENM, thereby substantially reducing the dimensionality of the associated deformation space. For instance, to capture 80% of ligand-coupled local motions *mc*ENM requires a median of 22 modes, whereas ENM needs 95. Being less constrained, *mc*ENM favors otherwise high-energetic modes that seem to be relevant to capture the function-related movement. Hence, these modes "shift" towards lower frequencies. Consequently, *mc*ENM reaches higher accuracy with fewer, but more relevant low-frequency modes because their individual contribution to the overlap is higher.

This mode shifting is further supported by the large decrease in rank of the best-overlapping mode of *mc*ENM compared to ENM as shown in Fig 5.8.

*mc*ENM not only captures the direction of this mode much more accurate, but also increases its contribution to the structural variance to a large extent. Interestingly, the degree of collectivity of the best-overlapping mode for proteins with localized functional transitions is much smaller when being analyzed by *mc*ENM instead of ENM. Hence, the best-overlapping *mc*ENM-mode must be more relevant for the local transition given its higher overlap and larger variance. A similar "shifting" effect was observed by other groups when analyzing molecular dynamics trajectories (Orellana et al., 2010, Rueda et al., 2007a) or conformational ensembles (Yang et al., 2008) by essential dynamics (ED). Fewer ED-modes captured more of the structural variance (i.e. relative amplitude of deformations) than ENM-modes. Hence, the absence of observed breaking contacts makes relevant deformations accessible.

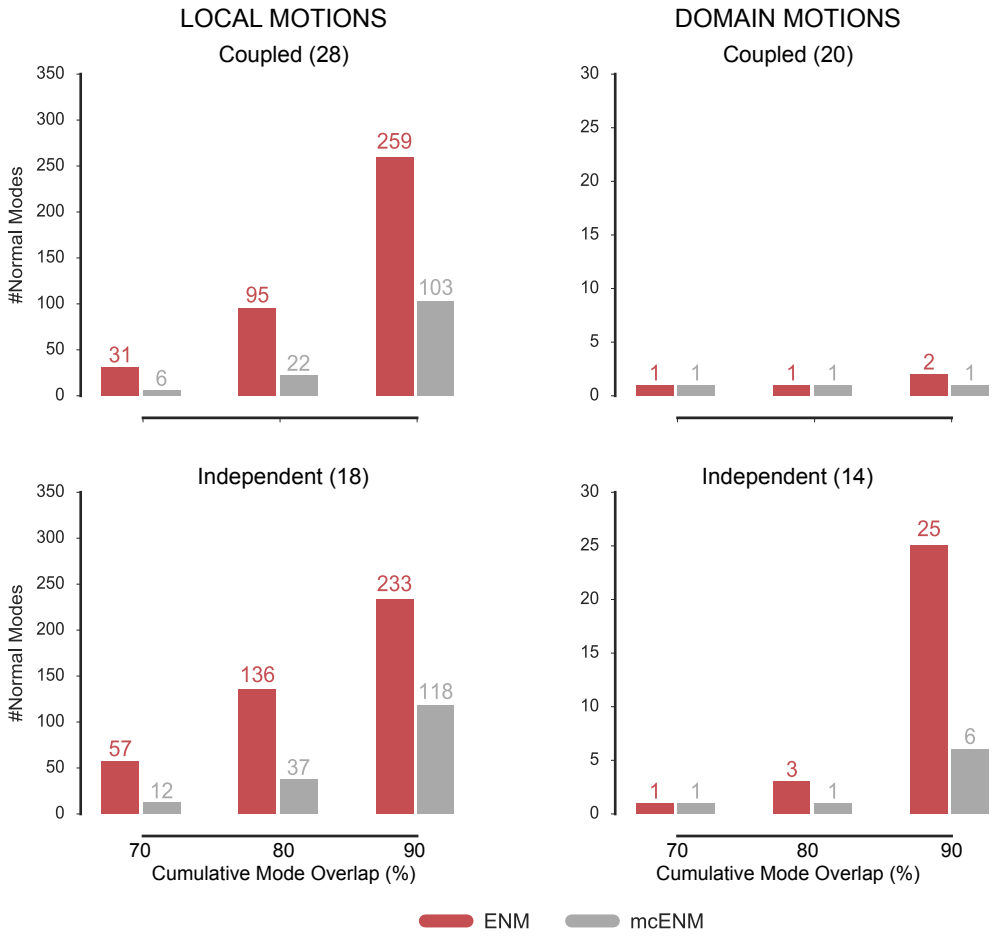

Figure 5.7: Dimensionality of deformation subspaces of *mc*ENM compared to ENM on subset of local and domain motions (80 proteins). The panels show the median number of normal modes (spanning the deformation subspace) required to explain between 70% and 90% of the functional transition (measured in cumulative mode overlap (%)). *mc*ENM consistently requires fewer modes to capture the same amount of conformational change as ENM. Figure adapted from: Putz and Brock (2017).

5.4.5 Relationship Between Observed Breaking Contact Occurrence and Effect on ENM Accuracy

To the best of our knowledge, *mc*ENM is the first approach to examine the effect of observed breaking contacts on ENM accuracy. Above we showed that they are a novel source of information, which helps to capture localized, functional transitions with ENMs. To further explore their importance, we now analyze how their occurrence and impact are linked depending on motion type, structural fold, and functional class of the proteins.

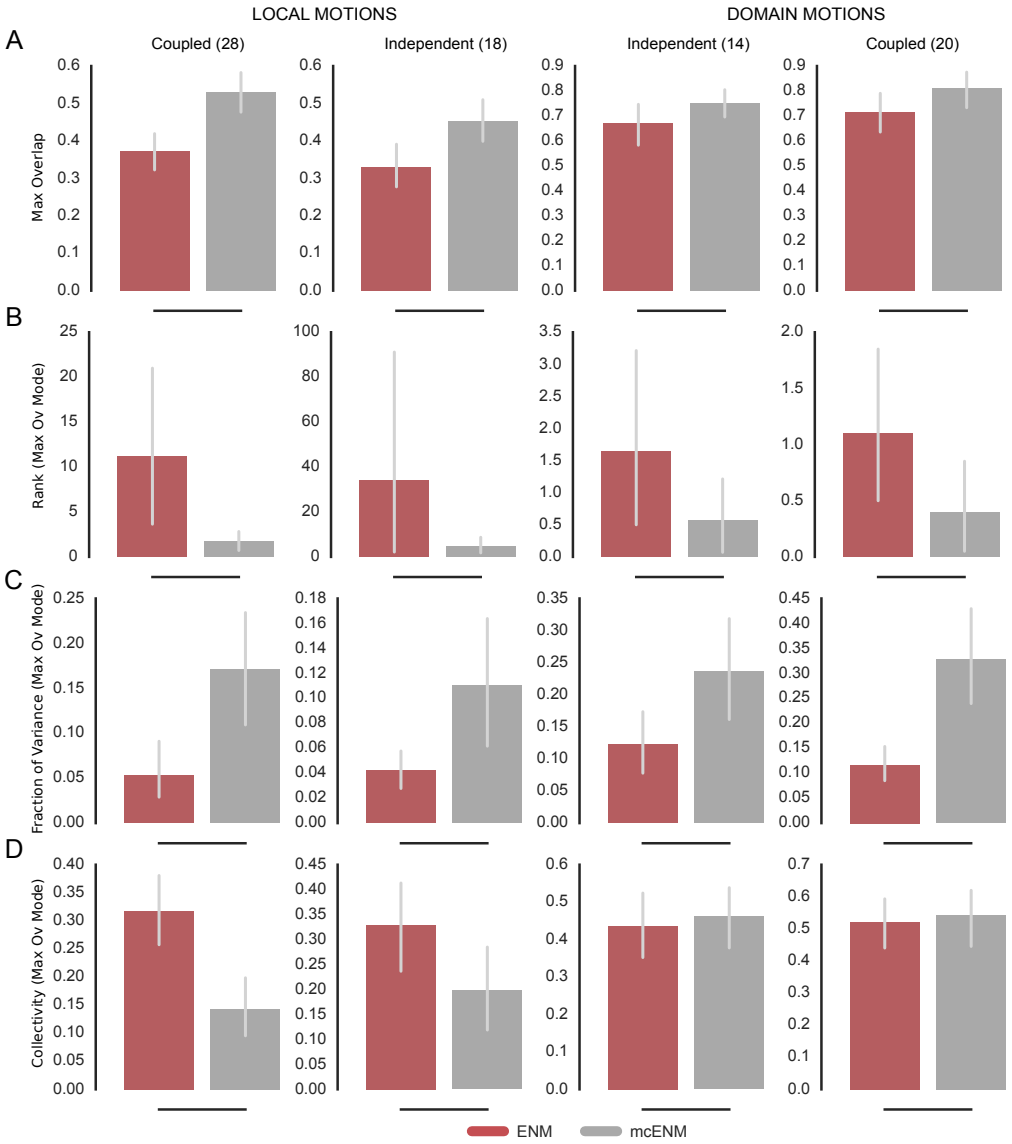

Figure 5.8: Accuracy of *mc*ENM w.r.t. maximum mode overlap related measures compared to baseline ENM on proteins grouped by motion type, subset of local and domain motions (80 proteins). (A) Maximum mode overlap of all modes. (B) Rank of best-overlapping mode. (C) Fraction of variance explained by best-overlapping mode. (D) Degree of collectivity of best-overlapping mode. Figure source: Putz and Brock (2017).

Dependence on Motion Type

Fig 5.9 relates average accuracy improvement of *mc*ENM over ENM to average amount of removed breaking contacts by considering the motion type of the studied proteins. For this analysis we include the categories *burying ligand* and *other types of motions* despite their few samples. The reason is that we want to examine whether removing breaking contacts has any effect on their predicted motions at all (see 4.1 for details on the motion classification of our data set).

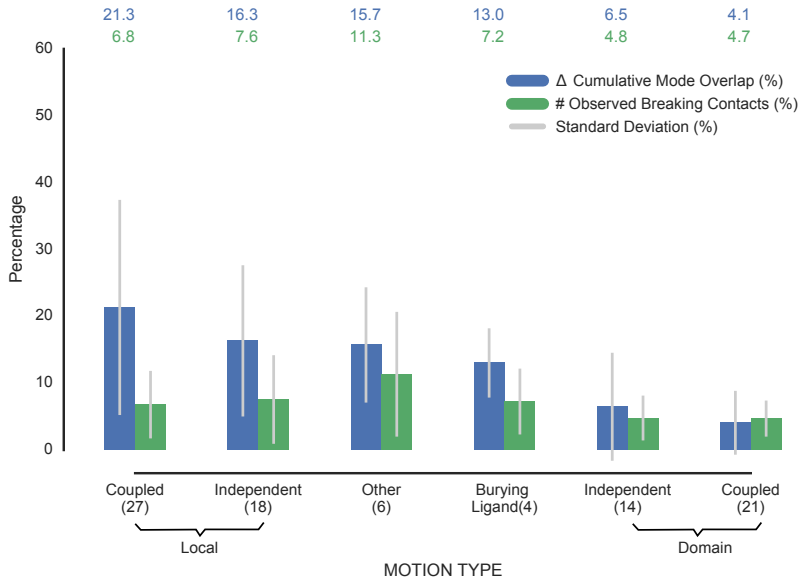

Figure 5.9: Accuracy improvement of *mc*ENM over ENM in relation to percent of observed breaking contacts on full data set (90 proteins) grouped by motion types. The blue bars depict the absolute accuracy improvement of *mc*ENM over ENM averaged over each group, whereas the green bars show the average amount of removed breaking contacts. The accuracy improvement is calculated by the difference between cumulative mode overlap of the first ten low-frequency modes of *mc*ENM and ENM. Figure adapted from: Putz and Brock (2017).

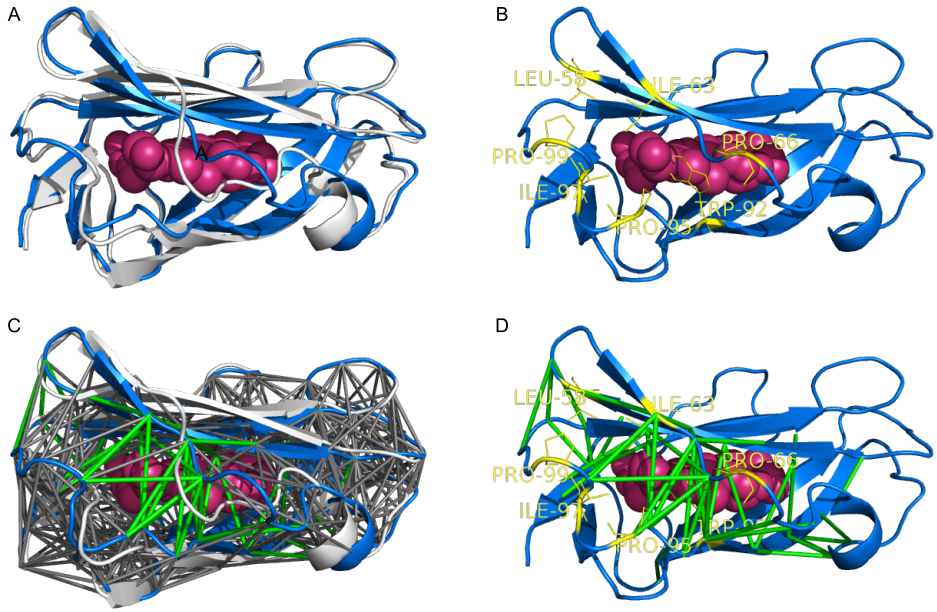

Figure 5.10: Observed breaking contacts in contact topology of house dust mite allergen Der f. (A) Unbound and bound conformation (PDB_IDs: 2f08D, 1xwvB) colored blue and white, respectively. Ligand is shown as magenta spheres. (B) The ligand is proposed to enter the binding site via a narrow tunnel opening at the left, where the contact density is lower. The residues assumed to form the tunnel opening are highlighted in yellow (Johannessen et al., 2005). (C) Observed breaking (green) and maintained (gray) contacts networks. (D) Observed breaking contacts locate around the proposed tunnel opening. Figure source: Putz and Brock (2017).

We note that *mc*ENM improves much more in accuracy for local motions than for domain motions given the amount of removed breaking contacts. Hence, individual breaking contacts seem to encode more information about motion when they belong to local movers than to domain movers.

Surprisingly, also proteins that bury a ligand in their end conformation benefit from removing observed breaking contacts. This is particularly interesting as these proteins show only subtle differences between unbound and bound conformation (< 1 Å RMSD). Hence, to facilitate a ligand's move into the binding the protein must transiently open the entry to the binding pocket. But there is no guarantee that the structural differences at the pocket entry are large enough to produce breaking contacts in the coarse-grained contact topology of the protein. We visually inspected the location of observed breaking contacts for all four proteins in this category whether it matches the entry of the binding site. Only one protein, depicted in Fig 5.10, shows observed breaking contacts at the assumed entry of the binding site (Johannessen et al., 2005). Ignoring those contacts improves *mc*ENM-accuracy by 16.3% compared to ENM for this protein.

Obviously, the occurrence of observed breaking contacts around a binding site entry strongly depends on the chosen parametrization of distance cutoff and allowed extension of contact distance, which is used to identify breaking contacts. Nonetheless, *mc*ENM improves prediction accuracy in all four cases. This indicates that observed breaking contacts should be ignored to predict the motions of proteins also in this category, despite their small structural differences.

The remaining proteins performing other types of motions improve on average about the same as independent local movers. The higher average amount of required breaking contacts is caused by one protein (PDB_ID: 1uorA). It is relatively large with 580 residues and an *all alpha* fold, where the 30% observed breaking contacts are distributed between α-helices over the whole structure. Without this protein, the average amount of observed breaking contacts is in the range of independent local movers. Hence, proteins with other types of motions benefit about the same as independent local movers from the refined network of *mc*ENM.

Dependence on Structural Fold

The motions of proteins are largely governed by their structural fold. Therefore, we analyzed if certain folds promote contact changes more than others and how this relates to the achieved accuracy improvement of *mc*ENM. Fig 5.11 summarizes the results for the proteins in our data set averaged over their SCOP classes, which we obtained from the Structural Classification of Proteins (SCOP) database (Murzin et al., 1995, Fox et al., 2014). The individual correlation between the observed breaking contact occurrence and improved accuracy for each fold class is shown in Fig 5.12. Each protein is colored whether it performs local, domain, or another type of motion.

Remarkably, the only membrane protein in our data set improves by almost 60% in cumulative overlap despite a relatively small amount of breaking contacts. We will analyze and discuss this protein in detail in case study 6.4.7 in the following chapter.

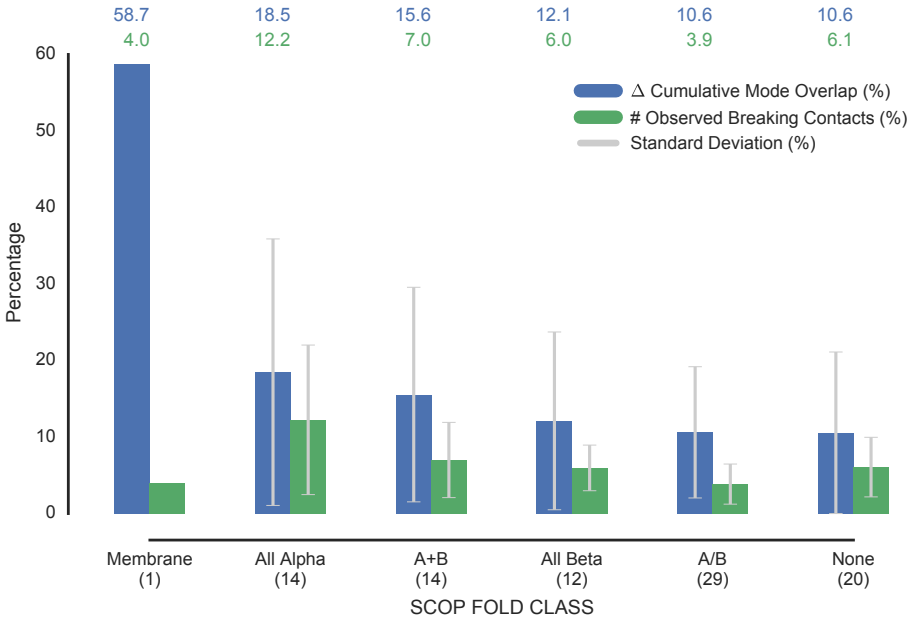

Figure 5.11: Accuracy improvement of *mc*ENM over ENM in relation to percent of observed breaking contacts on full data set (90 proteins) grouped by SCOP fold class. The blue bars depict the absolute accuracy improvement of *mc*ENM over ENM averaged over each group, whereas the green bars show the average amount of removed breaking contacts. The accuracy improvement is calculated by the difference between cumulative mode overlap of the first ten low-frequency modes of *mc*ENM and ENM. Figure adapted from: Putz and Brock (2017).

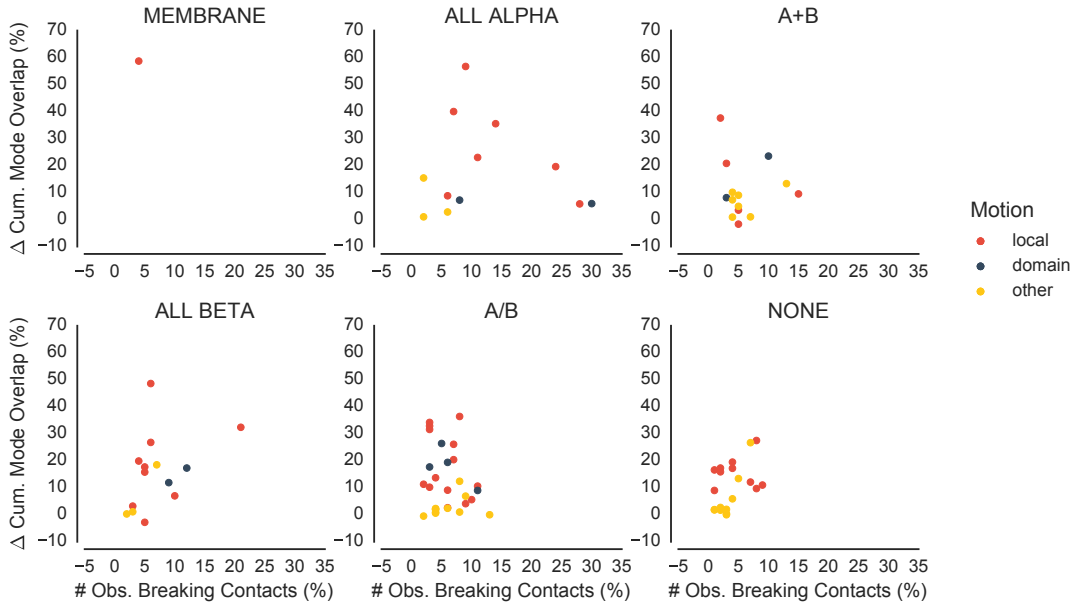

Figure 5.12: Correlation between occurrence of observed breaking contacts and achieved accuracy improvement of *mc*ENM over the baseline ENM by considering structural fold and motion type of the studied proteins. The accuracy improvement is calculated as the difference between cumulative mode overlap of the first ten low-frequency modes of *mc*ENM and ENM. The motion classification is simplified to local (coupled and independent), domain (coupled and independent), and other motions (burying ligand and other types of motion).

We also find that *all alpha* proteins benefit more from removing breaking contacts than the remaining classes, although individual breaking contacts seem to have less impact than for the other classes. This may be due to the relatively high structural flexibility of *all alpha* proteins. Thus, breaking contacts may even occur in regions not necessarily related to the functional transition, making them less relevant. This is supported by the observation that a higher number of observed breaking contact does not always lead to greater improvement of *mc*ENM-accuracy (Fig 5.12).

In contrast, folds strongly stabilized by a central beta sheet or beta barrel as in the *all beta, a/b*, or *a+b* classes appear to be more robust towards changes in the contact topology. This may lead to fewer, but larger clusters of breaking contacts, whose absence have larger impact on the accuracy of *mc*ENM. Further investigation of these hypotheses is beyond the scope of this thesis and is left our for future research.

Dependence on Functional Class

We also evaluated to what extent the accuracy improvement of *mc*ENM depend on the functional class of the proteins. Two third of the proteins in our data set are enzymes belonging to six classes, which we obtained from the PSCDB (Amemiya et al., 2011): Hydrolases (26), Transferases (15), Oxidoreductases (9), Lyase (4), Isomerase (2), Ligase (1). For the remaining proteins (33), which are no enzymes, we do not further distinguish between their functional classes due to their functional diversity[1].

Fig 5.14 shows the relation between averaged occurrence of breaking contacts and their averaged impact on ENM accuracy. The correlation between the observed breaking contact occurrence and improved accuracy per function class is depicted in Fig 5.14. Each protein is colored according to whether it performs local, domain, or another type of motion.

Oxidoreductases catalyze the pass of electrons from one molecule to another one (May, 1999). They have various applications, for instance in medical diagnostics, quality control, or in the production of agrochemicals, pharmaceuticals, cosmetics, or biofuels (Martinez et al., 2017, Xu, 2005). Despite a relatively small amount of breaking contacts, they show the largest improvement of accuracy of *mc*ENM because they mostly perform local motions.

Hydrolases build the largest enzyme group in our data set. In the presence of water they bind smaller molecules at their surface to break chemical bonds, which often requires the enzymes to deform locally (Koike et al., 2014). Because they function without the need of a cofactor and bind to various substrates they dominate research and applications in the field of biotransformations (Faber, 2018). Surprisingly, only half of the hydrolases in our data set are classified as local movers, whereas for the other half domain motions seem to dominate the movements. Nonetheless, hydrolases with localized movements are captured much more accurately when observed breaking contacts are removed in *mc*ENM.

[1]Based on functional annotations retrieved for each unbound conformation from the Protein Database (PDB) they fall into classes, such as "Binding Proteins", "Transport Proteins", "Enzyme Inhibitors", "Hormones", "Toxins", "Viral Proteins", "Signaling Proteins" or "Membrane Proteins".

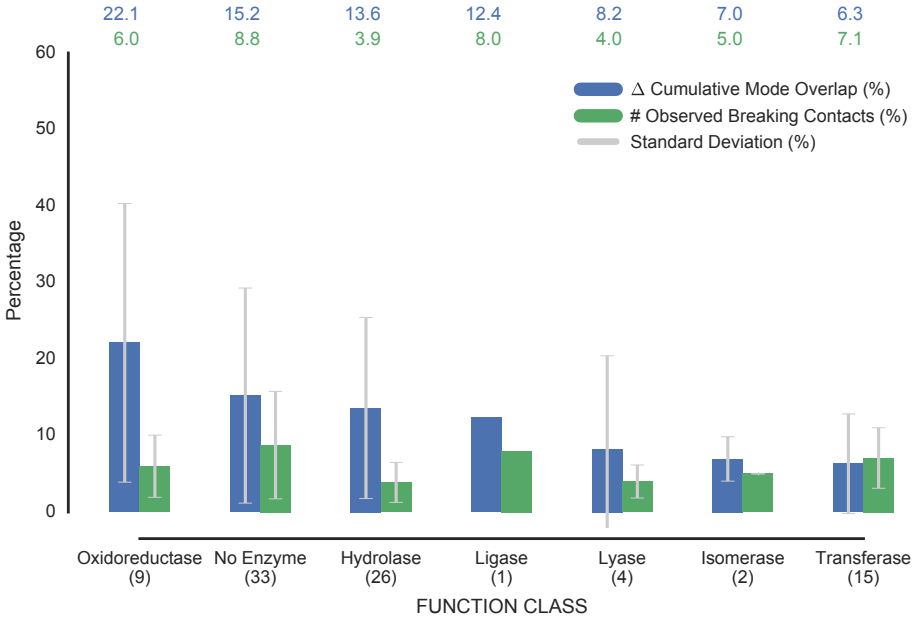

Figure 5.13: Accuracy improvement of *mc*ENM over ENM in relation to percent of observed breaking contacts on full data set (90 proteins) w.r.t the functional class of the proteins. The blue bars depict the absolute accuracy improvement of *mc*ENM over ENM averaged over each group, whereas the green bars show the average amount of removed breaking contacts. The accuracy improvement is calculated as the difference between cumulative mode overlap of the first ten low-frequency modes of *mc*ENM and ENM.

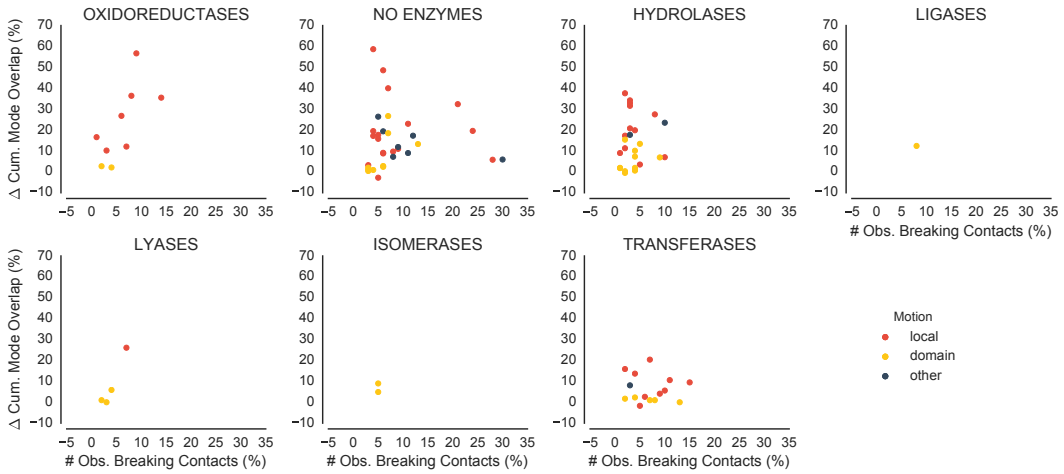

Figure 5.14: Correlation between occurrence of observed breaking contacts and achieved accuracy improvement of *mc*ENM over the baseline ENM by considering function class and motion type of all studies proteins (90 proteins). The accuracy improvement is calculated as the difference between cumulative mode overlap of the first ten low-frequency modes of *mc*ENM and ENM. The motion classification is simplified to local (coupled and independent), domain (coupled and independent), and other motions (burying ligand and other types of motion).

The second largest group in our data set are transferases. They bind two molecules at the same time to enable the transfer of chemical groups between them. Therefore, their binding site

is usually rather deep and accessible from two sides. About two third of the transferases in our data set are associated with local conformational changes. However, transferases are suspected to undergo more complicated conformational transitions that also could involve a hierarchy of deformation steps (Koike et al., 2014). This may be one of the reasons why a relatively large amount of breaking contacts causes rather small improvement in accuracy for MC-ENM. In addition, it may be relevant in which order the contacts "break" during the transaction steps.

The remaining three enzyme classes (lyases (4), isomerases (2), ligase (1)) are not really representative as they only consist of few proteins.

Another 33 proteins are no enzymes but represent different functional annotations and motion types. They vary widely in terms of accuracy improvement of *mc*ENM and occurrence of breaking contacts. As for the other classes *mc*ENM is more effective for local movers among these proteins.

5.5 Conclusion

5.5.1 Summary

Our results in this chapter demonstrate that ENMs are indeed capable of predicting local and uncorrelated functional motions **if** they are allowed by the underlying network. *mc*ENM, the elastic network model based on maintained contacts, naturally meets this condition because it ignores springs associated with contacts that are observed to break during functional movements. As a consequence, previously over-constrained local areas in the network get the required mobility to capture the localized functional motions.

In addition, we have seen that *mc*ENM also overcomes the problem of not knowing how many and which modes should be considered to explain localized function-related motions (Cavasotto et al., 2005). In most cases, the relevant *mc*ENM-modes are included in the dominant low-frequency modes. This reduces computational costs for subsequent applications of ENMs, such as normal-mode-guided conformational exploration using molecular dynamics, because a lower dimensional space has to be sampled.

By analyzing the relation between breaking contact occurrence and *mc*ENM-accuracy improvement we found that *mc*ENM is effective for a wide range of motion types and structural folds of proteins. When considering the enzyme class of the studied proteins, we found that hydrolases and oxidoreductases benefit the most from the refined network of *mc*ENM, in particular, if they belong to the category of local movers. We cannot draw such a clear picture for transferases. While some transferases with local motions gain some improvement, others do not. Same is true for transferases that perform domain motions. This may be overcome by a time-resolved ENM, which considers in which order the observed breaking contacts need to break in order to facilitate the complex hierarchical motions found in transferases (Koike et al., 2014).

Our results also show that improving the general applicability of ENMs is possible without the need to change model resolution and potential function. *mc*ENM builds upon the same contact network between C_α-atoms representing a protein's structural connectivity and relies on uniform spring stiffness as the baseline ENM. Hence, the key to expand the range of motions that can be

captured by ENMs is to leverage information about dynamic changes in their underlying simplified model, i.e. contacts that break throughout a protein's motion. Of course, by optimizing spring stiffness as proposed by several other approaches (Orellana et al., 2010, Lezon and Bahar, 2010, Kovacs et al., 2004, Hinsen et al., 2000) the performance of *mc*ENM may be further improved as indicated by preliminary results. However, if more than one conformation of a protein is known, *mc*ENM has demonstrated to be a valuable alternative to the widely used, distance-cutoff based ENMs.

5.5.2 Limitations

*mc*ENM has one important limitation. To identify the erroneous restrictions, i. e. observed breaking contacts, it requires a known end conformation, which is usually not available. But as we will see in the next chapter it is possible to *predict* breaking contacts instead of observing them. To do so we leverage information about their local embedding into a protein's contact network and its physicochemical and topological properties.

6

Elastic Network Model of Learned Maintained Contacts (*lmc*ENM)

6.1 Introduction

The main hypothesis of this thesis is that in order to enable elastic network models to predict localized, function-related protein motions they need to account for dynamic changes in the contact topology of proteins. In the previous chapter we have seen that removing the springs associated with *observed* breaking contacts enables ENMs to capture localized functional transitions with low degree of collectivity. But to identify the breaking contacts we need to know start **and** end conformation representing a protein's functional transition, which are usually not available. Hence, to employ ENMs in the standard case when only a single protein conformation is known, we must be able to *predict* these breaking contacts given that single conformation.

In this chapter we present the core contribution of this thesis: the ability to *predict* the dynamic behavior of contacts (whether they break or are maintained). To do so, we leverage information encoded in the physicochemical characteristics of local parts of the protein structure. These parts largely maintain their structural shape but move with respect to each other controlled by the strength of their physicochemical interactions. As a consequence, some contacts break in the underlying contact topology depending on the type of movement. We predict these breaking contacts using machine learning based on a graph-based representation of their structural context.

