IN SILICO PREDICTION OF PROTEIN STRUCTURES AND ENSEMBLES

By

Axel Walter Fischer

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

January 31, 2018

Nashville, Tennessee

Approved:

Jens Meiler, Ph.D.

Hassane S. Mchaourab, Ph.D.

Michael P. Stone, Ph.D.

Carlos F. Lopez, Ph.D.

Terry P. Lybrand, Ph.D.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Jens Meiler for his continuous support during my graduate career. He convinced me to switch the focus of my work from computer engineering and software development to computational structural biology and therefore apply my prior knowledge of computer science, physics, and mathematics to a new set of problems. This experience has substantially broadened my scientific horizon and encouraged my strive for knowledge.

Working and studying in the Meiler Laboratory at Vanderbilt University showed me the interdisciplinary nature of modern research — how people with different scientific backgrounds can complement each other's knowledge to solve scientific problems that could not be solved with knowledge about one scientific field alone. This scientific diversity of the Meiler Laboratory enabled insightful discussions leading to innovative solutions. In this context, I especially want to thank Dr. Soumya Ganguly, Dr. Sten Heinze, Dr. Daniel K. Putnam, Dr. Rocco Moretti, Dr. Gregory Sliwoski, Jeffrey Mendenhall, Yan Xia, Alexander R. Geanes, Marion F. Sauer, and Diego del Alamo for always being available and interested in discussing scientific problems and possible solutions. I especially want to thank my committee members, Drs. Hassane S. Mchaourab, Michael P. Stone, Carlos F. Lopez, and Terry P. Lybrand for their constant guidance and support. In particular, I want to thank Drs. Hassane S. Mchaourab, Reza Dastvan, and Richard Stein for outstanding collaborations and very valuable advise. I would also like to thank Heather Darling who provided exceptional organizational support by taking care of scheduling meetings, booking rooms, and providing general administrative advise.

Deciding to attend Graduate School in the United States also meant that I would not be able to visit my family frequently. Despite this, my family has always been very understanding and supportive during my graduate studies, which deserves my deepest gratitude.

Conducting my research required access to substantial amounts of computational resources. Most of the computations were run on the *Advanced Computing Center for Research & Education* (ACCRE) at Vanderbilt University and on the *Oak Ridge Leadership Computing Facility* at Oak Ridge National Laboratory. Without continued support from the staff of ACCRE, Roy Hoffman, William C. Riner, and Sabuj Pattanayek from the *Center for Structural Biology* at Vanderbilt University, and the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725, I would not have been able to conduct my research. Also, my research heavily relied on the free and/or open source software packages R,[a] Inkscape,[b] Emacs,[c] and LuaTeX,[d] which are developed by volunteers and made available free of charge.

[a] https://www.r-project.org

[b] https://inkscape.org

[c] https://www.gnu.org/s/emacs

[d] http://www.luatex.org

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BAX Bcl-2-associated X protein. xvii, xix, 83–87, 89–97, 129, 183–186, 188, 189

BCL BioChemical Library. xvii–xix, 8, 12, 15, 21, 25, 26, 30, 39, 42, 43, 47, 48, 56, 68, 84, 85, 90–93, 95, 97, 100, 113, 159, 169, 176, 178, 179, 181, 186, 188–190, 194

CASP Critical Assessment of Protein Structure Prediction. xvii, xviii, 8, 11–14, 16, 17, 19–21, 24–28, 67, 126, 155

CCD cyclic coordinate descent. xvii, xix, 17, 67–70, 72, 73, 76–78, 80, 81, 103, 132, 162, 178, 179, 195

CO contact order. 14, 15, 19, 20, 155

CONE motion-on-a-cone. 9, 30, 35, 42, 45, 46, 86, 87, 89, 103, 104, 114, 128, 131, 133, 169, 183, 191

CPU central processing unit. xvii, 11, 12, 68, 70, 76–78, 80, 127, 132, 161

CW continuous wave. 113, 118, 121, 200

DDM dodecyl maltoside. xix, 112, 113, 115, 121, 199–201

DEER double electron-electron resonance. xix, 5, 30, 42, 83, 86, 89, 90, 100, 106, 109, 111–113, 115, 118, 119, 123, 191, 192, 198, 200

DSSP Dictionary of Secondary Structure of Proteins. 19, 70, 73, 75, 178, 179

EM electron microscopy. 4, 5, 8–10, 13, 49, 111, 112, 130

EmrE efflux-multidrug resistance protein. xviii, xix, 1, 11, 111–115, 117, 119, 121–123, 125, 131, 198–200

EPR electron paramagnetic resonance. xvii–xix, 4, 5, 8–11, 13, 29–31, 33–47, 49, 64, 83–87, 89, 90, 92–109, 113, 114, 118, 121–123, 126, 128–133, 163–167, 169, 183–186, 188, 189, 191, 194, 198, 200

ExoU exotoxin U. xviii, xix, 11, 99–102, 104–109, 131, 191–194

GDT_TS global distance test total score. 12, 14, 18–27, 156–158

GPU graphics processing unit. 132, 133

KIC kinematic closure. 68

LC liquid chromatography. 49

MC Monte Carlo. 12, 15, 16, 25, 35, 36, 53, 54, 67, 77, 85–87, 102, 103, 159, 181, 186, 188, 195

MCM Monte Carlo Metropolis. 7, 15, 16, 25, 29, 35, 37, 54, 55, 68–70, 72, 75, 78, 97, 100, 102, 103, 107, 127, 133, 159, 169, 170, 181, 188, 189, 194, 195

MD molecular dynamics. 3, 7, 9, 10, 12, 13, 16, 17, 22, 27, 108, 126, 127, 130, 159

MMM Multiscale Modeling of Macromolecular systems. 90, 113, 114, 117, 131, 132, 202

MOM mitochondrial outer membrane. 84, 89

MS mass spectrometry. xvii, 11–14, 17, 19, 24, 25, 48, 49, 52, 53, 60, 62–66, 126–130, 158, 171, 176

MW molecular weight. 52, 55–58, 63, 66, 130

NMR nuclear magnetic resonance. xvii, 1, 4, 5, 9, 10, 12–14, 17, 19, 23–25, 28–31, 36, 48, 49, 62, 64, 66, 83, 84, 87, 89–98, 112, 117, 126–128, 130, 131, 158, 184

NOE nuclear overhauser effect. 12–14, 17, 19, 24, 25, 28, 126, 127, 158

OCTOPUS obtainer of correct topologies for uncharacterized sequences. 37, 45, 169

PAGE polyacrylamide gel electrophoresis. 113

PDB Protein Data Bank. 4, 5, 7, 9, 15, 25, 26, 29, 31, 42, 47, 48, 50, 52, 53, 56, 60, 65, 67–70, 73, 75, 77, 78, 80, 81, 84, 85, 87, 89–93, 96, 99, 102, 104, 105, 108, 109, 113, 114, 117, 127, 130, 132, 133, 161, 176, 178, 179, 181, 190, 191

PELDOR pulsed electron paramagnetic resonance. 83, 86, 115

PISCES protein sequence culling server. 50, 70, 75, 178

PSIPRED PSI-blast based secondary structure PREDiction. 16, 54, 55, 72, 85, 102, 159, 176, 186, 188, 194

RMSD root-mean-square-distance. 17, 57, 75, 78, 88, 90, 102, 103, 114, 127, 195, 202