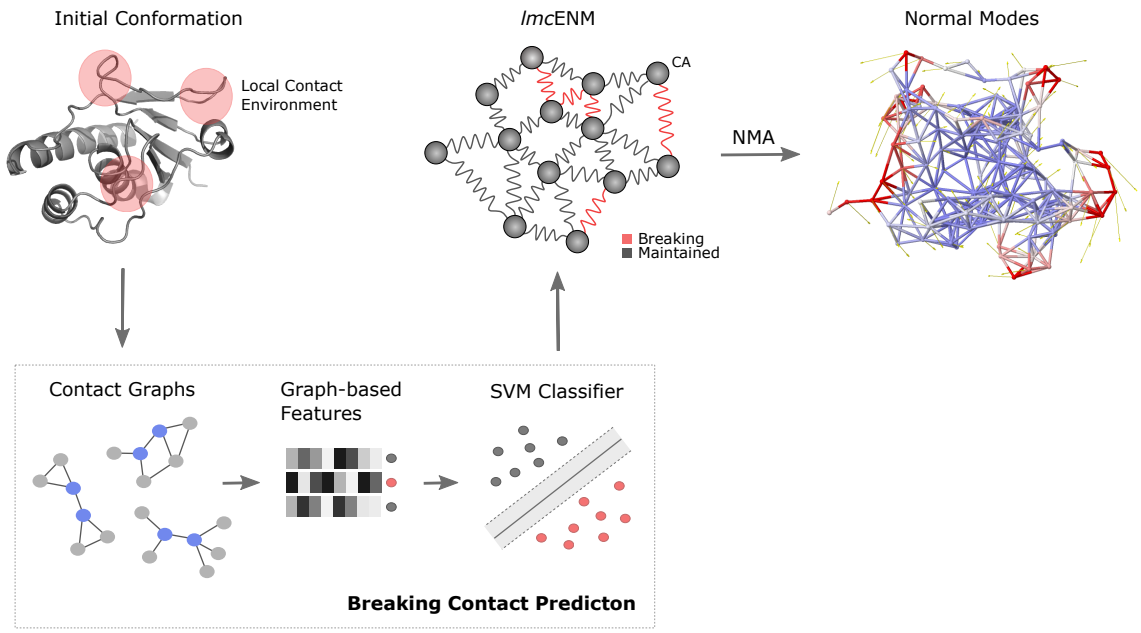

Figure 6.1: Flowchart overview of *lmc*ENM construction and analysis. The first step to build *lmc*ENM is to predict the breaking contacts by: (i) constructing the contact graphs from the local contact environments, (ii) deriving features from these graphs that characterize the physicochemical properties of a contact's structural context, and (iii) classify contacts as breaking or maintained. Next, we remove the springs associated with the predicted breaking contacts from the initial elastic network model of the start conformation, resulting in the network of learned maintained contacts (*lmc*ENM). Last, we analyze *mc*ENM using normal mode analysis (NMA) to predict the intrinsic deformations (normal modes) of the protein (image generated with ANM 2.0 web server (Eyal et al., 2015)). The illustration of the breaking contact prediction is inspired by Schneider and Brock (2014).

Based on the predicted contact changes we propose a novel **elastic network model of learned maintained contacts (*lmc*ENM, see Fig 6.1)**. It learns how to adjust its network without increasing the complexity of the original ENM approach, i.e. resolution and potential function remain the same. Instead, *lmc*ENM encodes information about dynamic changes of the contact topology by ignoring the *predicted* breaking contacts in the same way that *mc*ENM does with *observed* breaking contacts.

In contrast to *mc*ENM presented in the previous chapter, *lmc*ENM does not need to know a target conformation to identify possibly erroneous constraints blocking localized functional movements. Thus, it is applicable to the standard case of predicting the structure-encoded motions of proteins where only a single conformation is known.

We evaluate the performance of *lmc*ENM on a set of 90 conformational pairs of proteins that perform different types of function-related motions, including highly collective domain motions as well as localized, uncorrelated movements. We will see that, in contrast to the widely used distance-cutoff ENM and three reference ENM variants, *lmc*ENM is capable of predicting functional transitions that are localized. We also show that *lmc*ENM is particularly suited to explain functional transitions that involve the binding of a ligand. These localized movements

remain largely underestimated by the other ENM variants, whereas the adjusted network of *lmc*ENM makes them accessible. This alleviates a major shortcoming of ENMs.

We will also see that *lmc*ENM mitigates the problem of not knowing how many and which modes to consider to effectively capture localized functional transitions. *lmc*ENM requires fewer modes than the other ENM variants because the relevant *lmc*ENM-modes become more dominant, i.e. already reside in the low-frequency range, due to the removed predicted breaking contacts. Furthermore, we evaluate *lmc*ENM in detail by presenting case studies of three biologically interesting proteins selected from our data set, the outer membrane transporter FecA, the fatty acids oxidizing enzyme Arachidonate 15-Lipoxygenase, and SopA–a salmonella effector protein. Finally, we analyze and discuss which features contribute the most to correctly differentiate breaking from maintained contacts.

6.1.1 Contributions

In this chapter, we make the following contributions:

Conceptual Contributions

- We propose to *predict* the dynamic behavior of contacts, i.e. whether they break or are maintained when the protein moves, by leveraging information from the protein's structure. This information is encoded in the physicochemical properties of local parts of the structure, which capture the relative motions between these parts as well as their deformability. Accounting for the associated dynamic changes in the contact topology of proteins expands the range of motion types that can be explained by elastic network models (ENMs).

Technical Contributions

- We present a novel machine-learning based classifier that *predicts* breaking contacts based on a graph-based encoding of their structural context. We developed a set of features that characterize the physicochemical, structural, and topological properties of this local contact environment and its embedding into the overall protein structure. The classifier outputs the likeliness of a contact to break given a protein's initial contact topology.

- We introduce a novel elastic network model of learned maintained contacts (*lmc*ENM) that accounts for these *predicted* dynamic changes in the contact topology of proteins. *lmc*ENM adjusts its initial network by removing the springs corresponding to the *predicted* breaking contacts. While preserving the simplicity of the original ENM, *lmc*ENM is better suited to capture localized, functional transitions of proteins. It can be applied to proteins, where only a single conformation is known.

Empirical Contributions

- We show that the prediction of breaking contacts by leveraging information about their structural context is possible and accurate enough to substantially improve ENM accuracy. Without the predicted breaking contacts *lmc*ENM requires only a small subset of low-frequency modes to explain localized functional transitions that would otherwise be barely accessible for the network. *lmc*ENM expands the range of motions that can be captured by ENMs, thereby increasing their practical relevance.
- We present evidence that the dynamic behavior of contacts, and thus protein motion, most likely results from the interplay of a broader set of features characterizing the properties of their structural context. Our approach provides a unified and extensible framework for exploring, using, and correlating additional features to advance our understanding of protein motion.

6.1.2 Outline

The rest of this chapter is organized as follows:

- **Section 6.2 Methods** introduces our approach to leverage information about the dynamic behavior of contacts, how this information is used to predict breaking contacts, and how *lmc*ENM, the network of learned maintained contacts, is built based on the predicted contacts. It also introduces how we assess the performance of our classifier and the evaluated ENMs as well as the used protein data set.
- **Section 6.3 Implementation** describes the implementation details of our algorithm. This includes the parametrization of *lmc*ENM and the reference ENMs used for validation of our approach, the external software used to generate features and to analyze and present our results, and the training procedure of the SVM-classifier.
- **Section 6.4 Results and Discussion** describes the experimental setup and analyzes and discusses the experimental results. It starts with an evaluation of the classifier performance followed by a thorough assessment of the performance of *lmc*ENM compared to the reference ENM variants.
- **Section 6.5 Relevance of Features to Predict Breaking Contacts** presents an analysis, which features contribute the most to accurately predicting breaking contacts, and discusses the results.
- **Section 6.6 Conclusion** summarizes the findings of this chapter, discusses its limitations, and establishes *predicted* breaking contacts as novel source of information to improve prediction accuracy and general applicability of ENMs.

6.2 Methods

In this section we introduce the algorithm to build the elastic network model of learned maintained contacts (*lmc*ENM). First, we explain how we leverage information about the dynamic behavior of contacts, which is captured in the physicochemical characteristics of their structural context. We introduce how the contact environment is encoded in a graph-based representation that allows us to derive features characterizing its properties. These features then serve as input to train the SVM-classifier. Its implementation details are described in 6.3.4.

Second, we introduce the three stages to construct the network of *lmc*ENM. We start by explaining how breaking contacts are predicted using the trained SVM-model. Next, we present the strategies to select a subset of highest scoring predicted breaking contacts as removal candidates. The last step is to remove the springs associated with the selected removal candidates from the initial network of *lmc*ENM.

6.2.1 Leveraging Information About the Dynamic Behavior of Contacts

In the previous chapter we found that the absence of *observed* breaking contacts enables ENMs to capture localized functional transitions of proteins. In contrast, adding *observed* forming contacts hurts ENM accuracy (see 5.2.2). This leaves us two types of contacts that need to be predicted: breaking and maintained contacts. Hence, we face a binary classification problem, which we tackle by using a support vector machine (SVM) (see 6.3.4).

In the following we introduce the graph-based representation of the local contact environment as well as its embedding into the overall protein structure. Next, we give an overview of the features that we developed to characterize the physicochemical, structural, and topological properties of this structural context of a contact. They serve as input to train the SVM in order to distinguish breaking from maintained contacts.

Contact Neighborhood Graph

To encode the local contact environment, we use the *immediate neighborhood graph* (IN_{ij}) of a contact (Schneider and Brock, 2014) that is depicted in Fig. 6.2. The graph consists of residues (nodes) and edges (between residues in contact). It captures the direct environment of the contact between residues i and j. This includes residues i and j and their first-shell neighbors, i.e. residues in direct contact.

In the neighborhood graph, nodes and edges are labeled. Node labels carry characteristics of individual residues, whereas edge labels characterize individual contacts. A detailed description of the labels can be found in the appendix (Tables A.1 and A.2). The labels are referred to as features and will be used to train and test the SVM classifier.

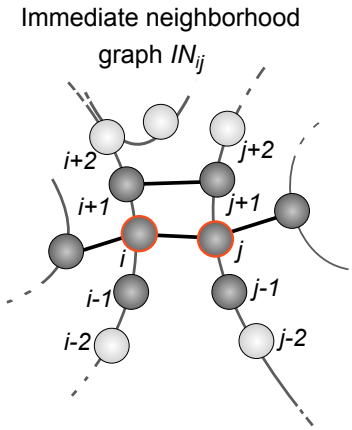

Figure 6.2: Definition of Immediate Neighborhood Graph of a Residue-Residue Contact. Nodes represent residues that are linked by an edge if the are *in contact*, i.e. they are within a pre-defined distance cutoff. The immediate neighborhood graph includes residues i and j (red stroke) and their direct neighboring residues in contact colored in dark gray. Figure adapted from: Schneider and Brock (2014).

Secondary Structure Graph

To characterize the embedding of a contact within the global structural topology of a protein, we define the *secondary structure element (SSE) graph*. Fig 6.3 shows an example for such a graph attributed by a small set of structural and physicochemical properties.

The nodes correspond to secondary structure elements, i.e. α-helices, β-strands, or loops with a minimum length of three residues. Two nodes are connected by an edge if the corresponding SSEs are in contact, i.e. they share at least one residue-residue contact. Node labels capture the characteristics of individual SSEs, whereas edge labels characterize the interface between two SSEs in contact. Based on the SSE-graph we distinguish between *intra*-SSE and *inter-SSE* contacts.

The structural context of a contact can now be characterized by a set of features derived from its neighborhood graph and the secondary structure graph of the protein. The features capture its physicochemical, structural, and topological properties and serve as input for our SVM-classifier. An overview of the designed features is given below.

Overview of Features

We use a set of 75 features to characterize the properties of the local contact environment and its embedding into the overall structural topology. We concatenate these features into a feature vector that is then used to train and test our classifier. Continuous features are encoded as single, real-valued inputs, whereas categorical features are specified as a set of binary values. In total, the feature vector is 170-dimensional.

In addition to novel features specifically tailored to our problem, we add or adapt some features used by Schneider and Brock (2014). Several of the latter features have been introduced by Cheng and Baldi (2007) and Li et al. (2012) (see 6.3.3 for details). The features are grouped into seven categories (see Table 6.1): pairwise, graph topology, graph spectrum, single node, node label statistics, edge label statistics, and whole protein features. Appendix A contains a detailed description of the individual features in each category and reports. Re-used or extended features

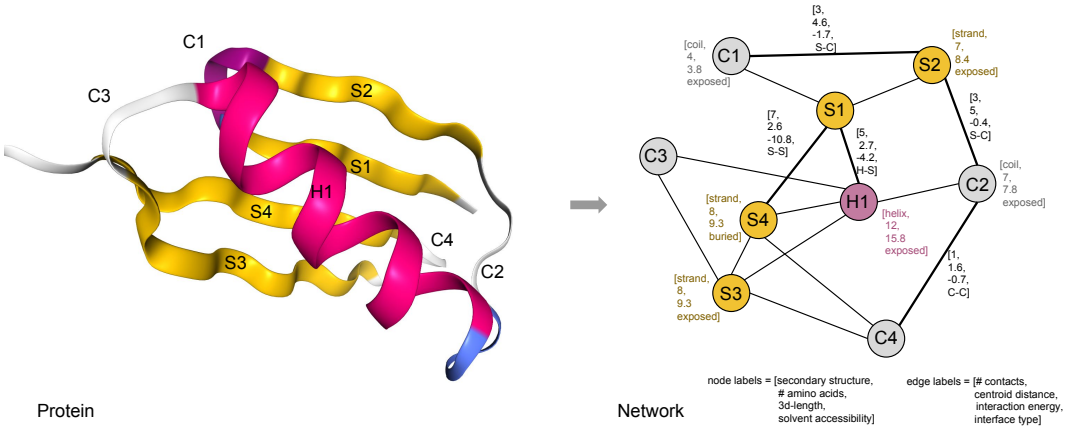

Figure 6.3: Example of a secondary structure element (SSE) graph of a protein structure. Nodes represent secondary structure elements. Edges link secondary structure elements that are in contact, i.e. at least one residue of element A is within a pre-defined distance with a residue in element B. Nodes and edges are characterized by structural and physicochemical properties.

are marked accordingly (Tables A.3-A.9). We now introduce each feature category with some examples.

Pairwise features encode properties of an individual contact. As contacts seldom change their distance in isolation, many of the pairwise features are defined on their associated secondary structure element(s) (SSEs). The features capture for instance SSE types, sequential and three-dimensional distance between the SSEs, hydrogen bonding between SSEs, closeness to empty pockets, or closeness to binding site.

To capture topological characteristics of the local contact environment we re-use the graph-topology, graph spectrum, and single node features from Schneider and Brock (2014), part of which have been introduced by Li et al. (2012). For instance, one feature captures that contacts embedded into a highly constrained neighborhood are less likely to change than contacts in sparsely connected local contact networks. The average number of neighbors of each node in the local contact environment can be characterized by the average degree centrality.

Node and edge label statistics encode properties of the contact's neighborhood not captured by its topology. For example, local contact networks with high symmetry coverage, i.e. where most residues belong to a symmetric segment of the protein, are likely to maintain their connectivity even when the protein moves. This can be measured by the normalized number of symmetric residues.

We further collect properties of the whole protein, such as the connectivity class based on the total number of contacts, and the distribution of secondary structure types. These features now serve as input to train and test our classifier, whose implementation details will be described below (see 6.3.4).

We will now explain how we adjust the network of the original ENM approach based on the predicted breaking contacts to build our novel elastic network model of learned maintained contacts, *lmc*ENM.

Table 6.1: Overview of used features. Table lists the features used by our classifier to predict function-related contact changes in the contact topology of proteins. Added or adapted features from Schneider et al. (Schneider and Brock, 2014) are marked. If all features in one category are added from Schneider and Brock (2014) the category is marked instead of the individual features. Appendix A describes individual features and their implementation in detail. Table adapted from: (Putz and Brock, 2017).

Group	Feature examples	Number of inputs
Pairwise	Secondary structure element (SSE) type;[1] sequence separation between SSEs, distance between SSE centroids, symmetry coverage of SSE(s), intra-SSE contact and intra-SSE topology descriptors, inter-SSE contact and inter-SSE interface descriptors, contact residues part of terminal SSEs, hydrogen bonding;[2] side-chain contact, contact with pocket and number of atom contacts with pocket, pocket descriptors (polarity, hydrophobicity, volume, drug score), contained in symmetric segments, distance to symmetry plane, 4-bin contact depth and residue depth difference classes, mutual information[1]	63
Graph topology[1]	Number of nodes, number of edges, average degree centrality, average closeness centrality, average betweenness centrality, graph radius, graph diameter, average eccentricity, number of end points, average clustering coefficient	10
Graph spectrum[1]	Largest two eigenvalues, number of different eigenvalues, sum of eigenvalues, energy of adjacency matrix	5
Single node[1]	Degree, closeness centrality, betweenness centrality, sequence separation from N/C-terminus, sequence conservation and sequence neighborhood conservation for i and j	12
Node label statistics	Chemical type of residues;[1] secondary structure descriptors;[1] solvent accessibility;[1] hydrogen bonding;[2] average free solvation energy;[1] 4-bin solvation energy distribution;[1] entropy of labels, neighborhood impurity degree;[2] average distance from centroid;[1] symmetry coverage, average degree of symmetry, average residue depth, 5-bin distribution of residue depth, average lower/upper half-sphere exposure, sequence conservation;[1] sequence neighborhood conservation[1]	57
Edge label statistics	Link impurity;[2] 5-bin mutual information distribution;[1] cumulative mutual information[1]	13
Whole protein	Secondary structure composition;[1] 5-bin connectivity class based on number of contacts, symmetry coverage	10
		170

[1] Added from Schneider et al. (Schneider and Brock, 2014).
[2] Adapted from Schneider et al. (Schneider and Brock, 2014).

6.2.2 Construction of *lmc*ENM

*lmc*ENM consists of three stages described in detail below:

1. scoring of each contact with its probability to break based on the initial contact topology,
2. selecting removal candidates from the top scoring breaking contacts, and
3. removing the selected candidates from the initial contact topology to build *lmc*ENM.

Prediction of Breaking Contacts

The SVM classifier scores all contacts in the initial contact topology of a protein between residues at least four sequence positions apart. We chose to exclude shorter-range contacts from the classification for two reasons: First, removing them caused network instabilities in several cases for both ENMs proposed in this thesis, *mc*ENM and *lmc*ENM (see 6.3.1). Second, the improvement in ENM accuracy was negligible for those networks that remained stable without them. The classifier outputs a rank-ordered list of contacts by decreasing confidence score, which indicates their likeliness to break.

Selection of Removal Candidates

To adjust the network of *lmc*ENM, we now seek a function to select how many top scoring predicted breaking contacts should be removed. We expect that the amount of breaking contacts depends on the collectivity of the function-related movement. We have seen in the previous chapter (see 5.4.5) that local, uncorrelated motions require on average more initial contacts to break than large-scale, collective motions. However, in most cases the nature of the functional transition is unknown a priori and furthermore depends on various properties of the protein. This makes it difficult to find such a function.

Therefore, we tested three simple strategies to select the subset of predicted breaking contacts to be removed, which are based on a:

constant cutoff that removes the top n predicted breaking contacts. It is based on the rationale that the amount of breaking contacts is limited in number and variance among different proteins. Given our observation that removing breaking contacts is highly relevant to capture localized, functional transitions (see 5.4.5), we would expect that they concentrate on particular regions of the protein. The spatial extent of these regions should be rather small and not necessarily depend on the protein's size.

relative cutoff that removes the top n percent of predicted breaking contacts. Opposite to the previous strategy we now assume that the amount of breaking contacts is affected by the total number of contacts of the protein.

score-dependent cutoff that removes all predicted breaking contacts with probability larger than a predefined cutoff score. Here, we assume that prediction accuracy of the classifier is comparable among different proteins.

We evaluated each strategy on a predefined set of cutoff values to empirically determine the most effective strategy and associated cutoff value in our setting (see 6.3.1).

Building the Network of Learned Maintained Contacts

Finally, we adjust the initial contact topology of a protein by removing the selected predicted breaking contacts. This results in the elastic network of learned maintained contacts, which we call *lmc*ENM.

In the following we introduce how we evaluate the performance of the SVM classifier and which evaluation measures we employ to assess the performance of *lmc*ENM compared to the reference ENMs.

6.2.3 Protein Data Set

To train and test our classifier we use the data set of 90 conformational protein pairs categorized by motion type that we introduced in detail in section 4.1. We also utilize these conformational pairs to evaluate the performance of the ENM variants w.r.t. to different types of function-related protein motions, such as collective domain motions or localized motions involved in ligand binding.

6.2.4 Evaluation of Binary Classifiers

A binary classifier is trained on examples of two classes, a positive and a negative one. For an unknown sample it predicts one of two possible outcomes, i.e. whether it belongs to the positive class (1) or not (0). Given a test data set we can evaluate a binary classifier using the so-called **confusion matrix** or **contingency table** (see Tab. 6.2). It compares the predicted output of a binary classifier with the actual value of the samples, which are known (gold standard) or may result from another reference classifier.

Table 6.2: Confusion matrix for a binary classification problem.

		Actual	
		+	-
Predicted	1	True Positive (TP)[1]	False Positive (FP)[2]
	0	False Negative (FN)[3]	True Negative (TN)[4]

[1] positive sample correctly classified as positive
[2] negative sample misclassified as positive
[3] negative sample correctly classified as negative
[4] positive sample misclassified as negative

The total size of the data set is given by the summing up all true and false positive and negative predictions. Based on counting the number of predicted samples in each of the above categories, different measures can be specified to evaluate the performance of a binary classifier.

In this thesis, we aim to differentiate breaking contacts (positive class) from maintained contacts (negative class). We define the gold standard by comparing the contact topologies of

start and end conformation of the proteins in our data set (see 5.3.1). Hence, the actual values of the predicted contacts are given by the *observed* breaking and maintained contacts. Our SVM classifier estimates the probability of each contact to break and outputs these predictions as a rank-ordered list. From this list we choose a subset of top-scoring predicted breaking contacts to build our novel elastic network model of learned maintained contacts, *lmc*ENM. Therefore, we evaluate the performance of our SVM classifier w.r.t. the chosen subset of predictions using the following common measures:

Precision measures the probability of a correct positive prediction, defined as Prec=TP/(TP+FP).

Coverage specifies the percentage of true positives captured by a subset of predictions. It is given by Cov = TP_{frac}/TP_{all}, where TP_{frac} refers to the number of true positive predictions in a selected fraction (subset) of all predicted contacts, and TP_{all} is the total number of true positives for a protein. This is a useful measure because we consider only a top scoring subset of all predicted breaking contacts (see 6.2.2). A coverage of 1 indicates that all true positives of a protein are contained in the selected fraction.

Area Under the Receiver Operator Characteristic (AUROC) (Fawcett, 2006) estimates the probability that a positive sample reaches higher score than a negative one if both are chosen randomly. The ROC curve visualizes the trade-off between true positive rate (TPR=TP/(TP+FN)) and false positive rate (FPR=FP/(FP+TN)) at different thresholds. A predictor with AUROC of 1 is considered perfect, whereas it is random at a value of 0.5.

6.2.5 Evaluation of Elastic Network Models

We evaluated all tested ENM variants in this thesis using a variety of common measures that are introduced in detail in section 4.2. For convenience, we provide a short summary here.

We assess the performance of all evaluated ENMs in terms of their biological accuracy and the dimensionality of their essential deformation space. Both properties are relevant to effectively use low-frequency ENM-modes as guidance for subsequent applications, such as conformational exploration or ensemble generation for protein docking. Biological accuracy indicates how much we can trust the guidance of ENMs. The dimensionality of the essential deformation space determines the computational cost for search.

Assessing the Biological Accuracy of the ENMs We use the **mode overlap** (Marques and Sanejouand, 1995, Tama and Sanejouand, 2001) and **cumulative mode overlap** (Yang et al., 2007) of the first ten low-frequency modes to measure the alignment between predicted motion directions and actual conformational displacement. We employ the **Pearson correlation coefficient** to assess the similarity between predicted and actual residue fluctuation profiles. We evaluate how much structural variance can be explained by individual or a subset of low-frequency modes by the **(cumulative) fraction of variance**. Finally, we use the **degree of collectivity** (Brüschweiler, 1995) to quantify the amount of residues involved in the protein's motion.

Assessing the Dimensionality of the Essential Deformation Space We analyze the number of modes that are required to capture 70%, 80%, and 90% of the actual conformational displacement. For the mode with maximum overlap among the first ten low-frequency modes we report its overlap as well as its rank, collectivity, and fraction of variance.

Comparison against Essential Dynamics of Conformational Ensembles Finally, we evaluate the agreement between mobility of the first ten low-frequency modes and the actual structural variance of conformational ensembles determined by Essential Dynamics (ED). Again we use the **Pearson correlation coefficient** to compare the predicted and actual fluctuation profiles. In addition, we evaluate the similarity of deformation spaces spanned by the first ten-low frequency ENM-modes and the first ten principal components identified by ED using the **root mean square inner product (RMSIP)** (Amadei et al., 1999) and its extension, the **root weighted square inner product (RWSIP)** (Carnevale et al., 2007).

6.3 Implementation

We now describe the implementation details of our novel *lmc*ENM and the reference ENM variants evaluated in this chapter. This includes their parametrization and references to the software packages that we used for their implementation and in the context of the SVM classification as well as additional software used to analyze and present the results in this thesis.

We evaluate *lmc*ENM with respect of two boundaries: First, a lower bound defined by the original performance of the widely, used distance-cutoff based anisotropic network model introduced in 3.3.3. We will refer to it simply as ENM or *baseline* ENM for the rest of this thesis. Second, an upper bound marked by the theoretical maximum improvement reached by *mc*ENM that we introduced in the previous chapter. We call it *theoretical* maximum improvement because *mc*ENM relies on the knowledge of a usually not available target conformation to identify *observed* breaking contacts.

A detailed description of the parametrization of the baseline ENM is presented in 4.3.1. We find that a cutoff distance of 10Å yields the best performance of the baseline ENM in terms of capturing the functional transitions in our data set while maintaining network stability by enforcing the four-neighbor-connectedness criterion. *mc*ENM relies on the same distance cutoff but ignores *observed* breaking contacts. Due to network instabilities only contacts between residues at least four sequence positions apart are removed (see 5.3.1).

6.3.1 Parametrization of *lmc*ENM

To build the initial network of *lmc*ENM we use the distance-cutoff 10Å as the baseline ENM. This allows us to solely focus on the effect of the changed contact topology when evaluating our approach. Above we introduced three strategies to determine the amount of top scoring predicted contacts to be removed from *lmc*ENM using a: (i) constant cutoff, (ii) a relative cutoff, and (iii) a score-dependent cutoff.

For each strategy we evaluated how it affects *lmc*ENM accuracy along a range of cutoff values:

- **constant cutoff** every 5th value between $[5 - 50]$ and every 10th value between $[60 - 200]$, where each value denotes the number of highest scoring breaking contacts that are removed
- **relative cutoff** every value between $[1 - 20]$ and every 5th value between $[25 - 50]$, where each value refers to the percentage of highest scoring breaking contacts that are removed
- **score-dependent cutoff** every score value between $[0.1 - 1.0]$, where 1.0 denotes maximum confidence of the classifier

To facilitate a fair comparison we determine the best cutoff for each strategy as the one that maximizes the average over all proteins in our data set. We empirically find that the top $n = 60$ (constant cutoff), top $n = 16\%$ (relative), and SVM score > 0.4 (score-dependent) work best in our setting.

6.3.2 Parametrization of Reference ENMs

In addition, we evaluate *lmc*ENM with respect to three ENM variants exploiting different sources of information to refine connectivity and stiffness of the network:

- **HCA** - a cutoff-free model with distance-dependent spring constants (Hinsen et al., 2000)
- **OFC-ENM** - a model analyzing structural properties of NMR ensembles to optimize force constants for secondary structure elements (Lezon and Bahar, 2010)
- **edENM** - a hybrid model using a combination of bond-cutoff strategy in the local sequential neighborhood and distance-dependent force constants to model remote interactions (Orellana et al., 2010).

A detailed description of their potential functions and parametrization can be found in subsection 4.3 of the methods chapter.

6.3.3 Used Software

We use several software packages to generate the features used to discriminate breaking from maintained contacts given their structural context. For features introduced or inspired by others we list the relevant publications. In addition, we credit all software used to analyze and visualize our results in this chapter.

Generation of Features

We detect pockets and cavities in protein structures and analyze their properties, such as volume, size, and druggability score with FPocket (Guilloux et al., 2009). We identify symmetric parts and symmetry axes using SymD (Kim et al., 2010). Residue depth is computed with Biopython (Cock et al., 2009).

We use or extend several of the pairwise features introduced by Cheng and Baldi (2007), which have also been used by Schneider and Brock (2014) to predict residue contacts in the context of protein structure prediction. For completeness, we also list the involved software packages and relevant publications to generate the features.

Solvent accessibility and free solvation energies are calculated using POPS (Cavallo et al., 2003). Secondary structure types and hydrogen bonds are assigned based on STRIDE (Frishman and Argos, 1995). Features capturing sequence conservation are calculated as proposed by Fischer et al. (2008).

The Python library NetworkX (Hagberg et al., 2008) is used to generate the graphs and to extract topological and spectral graph features as well as label statistics, many of which have been introduced by Li et al. (2012).

Finally, to predict breaking contacts, we use the SVM library of scikit-learn (Pedregosa et al., 2011) that internally builds on LIBSVM (Chang and Lin, 2011).

Analysis and Visualization of Results

We implemented and analyzed all evaluated ENM variants using the open-source Python framework ProDy in version 1.8.2 (Bakan et al., 2011), which provides various tools and methods to analyze protein structural dynamics.

To produce the figures, tables, and plots presented in this chapter we used ProDy (Bakan et al., 2011), Matplotlib (Hunter, 2007), Seaborn (mwa), Pandas (McKinney et al., 2010), IPython (Pérez and Granger, 2007), Jupyter (Kluyver et al., 2016), and Pymol (Schrödinger, LLC, 2015).

6.3.4 SVM Learning

We train a support vector machine (SVM) to differentiate breaking from maintained contacts, given the features described above (see 6.2.1). This classifier builds upon an in-house contact prediction framework (Schneider and Brock, 2014). For a detailed introduction into support vector machines please refer to the background chapter (see 3.2.1). In the following, we describe the handling of class imbalance in our training data, the estimation of probabilities for the predicted breaking contacts, and the training and hyperparameter tuning of the SVM classifier.

Handling Imbalanced Data

The number of observed breaking contacts per protein is rather low in our data set (on average 4.5% of the total number of contacts in a protein). Thus our labeled input data to train the SVM is highly imbalanced because we have much more maintained examples (negative class) than breaking ones (positive class). A standard approach to tackle this problem is random undersampling of the majority class (He and Garcia, 2009).

We have empirically found that taking all observed breaking contacts as positive samples, while randomly picking three times as many maintained contacts as negatives maximizes the prediction accuracy in our data set. In addition, we adjust the cost (C) for misclassification by a class-dependent weighting factor ($w = \{\text{breaking} : 3, \text{maintained} : 1\}$), which penalizes the misclassification of breaking contacts more than that of maintained contacts. This increases the importance of correctly classifying positive samples (Pedregosa et al., 2011).

Estimating Probabilities

The SVM classifier yields a probability score for each sample to belong to the positive class, i.e. to be a breaking contact. This probability is estimated using Platt's scaling method (Wu et al., 2004). It performs logistic regression on binary classification scores of the SVM using additional cross-validation as implemented in scikit-learn (Pedregosa et al., 2011).

SVM Training, Kernels, and Tuning of Hyperparameters

We tested two different common kernels to train the SVM classifier on our data, the linear kernel and the Gaussian radial basis function (RBF) kernel using leave-one-out-cross-validation (LOOCV) on our data set. Both kernels yield similar performance with slight advances for the RBF kernel in our setting (see 6.5 for details on their performance). Thus, we chose the SVM with RBF-kernel to implement our classifier.

The generalization performance of SVMs depends on the choice of its hyperparameters. They have to be tuned in order to avoid overfitting. The Gaussian RBF-kernel has two hyperparameters, cost C and the kernel parameter γ, which controls the width of the Gaussians.

A standard approach is to tune these parameters w.r.t. to an optimization objective using grid search. We chose the precision (Prec=TP/(TP+FP))[1] of the $L/5$ contacts with highest SVM probability as optimization objective, where L refers to the length of the protein. We find that cost $C = 100$ and kernel width $\gamma = 0.00001$ determined in leave-one-out-cross validation perform best in our setting after evaluating cost values ranging between $[0.1, 1, 10, 100, 1000, 10000]$ and gamma values in $[0.01, 0.001, 0.0001, 0.00001, 0.000001]$ using grid search.

6.3.5 Experimental Setup

For all tested ENM variants we focus on analyzing the intrinsic motions based on the unbound (start) conformation of the proteins in our data set. It has been shown that unbound conformations generally capture more of the functional transition than the more compact bound (end) conformation (Tama and Sanejouand, 2001, Yang et al., 2007, Frappier and Najmanovich, 2014). However, our third case study (see 6.4.7) below indicates that *lmc*ENM is also well suited to explain the backwards movements starting from the more compact bound conformation. Further analysis is left out for future research.

We compare the predicted ENM-motions to the actual structural displacement between the two conformations of each protein. In addition, we compare the essential deformation space of the ENMs with essential dynamics of conformational ensembles.

Section 4.2 introduces the measures used in this thesis to assess biological accuracy, dimensionality of the essential deformation space, and agreement with essential dynamics of conformational ensembles.

6.4 Results and Discussion

*lmc*ENM is built in three steps: (i) we predict the most likely breaking contacts with our machine learning based classifier, (ii) we choose a highest scoring subset of contacts, and (iii) we remove them from the initial contact network of the unbound conformation.

Therefore, we structure the evaluation of *lmc*ENM as follows: First, we identify the best strategy to select an appropriate subset of top-scoring predicted breaking contacts. Given this subset of contacts we then evaluate the ability of our classifier to identify correct and, most importantly, relevant breaking contacts. Next, we assess the performance of *lmc*ENM w.r.t the baseline ENM, the theoretical upper bound reached by *mc*ENM, and three reference ENM variants. To do so, we compare the predicted mobility to the actual mobility captured by conformational pairs as well as conformational ensembles. Then, we present three detailed case studies selected from our data set. Last, we analyze which features contribute the most to a correct classification.

[1] TP stands for true positives and refers to predicted breaking contacts that have been observed, whereas FP denotes false positive predictions, where the predicted breaking contacts are actually observed maintained contacts.

Please note that starting with the previous chapter 5 we use "baseline ENM" or simply "ENM" interchangeably to refer to the original, distance-cutoff based ANM on which our approach is based on.

6.4.1 Choosing How Many Top Scoring Predicted Contacts to Remove

We tested three simple selection strategies to select an appropriate subset of highest scoring predicted breaking contacts to be removed from the network of *lmc*ENM (see 6.2.2 for details). In the following, we analyze and discuss the performance of the selection strategies based on: (i) a constant cutoff, (ii) a relative cutoff (percent), and (iii) a score-dependent cutoff. We also performed a control experiment that removes the same amount of contacts as the best performing selection strategy but chooses them randomly.

Performance of Selection Strategies Fig 6.4 shows the accuracy distribution of each strategy grouped by motion type. For each strategy we choose the cutoff value that maximizes *lmc*ENM-accuracy averaged over all proteins in our data set (see 6.3.1 for details). We found that top $n = 60$ (constant cutoff), top $n = 16\%$ (relative), and SVM score > 0.4 (score-dependent) work best in our setting.

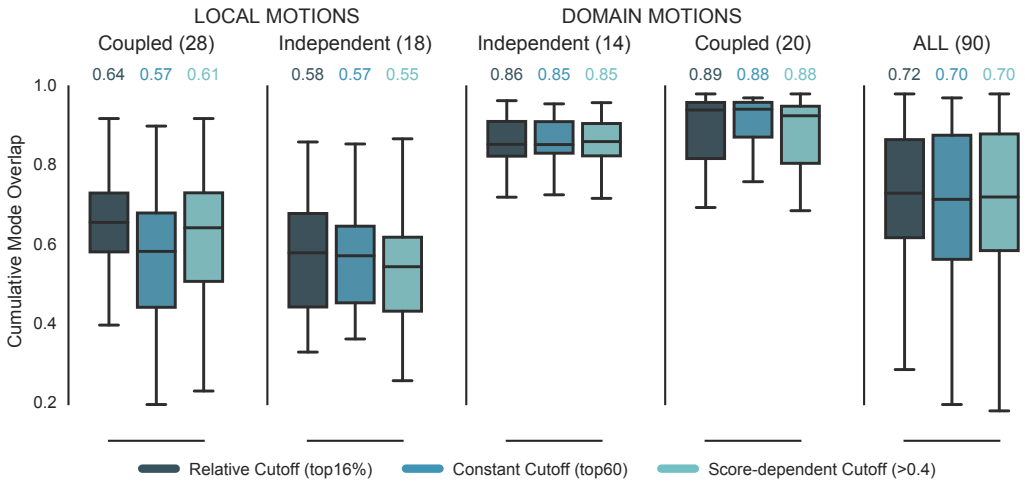

Figure 6.4: Effect of breaking contact selection strategies on *lmc*ENM accuracy for proteins grouped by motion type. For each strategy we chose the cutoff value that maximizes the accuracy of *lmc*ENM averaged over all proteins in our data set (90). The panels show the distributions as box plot, where boxes show the quartiles of the data. The numbers above each box report the mean. The selection strategies achieve similar performance with small advances for the relative cutoff strategy. Figure adapted from: Putz and Brock (2017).

Overall, the strategies perform similar with small advances for the relative-cutoff strategy. Except, for proteins with independent motions, we observe that the constant-cutoff strategy performs considerably worse than the other two. We attribute this to the fact that, on average, fewer breaking contacts are removed than with the relative cutoff. This indicates that either not

enough breaking contacts have been removed to reach the "critical mass" or that some relevant breaking contacts simply were missed due to the smaller amount of removed contacts.

For coupled domain movers the constant-cutoff strategy results in smaller variance as compared to the other two strategies. Given that domain movers typically have fewer *observed* breaking contacts (see 5.4.5) also removing fewer *predicted* breaking contacts by choosing the constant-cutoff strategy seems to be better. This is supported by our finding below (see 6.4.2) that for domain movers a considerably lower relative cutoff (around top5%) would be optimal than the one optimized over all proteins (top16%)).

Given its better overall performance, we chose the relative-cutoff strategy to select the removal candidates and build *lmc*ENM by removing the top16% predicted breaking contacts. For a detailed report on the contact statistics for each protein, such as initial number of contacts and removed breaking contacts for both, *lmc*ENM and *mc*ENM, we refer to the supplementary information of our paper (Putz and Brock (2017), Table B in the supplementary file S2).

Control Experiment We also performed a control experiment by removing the same amount of randomly selected contacts from the initial contact topology of the proteins as with the chosen relative-cutoff strategy above. We refer to this network as *rmc*ENM.

Table 6.3 lists the prediction accuracy of *rmc*ENM compared to the baseline ENM and *lmc*ENM. As expected we find no accuracy improvement over the baseline ENM (detailed results for each protein are reported in Table C of supplement S2 in Putz and Brock (2017)).

Table 6.3: Performance of *rmc*ENM without randomly selected breaking contacts compared to baseline ENM and *lmc*ENM on the full data set (90 proteins). Cumulative overlap of the first ten low-frequency modes (mean and median) are reported for the proteins grouped by their motion types. The number of proteins in each category is given in brackets after the motion labels. The last row reports the average values for all proteins. *rmc*ENM does not improve over the baseline ENM, which shows that removing actually relevant breaking contacts matters.

Motion Type	ENM	*rmc*ENM	*lmc*ENM
Coupled Local Motions (28)	0.53/0.52	0.51/0.52	0.66/0.64
Independent Local Motions (18)	0.48/0.53	0.47/0.53	0.58/0.58
Coupled Domain Motions (20)	0.94/0.88	0.94/0.88	0.94/0.89
Independent Domain Motions (14)	0.85/0.83	0.85/0.83	0.85/0.85
Burying Ligand Motions (4)	0.75/0.75	0.75/0.76	0.75/0.76
Other Types of Motions (9)	0.62/0.61	0.63/0.61	0.65/0.60
All (90)	0.69/0.67	0.69/0.66	0.73/0.72

This experiment demonstrates that the predicted breaking contacts indeed carry relevant information to improve the prediction accuracy of ENMs. It clearly matters which contacts are removed from the network to do so.

To summarize, our results indicate that finding a good selection strategy most likely depends on more factors besides protein motion type and classifier performance. Nonetheless, even such a simple strategy as our chosen one already leads to substantial accuracy improvements of *lmc*ENM.

6.4.2 SVM Predicts Correct and Relevant Breaking Contacts

Given the above chosen fraction of top-scoring breaking contacts (top16%), we now can evaluate the SVM classifier. We use the common measures precision (Prec=TP/(TP+FP)) and coverage (Cov = $\mathrm{TP}_{frac}/\mathrm{TP}_{all}$), where TP denotes true positive and FP false positive predicted breaking contacts. TP_{frac} are the true positives among the selected fraction, whereas TP_{all} is the total number of true positives for a protein. Furthermore, we report the area under the receiver operator characteristic (ROC) curve (AUROC) (Fawcett, 2006). It estimates the probability of scoring a positive sample higher than a negative one if both are chosen randomly. An AUROC of 1 indicates a perfect predictor, a value of 0.5 refers to a random predictor.

Fig 6.5A shows the prediction performance of the classifier along the protein motion types. The results for the proteins grouped by motion type are listed in Table 6.4, whereas individual results can be found in Table A of supplement S2 of our paper (Putz and Brock, 2017).

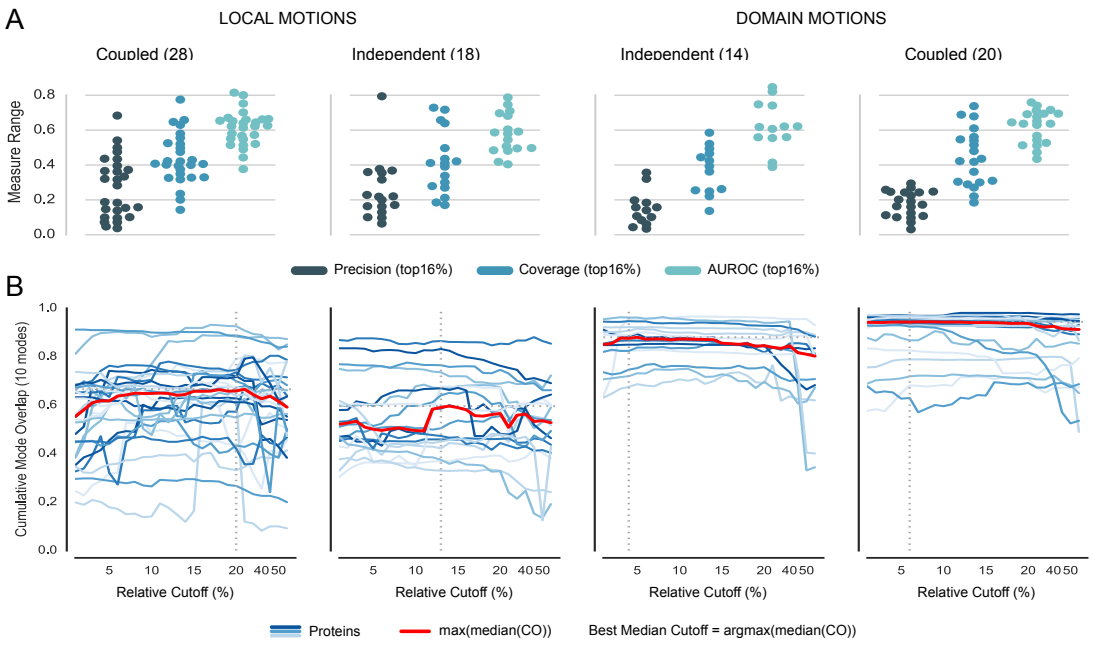

Figure 6.5: Classifier performance and sensitivity analysis of breaking contacts selection strategy, subset of local and domain motions (80 proteins). (A) Performance evaluation of classifier based on top16% predicted breaking contacts. The panels show precision, coverage, and area under receiver operator characteristic (AUROC) as swarmplot for each motion category. (B) Dependence of *lmc*ENM accuracy on removed topN% predicted breaking contacts ranked by decreasing SVM score. The blue lines depict how the *lmc*ENM-accuracy evolves for individual proteins when gradually removing more breaking contacts from their network. The cumulative mode overlap of protein with local motions often "jumps" upwards, which indicates that the removed breaking contacts causing the accuracy improvement are more relevant compared to the previously removed ones. Accuracy drops if too many breaking contacts have been removed. Figure source: Putz and Brock (2017).

Overall, the precision of the classifier is rather low. However, proteins with coupled local motions show higher precision on average. Interestingly, for some proteins–mostly domain movers–

Table 6.4: SVM performance overview of the top16% predicted breaking contacts on the full data set (90 proteins). Different performance measures (mean and median) are reported for the proteins grouped by their motion types. The number of proteins in each category is given in brackets after the motion labels. The last row reports the average values for all proteins. Table source: Putz and Brock (2017).

Motion Type	Precision	Coverage	AUC[1]
Coupled Local Motions (28)	0.24/0.27	0.41/0.44	0.62/0.61
Independent Local Motions (18)	0.22/0.25	0.40/0.41	0.56/0.58
Coupled Domain Motions (20)	0.18/0.18	0.43/0.44	0.64/0.62
Independent Domain Motions (14)	0.15/0.16	0.39/0.37	0.62/0.63
Burying Ligand Motions (4)	0.13/0.19	0.32/0.30	0.49/0.51
Other Types of Motions (9)	0.20/0.30	0.25/0.28	0.55/0.55
All (90)	0.19/0.23	0.41/0.41	0.61/0.60

[1] Area under curve (AUC) of receiver operator characteristic (ROC)

coverage is good despite a low precision. The fact that these proteins possess rather few observed breaking contacts might increase the chances of a TP among the top16% selected contacts.

We also performed a sensitivity analysis to test whether some predicted breaking contacts are more relevant for capturing the functional transition than others. Starting from the top1% until the top50% breaking contacts, we gradually removed more predicted contacts, while evaluating the reached accuracy.

Fig 6.5B shows the results for proteins with local and domain motions. Most steps yield only small accuracy improvements. But sometimes they cause a "jump" to a significantly higher or drop to a substantially lower value. This indicates that the associated breaking contacts are either more relevant than the previously removed ones or that they were required to reach the "critical mass" to be effective. In particular, proteins with coupled local motions show the largest jumps in *lmc*ENM accuracy. We also find substantial drops in accuracy, which most likely result from removing too many false positive predicted breaking contacts.

Hence, despite its deficiencies in precision and coverage, our classifier seems to be able to identify breaking contacts that are not only correct but also relevant to improve *lmc*ENM accuracy.

6.4.3 Predicted Breaking Contacts Matter

The only difference between the original, distance-cutoff based ENM and *lmc*ENM is that the latter ignores the predicted breaking contacts. We now evaluate the impact of removing these contacts on ENM accuracy by comparing the performance of *lmc*ENM to the baseline ENM and the theoretical maximum reached by *mc*ENM.

Fig 6.6 summarizes the results for the proteins binned by cumulative mode overlap of the first ten low-frequency modes (extended version of Fig 5.4, see Table C of supplement S2 in Putz and Brock (2017) for individual results).

Overall, *lmc*ENM substantially outperforms the baseline ENM in accuracy, in particular, for proteins poorly captured by the baseline ENM. For these proteins *lmc*ENM achieves on average

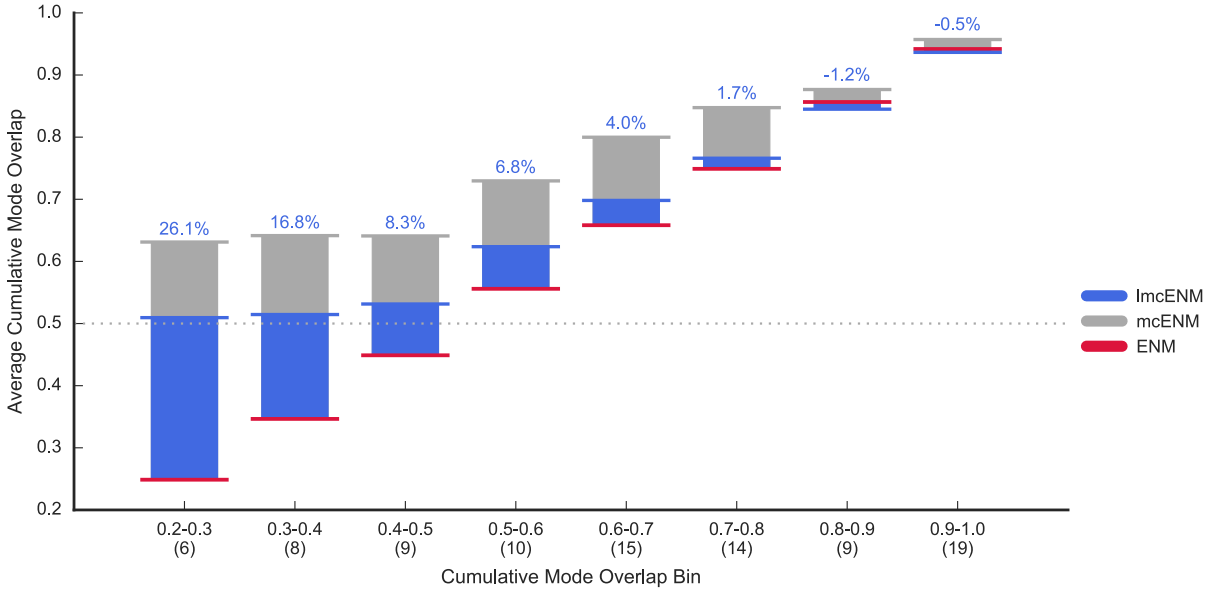

Figure 6.6: Accuracy of *lmc*ENM (our method) compared to ENM (baseline) and *mc*ENM (theoretical upper bound) on our data set (90 proteins). The accuracy is measured by the cumulative mode overlap of the first ten low-frequency normal modes ($CO(10)$). Proteins are binned based on the cumulative mode overlap reached by ENM (number of proteins per bin is given in brackets). The horizontal blue, gray and red lines mark the average accuracy per bin of *lmc*ENM, *mc*ENM, and ENM, respectively (numbers above each bin denote the absolute improvement of *lmc*ENM over ENM in percent). *lmc*ENM is most effective for proteins that largely remain elusive for ENM ($CO(10) < 0.6$). It is on par with ENM for the remaining proteins that are already accurately explained by ENM. Figure source: Putz and Brock (2017).

more than 60% of the improvement reached by *mc*ENM (theoretical maximum). Individual accuracy improvements range between 1.5% up to 59.8% and sometimes even exceed the theoretical maximum reached by *mc*ENM (see Table C of supplement S2 in Putz and Brock (2017)). As expected, proteins well captured by the original ENM benefit less from *lmc*ENM.

We also find that *lmc*ENM substantially increases the number of proteins reaching 60% coverage of the functional transition with only the ten lowest-frequency normal modes, albeit not as much as *mc*ENM (theoretical upper bound) (*lmc*ENM: 78% of proteins, ENM: 63%, *mc*ENM: 92%; see Table C of supplement S2 in Putz and Brock (2017)).

The overall improvement of *lmc*ENM by 5.5% on average (4.5% median) over the baseline ENM might appear small given the computational overhead of the machine-learning based classifier (see Table C of supplement S2 in Putz and Brock (2017)). However, in relation to the performance of the reference ENMs (OFC-ENM: 0.95%/-1.35%(mean/median), edENM: 0.83%/-1.0%, HCA: 1.24%/-0.10%) on our data set it becomes evident that general applicability of ENMs might require such additional computational costs.

For eight proteins, *lmc*ENM accuracy drops notably below the baseline (more than -5.0%; see Table 6.5 and Table C of supplement S2 in Putz and Brock (2017)).

Two of them (PDB_IDs: 2v8iA, 1lfhA) are domain movers. In both cases, *lmc*ENM removes too many contacts due to the chosen selection cutoff (top 16% predicted breaking contacts). With

Table 6.5: Proteins, where *lmc*ENM drops by more than 5% in cumulative overlap of the first ten low-frequency modes compared to ENM (baseline). For each ENM variant the difference w.r.t. ENM is shown in percent (Δ). In some cases also other ENM variants perform worse than ENM albeit to a lesser extent. Table source: Putz and Brock (2017).

Unbound	ENM	Δ (%) *lmc*ENM	Δ (%) OFC-ENM	Δ (%) edENM	Δ (%) HCA	Δ (%) *mc*ENM	Motion label	RMSD	#Residues	Scop class	#Domains
1dx9C	0.340	-10.20	2.90	0.50	4.70	26.40	OTM	1.690	168	a/b	1
2dh3B	0.439	-10.80	-0.70	-2.80	-0.50	9.70	ILM	1.730	416	None	2
1gohA	0.467	-11.20	-0.20	-3.60	-2.70	26.80	ILM	1.890	639	all beta	1
1a8dA	0.523	-8.30	2.90	-6.60	-1.10	19.90	ILM	1.870	451	all beta	2
2jepB	0.635	-17.70	3.10	-11.30	-0.10	17.30	ILM	1.140	359	None	1
1kp9A	0.739	-8.70	0.70	-4.10	-1.80	5.60	CLM	4.010	270	a/b	1
2v8iA	0.910	-5.60	0.50	-0.40	-0.20	5.90	IDM	2.000	535	None	3
1lfhA	0.931	-10.30	0.10	-1.80	-0.50	1.90	CDM	6.510	691	a/b	3

an optimal selection cutoff removing fewer contacts, *lmc*ENM would perform as good as the baseline ENM (Table 6.6). Nonetheless, the performance of *lmc*ENM is still good (above 0.85 CO(10) for both proteins).

Table 6.6: Optimal selection cutoff (topN percent) of *lmc*ENM and corresponding cumulative overlap of the first ten low-frequency modes for proteins, where *lmc*ENM performs significantly worse than ENM (baseline). For ease of comparison, also the cumulative overlap of ENM and *lmc*ENM based on the chosen selection cutoff of top16% as well as corresponding precision and coverage of the SVM is shown. Fig 6.7 shows how the cumulative overlap for these proteins evolves, when gradually removing more predicted breaking contacts. Table source: Putz and Brock (2017).

Unbound	ENM	*lmc*ENM	$CO10_{best}$	Cutoff_{best} (%)	SVM Precision	SVM Coverage
1dx9C	0.340	0.238	0.328	3	0.153	0.308
2dh3B	0.439	0.331	0.384	4	0.218	0.300
1gohA	0.467	0.355	0.469	1	0.101	0.185
1a8dA	0.523	0.440	0.547	3	0.164	0.411
2jepB	0.635	0.458	0.618	1	0.064	0.338
1kp9A	0.739	0.652	0.738	1	0.434	0.429
2v8iA	0.910	0.854	0.912	1	0.043	0.137
1lfhA	0.931	0.828	0.932	4	0.108	0.289

Also for three other cases (PDB_IDs: 1gohA, 1a8dA, 1kp9A)–all local movers–the optimal selection cutoff would yield comparable performance of *lmc*ENM. Notably, 1kp9A, is the only case with SVM precision and coverage above average of the motion category. Yet even with an optimal selection cutoff it would not improve over ENM. Given that *mc*ENM improves over ENM by 5.9%, *lmc*ENM most likely predicted breaking contacts that were correct but not relevant. Fig 6.7 (leftmost panel) supports this view.

Gradually removing more predicted breaking contacts yields continuously decreasing cumulative overlap. In particular for proteins with independent local motions or domain motions a better selection strategy may help to reduce the overall amount of removed breaking contacts, thereby decreasing the number of removed false-positive contacts (see the marked best median cutoff for individual motion types in Fig 6.5).

For the remaining three cases (PDB_IDs: 1dx9C, 2dh3B, 2jepB) even the optimal selection cutoff yields between 1.2% and 5.5% lower cumulative mode overlap than the baseline ENM. Fig 6.8 shows the networks with breaking and maintained contacts for 2dh3B accompanied by a plot depicting the fluctuation profiles of the different ENM variants scaled to the observed displacements.

Although *lmc*ENM partially captures true-positive breaking contacts, it misses observed ones (indicated by the dark arrows) in particular at the interface between two helices in the center performing a shear motion as well as between their connecting loop and the right helix (arrow a2). Consequently, the flexibility of these regions is underestimated (mostly around the most flexible center of the loops), whereas it is largely overestimated around two solvent-exposed loops (arrow 4), where only few breaking contacts have been observed. Hence, our feature capturing the location (border vs center) of a contact on a loop seems to be not discriminative enough.

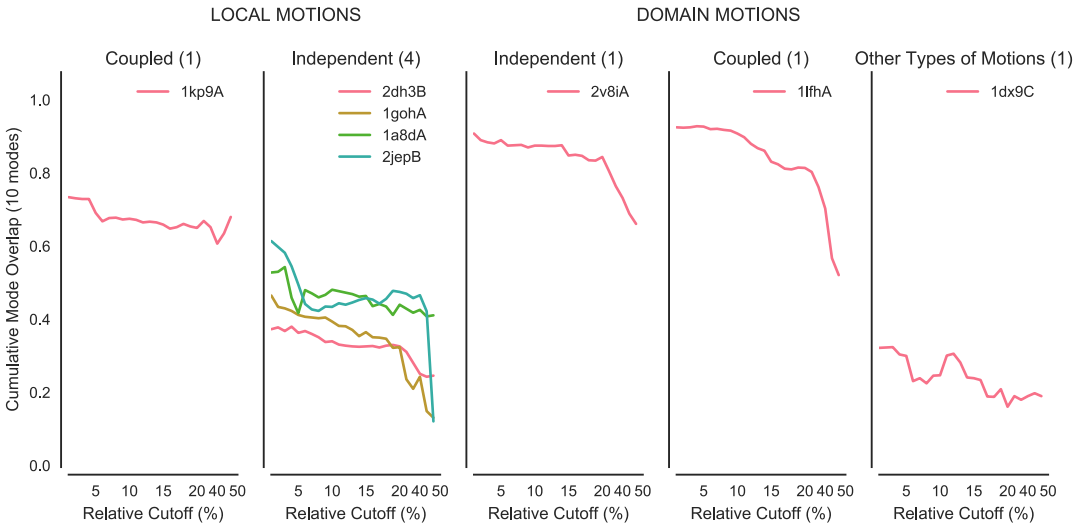

Figure 6.7: Sensitivity analysis of *lmc*ENM-selection cutoff (topN percent) for the eight proteins, where *lmc*ENM drops by more than 5% in accuracy compared to ENM (baseline). Dependence of *lmc*ENM accuracy on removed topN% predicted breaking contacts ranked by decreasing SVM score for the eight proteins grouped by their motion type. The lines depict how *lmc*ENM-accuracy evolves for individual proteins when gradually removing more breaking contacts from their network. In all cases the accuracy drops almost starting from the beginning. The optimal topN percent cutoff for each protein is reported in Table 6.6 above. Figure source: Putz and Brock (2017).

The situation for the other two proteins is highly similar. We also note that four out of the eight cases are proteins with independent local motions, i.e. not coupled to a ligand. For such proteins designing better features or training an ensemble of SVMs may help to improve the performance of the classifier. When using an SVM ensemble, each SVM could be trained to capture specific properties of a single motion category, which are then combined into an ensemble of classifiers for prediction. Such ensemble classifiers have been successfully applied in the context of protein contact prediction (Schneider and Brock, 2014), for instance.

In addition to the mode overlap, we also evaluated other metrics that are commonly used to assess the performance of ENMs. The performance of *lmc*ENM compared to all other ENM variants w.r.t. to these metrics is summarized in Table 6.7.

Table 6.7: Evaluated similarity measures for *lmc*ENM compared to ENM (baseline), *mc*ENM (theoretical upper bound) and three reference ENM variants. Median/mean values are reported for each measure. For several measures we consider only the subset of the first ten low-frequency modes indicated by (10) after the measure's name. Except for Rank (Max Overlap) and Collectivity higher values are better. A lower rank of the best overlapping mode with the observed displacement vector indicates that the most relevant motion captured by the elastic network also is more dominant. In terms of Degree of Collectivity, we find that lower values indicate that less collective, localized functional transitions are better captured (see next paragraph for more details). Table source: Putz and Brock (2017).

Measure	ENM	OFC-ENM	edENM	HCA	*lmc*ENM	*mc*ENM
Cumul. Mode Overlap (10)	0.69/0.66	0.67/0.67	0.68/0.67	0.68/0.68	0.73/0.72	0.82/0.80
Cumul. Fraction of Variance (10)	0.35/0.38	0.40/0.43	0.57/0.59	0.34/0.36	0.60/0.60	0.57/0.59
CorrCoeff Fluctuations - Displacements (10)	0.52/0.50	0.52/0.52	0.52/0.50	0.52/0.50	0.58/0.56	0.81/0.78
CorrCoeff Temperature Factors - Betas (10)	0.40/0.40	0.41/0.41	0.48/0.46	0.45/0.44	0.41/0.40	0.41/0.40
Max Overlap	0.47/0.50	0.46/0.51	0.44/0.50	0.46/0.50	0.45/0.50	0.60/0.62
Rank (Max Overlap Mode)	1.00/11.08	1.00/15.71	1.00/49.72	1.00/32.81	1.00/2.80	0.00/1.93
Degree of Collectivity (Max Overlap Mode)	0.38/0.39	0.40/0.40	0.45/0.41	0.40/0.40	0.33/0.34	0.27/0.31
Fraction of Variance (Max Overlap Mode)	0.05/0.08	0.06/0.08	0.09/0.12	0.05/0.06	0.08/0.13	0.12/0.20
#Modes Cumul. Mode Overlap (70%)	11.00/34.51	12.50/29.72	11.00/30.09	10.00/29.29	7.00/23.37	3.00/6.92
#Modes Cumul. Mode Overlap (80%)	35.00/79.07	34.00/66.62	31.50/70.99	31.50/65.76	22.50/54.08	7.50/19.41
#Modes Cumul. Mode Overlap (00%)	164.50/200.60	119.00/165.86	128.00/187.02	105.50/175.33	100.50/156.79	39.00/75.56

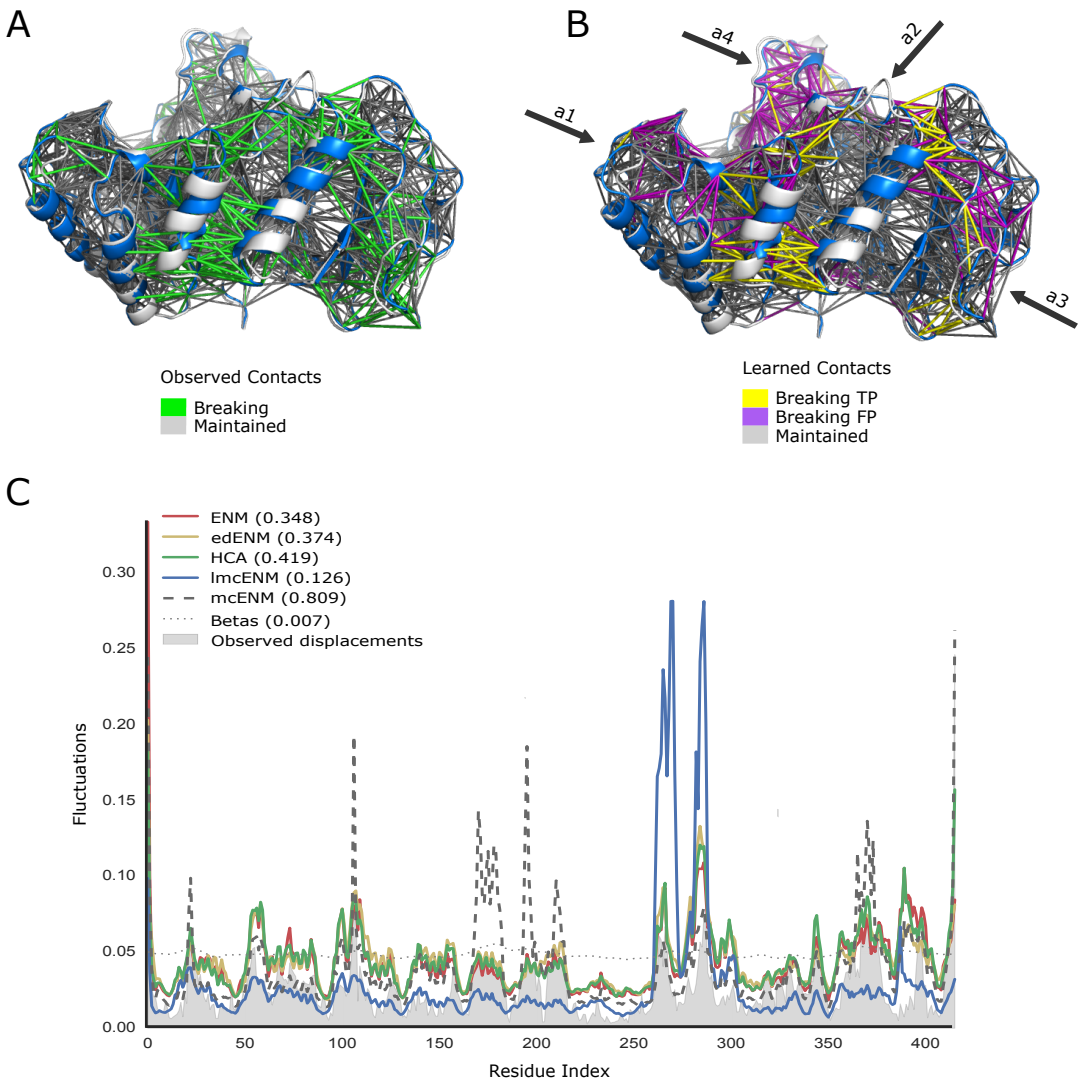

Figure 6.8: Example protein (2dh3B), where *lmc*ENM performance significantly drops below ENM (baseline). (A) Observed breaking and maintained contact networks. Unbound and bound conformation colored blue and white, respectively. (B) Predicted true-positive (TP), false-positive (FP) breaking, and maintained networks. (C) Fluctuation profiles of all ENM variants scaled to observed displacements. The dark arrows point to parts, where *lmc*ENM substantially underestimates the flexibility between residues 106-109 (a1), 160-220 (a2: helix-loop-helix), 360-380 (a3) because relevant observed breaking contacts mostly constraining flexible loops have not been predicted. Between residues 260-290 (a4) it largely overestimates flexibility due to the removal of too many false-positives. Figure adapted from: Putz and Brock (2017).

We find that *lmc*ENM consistently outperforms all other ENM variants (apart from *mc*ENM (theoretical upper bound)) in all metrics except for the correlation between temperature factors and maximum overlap (considering all modes). Detailed results are given in Table C of supplement S2 of our paper (Putz and Brock, 2017). In the following we will discuss these results in more detail.

*lmc*ENM improves over the other ENM variants in capturing motion directions (overlap, structural variance, number of modes to explain up to X percent cumulative mode overlap) as well as motion amplitudes (correlation between fluctuation profiles) of the functional transition.

edENM reaches a comparable cumulative fraction of variance (10 lowest-frequency modes) and is the best method to explain experimental b-factor profiles with predicted temperature factors (squared residue fluctuations of the first ten low-frequency modes scaled to b-factors). We attribute this to the carefully optimized stiffness constants of edENM based on MD simulations.

In terms of maximum overlap, all ENM variants reach similar values. A closer look at the results for different motion types reveals more variation as we will see in the following (see 6.4.4). Considering the fraction of variance explained by the best-overlapping mode *lmc*ENM and edENM perform the best on average. Although the median rank of the best-overlapping mode is 1 for all ENM variants, the average rank shows that *lmc*ENM effectively shifted the best-overlapping mode towards lower frequencies (*lmc*ENM: 2.8 (best), ENM: 11.1 (2nd best), *mc*ENM: 1.9).

Interestingly, *lmc*ENM and *mc*ENM yield much lower degree of collectivity for the best overlapping mode, whereas the other ENM variants reach higher values compared to the baseline ENM. Hence, elastic networks without observed/predicted breaking contacts seem to better capture localized transitions with lower degree of collectivity. We will further investigate this observation in the following (see 6.4.4).

To summarize, our results show that the selected, learned breaking contacts in fact contain valuable information to improve ENM accuracy. *lmc*ENM is most effective for proteins that are poorly captured by ENM, suggesting that it helps where it is most needed.

6.4.4 *lmc*ENM is Most Effective For Coupled Localized Functional Transitions

In the previous chapter 5 we showed that *observed* breaking contacts matter to capture localized functional transitions. To evaluate whether this holds true also for the chosen *predicted* breaking contacts we analyze the performance of *lmc*ENM considering the motion type of the proteins. Fig 6.9 shows the results. Median and mean values for each motion type are listed in Table B.1.

*lmc*ENM consistently outperforms ENM in accuracy regardless of the depicted motion type, being most effective for proteins with ligand-coupled local motions (*lmc*ENM: 12% improvement, *mc*ENM: 21%, HCA: 2%, edENM and OFC-ENM: 1% on average). Proteins with independent local motions improve less due to lower classification accuracy (see Fig 6.5A).

We also find that *lmc*ENM captures domain motions slightly better than other ENM variants or is on par despite the relatively poor classifier accuracy (see Fig 6.5A). We attribute this to the fact that proteins performing domain motions are structurally more rigid than proteins with local motions. Hence, the former seem to be more robust against removing false positive predictions, which have higher chances to be a redundant constraint that has no influence on the overall motion of the protein.

Considering the total variance captured by the first ten low-frequency modes, *lmc*ENM largely improves over the other ENM variants, closely followed by edENM (Fig 6.10).

We attribute this to the fact that by removing predicted breaking contacts *lmc*ENM effectively compensates for the overestimated rigidity in the baseline ENM (Orellana et al., 2010). Hence, the *lmc*ENM-modes with more relevance–as indicated by the larger cumulative mode overlap

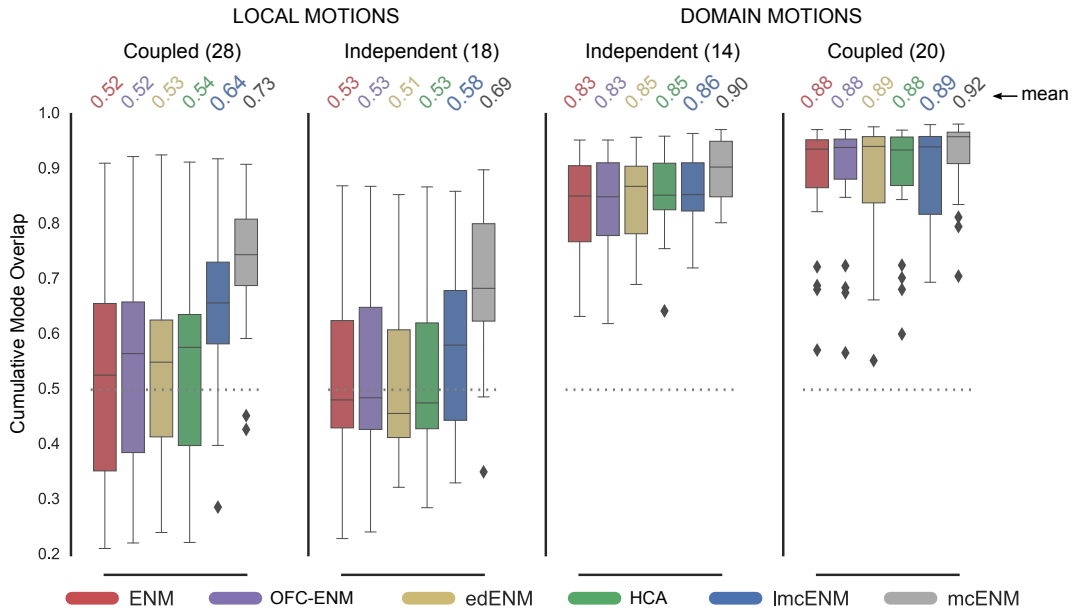

Figure 6.9: Dependence of accuracy of evaluated ENM variants on motion type of protein, subset of local and domain motions (80 proteins). Accuracy is measured by the cumulative mode overlap of the first ten low-frequency normal modes ($CO(10)$). *lmc*ENM consistently improves over ENM in each motion category, being particularly effective for proteins with coupled localized functional transitions. Figure adapted from: Putz and Brock (2017).

above–become easier accessible and contribute more to the total variance of the system. In the other ENM variants these modes are spread among a wider range, which decreases their individual contribution as well as their captured total variance. Given that removing breaking contacts is a purely topological change, our work supports the findings by Orellana et al. (2010) that such an effect cannot be achieved by refining spring stiffness alone.

Taking into account the correlation coefficients between predicted and observed fluctuations, only coupled local and independent domain motions are better captured by *lmc*ENM, while it is on par with the other ENM variants for the remaining motion types (Fig 6.10(B)). Experimental b-factors are best explained by edENM followed by HCA, whereas *lmc*ENM does not improve over the baseline ENM (Fig 6.10(C)). This can be explained by the fact that *lmc*ENM only adjusts the network topology without refining the stiffness of the springs that is typically tuned for ENMs to better match B-factor profiles. Also, larger distance cutoffs (>16 Å) are usually required to gain better agreement with experimental B-factors thereby increasing structural stiffness and collectivity of motion (Kondrashov et al., 2007). Given our aim to improve the prediction accuracy of ENMs for localized functional transitions with low degree of collectivity, this metric is of limited use in our context.

We also note that edENM improves little over the baseline ENM for coupled local motions and even drops below it for independent local motions. This is unexpected given the reported performance of edENM in the original publication (Orellana et al., 2010). The main difference between *lmc*ENM and edENM is the protein-size dependent cutoff used by the latter to identify remote interactions. edENM also scales the stiffness constants depending on sequence or spatial

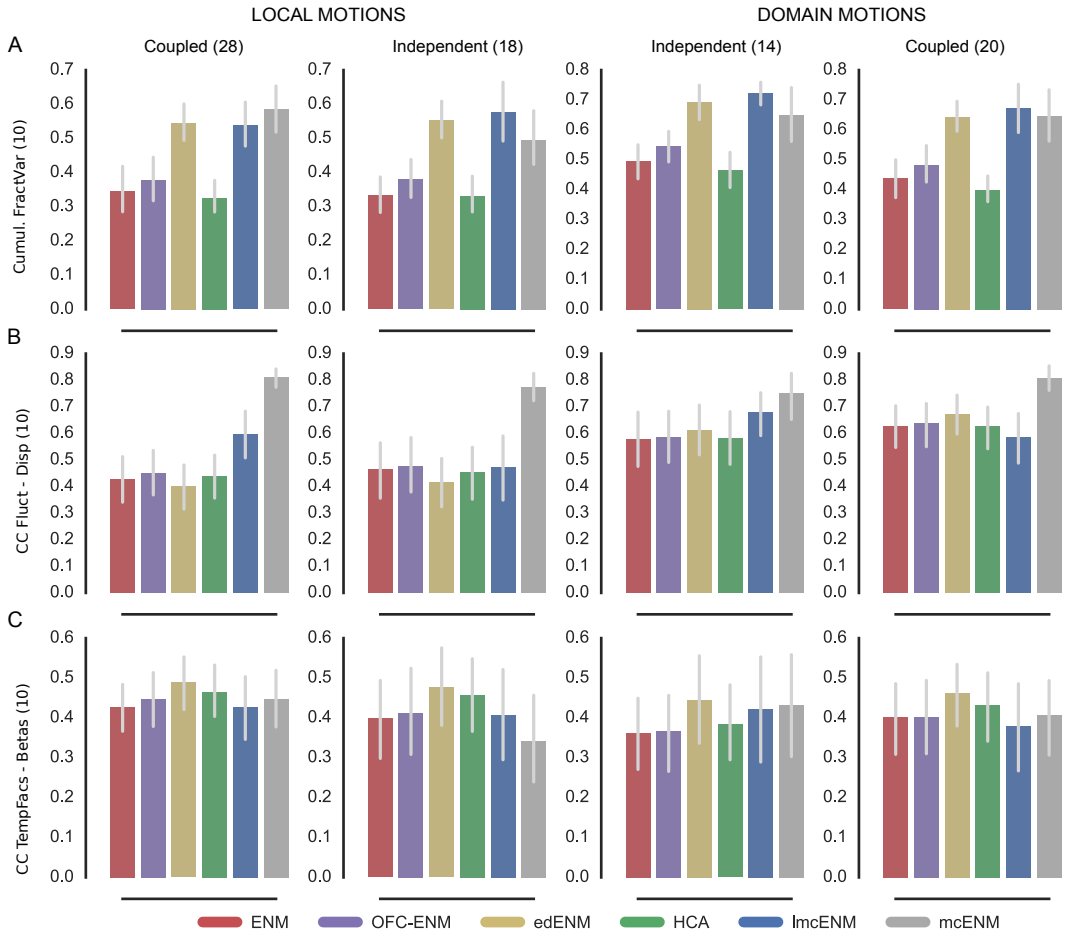

Figure 6.10: Accuracy of *lmc*ENM compared to reference ENM variants using additional metrics on our protein data set grouped by motion type. (A) Cumulative Fraction of Variance (10 modes). *lmc*ENM and edENM consistently capture by far largest amount of structural variance with the lowest frequency modes, with slight advances for *lmc*ENM except for coupled local motions. They perform as good or better than *mc*ENM (theoretical upper bound). (B) Correlation coefficient between predicted residue fluctuations and observed displacement magnitudes (10 modes). For coupled local and independent domain motions *lmc*ENM reaches largest agreement between predicted and observed fluctuation profiles. For the other two motion types *lmc*ENM performs as good as the other ENM variants or slightly worse. (C) Correlation coefficient between predicted Temperature factors and experimental Beta factors (10 modes). Considering the similarity of temperature factor profiles *lmc*ENM and ENM perform roughly the same. The largest agreement for all motion types is achieved by edENM. Figure source: Putz and Brock (2017).

distance. But this cannot explain the large difference in cumulative mode overlap between edENM and *lmc*ENM. Thus, we analyzed how well this protein-size dependent cutoff matches the (theoretical) optimum cutoff of the baseline ENM per protein.

Fig 6.11 compares the best cutoff yielding largest cumulative mode overlap of the first ten low-frequency modes with the protein-size dependent cutoff of edENM. Most of the proteins do not follow the proposed logarithmic function. Consequently, the protein-size dependent cutoff seems to largely over-constrain the network for most proteins in our data set compared to the distance-cutoff used by the baseline ENM, *lmc*ENM, and *mc*ENM. This explains why edENM on

average does not improve in cumulative mode overlap over the basic ENM for proteins with local function-related movements (see Table C of supplement S2 in Putz and Brock (2017)).

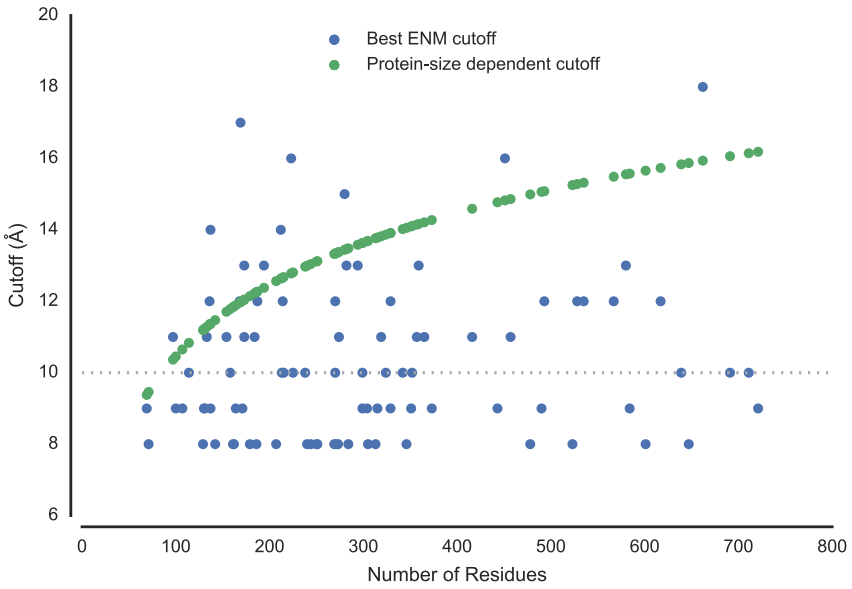

Figure 6.11: Scatter plot of cutoff distance against protein length. In blue the best cutoff values (in range 8-18Å) of the baseline ENM are shown, which yield the largest cumulative mode overlap considering the first ten low-frequency modes. In green the protein-size dependent cutoff is shown as proposed by Orellana et al. (2010). The dotted horizontal line indicates the median best ENM cutoff. In our set of proteins we find no correlation between optimum cutoff and protein size. The protein-size dependent cutoff largely over-constrains the network for most proteins in our data set. Figure source: Putz and Brock (2017).

To summarize, our results demonstrate that the predicted breaking contacts are in fact relevant to capture localized functional transitions, in particular if they are coupled to the binding of a ligand.

6.4.5 *lmc*ENM Reduces Dimensionality of Essential Deformation Space

In the previous chapter we showed that *mc*ENM substantially narrows down the essential deformation space of proteins used for subsequent fine-grained exploration (see 5.4.4). We cannot expect such a drastic dimensionality reduction for *lmc*ENM because the classifier only partially covers the observed breaking contacts and additionally outputs many false positives. Nonetheless, there should be some effect.

Fig 6.12 shows the median number of modes required for each ENM variant to reach a cumulative overlap of 70%, 80%, and 90%. As expected, *lmc*ENM cannot compete with *mc*ENM. But regardless of motion type, *lmc*ENM requires a considerable smaller amount of low-frequency normal modes than the baseline ENM to capture up to 90% of the functional transition with one

exception. To reach 90% overlap for proteins with independent motions the baseline ENM needs on average fewer modes than *lmc*ENM.

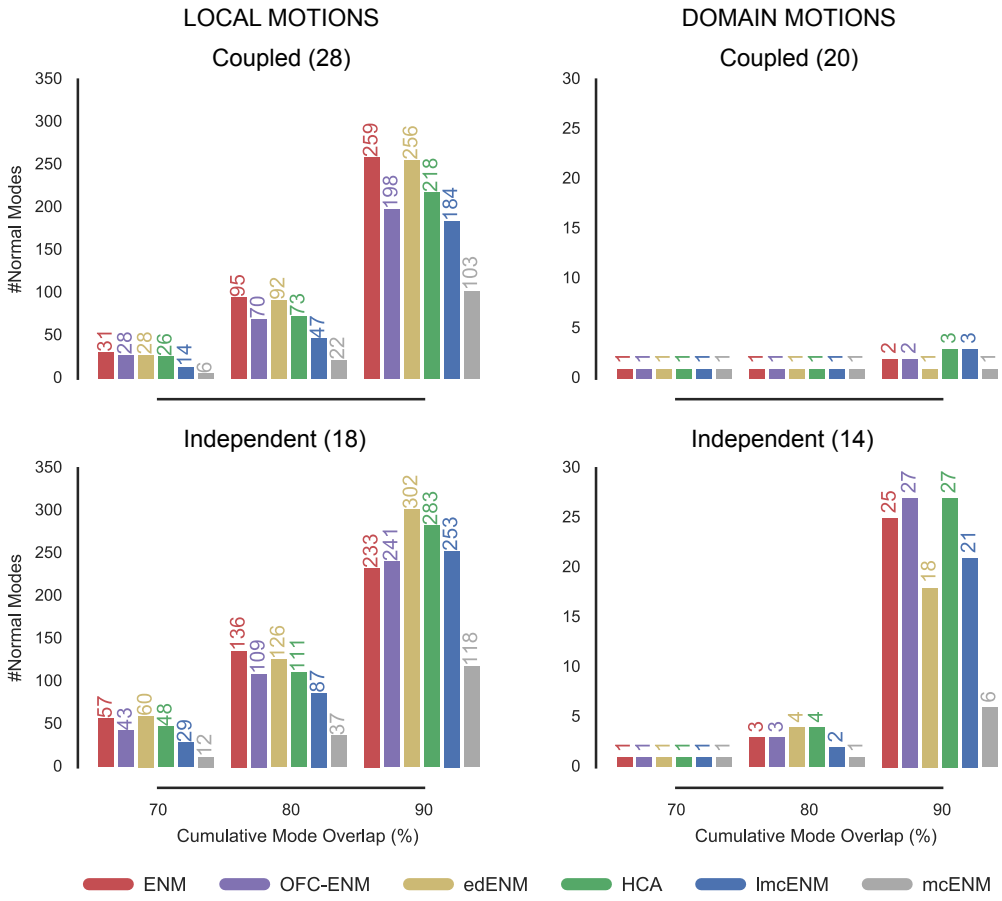

Figure 6.12: Dependence of dimensionality of deformation subspaces of ENM variants on motion type of protein, subset of local and domain motions (80 proteins). The panels show the median number of normal modes (spanning the deformation subspace) required to explain between 70% and 90% of the functional transition (measured in cumulative mode overlap (%)). *lmc*ENM consistently requires fewer modes to capture the same amount of conformational change as ENM. Figure adapted from: Putz and Brock (2017).

Also the other ENM variants are able to reduce the number of required modes compared to the baseline ENM, albeit not as much as done by *lmc*ENM in most cases. For instance, to capture coupled local motions with 80% overlap *lmc*ENM needs about half as much modes as the baseline ENM (*lmc*ENM: 47, ENM: 95), whereas the next best other ENM variant is OFC-ENM with 70 modes.

Reaching a desired overlap with fewer modes only works if individual modes capture more of the conformational transition. Hence, *lmc*ENM is able to reveal actually relevant modes, particularly for coupled localized functional transitions.

Another way to investigate this is to look at the best overlapping mode out of all modes. For highly collective protein motions usually a single low-frequency mode captures the movement

quite well. Thus, a large overlap together with a low rank of this mode indicates that the ENM is able to accurately explain the movement.

However, localized functional transitions with low degree of collectivity (i.e. fewer residues are involved in the movement) require more modes (usually less than 10) to be captured (Cavasotto et al., 2005). These modes are often spread among higher frequencies yielding rather low overlaps in the low-frequency mode spectrum. Hence, apart from higher overlap and lower rank also lower collectivity of the best-overlapping mode is desirable in this case. This is because lower collectivity indicates that an actually relevant mode has been successfully shifted towards lower frequencies. We evaluate this by reporting the reached maximum overlap, the rank of this mode among all modes, the fraction of variance explained by this mode as well as its degree of collectivity (see 4.2 for details). Fig 6.13 shows the results.

In particular for localized transitions, *lmc*ENM improves in maximum overlap over the other ENM variants (except *mc*ENM). The best overlapping modes of *lmc*ENM have not only much lower rank but also contribute more to the structural variance compared to the other ENM variants. This is because they more likely represent a localized transition due to their lower degree of collectivity.

We also find that the best overlapping mode of *mc*ENM has even lower collectivity, although it reaches a higher overlap. This indicates that *lmc*ENM missed to capture some of the localized transitions. In contrast, for domain motions *lmc*ENM shows smaller maximum overlap. Especially for independent domain movers the best overlapping *lmc*ENM-modes are less collective compared to the other ENM variants, which is not desired for this motion type. Due to the chosen selection cutoff *lmc*ENM removes much more predicted breaking contacts than it would be optimal for this class of proteins (see Fig 6.5B, not for coupled domain movers). As a consequence, actually irrelevant movements with low degree of collectivity become accessible and may contribute more to the predicted deformability than the relevant collective ones. However, despite this smaller maximum mode overlap *lmc*ENM still outperforms the other ENM variants in cumulative mode overlap as shown above (see 6.4.4).

Taken together, these results show that *lmc*ENM effectively "shifts" modes that are relevant to explain localized motions towards lower frequencies. Nonetheless, in several cases *lmc*ENM still requires more than 100 modes to capture 70% of the conformational change (see for instance 1a8dA (*lmc*ENM: 102, ENM: 136, OFC-ENM: 112, edENM: 121, HCA: 108, *mc*ENM: 33) or 1bsqA (*lmc*ENM: 126, ENM: 219, OFC-ENM: 155, edENM: 114, HCA: 179, *mc*ENM: 21) in Table C of supplement S2 in Putz and Brock (2017)). In such cases neither of the evaluated ENM variants is able to significantly narrow down the essential deformation space, which is the actual advantage of ENMs over other prediction methods in particular for proteins with collective motions.

However, *mc*ENM (based on the removal of *observed* breaking contacts) clearly demonstrates that this advantage does exist also for the cases that are difficult to capture by standard ENM. For 83/90 proteins, *mc*ENM needs less than 20 modes to capture 70% of the conformational change. *lmc*ENM is able to do so for 67/90 proteins, which is an improvement of 11% over the second best method, OFC-ENM, that is successful in 57/90 cases. Thus, we believe that

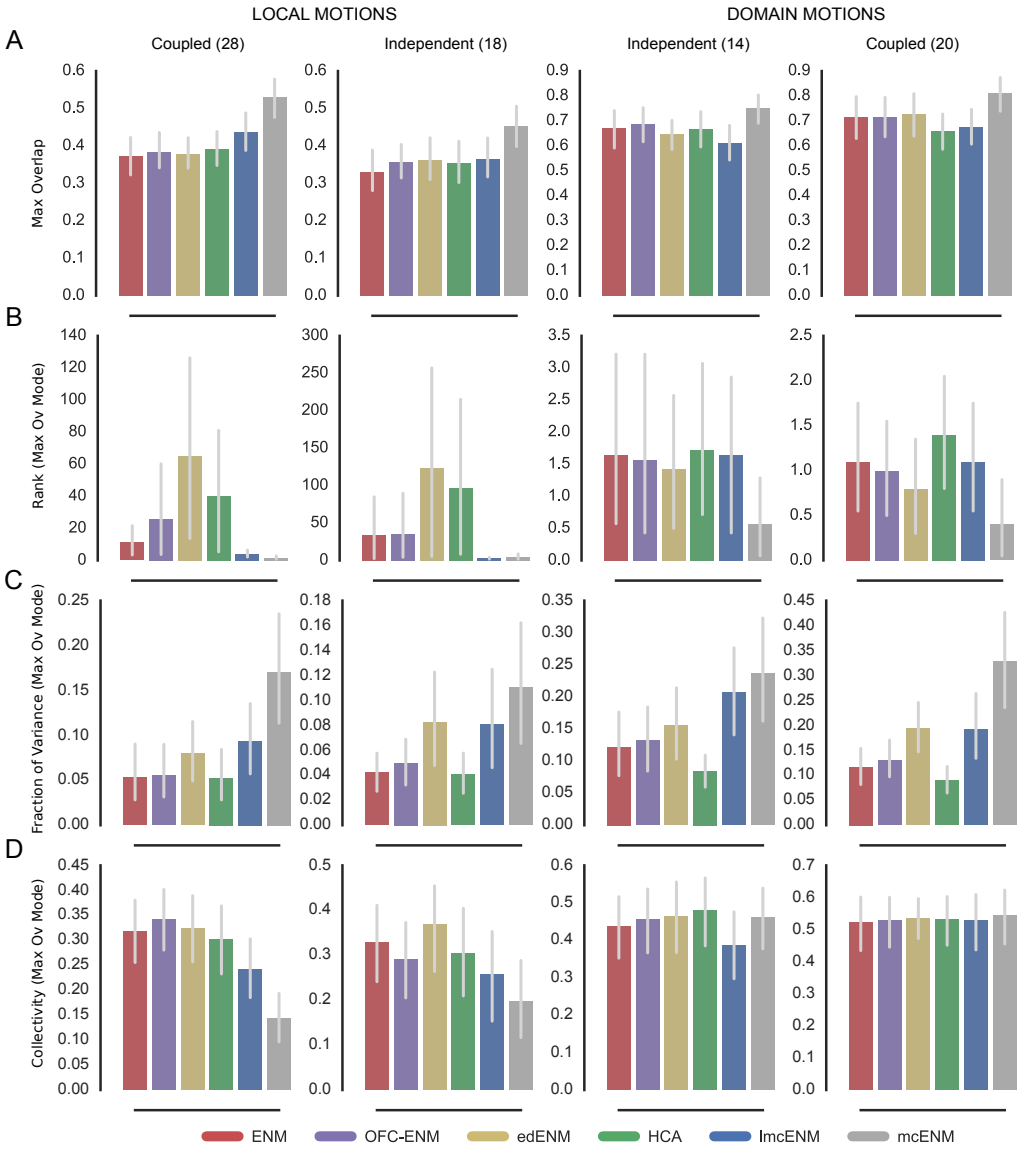

Figure 6.13: Accuracy of *lmc*ENM w.r.t. maximum mode overlap related measures compared to reference ENM variants on LMC_all data set grouped by motion type (A) Maximum mode overlap of all modes. (B) Rank of best-overlapping mode. (C) Fraction of variance explained by best-overlapping mode. (D) Degree of collectivity of best-overlapping mode. Figure source: Putz and Brock (2017).

*lmc*ENM–despite its current limitations–provides the necessary means to advance the predictive power of ENMs for yet poorly captured proteins.

6.4.6 Validating Against Essential Dynamics of Conformational Ensembles

Finally, we validate our method against the structural flexibility captured by redundant conformational ensembles. Due to the rapid growth of the Protein Data Bank (PDB) such ensembles recently emerged as valuable source characterizing the conformational diversity around the native state (Best et al., 2006, Burra et al., 2009, Monzon et al., 2016).

Amongst others, the CoDNaS 2.0 database (Monzon et al., 2016) provides such a redundant collection of conformers obtained under different conditions for the requested protein. To adequately capture the native conformational diversity a minimum ensemble size of ten is recommended (Best et al., 2006, Monzon et al., 2016). For 35 proteins in our dataset we could retrieve an ensemble with at least ten conformers (see Table D of supplement S2 in Putz and Brock (2017)). Principal component analysis (PCA) identifies the Essential Dynamics (ED) captured by the conformational ensembles, which can be compared to the normal modes of ENMs (David and Jacobs, 2014). A basic introduction into PCA is provided in the background chapter (see 3.2.3).

Fig 6.14 shows how well *lmc*ENM explains the native conformational diversity compared to the other ENM variants. Detailed results for each protein can be found in Table E of supplement S2 of our paper (Putz and Brock, 2017).

We measure the similarity of PCA space and ENM spaces by (i) comparing fluctuation profiles of the first ten low-frequency modes (subfigure A), (ii) the subspace overlap (also called RMSIP10) of the same mode set (subfigure B), and the weighted overlap (RWSIP) of both spaces (subfigure C). While the first two measures only consider the agreement in either magnitudes or directions of motion, respectively, the latter accounts for their interplay. In addition, RWSIP has no limit on the size of the compared spaces. Despite the wide use of RMSIP, RWSIP is considered the more comprehensive measure to assess vector space similarity (Carnevale et al., 2007, Fuglebakk et al., 2012). For more details on these measures we refer to 4.2.3 in the methods chapter.

In fact, all ENM variants reach comparable subspace overlap thereby limiting the information gain of RMSIP. The comparison of slow-frequency fluctuation profiles reveals rather small advances for *lmc*ENM except for coupled domain motions, where *lmc*ENM performs even slightly worse than the baseline ENM. However, in terms of RWSIP *lmc*ENM clearly performs the best, followed by edENM and OFC-ENM.

These results confirm the outcome of the previous experiments. Conformational ensembles, e.g. from X-Ray, NMR, or MD simulations, provide another rich source of information about the dynamic behavior of inter-residue contacts that could be encoded in additional features. For instance, simple counting of contact occurrence in predicted candidate protein structures is a very successful method for *ab initio* protein contact prediction (Eickholt et al., 2011). Training the classifier with these additional features may improve its prediction accuracy, which will in turn positively impact the performance of *lmc*ENM.

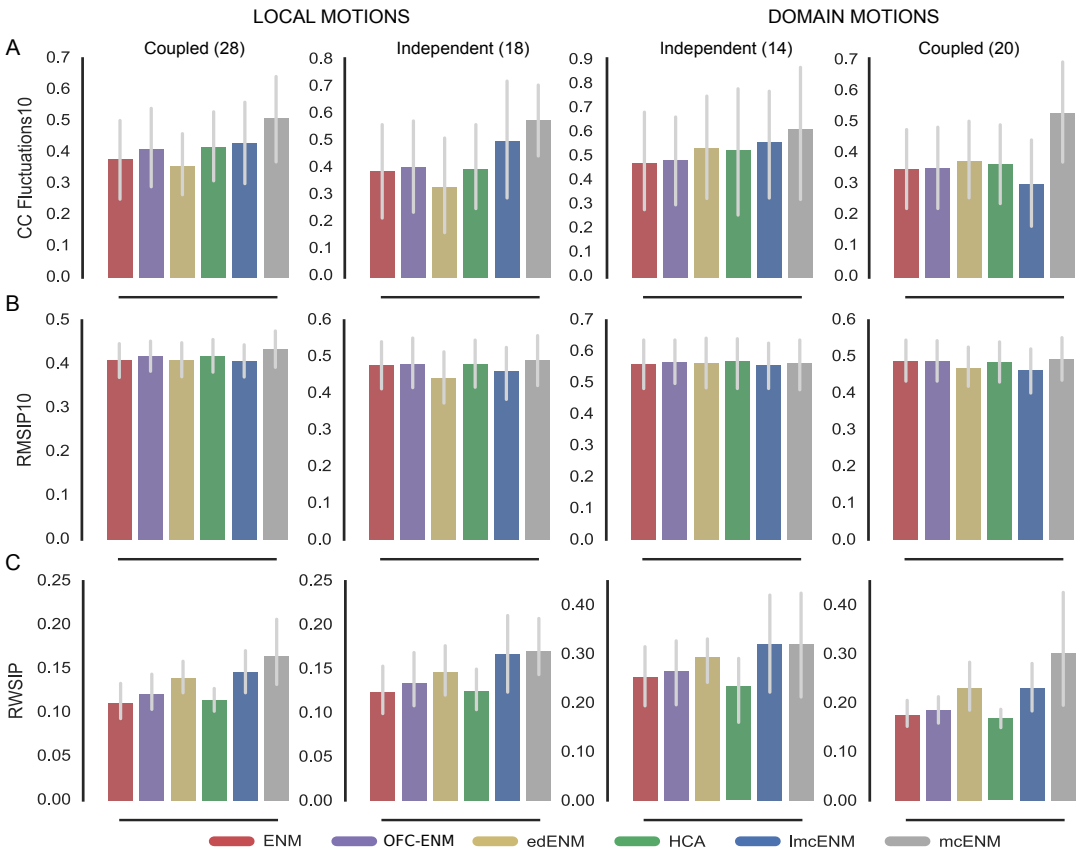

Figure 6.14: Ability of *lmc*ENM to capture structural flexibility of conformational ensembles compared to ENM (baseline), *mc*ENM (theoretical upper bound) and three other ENM variants on subset of 35 proteins having at least 10 conformational states. The panels grouped by motion type show the similarity of fluctuation profiles (magnitudes) considering the first ten low-frequency modes (A), the subspace overlap (directions) of the same mode set (B), and the weighted overlap (directions and magnitudes) of both spaces (C). *lmc*ENM clearly outperforms the other ENM variants in the most robust measure, which is the weighted overlap (RWSIP) that takes into account motion directions and magnitudes captured by the full deformation spaces. The 2nd best method is edENM, which performs slightly worse for coupled local motions and comparable for coupled domain motions. Figure source: Putz and Brock (2017).

6.4.7 Case Studies

In the following we discuss the performance of *lmc*ENM in more detail on three biologically interesting proteins selected from our data set: the outer membrane transporter FecA, the fatty acids oxidizing enzyme Arachidonate 15-Lipoxygenase, and SopA–a salmonella effector protein.

FecA - an outer membrane transporter protein

The outer membrane protein FecA has two main functions: First, to actively transport iron (ferric citrate) into the cells of *Escherichia coli* through their outer membrane (Ferguson et al., 2002). Second, to trigger the transcription of genes responsible for the iron uptake. Fooling this iron-transport mechanism allows to infiltrate antibiotics into the cells of multi-drug resistant bacteria, which makes FecA a biologically interesting target (Górska et al., 2014). We picked FecA for this case study because *lmc*ENM captures its functional transition almost 40% more accurate than ENM although it is the only membrane protein in our data set.

FecA is a three-domain protein (Ferguson et al., 2002) consisting of (i) a β-barrel spanning the membrane, (ii) a "plug" domain comprised by a mixed four-stranded β-sheet blocking direct diffusion through the barrel, and (iii) an NH-domain in the periplasm (not resolved in the crystal structure). Fig 6.15, A and D, depict the functional transition of FecA marked by unbound and ligand-bound conformation (PDB-ids: 1pnzA (Yue et al., 2003) and 1kmpA (Ferguson et al., 2002)). Two large extracellular loops (7 and 8) of the β-barrel dominate the transition by covering the ligand in the binding site (Ferguson et al., 2002, Piggot et al., 2013). Being propagated through the plug-domain these movements then cause an unwinding of the H1-helix ("switch" helix) to trigger the gene transcription process.

Fig 6.15, B and E, reveal that initially both loops are tightly constrained within the contact network of the unbound conformation. However, most of these surrounding contacts are observed to break (highlighted in green) to facilitate the major conformational changes of loops 7 and 8. Remarkably, the learned breaking contacts (true positives (TP), yellow) closely resemble the observed ones in the most relevant core region of both loops. Only towards their less flexible anchor points fewer contacts have been predicted to break (Fig 6.15, C and F).

However, we also notice many false positive predicted breaking contacts (FP, violet) that have not been observed, for instance around loops 3, 4, and 5 (Fig 6.16(A)). Interestingly, there is a single observed breaking contact between loops 4 and 5, which indicates that a more strict extension threshold would have identified more contacts as breaking around these loops (Fig 6.15B). In fact, for this protein the optimal extension threshold to identify observed breaking contacts for *mc*ENM would be 3% (CO10: 0.839) instead of the used 9% (CO10: 0.809), which is the optimal threshold averaged over the whole data set (tested for distance cutoff within 8-18Å and extension thresholds between 3 and 25%). Based on this optimal extension threshold the agreement between predicted and observed breaking contacts would improve, in particular in loops 4 and 5 (Fig 6.17).

Hence, the predicted increased flexibility for these loops may be actually correct. This hypothesis is supported by experimentally observed structural differences between these two loops (Ferguson

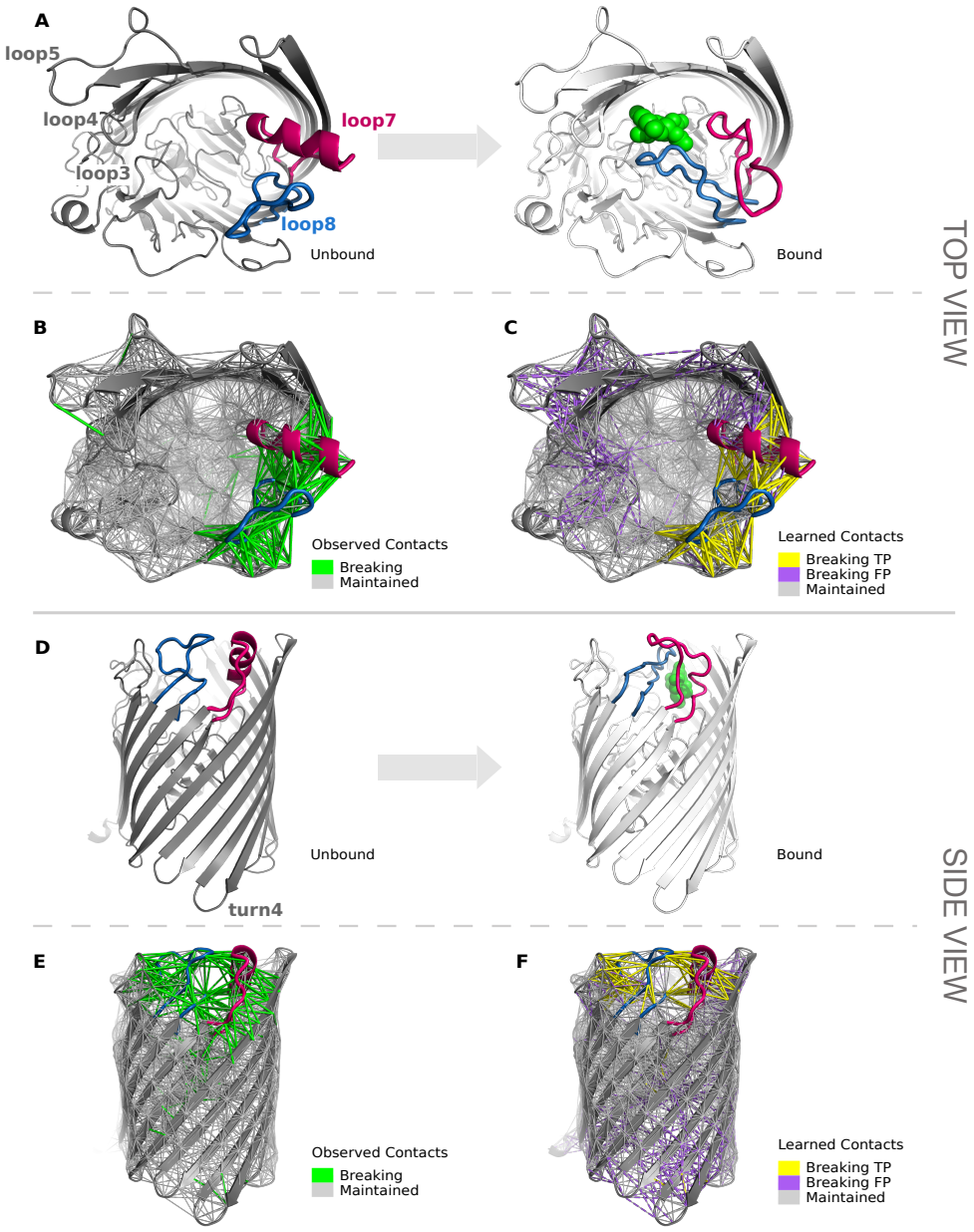

Figure 6.15: Conformational transition of outer membrane transporter FecA compared to observed and learned changes in its contact topology. (A,D): Function-related movement from unbound to bound conformation. The highlighted loops 7 (red) and 8 (blue) move the most to cover the ligand (green spheres) in the binding pocket. (B,E) *Observed* contact network of the unbound conformation mostly residing around the two highlighted loops. (C,F) *Learned* contact network. True positive (TP) predicted breaking contacts accurately match the observed ones around loop 7 and 8. The top view (C) reveals a cluster of false positive (FP, violet) predictions around loops 3, 4, and 5. Between loop 4 and 5 a single breaking contact is observed, which is not predicted. Some more FP breaking contacts are predicted around the plug domain within the β-barrel and turn 4 at the bottom of the barrel (F). For clarity, we omit drawing short-range contacts (sequence separation < 4 residues). Figure source: Putz and Brock (2017).

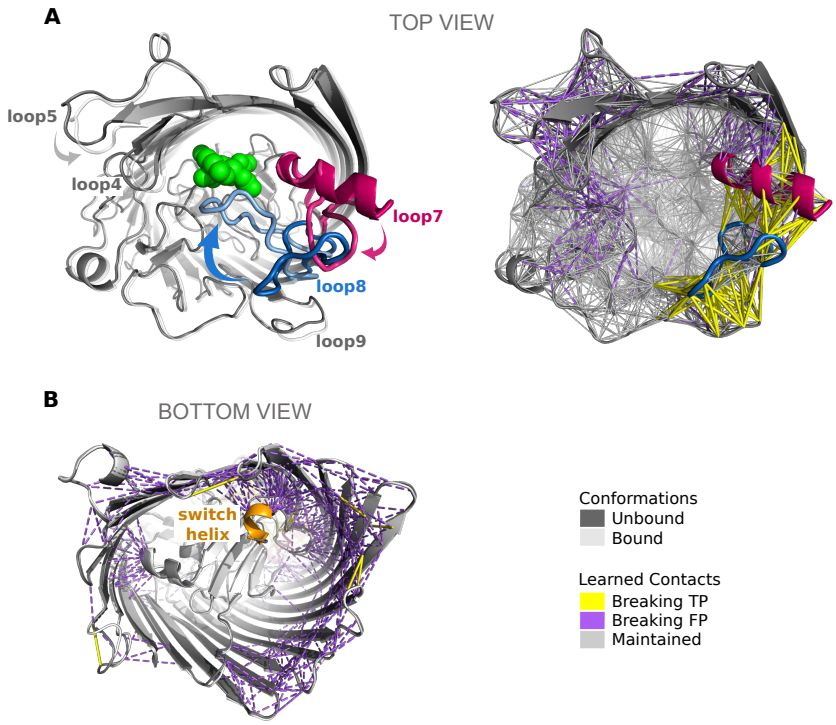

Figure 6.16: Outer membrane transporter FecA: False positive predicted breaking contacts. (A) Top view of the transition between unbound and bound conformation (left) compared to the location of predicted breaking contacts (right). Loops 7 and 8 dominate the conformational change. Most of the predicted breaking contacts around these loops are true positives (TP). Some more breaking contacts are predicted at the opposite side of these loops, but considered false positive predictions (FP). (B) Bottom view of FecA in unbound and bound state. Only the learned breaking contacts are shown for clarity. Many false positive predicted breaking contacts locate around the switch helix (orange). Although the helix retains its shape between the two conformations, MD simulations revealed reversible unwinding of the helix that is involved in the functional behavior of FecA. Hence, these FP predictions may be correct. Some more FP breaking contacts reside between the plug domain and the surrounding β-barrel. Figure source: Putz and Brock (2017).

et al., 2002) as well as fast fluctuations of loop 5 in order to interact with membrane environment and ligand, which have been revealed by MD simulations (Piggot et al., 2013).

Additional false positive predicted contacts locate between plug-domain and β-barrel, as well as around the switch-helix (Fig 6.16(B)). To enable the passage of the ligand through the protein the plug-domain is supposed to move within the β-barrel (Ferguson et al., 2002), yet MD simulations revealed only small positional changes (Piggot et al., 2013). Also, the switch-helix, not captured in the bound conformation, transiently unfolded in MD simulations (Piggot et al., 2013). Taken together, our results suggest that our classifier might generalize much better than indicated by its relatively low prediction accuracy over the full data set (Table 6.4).

We also analyzed how accurate *lmc*ENM predicts the motion directions compared to the other ENM variants. Fig 6.18A shows the cumulative mode overlap of the top 50 lowest-frequency modes. With the first ten modes *lmc*ENM explains more than 60% of the functional transition, an improvement of 40% compared to the baseline ENM and other ENM variants. Only edENM

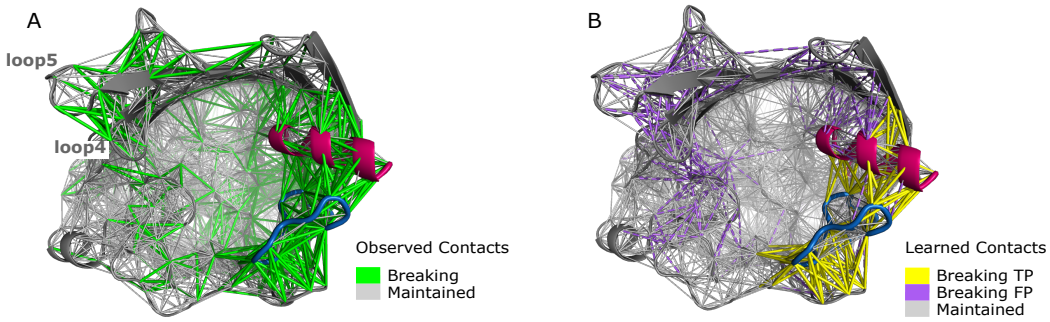

Figure 6.17: Contact networks for outer membrane transporter FecA based on the optimal extension threshold determined for this protein only. (A) Top view of the observed breaking contacts identified based on the optimal extension threshold of 3% that maximizes the cumulative mode overlap of the first ten low-frequency modes of this protein. Please note that the optimal extension averaged over our full data set is 9%. (B) Location of predicted breaking contacts of *lmc*ENM. Given this stricter extension threshold observed and predicted breaking contacts would agree much better, especially around loops 4 and 5. Many of the false positive predictions actually match observed breaking contacts in this case. This indicates that the classifier may has correctly predicted more flexibility in these regions, which is supported by the observed fluctuations for loop 5 in MD simulations. Figure source: Putz and Brock (2017).

captures almost 40% of the movement, but eventually aligns with the other ENMs significantly below *lmc*ENM when considering more modes. This shift of relevant modes towards lower frequencies also becomes evident w.r.t. the lower rank of the best-overlapping mode (*lmc*ENM: 6, ENM: 15, edENM: 7, *mc*ENM: 0) and reduced number of modes required to capture, for instance, 70% of the cumulative overlap (*lmc*ENM: 13, ENM: 187, edENM: 82, *mc*ENM: 3). The improvement of *lmc*ENM over the baseline ENM and the reference ENMs is consistent for all evaluated measures (Table C of supplement S2 in Putz and Brock (2017)).

While the mode overlap of *lmc*ENM seems to be robust against false positive predicted contacts, they clearly have negative impact on the correlation of predicted and observed fluctuation patterns. Fig 6.18B shows that only *mc*ENM is able to capture the observed displacement magnitudes. All other ENM variants, including *lmc*ENM, reach poor agreement with the observed fluctuations. In particular, turns 4 and 3 connecting the strands at the bottom of the β-barrel become way too flexible due to the removed false positive predicted breaking contacts in *lmc*ENM. Also the other ENM variants overestimate the flexibility of these turns. Despite the many true positive breaking contacts around loop 7 the SVM classifier missed relevant observed ones towards the anchor points (Fig 6.15F) and within the helical part, which unfolds completely in the bound conformation. Such an unfolding of helical parts of a loop is currently not explicitly captured by our features. Instead the classifier treats the helix-like part as rather stable.

However, *lmc*ENM has lower tendency to overestimate the flexibility of the other loops and turns than the other ENMs. This together with the closer match of the highly flexible extracellular loop 8 accounts for the slightly higher correlation of *lmc*ENM with the observed fluctuations. One way to reduce the amount of false positive predictions could be to filter the predicted contacts using corroborating evidence. The idea is that predicted breaking contacts close to each other

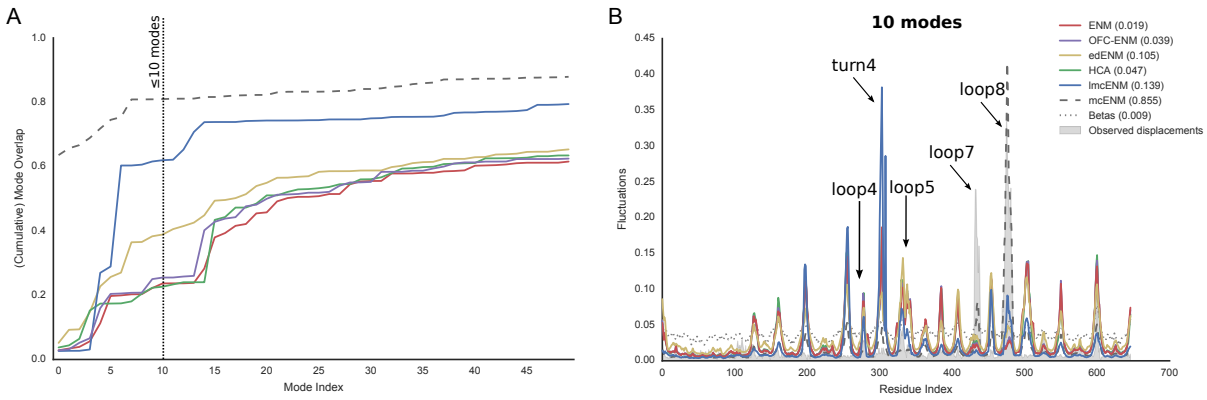

Figure 6.18: Performance of ENM variants for the outer membrane transporter FecA. (A) Reached cumulative overlap (curves) of the first 50 normal modes with the conformational transition. The bars depict how much of the movement individual modes capture. *lmc*ENM largely outperforms the baseline ENM and the reference ENM variants (color coding is the same as in panel B). The vertical dotted line marks the cumulative mode overlaps reached with the first ten low-frequency modes. (B) Residue fluctuations along the first ten low-frequency modes scaled to fit the observed displacement magnitudes (filled gray curve) between the two conformations. The Pearson correlation coefficient is given in brackets behind the ENM labels. *lmc*ENM resembles the higher flexibility of loop 8 more accurately than ENM and other ENM variants, but largely underestimates the flexibility of loop 7. Also, loops 4 and 5 are captured well by *lmc*ENM. But due to the removal of too many false positive predicted breaking contacts (see Fig 6.15F), *lmc*ENM largely overestimates the flexibility of turns 4 and 3 connecting the strands at the bottom of the β-barrel. Figure source: Putz and Brock (2017).

increase their individual likelihood to be a correct prediction. The SVM classifies each contact individually without knowing whether contacts in the neighborhood have been assigned a high probability to break. Such an approach has been successful to filter predicted contacts in an *ab initio* contact prediction approach (Bohlke-Schneider, 2016).

Nonetheless, the overall performance of *lmc*ENM for FecA w.r.t. to all other metrics is remarkable given that it is the only membrane protein in our data set. Even though our SVM-classifier was not specifically trained on membrane proteins it correctly predicted relevant breaking contacts. This indicates that proteins may share similar local structural parts that are involved in similar movements although they differ in their overall structure. In fact, previous work proposed that protein dynamics and deformation patterns may be evolutionary conserved and shared among proteins (Marsh and Teichmann, 2014, Micheletti, 2013, Hensen et al., 2012, Liu and Bahar, 2012). However, further research is required to confirm this hypothesis.

Arachidonate 15-Lipoxygenase - a fatty acids oxidizing enzyme

Arachidonate 15-Lipoxygenase (15S-LOX1) belongs to a class of fatty acids oxidizing enzymes that are involved in inflammatory diseases. Understanding how these enzymes move may advance successful inhibitor design (Choi et al., 2008). 15S-LOX1 is a two-domain protein exhibiting domain *and* local conformational changes. But only the local motions within the larger, catalytic domain enable the ligand binding (Choi et al., 2008). We selected this enzyme for the second case study because *lmc*ENM explains this functional transition even more accurate than *mc*ENM (theoretical maximum) with the first ten low-frequency modes, thereby substantially outperforming all other ENM variants.

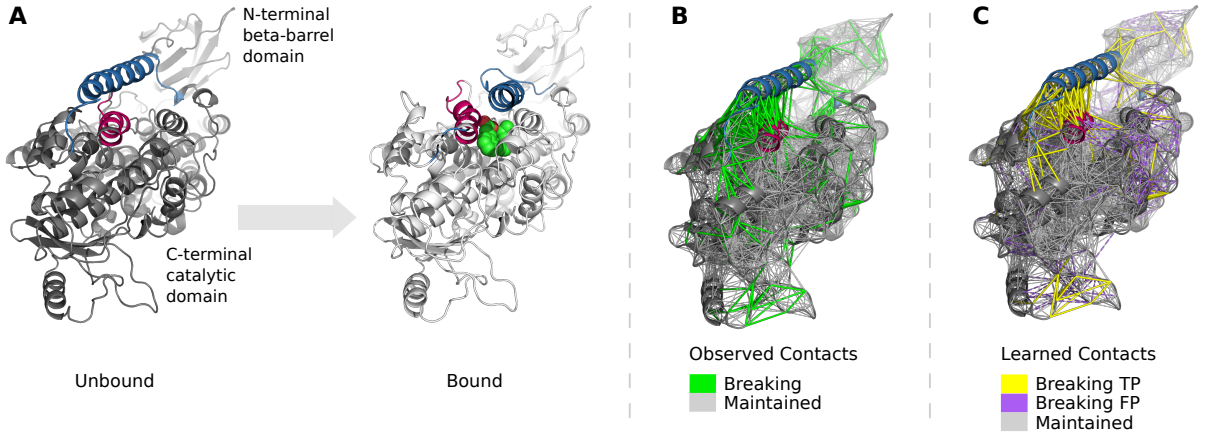

Figure 6.19: Conformational transition of Arachidonate 15-Lipoxygenase compared to observed and learned changes in the contact topology. (A) To accomodate the ligand (green spheres) in the binding site mostly the two highlighted helices (blue and magenta) move between unbound and bound conformation. (B) Most *observed* breaking contacts reside at the interface of the α2-helix (blue) to the rest of the structure. (C) The *learned* breaking contacts match most of the observed ones near the two helices. Most false positive contacts are predicted between the two domains, which seems actually be correct given the high mobility of the N-terminal domain in MD simulations. Figure source: Putz and Brock (2017).

Fig 6.19A depicts unbound and bound conformation (PDB-ids: 2p0mA and 2p0mB (Choi et al., 2008)) of the functional transition. Accomodating the ligand in the narrow pocket mainly requires movement and partial unfolding of the two highlighted helices (proposed induced-fit mechanism) (Choi et al., 2008). Not surprisingly, most observed breaking contacts reside around these helices (Fig 6.19B). Our method correctly predicts most of the observed breaking contacts, but overestimates the occurrence of breaking contacts (false positives, FP) in other parts of the network (Fig 6.19C) and in particular between the two domains (Fig 6.20).

Although the domain motion is not captured by the X-ray conformations, MD simulations reveal large inter-domain movement. Hence, the FP breaking contacts between the two domains seem to be correct. The other false positives around solvent exposed loop regions indicate that our classifier may overemphasize the relevance of such loops.

Fig 6.21A shows the cumulative mode overlap of *lmc*ENM of the first 50 low-frequency normal modes compared to ENM (baseline) and *mc*ENM (theoretical maximum). The first ten *lmc*ENM-

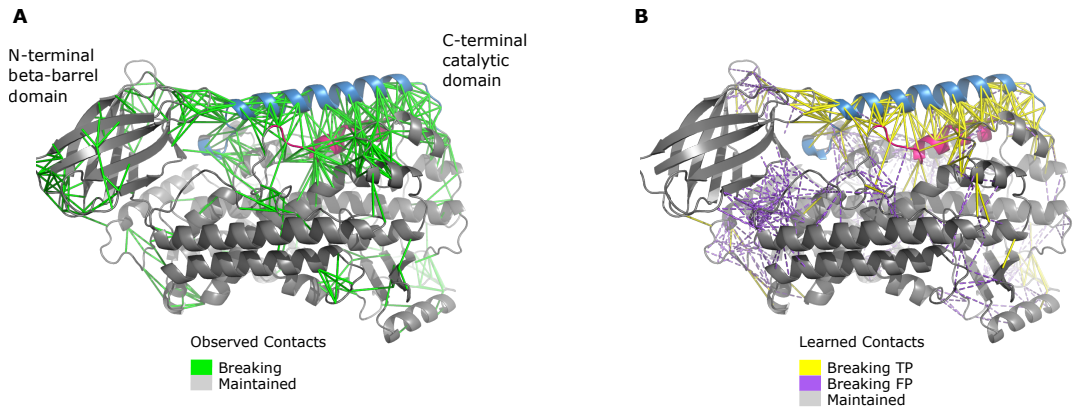

Figure 6.20: Observed and predicted breaking contacts of Arachidonate 15-Lipoxygenase (side view). (A) Most observed breaking contacts reside at the interface of the α2-helix (blue) to the rest of the structure. (B) The learned breaking contacts match most of the observed ones near the helices. Experimental studies indicate that the false positive predictions between the highly flexible N-terminal β-barrel domain and the catalytic C-terminal domain may actually be correct. Figure source: Putz and Brock (2017).

modes capture 89% of the functional transition. With the same number of modes, ENM explains only 29%, *mc*ENM 86% overlap. edENM (43%) and HCA (40%) slightly improve over the baseline ENM. Hence, *lmc*ENM substantially improves over the baseline and the reference ENMs even when considering up to 50 modes.

*lmc*ENM even outperforms *mc*ENM (theoretical upper bound) w.r.t. to the first ten modes. This is surprising because *mc*ENM contains not only the removed false-positive predicted breaking contacts in *lmc*ENM but also lacks observed breaking contacts that have not been detected by *lmc*ENM. The reason is that the three most relevant *lmc*ENM-modes are spread among modes 1, 2, and 4, which account for translation and upwards swinging of the α-helix. The corresponding *mc*ENM-modes distribute among modes 1, 3, and 10. Thus, *lmc*ENM seems to capture the network topology around this helix slightly more accurate than *mc*ENM, maybe due to the missed breaking contacts between the shorter helix (red) and a larger helix (Fig 6.20). As a result *lmc*ENM-modes focus more on the movement of the large helix (blue). Nonetheless, both methods perform about the same when considering more than ten modes.

Fig 6.21B shows that the changed contact topology of *lmc*ENM also accounts for a much better match between predicted and observed residue fluctuations, in particular for the most flexible helix (α2-helix). The other ENM variants, including the baseline ENM, largely underestimate the flexibility of this helix. *lmc*ENM consistently improves over the other ENM variants also w.r.t. all other measures (Table C of supplement S2 in Putz and Brock (2017)).

15S-LOX1 is not the only protein, where *lmc*ENM is more accurate than *mc*ENM. Overall, eight of the 90 proteins in our data set are better captured by *lmc*ENM than by *mc*ENM (see Table C of supplement S2 in Putz and Brock (2017)). This further underlines the potential of our method to explain functional transitions that can not be captured otherwise.

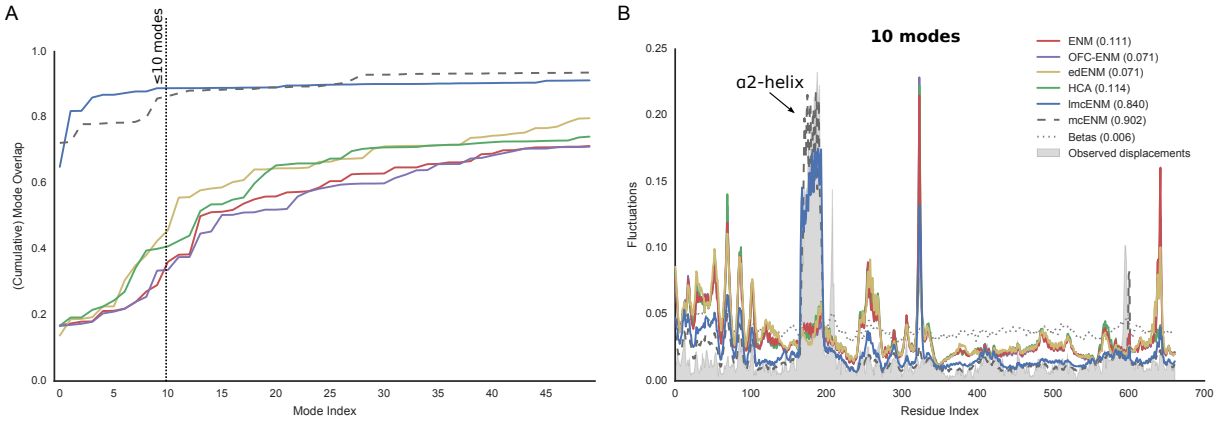

Figure 6.21: Performance of ENM variants for 15S-LOX1 - a fatty acids oxidizing enzyme. (A) Reached cumulative overlap (curves) of the first 50 normal modes with the conformational transition. The bars depict how much of the movement individual modes capture. *lmc*ENM largely outperforms the baseline ENM and the reference ENM variants (color coding is the same as in panel B). The vertical dotted line marks the cumulative mode overlaps reached with the first ten low-frequency modes. (B) Residue fluctuations along the first ten low-frequency modes scaled to fit the observed displacement magnitudes (filled gray curve) between the two conformations. The Pearson correlation coefficient is given in brackets behind the ENM labels. Figure source: Putz and Brock (2017).

SopA - a salmonella effector protein

Another interesting case is the conformational transition of SopA (Diao et al., 2008), a salmonella effector protein (PDB-ids: 2qzaA, 2qyuA). When analyzing this pair of conformations we found that it describes the transition from bound to unbound state instead of the expected transition from unbound to bound. The reason is that 2qyuA–the actual native unbound conformation (Diao et al., 2008)–has been erroneously labeled as the bound conformation in the PSCDB database[1]. While the open-to-closed transition from 2qyuA (unbound) to 2qzaA (bound) is a classical hinge motion, accurately captured by ENMs, our analyzed direction from closed to open is much more difficult to explain by ENMs. This is because ENMs based on closed conformations tend to be overconstrained due to their compact structure. As a result, they often underestimate the mobility of structural parts involved in the opening of the binding pocket to allow the entry of a ligand. Therefore, the closed-to-open transition from 2qzaA to 2qyuA is an interesting test case.

Fig 6.22 shows the conformational transition of SopA together with the different contact networks. Most observed breaking contacts reside in the interface between the moving C-terminal domain and the other two domains (Fig 6.22B). Some more breaking contacts are observed in the lower right corner of the central domain. The predicted breaking contacts capture most of the observed ones, in particular, at the interface between the domains (Fig 6.22C). These true positive breaking contacts are surrounded by false positives. However, these false positives contribute little to the overall deformability of the protein due to the globular and compact shape of the individual domains. Additional false positives are found in less constrained regions of the protein.

[1] http://idp1.force.cs.is.nagoya-u.ac.jp/pscdb/093.html

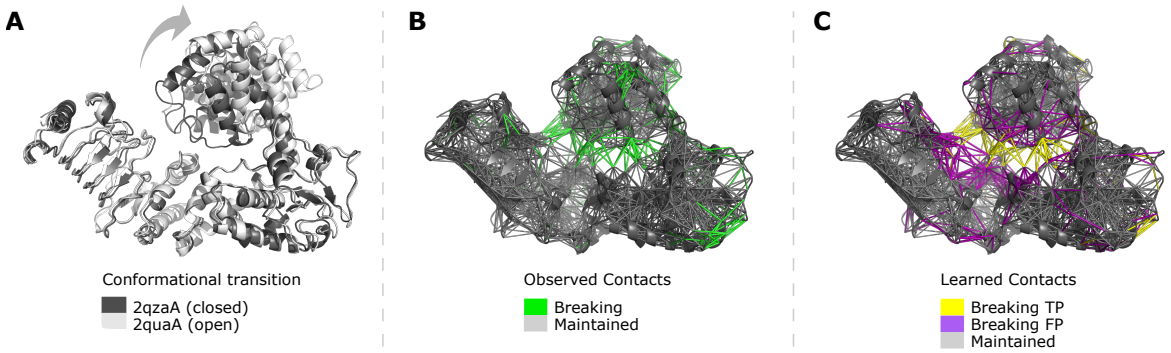

Figure 6.22: Conformational transition from SopA, a salmonella effector protein, and networks with observed and predicted breaking contacts. (A) Hinge-like opening motion from 2qzaA (closed) to 2quaA (open) indicated by the arrow (arrow icon by Michael Kussmaul, Noun Project). (B) Network of maintained contacts with highlighted observed breaking contacts, mostly at the interface between the moving and the other two domains. (C) Network of maintained contacts with highlighted predicted breaking contacts. True positive breaking contacts closely resemble the most relevant observed ones in the domain interface. False positives surround the true positives at the domain interface. However, their effect on the overall deformability is negligible due to the globular and compact structure of the individual domains.

Overall, *lmc*ENM performs as good as *mc*ENM (theoretical upper bound), followed by edENM and HCA. The cumulative mode overlaps of the different ENM variants in Fig 6.23A reveal that eventually all are able to capture the observed conformational change when enough modes are considered (around 20). But we also note that the first two low-frequency modes of *mc*ENM and *lmc*ENM are sufficient to reach an overlap above 0.8, whereas the other ENM variants need more than 13 modes to do so.

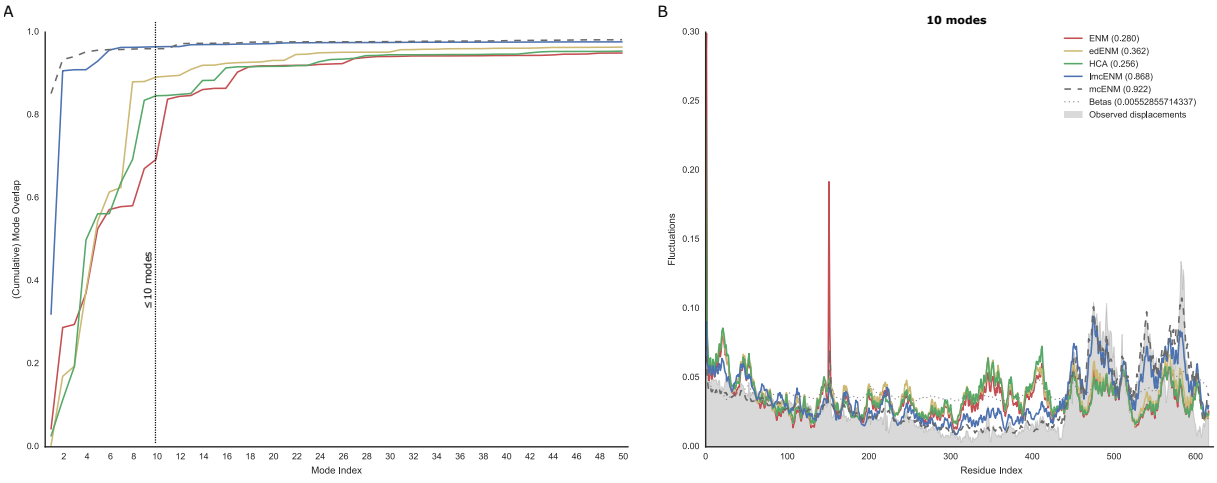

Figure 6.23: Performance of ENM variants for SopA, a salmonella effector protein. (A) Comparison of cumulative mode overlap of the first 50 low-frequency modes (lines). Individual mode overlaps are depicted by the bars. (B) Comparison of fluctuation profiles of the ENM variants with the magnitudes of the observed displacements considering the first ten low-frequency modes.

We also find that *lmc*ENM largely outperforms the other ENM variants considering the match between predicted and actual residue fluctuations (Fig 6.23B). While the other ENM variants over-

or underestimate the fluctuations in large parts, *lmc*ENM resembles them much more accurately. Detailed results for this protein can be found in Table C of supplement S2 in Putz and Brock (2017).

Overall, *lmc*ENM is clearly the best method among the tested ENM variants to capture the conformational transition of SopA. Removing the predicted breaking contacts makes the most relevant hinge-like opening motion of this transition accessible with a single mode, well separated from the others, as it is typical for hinge-like motions. These results suggest that *lmc*ENM is well suited to accurately capture conformational transitions in both directions, i.e. from open to closed **and** vice versa. Furthermore, the direction of the conformational change seems to be irrelevant for the classifier in order to distinguish between potentially breaking and maintained contacts. The required information to correctly predict the relevant breaking contacts is captured by the properties of the local contact environment of either conformation.

6.5 Relevance of Features to Predict Breaking Contacts

Our classifier relies on a broad range of features to differentiate breaking from maintained contacts. As mentioned before, these features characterize the physicochemical, structural and topological properties of the structural context of a contact and its embedding in the protein's structure. Hence, the question arises, which features contribute the most to a correct classification. This section aims to present initial answers to this question based on the most relevant features identified by a feature selection method.

6.5.1 Experimental Setup

One of the fastest methods to select relevant features using SVMs is to rank them by the weights obtained after the classifier was trained on all features (Guyon et al., 2002). This works well for SVMs with linear kernel but not for non-linear Gaussian radial-basis-function (RBF) kernel SVMs, which we used in our approach. However, we found that a linear-kernel SVM as implemented in scikit-learn (Pedregosa et al., 2011, Fan et al., 2008) trained and tested on our problem by Leave-One-Out Cross-Validation performed only slightly worse than the RBF-kernel SVM. Table 6.8 shows the classification performance in terms of precision and coverage for linear and RBF-kernel SVM.

The impact of the different kernels on *lmc*ENM-accuracy is shown in Table 6.9. *lmc*ENM based on the linear SVM performs almost as good as *lmc*ENM based on the RBF-kernel SVM. Thus, the feature weights of the linear SVM should be a reasonable indicator of feature importance in our classification problem.

Table 6.8: Performance of linear SVM (cost=100) and RBF-kernel SVM (cost=100, γ=0.00001) on the full data set (90 proteins). Performance is measured by precision and coverage of the $L/5$ contacts with highest SVM score, where L refers to the length of the protein. The linear SVM performs slightly worse than the lRBF-kernel SVM. Table source: Putz and Brock (2017).

	Precision	Coverage
linear SVM	0.294	0.455
RBF-kernel SVM	0.307	0.470

Table 6.9: Performance of *lmc*ENM based on linear SVM and *lmc*ENM based on RBF-kernel SVM (our presented approach) compared to ENM (baseline), and *mc*ENM (theoretical upper bound) on the full data set (90 proteins). Performance is measured by the cumulative mode overlap of the first ten low-frequency modes. Both *lmc*ENM-variants reach largest overlap when removing the top16% predicted breaking contacts. *lmc*ENM based on linear SVM performs slightly worse. Table source: Putz and Brock (2017).

	ENM	*lmc*ENM (linear SVM)	*lmc*ENM (RBF-kernel SVM)	*mc*ENM
Cumul. Mode Overlap (10)	69/0.66	0.72/0.71	0.73/0.72	0.82/0.80

6.5.2 Results and Discussion

Fig 6.24 shows the 20 features with largest (top) and lowest (bottom) weights. Features with positive weight contribute to identify breaking contacts, whereas features with negative weight help to classify maintained contacts. The magnitude of the weights indicates the importance of the feature. The majority of selected features characterizes topology, spectrum, or label statistics of the neighborhood graph capturing the local context of an individual contact (see Tables A.3-A.9 in appendix A for detailed feature description).

On the extremes of the importance spectrum are number of nodes (positive end) and number of edges (negative end). Contacts in larger but weakly connected (sparse) local neighborhoods are more likely to break than contacts in highly constrained (dense) regions, which have higher probability to be maintained. The latter is also supported by the feature with second largest negative weight, the energy of the immediate neighborhood graph. High graph energy seems to be an important property of maintained contacts because it is usually larger for dense graphs than for sparser ones (Shatto and Çetinkaya, 2017, Li et al., 2012).

The large positive weight of largest and second largest eigenvalue may be interpreted in terms of their gap (Lovász, 2007). Taken individually, they provide not much information, however their gap may hold relevant information about the graph connectivity. Although we did not include this gap as explicit feature, the SVM classifier may have exposed an implicit relation between both pointing towards breaking contacts.

Also, high degree of solvent accessibility and exposure indicate breaking contacts, especially when the impurity degree in the local context is higher. Further, long helices (3D length), a larger amount of turn residues, low amount of hydrogen bonds in the neighborhood, as well as a larger

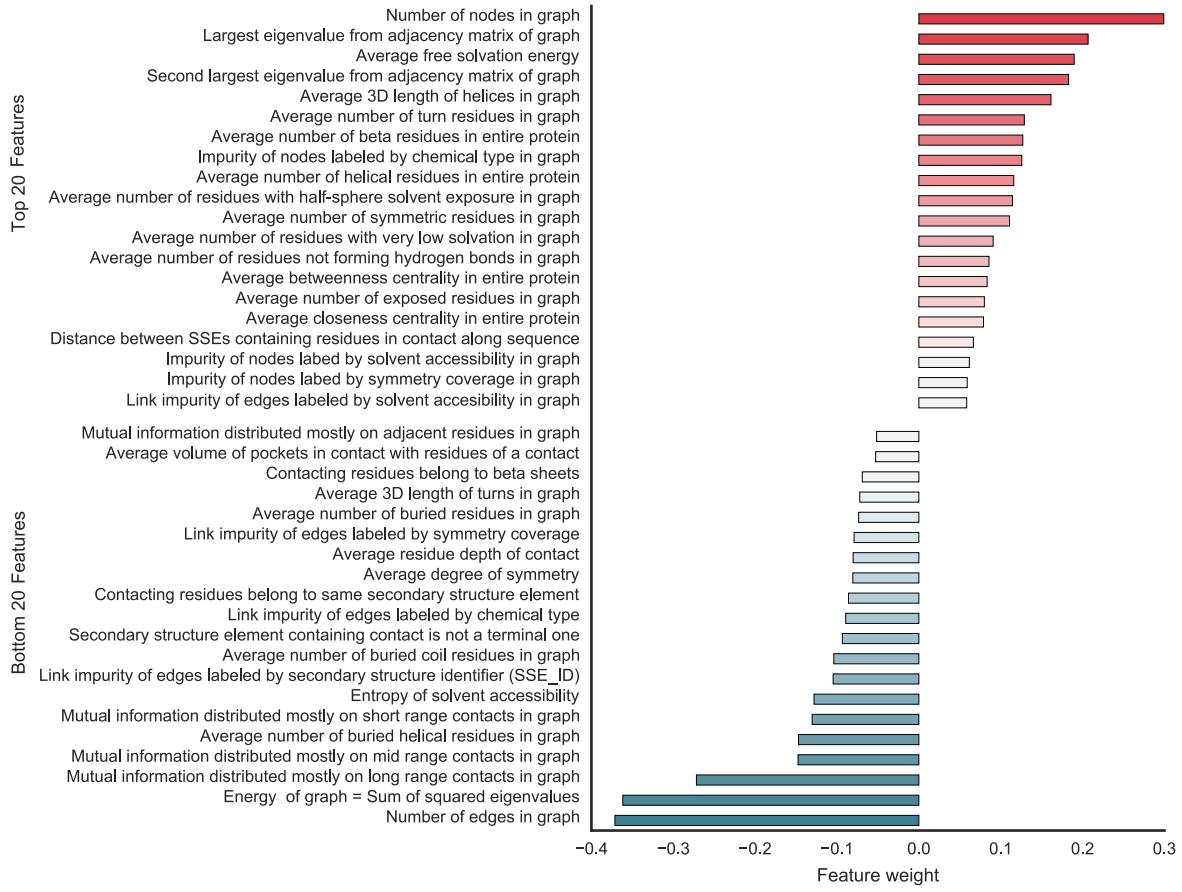

Figure 6.24: Top20 and Bottom20 features ranked by weight of the linearSVM. Features with largest weight are most important to classify breaking contacts, while features with minimum negative weight serve to identify maintained contacts. The graph refers to the neighborhood graph defining the local context of a single contact (see 6.2.1). Features characterizing the different properties of this graph seem to dominate the classification. Figure adapted from Putz and Brock (2017).

sequential distance of the secondary structure elements holding the contact seem to promote its breaking.

On the contrary, maintained contacts seem to populate rather buried neighborhoods (number of buried (helical/coil) residues, entropy of solvent accessibility, residue depth) with high degree of sequence conservation (mutual information distribution). In fact, Liu and Bahar (2012) have shown that there is a strong link between sequence conservation and intrinsic deformability for enzymes. Although some sequence correlations may be irrelevant for protein dynamics, certain amino acids involved in substrate recognition tend to be both, more mobile, while also coevolve more often. This points towards a breaking contact, whereas high sequence conservation rather characterizes maintained contacts.

We also find that a high degree of symmetry leads to enhanced structural stability (maintained contacts) in the symmetric parts, while weakly attached parts are more likely to move to facilitate a functional transition. The outer membrane transporter, FecA, presented in the case study

above (6.4.7) exemplifies the effect of stable symmetric core allowing motion within the barrel as well as at the entrances.

The average 3D length of turns intuitively measures the spatial extension of a turn. Largely extended turns or coils are restricted in their mobility due to stronger interactions with the neighborhood along their full length, which indicates rather maintained contacts.

Being in contact with pockets of larger volume, also seems to be associated with maintained contacts. A contact with a pocket is established if at least one of the contacting residues touches the surface of one of the alpha spheres characterizing the pocket's shape as determined by fPocket (Guilloux et al., 2009). A possible explanation could be that large pockets may tend to maintain their shape and hence the contact topology. Breaking contacts are more likely to be found at the pocket entrance to accommodate for ligand binding.

One might argue that several of our features are captured by other approaches. For instance, residue depth or solvent exposure of a contact are implicitly modeled by its embedding into a highly or weakly constrained part of an ENM, respectively. Also the influence of contact order, secondary structure type, and hydrogen bonding have been used to refine ENMs (see (Lezon and Bahar, 2010, Orellana et al., 2010, Jeong et al., 2006), for instance). However, Fig 6.24 reveals that only the topmost feature as well as the three bottommost features are clearly separated from the other features in terms of their weight/importance, whereas the importance of the other features shows a much smaller spread. In fact, the ranking and weight of these features slightly varies for the different motion categories (Table A.7). This has two implications: First, to reliably predict dynamic changes in the coarse-grained model of a protein and thereby its motions a broader set of features should be considered instead of only a few ones. Second, depending on the protein this specific feature/property combination may also vary. Both effects may be difficult to capture implicitly by modeling specific interactions, such as hydrogen bonds or disulfide bridges.

Overall, the strongest of our features seems to be the graph-based encoding of the local contact environment itself. With the presented feature set it holds valuable information about protein dynamics and can easily be extended by additional features. Yet, to improve the classifier by removing irrelevant features and to gain deeper understanding about features driving protein motion more advanced methods for features selection such as recursive feature elimination (SVM-RFE) (Guyon et al., 2002) could be used. Such methods also provide information about the importance and interplay of feature groups as opposed to their individual importance, which was analyzed by our approach. Nonetheless, the interplay of features is partially captured by our approach. We use the neighborhood graph of a contact to combine individual properties of its environment into aggregated features.

6.6 Conclusion

6.6.1 Summary

In the previous chapter we demonstrated that ENMs are able to capture localized, function-related motions if they account for dynamic changes in the contact topology of proteins. In particular, we showed that the absence of springs associated with observed breaking contacts makes localized functional transitions accessible for ENMs. The goal in this chapter was to present a way to *predict* the dynamic behavior to be able to adjust the ENM network in the standard case, when only a single protein conformation is known.

We proposed to *predict* the dynamic behavior of contacts, i.e. whether they break or are maintained when the protein moves, by leveraging information from their structural context. This context is built by the local contact environment and its embedding into the overall protein structure. We demonstrated that differentiating breaking from maintained contacts is possible by using a graph-based encoding of their structural context. We introduced an SVM classifier to predict breaking contacts using features derived from this graph-based representation. They characterize the physicochemical, structural, and topological properties of a contact's structural context. Our results show that the predicted breaking contacts closely resemble the observed ones, especially for proteins with localized function-related movements.

We also proposed that accounting for these *predicted* dynamic changes in the contact topology of proteins expands the range of motions that can be modeled by ENMs. We demonstrated this with *lmc*ENM, a novel elastic network model of *learned* maintained contacts, which remain after removing the predicted breaking ones from the initial network. Our results show that the predicted breaking contacts are relevant and accurate enough to substantially improve the accuracy of *lmc*ENM over the reference ENM variants, in particular for proteins that require localized movements to perform their function.

We have also seen that due to the absence of predicted breaking contacts *lmc*ENM requires a substantially smaller subset of low-frequency modes to accurately capture localized functional transition than the reference ENMs. This has two implications: First, searching a lower-dimensional space reduces computational costs for normal-mode-guided conformational exploration or ensemble generation. Second, the essential deformation space of *lmc*ENM more likely guides towards the right direction, in particular when a protein performs localized functional transitions with low degree of collectivity. This is of high practical relevance because the type of motion exhibited by proteins is usually unknown a priori.

*lmc*ENM confirms the findings of chapter 5 that accounting for dynamic contact changes, i.e. the absence of predicted breaking contacts, is key to capture localized functional transition with ENMs. There is no need to adjust network resolution or potential function. Nonetheless, preliminary results indicate that the accuracy of *lmc*ENM may be further improved by optimizing spring stiffness similar to other approaches (Orellana et al., 2010, Lezon and Bahar, 2010, Kovacs et al., 2004, Hinsen et al., 2000).

In contrast to *mc*ENM introduced in the previous chapter, *lmc*ENM can be applied in the standard prediction case, where only a single conformation of a protein is known. Overall, our results demonstrate that without increasing the complexity of the underlying model, *lmc*ENM offers a promising route towards improving the general applicability of ENMs and thereby their practical relevance.

Finally, we presented evidence that the dynamic behavior of contacts, and thus protein motion, most likely results from the interplay of a broader set of features characterizing the properties of their structural context, which may be difficult to be encoded into the ENM model implicitly. We introduced an easily extensible framework for exploring additional features to further advance our understanding of protein motion.

6.6.2 Limitations

*lmc*ENM relies on the same model resolution and uniform spring stiffness as the classical distance-cutoff based ENM. Hence, the computational costs to analyze their deformations, i.e. the intrinsic motions of a protein, are the same. However, *lmc*ENM requires additional computation to predict the breaking contacts to adjust its contact network. The actual amount of computation largely depends on the protein's size and summarizes over two steps: feature generation and contact prediction.

Feature generation ranges from a few minutes for small proteins (> 100 residues) up to half an hour for our largest protein, FecA, with 647 residues on a single CPU. The prediction step is much faster taking seconds for small proteins up to six minutes for FecA. Nonetheless, the gain in accuracy of *lmc*ENM should compensate for these additional computational costs. Only training of the classifier is computationally more intense. But in principle, it has to be done only once and runs parallelized. A web service to run *lmc*ENM for single-chain proteins is currently in preparation.

Furthermore, the effectiveness of *lmc*ENM is currently mostly focused on proteins with local motions *coupled* to ligand binding as shown in Fig 6.25. It reaches about two-third of the theoretical maximum accuracy achieved by *mc*ENM (see also Table N of supplement S2 in Putz and Brock (2017)). For proteins with independent local motions, *lmc*ENM is able to capture about half of them better than ENM, whereas the other half reaches only small if any improvement over ENM (baseline). Also, domain movers cannot benefit from *lmc*ENM to the extent as local movers, mostly due to the removal of too many predicted breaking contacts.

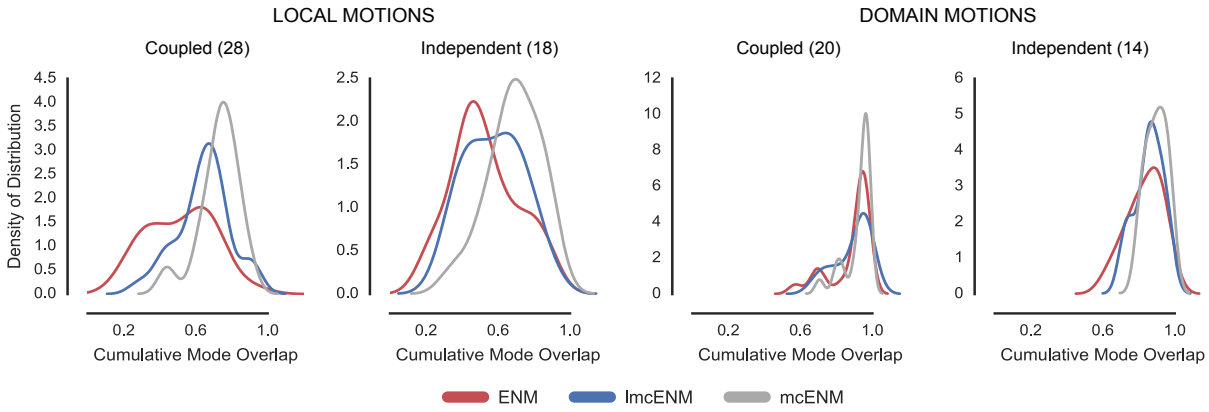

Figure 6.25: Distribution of *lmc*ENM-, *mc*ENM-, and ENM-accuracy, subset of local and domain motions (80 proteins). *lmc*ENM closely resembles the accuracy distribution of *mc*ENM (theoretical upper bound) for proteins with coupled local motions and domain motions. But, it only slightly improves in accuracy for proteins with independent local motions. Nonetheless, *mc*ENM clearly demonstrates that also the latter type of motions can be captured with high accuracy with a refined contact topology.

Despite these limitations, our results clearly demonstrate that the absence of predicted breaking contacts enables *lmc*ENM to explain otherwise poorly captured localized functional transitions. This further underlines the potential of our approach to further expand the range of motion types that can be modeled by ENMs.

7 Conclusion

7.1 Summary of Main Findings

We presented in this thesis a novel elastic network model based on learned maintained contacts, *lmc*ENM (chapter 6), which addresses a major shortcoming of elastic network models (ENMs). ENMs exploit the fact that a protein's motions are largely encoded in its contact topology. While ENMs accurately explain functional transitions of proteins that are large-scale and collective, they fail to capture localized, uncorrelated ones. Hence, the movements predicted by an ENM may be wrong or misleading, which limits the practical relevance of ENMs because the motion type of proteins is in general unknown a priori.

*lmc*ENM overcomes this limitation by leveraging a novel source of information, i.e. dynamic changes in the contact topology of a protein. *lmc*ENM refines its initial network by removing springs associated with contacts that have been predicted to break during the motion. To predict these contacts we developed a machine-learning based classifier that differentiates breaking from maintained contacts by leveraging information about their structural context, which influences their dynamic behavior. Our approach is a first step towards a "deformation-invariant" contact topology to study protein motions of any type on a coarse-grained scale.

Our approach is based on two key insights: First, the ability of ENMs to capture function-related transitions critically depends on a contact topology that remains maintained throughout the movement. While this is naturally fulfilled for highly collective movements, localized functional transitions often cause substantial changes in the contact topologies between start and end conformation. We showed that ENMs can accurately capture these localized movements if *observed* breaking contacts are removed from their initial contact topology (chapter 5). But, to *predict* protein motions with ENMs we also need to *predict* these breaking contacts.

Second, the additional information required to predict breaking contacts is hidden in the physicochemical characteristics of local parts of the protein structure. These characteristics capture how tightly different parts of the protein are bound to each other, how this affects their

movements, and ultimately their contact topology. We presented a way to access this novel source of information using a graph-based encoding of a contact's structural context and features that characterize the properties of this environment.

We showed that *lmc*ENM predicts function-related protein motions more accurate than the classical, distance-cutoff based ENM and three other reference ENM variants. *lmc*ENM is particularly effective in capturing ligand-coupled localized functional transitions that remain largely unexplained by all reference ENMs.

Furthermore, we showed that *lmc*ENM reduces the complexity of the deformation space relevant to capture function-related movements. This has also implications for subsequent applications, such as generating conformational ensembles for protein-ligand docking, which often involves localized, functional transitions. These applications utilize the deformation space spanned by the lowest-frequency modes as guidance. Hence, they may benefit from a lower dimensional space that reduces the computational costs for sampling.

We presented further evidence that protein motion likely results from the interplay of a broader set of properties/features characterizing the mobility of local structural parts. We also believe that combining different information sources (e.g. conformational ensembles obtained by MD, NMR, X-ray, or other experimental methods) will make the identification of relevant properties even more robust and accurate than relying on a single source alone. With our presented approach we provide a novel, unified, and extensible way to examine, exploit and relate additional features captured by each of these information sources in order to further advance our understanding of protein motion.

Last, my thesis unlocks breaking contacts, or generally dynamic contact changes, as a novel source of information that has proven valuable in coarse-grained prediction of protein motion. Because they are defined on a simplified model of the structural connectivity of a protein, they are insensitive to structural details that would otherwise make their identification and prediction more difficult. This makes them a valuable target for future research that aims to improve coarse-grained prediction of protein motion.

7.2 Future Work

In the following we discuss ideas to further advance the approaches presented in this thesis and potential applications that would benefit from the additional information leveraged by our methods.

7.2.1 Advancing the Proposed Methods

We now propose several ideas to improve the proposed methods in this thesis.

Additional Graph Features

To predict breaking contacts we use graph features that characterize the physicochemical properties of their structural context. In our current implementation the structural context of a contact considers its local environment build by its immediate neighbors, and its embedding within the overall structure captured by the secondary structure elements graph. However, proteins can share the same fold, i.e. similar arrangement of secondary structure, which influences its structural stability, mobility, and function. Features based on the type of secondary structure arrangement could help to better distinguish maintained from breaking contacts.

Another idea is to devise features that capture the "freedom" of secondary structure elements to move or deform. Obviously, elements in the core of the protein have less freedom to move than elements at the surface of the protein. Incorporating these or similar features would improve the characterization of breaking and maintained contacts and thereby lead to more accurate predictions.

Additional Information Sources

We used start and end conformation of a functional transition to identify breaking and maintained contacts. Thereby we ensured that the conformational change is functionally relevant and large enough to affect the contact topology of the protein. Nonetheless, conformational ensembles obtained by experimental methods (X-ray or NMR) or sampled by computational approaches (MD, Monte Carlo, geometric sampling (Greener et al., 2017)) could make the identification of breaking contacts even more robust, for instance by incorporating occurrence statistics (Eickholt et al., 2011) as additional features. However, care must be taken to preserve the discriminative power of the current implementation. Contacts wrongly labeled as breaking could reduce the prediction performance of the classifier.

Representation Learning on Graphs

With the advent of deep learning representation learning on graphs became highly popular (Goodfellow et al., 2016, Xie et al., 2016, Hamilton et al., 2018). Instead of time-consuming, error-prone manual feature engineering, such algorithms learn a lower-dimensional embedding of the structure of a graph on their own. An additional advantage of these algorithms is that they can be used for semi-supervised learning, where only a small portion of training instances are labeled. This reduces the risk of over-fitting, which is often observed for supervised learning approaches. Graph convolutional neural networks have recently be applied in the context of disease predictions (Parisot et al., 2018), e.g. Alzheimer's.

Corroborating Evidence

Our classifier predicts breaking contacts in isolation, i.e. it has no knowledge whether the contacts close to this contacts will also be classified as breaking or not. As we have seen in chapter 6 breaking contacts are only effective if they occur together and reach a "critical mass" that results in a substantial change in the network topology of ENMs. One way to tackle this is to use *corroborating evidence* between predicted breaking contacts to filter out false positive predictions. Such an approach has been used in the context of contact prediction for protein structure prediction and to refine restraints from cross-linking experiments (Bohlke-Schneider, 2016).

Optimizing Spring Stiffness

The two ENMs proposed in this thesis (see chapter 5 and 6) purely alter their underlying contact topology by removing observed/predicted breaking contacts. This choice was made on purpose to demonstrate that the ability to explain localized funtion-related changes by ENMs depends on an accurate contact topology. The substantial improvement of our ENMs compared to the reference ENMs showed that this cannot be achieved by optimizing spring stiffness. Nonetheless, preliminary experiments revealed that combining our ENMs with, for instance, edENM (Orellana et al., 2010) would lead to smaller improvements w.r.t. their current prediction accuracy.

Larger Data Set and Multimers

Although *lmc*ENM substantially improves in accuracy based on the predicted contact changes, the prediction accuracy of the classifier itself is rather low (see chapter 6). This has two obvious reasons: First, the classifier accuracy is measured on the selected top-scoring predicted breaking contacts based on a cutoff that is not necessarily optimal for domain proteins. This has a negative effect on the overall accuracy of the classifier. Second, our data set is relatively small due to specific requirements, such as significant conformational change, annotation by motion type, or restriction to single chain proteins (see 4.1). While this ensures high quality of the data set and low bias towards either local or domain motions, more data, including multimers, would certainly help to improve the accuracy of the classifier, to reduce bias, and to expand the applicability of our approach. Nonetheless, constructing such a data set is a laborious and time-consuming step and there is no guarantee that the test set fits to the evaluation set (Jonschkowski, 2018).

7.2.2 Potential Applications of *lmc*ENM

Above we showed that *lmc*ENM alleviates a major shortcoming of ENMs being less suited to capture localized functional transitions with low degree of collectivity. By removing predicted breaking contacts, *lmc*ENM substantially improves the prediction accuracy for proteins performing local function-related movements. As a result, *lmc*ENM largely increases the chances that a protein's motion is accurately modeled no matter if it performs a local or domain motion, thereby expanding the practical relevance of ENMs. In the following we will discuss some potential applications of ENMs that could benefit from using *lmc*ENM and the predicted breaking contacts.

Generating Conformational Ensembles for Protein Ligand Docking

A logical first step would be to apply *lmc*ENM in the context of protein ligand docking. The ability of ENM to capture collective protein motions with only a few modes allows to narrow down the accessible deformation space of the unbound conformation (Bahar et al., 2010a, López-Blanco and Chacón, 2016). Hence, conformational sampling in this reduced space not only requires less computation, but also increases the chances to sample good candidate conformations for the actual docking. Kurkcuoglu and Doruker (2016), for instance, generate such a pool of candidate conformations by applying an iterative scheme of deforming a protein structure along most dominant normal modes and subsequently minimizing its energy to reduce unrealistic structural distortions in order to prepare it for the next round of NMA analysis.

However, Dietzen et al. (2012) showed that in small-protein docking conformational ensembles generated by sampling along ENM-modes often yield no improvement. The major obstacle seems to be that existing ENM variants often fail to capture the localized movements associated with ligand binding by the first few low-frequency modes. Although usually a few modes (less than ten) suffice to explain local transitions they are often spread among higher frequencies (Cavasotto et al., 2005). This makes it difficult to decide how many modes should be included to accurately sample the relevant deformation space.

Our results show that *lmc*ENM effectively reduces the essential deformation space for localized functional transitions in most cases. Thus, it would be interesting to see whether a subset of for instance the first 20 low-frequency modes of *lmc*ENM would improve small-molecule docking. In addition, *lmc*ENM may also be helpful for the most difficult cases involving induced-fit movements that are triggered by the presence of a ligand. Training a SVM classifier specifically on such protein pairs may help to shift the most relevant *lmc*ENM-modes toward lower frequencies. This would alleviate the problem of identifying the relevant modes for a specific ligand because they already reside in the low-frequency mode spectrum.

Guiding Conformational Sampling

For the same reasons, *lmc*ENM could also provide more accurate guidance for more fine-grained conformational sampling, especially for larger proteins performing localized functional transitions. Gur et al. (2013), for instance, sample candidate structures in the deformation space spanned by the first few low-frequency normal modes. The candidate structures serve as starting points for MD runs that generate physically accurate conformations at full atomic detail.

This guided sampling can be used, for instance, to explore the conformational space accessible to the unbound conformation, to predict transition pathways between two end points of a function-related movement, or to refine low-resolution experimental model (Costa et al., 2015). Alternatively, ENM-modes have also been used to guide robotics-based sampling methods (Kirillova et al., 2008, Al-Bluwi et al., 2013, Shehu and Plaku, 2016) to explore the conformational space with reduced computational costs.

The quality of the guidance obviously depends on the accuracy of the predicted lowest-frequency modes. *lmc*ENM offers a way to improve this guidance for proteins exhibiting localized functional transitions that are difficult to capture by existing ENM variants.

Constructing Multi-Scale Models

The predicted breaking contacts to construct *lmc*ENM may also be useful when constructing multi-scale ENMs, such as RCNMA (Ahmed and Gohlke, 2006), to predict motions of large proteins or complexes at a coarse-grained scale. While the occurrence of predicted breaking contacts reveals parts of the network requiring higher resolution, their absence indicates parts that could be further simplified. This would help to analyze only relevant parts of the protein and their motions in more detail, thereby reducing computational demands. However, to explore this further we would first need to extend our SVM prediction framework to accept multi-chain proteins and optimize the feature generation part in our pipeline to reduce computation time of breaking contact prediction.

Predicting Targets for Elastic Network based Interpolation of Motion Pathways

Another interesting application for *lmc*ENM would be in the context of predicting pathways between start and end of functional transitions with a two-state ENM such as proposed by Das et al. (2014). Based on the ENMs of the two endpoints they construct a combined potential that allows to transition from one state to the other via an low-energy path. Such methods obviously require the knowledge of start and end conformation of a functional transition. However, in case the actual target conformation is unknown, the predicted network of learned maintained contacts of *lmc*ENM could be used as an estimate of the coarse-grained representation of the target conformation. Nonetheless, the prediction accuracy of the current SVM classifier may need to be improved before attempting such an experiment.

7.3 Epilogue

Leveraging information about dynamic changes in the contact topology substantially impacts the prediction capabilities of elastic network models: it expands the range of protein motions that can be explained by elastic network models, thereby improving their practical relevance.

Key to leveraging relevant information is that its extraction must be guided by the right problem domain, which means in our case: To improve a simplified model that aims to predict protein motions we need to exploit information about its dynamic changes. Or more concrete in the context of elastic network models and their underlying contact topology: Which properties of local parts of the protein structure promote breaking contacts when the protein moves and which do not.

In this thesis we identified breaking contacts and the physicochemical characteristics of their structural context as valuable source of information to incorporate knowledge about their dynamic behavior into elastic network models. We believe that further sources of information are readily available and should be studied.

This thesis provides a novel, unified, and extensible way to examine, exploit and relate additional features captured by each of these information sources. This will help to further improve coarse-grained prediction of protein motion, which ultimately advances our understanding of protein motion.

A

Appendix - Features for Breaking Contact Prediction

The following detailed description of features used to predict the dynamic behavior of inter-residue contacts, i.e. whether they break or are maintained when the protein moves, has been previously published in the supporting material of the following paper:

Putz I, Brock O (2017) *Elastic network model of learned maintained contacts to predict protein motion.* PLOS ONE 12(8): e0183889. https://doi.org/10.1371/journal.pone.0183889

A.1 Graphs for modeling physicochemical context

To model the local contact environment of breaking and maintained contacts we use the graph-based encoding developed by Schneider and Brock (2014). Nodes in the contact graph refer to residues, which are connected by edges if they are in contact. Schneider and Brock (2014) predict native-like contacts from ab *initio* predicted candidate structures (decoys) using physicochemical information. They rely on a quite general definition of node and edge labels that also applies to our problem domain. However, we extend these labels by more specific ones in our context.

Tables A.1 and A.2) summarize the node and edge labels and mark which labels are reused, extended, or novel. For convenience, we also recap the explanations of the reused and extended labels introduced in Schneider and Brock (2014).

Table A.1: Summary of node labels. Table source: Putz and Brock (2017).

Node label	Possible labels
Chemical type[a]	Discrete value
Secondary structure[a]	Discrete value
Solvent accessibility[a]	Discrete value
Free solvation energy[a]	Continuous value
Secondary structure length[a]	Discrete value
Secondary structure 3D length[a]	Continuous value
Secondary structure buried[a]	Continuous value
Secondary structure exposed[a]	Continuous value
Hydrogen bonding[b]	Discrete value
Distance to the centroid[a]	Continuous value
Sequence conservation[a]	Continuous value
Sequence neighborhood conservation[a]	Continuous value
Secondary structure unique ID	Continuous value
Part of symmetric element	Discrete value
Depth	Continuous value

[a]Reused node label from Schneider and Brock (2014).
[b]Extended node label from Schneider and Brock (2014).

A.1.1 Node labels

Chemical type: A residue can be non-polar, polar, acidic, or basic.

Secondary structure: A residue can be part of a helix, sheet, turn, or coil.

Solvent accessibility: distinguishes between buried and exposed residues. The former have a relative solvent accessibility (calculated by POPS (Cavallo et al., 2003)) $\leq 25\%$, while the latter are above this cutoff.

Free solvation energy: calculated by POPS (Cavallo et al., 2003).

Secondary structure length: Length of the secondary structure element associated with the residue, measured along the sequence.

Secondary structure 3D length: Three-dimensional distance (in Å) between first and last residue of the secondary structure element calculated between their C_α atoms.

Secondary structure buried: Specifies how buried the residue's secondary structure element is based on its average number of buried residues.

Secondary structure exposed: Specifies how exposed the residue's secondary structure element is based on its average number of exposed residues.

HYDROGEN BONDING: Residue can be donor, acceptor or not part of a hydrogen bond.

DISTANCE TO THE CENTROID: Three-dimensional distance (in Å) between the C_α atom of the residue and the centroid of the protein structure.

SEQUENCE CONSERVATION: Specifies the degree of sequence conservation of the residue obtained from a multiple-sequence alignment (based on (Janda et al., 2013, Fischer et al., 2008)).

SEQUENCE NEIGHBORHOOD CONSERVATION: Specifies the degree of sequence conservation within the local neighborhood of the residue up to three sequence positions away ($i-3$, $i-2$ $i-1$, $i+3$, $i+2$ $i+1$) as in (Janda et al., 2013, Fischer et al., 2008).

Secondary structure unique ID: Unique identifier of the secondary structure element the residue belongs to.

Part of symmetric segment: Specifies whether the residue is part of symmetric segment in protein (calculated by SymD (Kim et al., 2010)).

Structural depth: Specifies the depth of the residue w.r.t. the solvent accessible surface by averaging the distance of its atoms to the surface vertices (calculated with BioPython (Cock et al., 2009)).

Half Sphere Exposure: Specifies the degree of exposure of a residue by counting the contacts within the upper and lower half-sphere (default radius 12Å) around the residue's C_α atom. The sphere is cut into two halves by a plane centered at the C_α-atom, which is perpendicular to the vector between the C_α- and a pseudo-C_α-atom. (calculated with BioPython (Cock et al., 2009) based on (Hamelryck, 2005)).

A.1.2 Edge labels

Table A.2: Summary of edge labels. Table source: Putz and Brock (2017).

Edge label	Possible labels
3D distance[a]	Continuous value
Mutual information[a]	Continuous value

[a]Reused node label from Schneider and Brock (2014).

3D distance: Specifies how far away the two residues of the contact are in 3D by calculating the distance between their C_α.

Mutual information: The mutual information in the multiple-sequence alignment between the two residue positions of the contact.

A.2 Features listing and implementation details

The dynamic behavior of contacts depends on their immediate local context as well as their embedding into the overall arrangement of local structural parts. The physicochemical interactions between these parts control their movement with respect to each other, which ultimately influences the contact topology between them. We use a set of features to characterize the properties of the local neighborhood of a contact as well as its associated secondary structure elements. Feature can be a single real-value input or encode categorical properties by a set of binary values. Unless stated otherwise, a categorical feature with k states is encoded by an k-dimensional binary input vector. The individual features are concatenated into a single vector that serves as input for the SVM to differentiate breaking from maintained contacts. This feature vector has a total length of 170 input values.

We designed features specific to our problem domain, as well as reuse or extend features used in previous work from our group (Schneider and Brock, 2014). The features fall into eight categories: Pairwise, graph topology, graph spectrum, single node, node label statistics, edge label statistics and whole protein features. In the following we introduce in detail the features in each category and mark all reused or extended features respectively.

A.2.1 Pairwise residue features

Pairwise features encode properties of an individual contact. As contacts seldom change their distance in isolation, many of the pairwise features are defined on their associated secondary structure element(s) (SSEs). Table A.3 lists the individual features together with their number of input values in the feature vector. For all features that are not self-explanatory a detailed description of the feature and its generation is given in the text.

Distance between secondary structures elements along protein chain (SSE): Distance between relative index positions of SSEs associated with residue i and j along the protein chain (1 inputs).

SSE-contact type: Contacts can be within a SSE (intra-SSE) or between (inter-SSE). While the dynamic behavior of intra-SSE contacts mostly depends on the intrinsic flexibility of the SSE, inter-SSE contacts are influenced by the strength of the interface (1 input). Intrinsic SSE-flexibility or SSE-interface strength are characterized by the following features.

SSE-interface contact position: Position of an inter-SSE contact within the SSE-interface. Contacts located at the border of the interface have a higher probability to break than more central ones. Contacts with at least one residue in the border region (outer 10% of interface length measured in residue position) are encoded by [1,0], core contacts by [0,1] (2 inputs).

SSE-interface hydrogen bonding: Fraction of hydrogen bonds in SSE-interface relative to total number of hydrogen bonds in the structure. 0 for intra-SSE contact. (1 input)

SSE-interface density: Density of SSE-interface indicating the degree of connectedness of the two SSEs. Strongly connected SSE-interfaces are likely to be maintained. The interface of two SSEs can be represented as a bipartite graph, where the two disjunct node sets refer to the interface-residues of the two SSEs. The density of the SSE-interface is then calculated as actual number of contacts between these two node sets divided by the maximal possible number of contacts of a fully connected bipartite graph. 0 for intra-SSE contact. (1 input).

SSE-interface degree: Averaged number of interface connections of residues i and j, respectively. Indicates how much the contacting residues contribute to the SSE-interface strength. 0 for intra-SSE contact. (1 input).

Table A.3: Pairwise features between contacting residues i and j or their associated secondary structure elements (SSEs). Table source: Putz and Brock (2017).

Feature	Description	Number of inputs
Distance between SSEs along protein chain	Relative distance between position of SSEs along protein chain	1
Centroid distance between SSEs	3D-distance between centroids of SSEs	1
SSE-contact type	Contact within same SSE (intra-SSE) or between different SSEs (inter-SSE)	1
SSE-interface hydrogen bonding	See text	1
SSE-interface contact position	See text	4[a]
SSE-interface density	See text	1
SSE-interface balanced	See text	1
SSE-interface degree	See text	1
SSE-interface redundancy	See text	1
SSE-intra hydrogen bonding	See text	1
SSE-intra degree	See text	1
Contact with highest ranked pocket	See text	3[a]
Contact with a pocket	See text	3[a]
Exposure to pocket	See text	1
Polarity of pocket	See text	1
Hydrophobicity of pocket	See text	1
Druggability score of pocket	See text	1
Volume of pocket	See text	1
Side chain contact	See text	1
Contact depth	See text	4[a]
Residue depth difference	See text	4[a]
SSE-symmetry coverage	See text	1
Contact symmetry coverage	See text	3
Distance to symmetry plane	See text	6
Contact between terminal SSEs	See text	5
Secondary structure type[b]	Secondary structure of the contacting residues: helix, sheet, turn, coil	10[a]
Hydrogen bonding[c]	See text	3[a]
Mutual information[b]	Sequence mutual information	1
Total inputs		63

[a]Binary inputs
[b]Reused from Schneider and Brock (2014).
[c]Extended from Schneider and Brock (2014).

SSE-INTERFACE REDUNDANCY: Fraction of other contacting residues of residues i or j in SSE-interface that would remain in contact in a one-mode projection even if i or j would be removed (Latapy et al., 2008). This could be viewed as a measure of the importance of individual SSE-interface residues to maintain the connectivity of the interface. 0 for intra-SSE contact (1 input).

SSE-interface balanced: Equal number of residues participating in SSE-interface on both sides or not. Indicates whether the SSE-interface has "exposed" contacts with rather long distance (imbalanced SSE-interface). Due to a lower degree of connectivity in their neighborhood such contacts are more likely to break than shorter, highly constrained ones. 0 for intra-SSE contact. (1 input).

Contact between terminal SSEs: Terminal regions of the protein chain often possess more flexibility. This feature captures if the contacting residues belong to one of the two terminal secondary structure element along the protein chain (5 binary inputs in total). We distinguish five cases: If the contact is an intra-SSE contact, the SSE can be terminal or not (2 binary inputs). If the contact is between different SSEs, both SSEs can be terminal, only one, or none (3 binary inputs).

SSE-intra hydrogen bonding: Fraction of hydrogen bonds within SSE relative to total number of hydrogen bonds of the protein. Hydrogen bonds increase the stability of a SSE-interface. 0 for inter-SSE contact (1 input).

SSE-intra degree: Averaged intra-SSE degree of residue i and j, i.e. the number of contacts within SSE of each residue. α-helices, for instance, have a lower probability to unfold compared to loops that have fewer internal constraints. 0 for intra-SSE contact (1 input).

Contact with highest ranked pocket: This feature captures if both residues, i and j, are "in-contact" with the highest ranked pocket, or only one of the residues, or none. (3 inputs). Location and properties of pockets used to generate all pocket-related features are calculated with FPocket (Guilloux et al., 2009). FPocket reports detected pockets ranked by a probability score to be the functional-active binding site. Contacts around the binding pocket have higher propensity to change. They may be involved in movements to accomodate the ligand in the binding site or to shield it from the solvent.

Contact with a pocket: This feature captures if both residues, i and j, are "in-contact" with any detected pocket, or only one of the residues, or none. (3 inputs). Being in touch with any pocket, not necessarily the binding pocket, increases the changes for secondary structure elements to move into this "free space". Such movements may be required to propagate, for instance, allosteric signals and may result in changes of the local contact topology.

Exposure to pocket: Sum of atom contacts to closest pocket of residues i and j. A high number of atom contacts indicates that the residue extends into the pocket, i.e. has a high degree of exposure into the pocket (1 input).

Polarity of pocket: Average polarity of pocket(s) in contact with residues i and j (1 input). The polarity of a pocket is a measure for its hydrophilicity (calculated with FPocket (Guilloux et al., 2009)).

Hydrophobicity of pocket: Average hydrophobicity of pocket(s) in contact with residues i and j (1 input). This feature measures the degree of hydrophobicity of a pocket (in contrast to the polarity above, calculated with FPocket (Guilloux et al., 2009)).

Druggability score of pocket: Average druggability score of pocket(s) in contact with residues i and j (1 input). This score estimates the probability of a pocket to bind small drug like molecules (calculated with FPocket (Guilloux et al., 2009))

Volume of pocket: Average volume of pocket(s) in contact with residues i and j (1 input, calculated with FPocket (Guilloux et al., 2009)).

Side chain contact: Captures if contact is also a side chain contact. Two residues are in side chain contact if at least one pair of heavy side chain atoms is within 4.5Åadistance to each other (1 input).

Contact depth: Captures the depth of a contact based on the normalized structural depth of its residues (4 inputs). The residue depth is binned into four states: really deep (lower than 0.25), deep (between 0.25 and 0.5), exposed (between 0.5 and 0.75) and very exposed (larger than 0.75). If the depth class of the contacting residues differs, the class of the deeper residue determines the contact depth. Deeply buried contacts are more likely to be maintained than contacts close to the surface.

Residue depth difference: Measures the binned difference in structural depth of the contacting residues. The normalized depth difference bins are: $\Delta_{\text{depth}} < 0.25, 0.25 <= \Delta_{\text{depth}} < 0.5, 0.5 <= \Delta_{\text{depth}} < 0.75, \Delta_{\text{depth}} > 0.75$. Depending on the actual contact depth, also the difference in depth of the contacting residues may influence the contact's dynamic behavior. A large depth difference increases the chances that a contact may break (4 inputs).

SSE-symmetry coverage: (Average) fraction of residues being part of symmetric structural parts of the SSE(s) associated with the contact (see section A.1.1). Symmetric parts of a protein structure often stabilize the overall fold. For instance, β-barrels, β-sheets, or mixed α-β-barrels (TIM-barrels), are strongly stabilized by hydrogen bonds. Hence, contacts between SSEs involved in symmetric arrangements are likely to be maintained. (1 input).

Contact symmetry coverage: Captures if both residues of a contact are part of symmetric segment (see section A.1.1), only one or none (3 inputs).

Distance to symmetry plane: This feature captures the distance of the contacting residues to the symmetry plane (6 binary inputs in total). Both residues can be far apart (normalized distance ≥ 0.7) from the symmetry plane either on the positive or negative side (2 binary inputs). Or both residues are on either positive or negative side, but only one is far away from the symmetry plane (2 binary inputs). Or one residue is on the positive and the other one on the negative side (1 binary input). If no symmetry plane exists all except the last input are 0. Contacts closer to the core of a symmetric part of the protein structure may benefit more from its higher stability. Hence, they are likely to be maintained. However, contacts close to the border of a symmetric part, such as a β-barrel, may have a higher chance to be involved in the functional activity of the protein. We discuss an example of a highly symmetric membrane protein in detail in case study II (see Results and Discussion in main document).

Secondary structure: The secondary structures types (helix, sheet, turn or coil) of the protein are obtained with STRIDE (Frishman and Argos, 1995) (10 inputs).

Solvent accessibility: The solvent accessibility of residues is classified into solvent exposed or buried (see section A.1.1) (3 inputs).

Hydrogen bonding: Residues of a contact can be bonded by an hydrogen bond, or be donor or acceptor of another hydrogen bond, or not involved in hydrogen bonding. A.1.1) (3 inputs).

Mutual information: The mutual information in the multiple-sequence alignment between positions i and j (1 input).

A.2.2 Graph features

Our work is based on the assumption that breaking contacts and maintained contacts show differences in the properties of their local neighborhood. To specify these differences we re-use the graph-topology (Tab. A.4), graph spectrum (Tab. A.5), and single node features (Tab. A.6) from Schneider and Brock (2014). These topological features and node/edge label statistics characterize the properties of the local context of a contact defined by its immediate neighborhood graph (see section 6.2.1) and help us to distinguish breaking from maintained contacts.

For convenience we list the re-used features in the following tables (Tables A.4, A.5,and A.6). We introduced the graph-theoretic basics of these features in section 3.1. Some more details can be found in the original publication (Schneider and Brock, 2014) as well as their source Li et al. (2012).

Table A.4: Graph topology features[a]. Table source: Putz and Brock (2017).

Feature	Description	Number of inputs
Number of nodes	Number of nodes in the graph	1
Number of edges	Number of edges in the graph	1
Average degree centrality	Average number of node neighbors indicating packing density of graph.	1
Average closeness centrality	Average reciprocal distance of each node to all other nodes in the graph.	1
Average betweenness centrality	Average number of shortest paths passing through each node of the graph indicating the degree of influence of individual nodes onto the network.	1
Average eccentricity	Average maximum distance between each node and all other nodes in the graph.	1
Graph radius	Smallest eccentricity in the graph.	1
Graph diameter	Largest eccentricity in the graph.	1
Number of end points	Number of nodes with only one neighbor.	1
Average clustering coefficient	Average number of actual neighbors divided by possible neighbors of each node in the graph measuring the degree of transitivity in the network.	1
Total inputs		10

[a]Reused from Schneider and Brock (2014).

Table A.5: Graph spectrum features derived from the adjacency matrix[a]. Table source: Putz and Brock (2017).

Feature	Description	Number of inputs
Largest eigenvalue	Largest eigenvalue	1
Second largest eigenvalue	Second largest eigenvalue	1
Number of different eigenvalues	Number of different eigenvalues	1
Sum of eigenvalues	Trace of the adjacency matrix	1
Energy	Sum of squared eigenvalues	1
Total inputs		5

[a]Reused from Schneider and Brock (2014).

Table A.6: Single node features[a]. Table source: Putz and Brock (2017).

Feature	Description	Number of inputs
Degree centrality	Number of node neighbors of node i and j.	2
Closeness centrality	Reciprocal average distance from nodes i and j to all other nodes in the graph.	2
Betweenness centrality	Number of shortest paths that pass through nodes i and j.	2
Sequence separation from N/C-terminus	Distance in sequence position of i to N-terminus and j to C-terminus	2
Sequence conservation	Conservation of residue position of i and j in multiple sequence alignment.	2
Sequence neighborhood conservation	Conservation of neighboring residues of i and j in multiple-sequence alignment.	2
Total inputs		12

[a]Reused from Schneider and Brock (2014).

Node label statistics

Node label statistics, listed in Table A.7, capture the frequency of different node labels in the graph.

Average degree of symmetry: The degree of symmetry for a single node is the fraction of neighbor nodes that belong to symmetric segments. Average degree of symmetry is the average over all nodes in the graph (1 input).

Neighborhood impurity degree: Normalized number of neighbor nodes with different labels in the graph. Schneider and Brock (2014) evaluated the neighborhood impurity degree for the node labels chemical type, secondary structure, solvent accessibility. We extend this node label list by unique secondary structure identifier (SSE_ID), symmetry coverage, large positive distance to symmetry plane, large negative distance to symmetry plane (7 inputs).

Edge label statistics

Edge label statistics, listed in Table A.8, capture the frequency of different edge labels in the graph.

Table A.7: Node label statistics. Table source: Putz and Brock (2017).

Feature	Description	Number of inputs
Symmetry coverage	Average number of nodes covered by symmetry	1
Average degree of symmetry	Average fraction of node neighbors that are covered by symmetry for all nodes in the graph	1
Residue depth	Average residue depth in graph	1
Residue depth distribution	5-bin distribution of residue depth in graph	5
Average half-sphere exposure	Average lower/upper half-sphere exposure in graph	2
Neighborhood impurity degree[b]	Average number of neighbors with different labels	7
Hydrogen bonding[b]	Average numbers of nodes that act as donor, acceptor or do not form hydrogen bonds	3
Label entropy[a]	Entropy of the different labels, calculated for chemical type, secondary structure, and solvent accessibility.	3
Chemical type[a]	Number of polar, non-polar, acidic, basic labels	4
Secondary structure distribution[a]	Number of nodes with helix, sheet, turn, coil labels	4
Secondary structure length[a]	Average length of secondary structure element in amino acids	4
Secondary structure 3D length[a]	Average 3D length of secondary structure element	4
Secondary structure buried[a]	Average number of buried residues per ss_type (helix, sheet, turn, coil)	4
Secondary structure exposed[a]	Average number of exposed residues per ss_type (helix, sheet, turn, coil)	4
Solvent accessibility[a]	Average number of exposed/buried nodes	2
Average solvation energy[a]	Average free solvation energy	1
Solvation energy distribution[a]	4-bin distribution of free solvation energy	4
Distance to centroid[a]	Average distance of nodes to the centroid	1
Sequence conservation[a]	Average sequence conservation of nodes	1
Sequence neighborhood conservation[a]	Average sequence neighborhood conservation of nodes	1
Total inputs		57

[a]Reused from Schneider and Brock (2014).
[a]Extended from Schneider and Brock (2014).

LINK IMPURITY: Normalized number of edges between nodes with different labels in the graph. Schneider and Brock (2014) evaluated the link impurity for the edge labels chemical type, secondary structure, solvent accessibility. We extend this edge label list by unique secondary structure identifier (SSE_ID), symmetry coverage, large positive distance to symmetry plane, large negative distance to symmetry plane (7 inputs).

Table A.8: Edge label statistics. Table source: Putz and Brock (2017).

Feature	Description	Number of inputs
Link impurity[b]	Number of edges connecting two nodes with different labels	7
Mutual information distribution[a]	5-bin distribution of mutual information	5
Cumulative mutual information[a]	Cumulative mutual information over all edges	1
Total inputs		13

[a]Reused from Schneider and Brock (2014).
[b]Extended from Schneider and Brock (2014).

Mutual information distribution: Fraction of edges between nodes with different ranges of sequence separation (adjacent, 2-6, 7-11, 12-23, >24), yielding a 5-bin distribution of the mutual information of the graph. (5 inputs).

A.2.3 Whole protein features

These features characterize global properties of the whole protein (Table A.9). We reused one feature from Schneider and Brock (2014) in this category, which is marked.

Table A.9: Whole protein features. Table source: Putz and Brock (2017).

Feature	Description	Number of inputs
Secondary structure composition[b]	Distribution of secondary structure types in protein (helix, sheet, turn, coil).	4[a]
Connectivity class	Binned number of contacts of protein (<500, 501-1000, 1001-2000,2001-3000 >3000).	5[a]
Symmetry coverage	Normalized number of residues in symmetric segments.	1
Total inputs		10

[a]Binary inputs
[b]Reused from Schneider and Brock (2014).

B

Appendix - Supporting Information

B.1 Supplementary Table

Table B.1: Performance overview of *lmc*ENM compared to baseline ENM and *mc*ENM (theoretical upper bound), as well as three other reference ENM variants on the whole protein data set (90 proteins). The performance is measured by the cumulative mode overlap (CO) of the first ten low-frequency modes. *lmc*ENM consists of the learned maintained contacts after removing the top16% predicted breaking contacts. For each ENM variant the median and mean CO is reported of the proteins grouped by their motion types (coupled/independent local motions (CLM and ILM), coupled/independent domain motions (CDM and IDM), burying ligand motions (BLM), and other types of motions (OTM)). The number of proteins in each category is given in brackets after the motion labels. The last row reports the average values for all proteins. While the other ENM variants perform about the same, *lmc*ENM clearly improves in capturing localized functional transitions over their common baseline, thereby reaching almost half of the improvement made by *mc*ENM. Table source: Putz and Brock (2017).

Motion types	ENM	OFC-ENM	edENM	HCA	*lmc*ENM	*mc*ENM
Coupled Local Motions (28)	0.53/0.52	0.57/0.54	0.55/0.53	0.58/0.54	0.66/0.64	0.74/0.73
Independent Local Motions (18)	0.48/0.53	0.49/0.53	0.46/0.51	0.48/0.53	0.58/0.58	0.68/0.69
Coupled Domain Motions (20)	0.94/0.88	0.94/0.88	0.94/0.89	0.94/0.88	0.94/0.89	0.96/0.92
Independent Domain Motions (14)	0.85/0.83	0.85/0.83	0.87/0.85	0.85/0.85	0.85/0.86	0.90/0.90
Burying Ligand Motions (4)	0.75/0.75	0.77/0.75	0.81/0.80	0.78/0.77	0.75/0.76	0.87/0.88
Other Types of Motions (6)	0.62/0.61	0.63/0.62	0.64/0.60	0.64/0.60	0.65/0.60	0.82/0.76
All (90)	0.69/0.66	0.67/0.67	0.68/0.67	0.68/0.68	0.73/0.72	0.82/0.80

Bibliography

mwaskom/seaborn: v0.8.1 (September 2017) | Zenodo.

Aqeel Ahmed and Holger Gohlke. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. *Proteins: Structure, Function, and Bioinformatics*, 63(4):1038–1051, 2006.

Aqeel Ahmed, Saskia Villinger, and Holger Gohlke. Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses. *Proteins: Structure, Function, and Bioinformatics*, 78(16):3341–3352, 2010.

Ibrahim Al-Bluwi, Thierry Siméon, and Juan Cortés. Motion planning algorithms for molecular simulations: A survey. *Computer Science Review*, 6(4):125–143, 2012.

Ibrahim Al-Bluwi, Marc Vaisset, Thierry Siméon, and Juan Cortés. Modeling protein conformational transitions by a combination of coarse-grained normal mode analysis and robotics-inspired methods. *BMC structural biology*, 13 Suppl 1:S2, 2013.

Réka Albert. Network Inference, Analysis, and Modeling in Systems Biology. *The Plant Cell*, 19(11): 3327–3338, 2007.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, New York, 4th edition edition, 2002.

Andrea Amadei, Marc A. Ceruso, and Alfredo Di Nola. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteinsṁolecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, 36(4):419–424, 1999.

Takayuki Amemiya, Ryotaro Koike, Sotaro Fuchigami, Mitsunori Ikeguchi, and Akinori Kidera. Classification and Annotation of the Relationship between Protein Structural Change and Ligand Binding. *Journal of Molecular Biology*, 408(3):568–584, 2011.

Takayuki Amemiya, Ryotaro Koike, Akinori Kidera, and Motonori Ota. Pscdb: a database for protein structural change upon ligand binding. *Nucleic acids research*, 40:D554–D558, 2012.

R. Anand, Kiran Dip Gill, and Abbas Ali Mahdi. Therapeutics of Alzheimer's disease: Past, present and future. *Neuropharmacology*, 76:27–50, 2014.

Christian B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, 1973.

Pierre Arnold, Dominique Peeters, and Isabelle Thomas. Modelling a rail/road intermodal transportation system. *Transportation Research Part E: Logistics and Transportation Review*, 40(3):255–270, 2004.

A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1):505–515, 2001.

Andrea Avena-Koenigsberger, Bratislav Misic, and Olaf Sporns. Communication dynamics in complex brain networks. *Nature Reviews Neuroscience*, 19(1):17–33, 2018.

Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

Ivet Bahar, Timothy R. Lezon, Ahmet Bakan, and Indira H. Shrivastava. Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chemical Reviews*, 110(3): 1463–1497, 2010a.

Ivet Bahar, Timothy R. Lezon, Lee-Wei Yang, and Eran Eyal. Global Dynamics of Proteins: Bridging Between Structure and Function. *Annual review of biophysics*, 39:23–42, 2010b.

Ivet Bahar, Mary Hongying Cheng, Ji Young Lee, Cihan Kaya, and She Zhang. Structure-Encoded Global Motions and Their Role in Mediating Protein-Substrate Interactions. *Biophysical Journal*, 109(6):1101–1109, 2015.

Ahmet Bakan, Lidio M. Meireles, and Ivet Bahar. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics*, 27(11):1575–1577, 2011.

Danielle S. Bassett, Nicholas F. Wymbs, Mason A. Porter, Peter J. Mucha, Jean M. Carlson, and Scott T. Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641–7646, 2011.

Ugo Bastolla. Computing protein dynamics from protein structure with elastic network models. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(5):488–503, 2014.

Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology. *PLoS computational biology*, 4(10):e1000173, 2008.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

Robert B. Best, Kresten Lindorff-Larsen, Mark A. DePristo, and Michele Vendruscolo. Relation between native ensembles and experimental structures of proteins. *Proceedings of the National Academy of Sciences*, 103(29):10901–10906, 2006.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN 978-0-387-31073-2.

Csaba Böde, István A. Kovács, Máté S. Szalay, Robin Palotai, Tamás Korcsmáros, and Péter Csermely. Network analysis of protein dynamics. *FEBS Letters*, 581(15):2776–2782, 2007.

Michael Bohlke-Schneider. *Leveraging novel information sources for protein structure prediction*. PhD thesis, 2016.

Karsten M. Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schönauer, S. V. N. Vishwanathan, Alex J. Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1): i47–i56, 2005.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

Christoph Bostedt, Sébastien Boutet, David M. Fritz, Zhirong Huang, Hae Ja Lee, Henrik T. Lemke, Aymeric Robert, William F. Schlotter, Joshua J. Turner, and Garth J. Williams. Linac Coherent Light Source: The first five years. *Reviews of Modern Physics*, 88(1):015007, 2016.

Ulrik Brandes. *Network Analysis: Methodological Foundations*. Springer Science & Business Media, 2005. ISBN 978-3-540-24979-5.

Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998.

B. Brooks and M. Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, 80:6571–6575, 1983.

Andries E. Brouwer and Willem H. Haemers. *Spectra of Graphs*. Springer Science & Business Media, 2011. ISBN 978-1-4614-1939-6.

Rafael Brüschweiler. Collective protein dynamics and nuclear spin relaxation. *The Journal of Chemical Physics*, 102(8):3396–3403, 1995.

Michal Brylinski and Jeffrey Skolnick. What is the relationship between the global structures of apo and holo proteins? *Proteins: Structure, Function, and Bioinformatics*, 70(2):363–377, 2008.

I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis. High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing. *Journal of Chemical Information and Modeling*, 50(3):397–403, 2010.

Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

Prasad V. Burra, Ying Zhang, Adam Godzik, and Boguslaw Stec. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10505–10510, 2009.

Vincenzo Carnevale, Francesco Pontiggia, and Cristian Micheletti. Structural and dynamical alignment of enzymes with partial structural similarity. *Journal of Physics: Condensed Matter*, 19(28):285206, 2007.

Luigi Cavallo, Jens Kleinjung, and Franca Fraternali. Pops: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic acids research*, 31:3364–3366, 2003.

Claudio N. Cavasotto. Normal Mode-Based Approaches in Receptor Ensemble Docking. In Riccardo Baron, editor, *Computational Drug Discovery and Design*, number 819 in Methods in Molecular Biology, pages 157–168. Springer New York, 2012. ISBN 978-1-61779-464-3 978-1-61779-465-0.

Claudio N. Cavasotto, Julio A. Kovacs, and Ruben A. Abagyan. Representing Receptor Flexibility in Ligand Docking through Relevant Normal Modes. *Journal of the American Chemical Society*, 127 (26):9632–9640, 2005.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Jianlin Cheng and Pierre Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC bioinformatics*, 8(1):113, 2007.

Jongkeun Choi, Jae Kyung Chon, Sangsoo Kim, and Whanchul Shin. Conformational flexibility in mammalian 15s-lipoxygenase: Reinterpretation of the crystallographic data. *Proteins: Structure, Function, and Bioinformatics*, 70(3):1023–1032, 2008.

Koollawat Chupradit, Sutpirat Moonmuang, Sawitree Nangola, Kuntida Kitidee, Umpa Yasamut, Marylène Mougel, and Chatchai Tayapiwatana. Current Peptide and Protein Candidates Challenging HIV Therapy beyond the Vaccine Era. *Viruses*, 9(10), 2017.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

Thomas E. Cope, Timothy Rittman, Robin J. Borchert, P. Simon Jones, Deniz Vatansever, Kieren Allinson, Luca Passamonti, Patricia Vazquez Rodriguez, W. Richard Bevan-Jones, John T. O'Brien, and James B. Rowe. Tau burden and the functional connectome in Alzheimer's disease and progressive supranuclear palsy. *Brain*, 141(2):550–567, 2018.

Thomas H. Cormen. *Introduction to Algorithms*. MIT Press, 2009. ISBN 978-0-262-03384-8.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Mauricio G. S. Costa, Paulo R. Batista, Paulo M. Bisch, and David Perahia. Exploring free energy landscapes of large conformational changes: molecular dynamics with excited normal modes. *Journal of chemical theory and computation*, 11:2755–2767, 2015.

Peter Csermely, Tamás Korcsmáros, Huba J. M. Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology & Therapeutics*, 138(3):333–408, 2013.

Avisek Das, Mert Gur, Mary Hongying Cheng, Sunhwan Jo, Ivet Bahar, and Benoît Roux. Exploring the Conformational Transitions of Biomolecular Systems Using a Simple Two-State Anisotropic Network Model. *PLOS Computational Biology*, 10(4):e1003521, 2014.

Charles C. David and Donald J. Jacobs. Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. *Methods in molecular biology (Clifton, N.J.)*, 1084:193–226, 2014.

Andrea Di Luca and Ville R. I. Kaila. Global collective motions in the mammalian and bacterial respiratory complex i. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1859(5):326–332, 2018.

Jianbo Diao, Ying Zhang, Jon M. Huibregtse, Daoguo Zhou, and Jue Chen. Crystal structure of SopA, a Salmonella effector protein mimicking a eukaryotic ubiquitin ligase. *Nature Structural & Molecular Biology*, 15(1):65–70, 2008.

Matthias Dietzen, Elena Zotenko, Andreas Hildebrandt, and Thomas Lengauer. On the Applicability of Elastic Network Normal Modes in Small-Molecule Docking. *Journal of Chemical Information and Modeling*, 52(3):844–856, 2012.

Ken A. Dill and Hue Sun Chan. From Levinthal to pathways to funnels. *Nature Structural & Molecular Biology*, 4(1):10–19, 1997.

Ken A. Dill and Justin L. MacCallum. The protein-folding problem, 50 years on. *Science (New York, N.Y.)*, 338:1042–1046, 2012.

Ying Ding. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, 5(1):187–203, 2011.

Sara E. Dobbins, Victor I. Lesk, and Michael J. E. Sternberg. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proceedings of the National Academy of Sciences*, 105(30):10390–10395, 2008.

Nicolas Dony, Jean Marc Crowet, Bernard Joris, Robert Brasseur, and Laurence Lins. SAHBNET, an Accessible Surface-Based Elastic Network: An Application to Membrane Protein. *International Journal of Molecular Sciences*, 14(6):11510–11526, 2013.

Jan Drenth. *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

Ron O. Dror, Robert M. Dirks, J. P. Grossman, Huafeng Xu, and David E. Shaw. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annual Review of Biophysics*, 41 (1):429–452, 2012.

Jacob D. Durrant and J. Andrew McCammon. Molecular dynamics simulations and drug discovery. *BMC Biology*, 9(1):71, 2011.

David K. Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

Ashkan Ebadi and Andrea Schiffauerova. How to become an important player in scientific collaboration networks? *Journal of Informetrics*, 9(4):809–825, 2015.

Jesse Eickholt, Zheng Wang, and Jianlin Cheng. A conformation ensemble approach to protein residue-residue contact. *BMC Structural Biology*, 11:38, 2011.

Eran Eyal, Lee-Wei Yang, and Ivet Bahar. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics*, 22(21):2619–2627, 2006.

Eran Eyal, Gengkon Lum, and Ivet Bahar. The anisotropic network model web server at 2015 (anm 2.0). *Bioinformatics (Oxford, England)*, 31:1487–1489, 2015.

Kurt Faber. Biocatalytic applications. In *Biotransformations in organic chemistry*, pages 31–313. Springer, 2018.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal Of Machine Learning Research*, 9:1871–1874, 2008.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Andrew D. Ferguson, Ranjan Chakraborty, Barbara S. Smith, Lothar Esser, Dick van der Helm, and Johann Deisenhofer. Structural Basis of Gating by the Outer Membrane Transporter FecA. *Science*, 295(5560):1715–1719, 2002.

J. D. Fischer, Christian E. Mayer, and Johannes Söding. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, 24(5):613–620, 2008.

Naomi K. Fox, Steven E. Brenner, and John-Marc Chandonia. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(D1):D304–D309, 2014.

Vincent Frappier and Rafael J. Najmanovich. A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations. *PLOS Computational Biology*, 10(4):e1003569, 2014.

Hans Frauenfelder, Stephen G. Sligar, and Peter G. Wolynes. The Energy Landscapes and Motions of Proteins. *Science*, 254(5038):1598–1603, 1991.

Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4): 566–579, 1995.

Edvin Fuglebakk, Julián Echave, and Nathalie Reuter. Measuring and comparing structural fluctuation patterns in large protein datasets. *Bioinformatics*, 28(19):2431–2440, 2012.

Edvin Fuglebakk, Nathalie Reuter, and Konrad Hinsen. Evaluation of Protein Elastic Network Models Based on an Analysis of Collective Motions. *Journal of Chemical Theory and Computation*, 9(12): 5618–5628, 2013.

Edvin Fuglebakk, Sandhya P. Tiwari, and Nathalie Reuter. Comparing the intrinsic dynamics of multiple protein structures using elastic network models. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1850(5):911–922, 2015.

Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. "O'Reilly Media, Inc.", 2017. ISBN 978-1-4919-6229-9.

Felipe L. Gewers, Gustavo R. Ferreira, Henrique Ferraz de Arruda, Filipi Nascimento Silva, Cesar H. Comin, Diego R. Amancio, and Luciano da F. Costa. Principal component analysis: A natural approach to data exploration. *CoRR*, abs/1804.02502, 2018.

Arun K. Ghosh, Heather L. Osswald, and Gary Prato. Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS. *Journal of medicinal chemistry*, 59(11): 5172–5208, 2016.

Christoph Globisch, Venkatramanan Krishnamani, Markus Deserno, and Christine Peter. Optimization of an Elastic Network Augmented Coarse Grained Model to Study CCMV Capsid Deformation. *PLOS ONE*, 8(4):e60582, 2013.

Pawel Gniewek, Andrzej Kolinski, Robert L. Jernigan, and Andrzej Kloczkowski. Elastic network normal modes provide a basis for protein structure refinement. *The Journal of Chemical Physics*, 136(19), 2012.

N. Go, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Sciences of the United States of America*, 80:3696–3700, 1983.

Chern-Sing Goh, Duncan Milburn, and Mark Gerstein. Conformational changes associated with protein-protein interactions. *Current Opinion in Structural Biology*, 14(1):104–109, 2004.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016. ISBN 0-262-03561-8 978-0-262-03561-3.

David B. Gorman, Gregory J. Catherine, Richard Peragine, Beverly Conrad, G. Duane Gearhart, and David Moy. System for intrusion detection and vulnerability analysis in a telecommunications signaling network, 2004.

Agnieszka Górska, Anna Sloderbach, and Michał Piotr Marszałł. Siderophore-drug complexes: potential medicinal applications of the 'Trojan horse' strategy. *Trends in Pharmacological Sciences*, 35(9): 442–449, 2014.

Joe G. Greener, Ioannis Filippis, and Michael J. E. Sternberg. Predicting Protein Dynamics and Allostery Using Multi-Protein Atomic Distance Constraints. *Structure*, 25(3):546–558, 2017.

Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10(1):168, 2009.

R. Guimerà, S. Mossa, A. Turtschi, and L. a. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799, 2005.

Mert Gur, Jeffry D. Madura, and Ivet Bahar. Global Transitions of Proteins Explored by a Multiscale Hybrid Methodology: Application to Adenylate Kinase. *Biophysical Journal*, 105(7):1643–1652, 2013.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1-3):389–422, 2002.

Pelin Guzel and Ozge Kurkcuoglu. Identification of potential allosteric communication pathways between functional sites of the bacterial ribosome by graph and elastic network models. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1861(12):3131–3141, 2017.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

Turkan Haliloglu and Ivet Bahar. Adaptability of protein structures to enable functional interactions and evolutionary implications. *Current Opinion in Structural Biology*, 35:17–23, 2015.

Turkan Haliloglu, Ivet Bahar, and Burak Erman. Gaussian Dynamics of Folded Proteins. *Physical Review Letters*, 79(16):3090–3093, 1997.

Thomas Hamelryck. An amino acid has two sides: A new 2d measure provides a different view of solvent exposure. *Proteins: Structure, Function, and Bioinformatics*, 59(1):38–48, 2005.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1025–1035, 2017.

William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2018.

Felix Haurowitz. Das Gleichgewicht zwischen Hämoglobin und Sauerstoff. *Hoppe-Seyler's Zeitschrift für physiologische Chemie*, 254(3-6):266–274, 1938.

Haibo He and E. A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Ulf Hensen, Tim Meyer, Jürgen Haas, René Rex, Gert Vriend, and Helmut Grubmüller. Exploring Protein Dynamics Space: The Dynasome as the Missing Link between Protein Structure and Function. *PLOS ONE*, 7(5):e33931, 2012.

Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172): 964–972, 2007.

Susanne M. A. Hermans, Christopher Pfleger, Christina Nutschel, Christian A. Hanke, and Holger Gohlke. Rigidity theory for biomolecules: concepts, software, and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(4):e1311, 2017.

Konrad Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33(3): 417–429, 1998.

Konrad Hinsen, Andrei-Jose Petrescu, Serge Dellerue, Marie-Claire Bellissent-Funel, and Gerald R. Kneller. Harmonicity in slow protein dynamics. *Chemical Physics*, 261(1-2):25–37, 2000.

Konrad Hinsen, Edward Beaumont, Bertrand Fournier, and Jean-Jacques Lacapère. From electron microscopy maps to atomic structures using normal mode-based fitting. In *Membrane Protein Structure Determination*, pages 237–258. Springer, 2010.

Steve Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer Science & Business Media, 2011. ISBN 978-1-4419-8819-5.

Yin-Chen Hsieh, Frédéric Poitevin, Marc Delarue, and Patrice Koehl. Comparative normal mode analysis of the dynamics of denv and zikv capsids. *Frontiers in molecular biosciences*, 3:85, 2016.

John D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3): 90–95, 2007.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer-Verlag, New York, 2013. ISBN 978-1-4614-7137-0.

Michal Jamroz, Andrzej Kolinski, and Daisuke Kihara. Structural features that predict real-value fluctuations of globular proteins. *Proteins: Structure, Function, and Bioinformatics*, 80(5):1425–1435, 2012.

Jan-Oliver Janda, Andreas Meier, and Rainer Merkl. CLIPS-4d: a classifier that distinguishes structurally and functionally important residue-positions based on sequence and 3d data. *Bioinformatics*, 29 (23):3029–3035, 2013.

Jay I. Jeong, Yunho Jang, and Moon K. Kim. A connection rule for -carbon coarse-grained elastic network models using chemical bond information. *Journal of Molecular Graphics and Modelling*, 24 (4):296–306, 2006.

Birthe R. Johannessen, Lars K. Skov, Jette S. Kastrup, Ole Kristensen, Caroline Bolwig, Jørgen N. Larsen, Michael Spangfort, Kaare Lund, and Michael Gajhede. Structure of the house dust mite allergen Der f 2: Implications for function and molecular basis of IgE cross-reactivity. *FEBS Letters*, 579(5):1208–1212, 2005.

Linda C. Johansson, Benjamin Stauch, Andrii Ishchenko, and Vadim Cherezov. A bright future for serial femtosecond crystallography with xfels. *Trends in biochemical sciences*, 42(9):749–762, 2017.

Rico Jonschkowski. *Learning robotic perception through prior knowledge*. PhD thesis, 2018.

W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 34(5):827–828, 1978.

U. Kang, Leman Akoglu, and Duen Horng Chau. Big graph mining for the web and social media: algorithms, anomaly detection, and applications. In *WSDM*, pages 677–678, 2014.

M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, 102(19):6679–6685, 2005.

Martin Karplus and J. Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652, 2002.

Martin Karplus and Gregory A. Petsko. Molecular dynamics simulations in biology. *Nature*, 347(6294): 631–639, 1990.

Burak T. Kaynak, Doga Findik, and Pemra Doruker. Respec incorporates residue specificity and ligand effect into elastic network model. *The Journal of Physical Chemistry B*, 2017.

Daniel A. Keedy, Lillian R. Kenner, Matthew Warkentin, Rahel A. Woldeyes, Jesse B. Hopkins, Michael C. Thompson, Aaron S. Brewster, Andrew H. Van Benschoten, Elizabeth L. Baxter, Monarin Uervirojnangkoorn, Scott E. McPhillips, Jinhu Song, Roberto Alonso-Mori, James M. Holton, William I. Weis, Axel T. Brunger, S. Michael Soltis, Henrik Lemke, Ana Gonzalez, Nicholas K. Sauter, Aina E. Cohen, Henry van den Bedem, Robert E. Thorne, and James S. Fraser. Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography | eLife, 2015.

Laila Khedher, Ignacio A. Illán, Juan M. Górriz, Javier Ramírez, Abdelbasset Brahim, and Anke Meyer-Baese. Independent component analysis-support vector machine-based computer-aided diagnosis system for alzheimer's with visual support. *International journal of neural systems*, 27(03):1650050, 2017.

Alexey G. Kikhney and Dmitri I. Svergun. A practical guide to small angle x-ray scattering (saxs) of flexible and intrinsically disordered proteins. *FEBS letters*, 589(19):2570–2577, 2015.

Changhoon Kim, Jodi Basner, and Byungkook Lee. Detecting internally symmetric protein structures. *BMC Bioinformatics*, 11:303, 2010.

Min Hyeok Kim, Sangjae Seo, Jay Il Jeong, Bum Joon Kim, Wing Kam Liu, Byeong Soo Lim, Jae Boong Choi, and Moon Ki Kim. A mass weighted chemical elastic network model elucidates closed form domain motions in proteins. *Protein Science*, 22(5):605–613, 2013.

Min Hyeok Kim, Byung Ho Lee, and Moon Ki Kim. Robust elastic network model: A general modeling for precise understanding of protein dynamics. *Journal of Structural Biology*, 190(3):338–347, 2015.

Svetlana Kirillova, Juan Cortés, Alin Stefaniu, and Thierry Siméon. An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins. *Proteins: Structure, Function, and Bioinformatics*, 70(1):131–143, 2008.

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks - a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.

Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-Grained Protein Models and Their Applications. *Chemical Reviews*, 116 (14):7898–7936, 2016.

Ryotaro Koike, Motonori Ota, and Akinori Kidera. Hierarchical Description and Extensive Classification of Protein Structural Changes by Motion Tree. *Journal of Molecular Biology*, 426(3):752–762, 2014.

Dmitry A. Kondrashov, Qiang Cui, and George N. Phillips. Optimization and Evaluation of a Coarse-Grained Model of Protein Motion Using X-Ray Crystal Data. *Biophysical Journal*, 91(8):2760–2767, 2006.

Dmitry A. Kondrashov, Adam W. Van Wynsberghe, Ryan M. Bannen, Qiang Cui, and George N. Phillips Jr. Protein Structural Variation in Computational Models and Crystallographic Data. *Structure*, 15(2):169–177, 2007.

D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences*, 44(2):98–104, 1958.

Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

Julio A. Kovacs, Pablo Chacón, and Ruben Abagyan. Predictions of protein flexibility: First-order measures. *Proteins: Structure, Function, and Bioinformatics*, 56(4):661–668, 2004.

Michael Kovermann, Per Rogne, and Magnus Wolf-Watz. Protein dynamics and function from solution state nmr spectroscopy. *Quarterly reviews of biophysics*, 49, 2016.

W. G. Krebs, Vadim Alexandrov, Cyrus A. Wilson, Nathaniel Echols, Haiyuan Yu, and Mark Gerstein. Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. *Proteins: Structure, Function, and Bioinformatics*, 48 (4):682–695, 2002.

Ozge Kurkcuoglu, Zeynep Kurkcuoglu, Pemra Doruker, and Robert L. Jernigan. Collective dynamics of the ribosomal tunnel revealed by elastic network modeling. *Proteins: Structure, Function, and Bioinformatics*, 75(4):837–845, 2009a.

Ozge Kurkcuoglu, Osman Teoman Turgut, Sertan Cansu, Robert L. Jernigan, and Pemra Doruker. Focused Functional Dynamics of Supramolecules by Use of a Mixed-Resolution Elastic Network Model. *Biophysical Journal*, 97(4):1178–1187, 2009b.

Zeynep Kurkcuoglu and Pemra Doruker. Ligand docking to intermediate and close-to-bound conformers generated by an elastic network model based algorithm for highly flexible proteins. *PloS one*, 11: e0158063, 2016.

Zeynep Kurkcuoglu, Ahmet Bakan, Duygu Kocaman, Ivet Bahar, and Pemra Doruker. Coupling between Catalytic Loop Motions and Enzyme Global Dynamics. *PLOS Computational Biology*, 8(9):e1002705, 2012.

Carsten Kutzner, Szilárd Páll, Martin Fechner, Ansgar Esztermann, Bert L. de Groot, and Helmut Grubmüller. Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. *Journal of Computational Chemistry*, 36(26):1990–2008, 2015.

Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, 2008.

Byung Ho Lee, Soojin Jo, Moon-Ki Choi, Min Hyeok Kim, Jae Boong Choi, and Moon Ki Kim. Normal mode analysis of zika virus. *Computational biology and chemistry*, 72:53–61, 2018.

Nicholas Leioatts, Tod D. Romo, and Alan Grossfield. Elastic Network Models Are Robust to Variations in Formalism. *Journal of Chemical Theory and Computation*, 8(7):2424–2434, 2012.

Eitan Lerner, Thorben Cordes, Antonino Ingargiola, Yazan Alhadid, SangYoon Chung, Xavier Michalet, and Shimon Weiss. Toward dynamic structural biology: Two decades of single-molecule förster resonance energy transfer. *Science*, 359(6373):eaan1133, 2018.

Christina Leslie and Rui Kuang. Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research*, 5(Nov):1435–1455, 2004.

C. Levinthal. How to fold graciously. In P. DeBrunner, J. Tsibris, and E. Munck, editors, *Mossbauer spectroscopy in biological systems*, Urbana, IL, 1969. University of Illinois Press.

M. Levitt, C. Sander, and P. S. Stern. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *Journal of molecular biology*, 181:423–447, 1985.

Timothy R. Lezon and Ivet Bahar. Using Entropy Maximization to Understand the Determinants of Structural Dynamics beyond Native Contact Topology. *PLOS Computational Biology*, 6(6):e1000816, 2010.

Geng Li, Murat Semerci, Bülent Yener, and Mohammed J. Zaki. Effective graph classification based on topological and label attributes. *Statistical Analysis and Data Mining*, 5(4):265–283, 2012.

Yunqi Li, Yaping Fang, and Jianwen Fang. Predicting residue-residue contacts using random forest models. *Bioinformatics*, 27(24):3379–3384, 2011.

Tu-Liang Lin and Guang Song. Generalized spring tensor models for protein fluctuation dynamics and conformation changes. *BMC Structural Biology*, 10(Suppl 1):S3, 2010.

Ying Liu and Ivet Bahar. Sequence Evolution Correlates with Structural Dynamics. *Molecular Biology and Evolution*, 29(9):2253–2263, 2012.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444, 2002.

José Ramón López-Blanco and Pablo Chacón. New generation of elastic network models. *Current Opinion in Structural Biology*, 37:46–53, 2016.

José Ramón Lopéz-Blanco, José Ignacio Garzón, and Pablo Chacón. iMod: multipurpose normal mode analysis in internal coordinates. *Bioinformatics*, 27(20):2843–2850, 2011.

José Ramón López-Blanco, Osamu Miyashita, Florence Tama, and Pablo Chacón. Normal Mode Analysis Techniques in Structural Biology. In John Wiley & Sons Ltd, editor, *eLS*. John Wiley & Sons, Ltd, Chichester, UK, 2014. ISBN 978-0-470-01590-2 978-0-470-01617-6.

László Lovász. Eigenvalues of graphs. 2007.

Douglas A. Luke and Jenine K. Harris. Network Analysis in Public Health: History, Methods, and Applications. *Annual Review of Public Health*, 28(1):69–93, 2007.

Buyong Ma and Ruth Nussinov. Protein dynamics: Conformational footprints. *Nature Chemical Biology*, 12(11):890–891, 2016.

Jianpeng Ma. Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. *Structure*, 13(3):373–380, 2005.

Shuangge Ma and Ying Dai. Principal component analysis based methods in bioinformatics studies. *Briefings in bioinformatics*, 12(6):714–722, 2011.

Swapnil Mahajan and Yves-Henri Sanejouand. On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins. *Archives of biochemistry and biophysics*, 567:59–65, 2015.

Olivier Mailhot, Vincent Frappier, Francois Major, and Rafael Najmanovich. The elastic network contact model applied to rna: enhanced accuracy for conformational space prediction. *bioRxiv*, page 198531, 2017.

Axel Maireder, Brian E. Weeks, Homero Gil de Zúñiga, and Stephan Schlögl. Big Data and Political Social Networks: Introducing Audience Diversity and Communication Connector Bridging Measures in Social Network Theory. *Social Science Computer Review*, 35(1):126–141, 2017.

Osni Marques and Yves-Henri Sanejouand. Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins: Structure, Function, and Bioinformatics*, 23(4):557–560, 1995.

Joseph A. Marsh and Sarah A. Teichmann. Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *Bioessays*, 36(2): 209–218, 2014.

Jose M. Martin-Garcia, Chelsie E. Conrad, Jesse Coe, Shatabdi Roy-Chowdhury, and Petra Fromme. Serial femtosecond crystallography: A revolution in structural biology. *Archives of Biochemistry and Biophysics*, 602:32–47, 2016.

Angel T. Martinez, Francisco J. Ruiz-Duenas, Susana Camarero, Ana Serrano, Dolores Linde, Henrik Lund, Jesper Vind, Morten Tovborg, Owik M. Herold-Majumdar, Martin Hofrichter, et al. Oxidoreductases on their way to industrial biotransformations. *Biotechnology advances*, 35(6):815–831, 2017.

Tatiana Maximova, Ryan Moffatt, Buyong Ma, Ruth Nussinov, and Amarda Shehu. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLOS Computational Biology*, 12(4):e1004619, 2016.

Andreas May and Martin Zacharias. Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins*, 70:794–809, 2008.

S. W. May. Applications of oxidoreductases. *Current opinion in biotechnology*, 10:370–375, 1999.

J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267 (5612):585–590, 1977.

Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

Lidio Meireles, Mert Gur, Ahmet Bakan, and Ivet Bahar. Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins. *Protein Science*, 20(10): 1645–1658, 2011.

R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine Learning: An Artificial Intelligence Approach*. Springer Science & Business Media, 2013. ISBN 978-3-662-12405-5.

Cristian Micheletti. Comparing proteins by their internal dynamics: Exploring structure-function relationships beyond static structural alignments. *Physics of Life Reviews*, 10(1):1–26, 2013.

Cristian Micheletti, Paolo Carloni, and Amos Maritan. Accurate and efficient description of protein vibrational dynamics: Comparing molecular dynamics and Gaussian models. *Proteins: Structure, Function, and Bioinformatics*, 55(3):635–645, 2004.

R. J. Dwayne Miller. Femtosecond Crystallography with Ultrabright Electrons and X-rays: Capturing Chemistry in Action. *Science*, 343(6175):1108–1116, 2014.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997. ISBN 978-0-07-115467-3.

Osamu Miyashita, Christian Gorba, and Florence Tama. Structure modeling from small angle x-ray scattering data with elastic network normal mode analysis. *Journal of structural biology*, 173: 451–460, 2011.

Alexander Miguel Monzon, Cristian Oscar Rohr, María Silvina Fornasari, and Gustavo Parisi. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database*, 2016, 2016.

Kei Moritsugu and Jeremy C. Smith. Coarse-Grained Biomolecular Simulation with REACH: Realistic Extension Algorithm via Covariance Hessian. *Biophysical Journal*, 93(10):3460–3469, 2007.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0-262-01802-0 978-0-262-01802-9.

Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

William S. Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565, 2006.

Itta Ohmura, Gentaro Morimoto, Yousuke Ohno, Aki Hasegawa, and Makoto Taiji. MDGRAPE-4: a special-purpose computer system for molecular dynamics simulations. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 372(2021), 2014.

Rebecca A. Oot, Li-Shar Huang, Edward A. Berry, and Stephan Wilkens. Crystal structure of the yeast vacuolar atpase heterotrimeric egc(head) peripheral stalk complex. *Structure (London, England : 1993)*, 20:1881–1892, 2012.

Laura Orellana, Manuel Rueda, Carles Ferrer-Costa, José Ramón Lopez-Blanco, Pablo Chacón, and Modesto Orozco. Approaching Elastic Network Models to Molecular Dynamics Flexibility. *Journal of Chemical Theory and Computation*, 6(9):2910–2923, 2010.

Modesto Orozco. A theoretical view of protein dynamics. *Chemical Society Reviews*, 43(14):5051–5066, 2014.

Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease. *Medical image analysis*, 48:117–130, 2018.

Niko Pavliček and Leo Gross. Generation, manipulation and characterization of molecules by atomic force microscopy. *Nature Reviews Chemistry*, 1(1):0005, 2017.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Fernando Pérez and Brian E. Granger. Ipython: a system for interactive scientific computing. *Computing in Science & Engineering*, 9(3), 2007.

Giorgio Pessot, Hartmut Löwen, and Andreas M. Menzel. Dynamic elastic moduli in magnetic gels: Normal modes and linear response. *The Journal of chemical physics*, 145(10):104904, 2016.

Paula Petrone and Vijay S. Pande. Can Conformational Change Be Described by Only a Few Normal Modes? *Biophysical Journal*, 90(5):1583–1593, 2006.

Thomas J. Piggot, Daniel A. Holdbrook, and Syma Khalid. Conformational dynamics and membrane interactions of the E. coli outer membrane protein FecA: A molecular dynamics simulation study. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1828(2):284–293, 2013.

Giovanni Pinamonti, Sandro Bottaro, Cristian Micheletti, and Giovanni Bussi. Elastic network models for rna: a comparative assessment with molecular dynamics and shape experiments. *Nucleic acids research*, 43(15):7260–7269, 2015.

Ines Putz and Oliver Brock. Elastic network model of learned maintained contacts to predict protein motion. *PLOS ONE*, 12(8):e0183889, 2017.

D. L. Rousseau, R. P. Bauman, and S. P. S. Porto. Normal mode determination in crystals. *Journal of Raman Spectroscopy*, 10(1):253–290, 1981.

Manuel Rueda, Pablo Chacón, and Modesto Orozco. Thorough Validation of Protein Normal Mode Analysis: A Comparative Study with Essential Dynamics. *Structure*, 15(5):565–575, 2007a.

Manuel Rueda, Carles Ferrer-Costa, Tim Meyer, Alberto Pérez, Jordi Camps, Adam Hospital, Josep Lluis Gelpí, and Modesto Orozco. A consensus view of protein dynamics. *Proceedings of the National Academy of Sciences*, 104(3):796–801, 2007b.

A. L. Samuel. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959.

Kannan Sankar, Sambit K. Mishra, and Robert L. Jernigan. Comparisons of Protein Dynamics from Experimental Structure Ensembles, Molecular Dynamics Ensembles, and Coarse-Grained Elastic Network Models. *The Journal of Physical Chemistry B*, 2018.

Craig Saunders, Alexei Vinokourov, and John S. Shawe-taylor. String kernels, fisher kernels and finite state automata. In *Advances in Neural Information Processing Systems*, pages 649–656, 2003.

Michael Schneider and Oliver Brock. Combining Physicochemical and Evolutionary Information for Protein Contact Prediction. *PLOS ONE*, 9(10):e108438, 2014.

Gunnar F. Schröder, Axel T. Brunger, and Michael Levitt. Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution. *Structure (London, England : 1993)*, 15(12):1630–1641, 2007.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. 2015.

John Scott. *Social Network Analysis*. SAGE, 2017. ISBN 978-1-5264-1225-6.

Tristan A. Shatto and Egemen K. Çetinkaya. Variations in graph energy: A measure for network resilience. In *Resilient Networks Design and Modeling (RNDM), 2017 9th International Workshop on*, pages 1–7. IEEE, 2017.

David E. Shaw, Ron O. Dror, John K. Salmon, J. P. Grossman, Kenneth M. Mackenzie, Joseph A. Bank, Cliff Young, Martin M. Deneroff, Brannon Batson, Kevin J. Bowers, Edmond Chow, Michael P. Eastwood, Douglas J. Ierardi, John L. Klepeis, Jeffrey S. Kuskin, Richard H. Larson, Kresten Lindorff-Larsen, Paul Maragakis, Mark A. Moraes, Stefano Piana, Yibing Shan, and Brian Towles. Millisecond-scale Molecular Dynamics Simulations on Anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC '09, pages 39:1–39:11, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-744-8.

David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li-Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk-Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the Bar for Performance and Programmability in a Special-purpose Molecular Dynamics Supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '14, pages 41–53, Piscataway, NJ, USA, 2014. IEEE Press. ISBN 978-1-4799-5500-8.

Amarda Shehu and Erion Plaku. A survey of computational treatments of biomolecules by robotics-inspired methods modeling equilibrium structure and dynamic. *Journal of Artificial Intelligence Research*, 57:509–572, 2016.

Jia Shi, Naim Kapucu, Zhengwei Zhu, Xuesong Guo, and Brittany Haupt. Assessing Risk Communication in Social Media for Crisis Prevention: A Social Network Analysis of Microblog. *Journal of Homeland Security and Emergency Management*, 14(1), 2017.

Lindsay I. Smith. A tutorial on principal components analysis. 2002.

Guang Song and Robert L. Jernigan. An enhanced elastic network model to represent the motions of domain-swapped proteins. *Proteins: Structure, Function, and Bioinformatics*, 63(1):197–209, 2006.

Amit Srivastava, Roee Ben Halevi, Alexander Veksler, and Rony Granek. Tensorial elastic network model for protein dynamics: Integration of the anisotropic network model with bond-bending and twist elasticities. *Proteins: Structure, Function, and Bioinformatics*, 80(12):2692–2700, 2012.

Nathaniel Stanley and Gianni De Fabritiis. High throughput molecular dynamics for drug discovery. *In Silico Pharmacology*, 3, 2015.

Amelie Stein, Manuel Rueda, Alejandro Panjkovich, Modesto Orozco, and Patrick Aloy. A Systematic Study of the Energetics Involved in Structural Changes upon Association and Connectivity in Protein Interaction Networks. *Structure*, 19(6):881–889, 2011.

Joseph N. Stember and Willy Wriggers. Bend-twist-stretch model for coarse elastic network simulation of biomolecular motion. *The Journal of Chemical Physics*, 131(7), 2009.

Kaustubh Supekar, Vinod Menon, Daniel Rubin, Mark Musen, and Michael D. Greicius. Network Analysis of Intrinsic Functional Brain Connectivity in Alzheimer's Disease. *PLOS Computational Biology*, 4(6):e1000100, 2008.

F. Tama and Y.-H. Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Engineering*, 14(1):1–6, 2001.

Florence Tama, Florent Xavier Gadea, Osni Marques, and Yves-Henri Sanejouand. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Structure, Function, and Bioinformatics*, 41(1):1–7, 2000.

Kaare Teilum, Johan G. Olsen, and Birthe B. Kragelund. Functional aspects of protein flexibility. *Cellular and Molecular Life Sciences*, 66(14):2231, 2009.

Monique M. Tirion. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, 77(9):1905–1908, 1996.

Chung-Jung Tsai, Sandeep Kumar, Buyong Ma, and Ruth Nussinov. Folding funnels, binding funnels, and protein function. *Protein Science*, 8(6):1181–1190, 1999.

Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. A primer on kernel methods. *Kernel methods in computational biology*, 47:35–70, 2004.

Victor L. Villemagne, Vincent Doré, Samantha C. Burnham, Colin L. Masters, and Christopher C. Rowe. Imaging tau and amyloid-β proteinopathies in Alzheimer disease and other conditions. *Nature Reviews Neurology*, 2018.

S. Vichy N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11(Apr):1201–1242, 2010.

Neil R. Voss. *Geometric studies of RNA and ribosomes, and ribosome crystallization.* PhD thesis, Yale University, 2007.

Guifang Wang, Ze-Ting Zhang, Bin Jiang, Xu Zhang, Conggang Li, and Maili Liu. Recent advances in protein NMR spectroscopy and their implications in protein therapeutics research. *Analytical and Bioanalytical Chemistry*, 406(9-10):2279–2288, 2014a.

Jinan Wang, Qiang Shao, Zhijian Xu, Yingtao Liu, Zhuo Yang, Benjamin P. Cossins, Hualiang Jiang, Kaixian Chen, Jiye Shi, and Weiliang Zhu. *The journal of physical chemistry. B*, 118:134–143, 2014b.

Jun Wang, Fang Li, and Chunlong Ma. Recent progress in designing inhibitors that target the drug-resistant M2 proton channels from the influenza A viruses. *Peptide Science*, 104(4):291–309, 2015.

Qi Wang, YangHe Feng, JinCai Huang, TengJiao Wang, and GuangQuan Cheng. A novel framework for the identification of drug target proteins: Combining stacked auto-encoders with a biased support vector machine. *PloS one*, 12:e0176486, 2017.

Yongmei Wang, A. J. Rader, Ivet Bahar, and Robert L. Jernigan. Global ribosome motions revealed with elastic network model. *Journal of structural biology*, 147(3):302–314, 2004.

Michael D. Ward, Katherine Stovel, and Audrey Sacks. Network Analysis and Political Science. *Annual Review of Political Science*, 14(1):245–264, 2011.

Robert G. Webster and Elena A. Govorkova. Continuing challenges in influenza. *Annals of the New York Academy of Sciences*, 1323(1):115–139, 2014.

Guanghong Wei, Wenhui Xi, Ruth Nussinov, and Buyong Ma. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chemical Reviews*, 116(11):6516–6551, 2016.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability Estimates for Multi-class Classification by Pairwise Coupling. *J. Mach. Learn. Res.*, 5:975–1005, 2004.

Wan-Lin Wu, Christopher Robert Grotefend, Ming-Ting Tsai, Yi-Ling Wang, Vladimir Radic, Hyungjin Eoh, and I.-Chueh Huang. δ20 IFITM2 differentially restricts X4 and R5 HIV-1. *Proceedings of the National Academy of Sciences*, 114(27):7112–7117, 2017.

Fei Xia, Dudu Tong, Lifeng Yang, Dayong Wang, Steven C. H. Hoi, Patrice Koehl, and Lanyuan Lu. Identifying essential pairwise interactions in elastic network model using the alpha shape theory. *Journal of Computational Chemistry*, pages n/a–n/a, 2014.

Kelin Xia. Multiscale virtual particle based elastic network model (mvp-enm) for normal mode analysis of large-sized biomolecules. *Physical chemistry chemical physics : PCCP*, 20:658–669, 2017.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. Representation learning of knowledge graphs with entity descriptions. In *AAAI*, pages 2659–2665, 2016.

Feng Xu. Applications of oxidoreductases: recent progress. *Industrial Biotechnology*, 1(1):38–50, 2005.

Pinar Yanardag and S. V. N. Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1365–1374. ACM, 2015.

Lee-Wei Yang, Eran Eyal, Ivet Bahar, and Akio Kitao. Principal component analysis of native ensembles of biomolecular structures (PCA_nest): insights into functional dynamics. *Bioinformatics*, 25(5): 606–614, 2009a.

Lei Yang, Guang Song, and Robert L. Jernigan. How Well Can We Understand Large-Scale Protein Motions Using Normal Modes of Elastic Network Models? *Biophysical Journal*, 93(3):920–929, 2007.

Lei Yang, Guang Song, Alicia Carriquiry, and Robert L. Jernigan. Close Correspondence between the Essential Protein Motions from Principal Component Analysis of Multiple HIV-1 Protease Structures and Elastic Network Modes. *Structure (London, England : 1993)*, 16(2):321–330, 2008.

Lei Yang, Guang Song, and Robert L. Jernigan. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences*, 106(30):12347–12352, 2009b.

Wyatt W. Yue, Sylvestre Grizot, and Susan K. Buchanan. Structural Evidence for Iron-free Citrate and Ferric Citrate Binding to the TonB-dependent Outer Membrane Transporter FecA. *Journal of Molecular Biology*, 332(2):353–368, 2003.

Michael T. Zimmermann and Robert L. Jernigan. Elastic network models capture the motions apparent within ensembles of rna structures. *RNA*, 20(6):792–804, 2014.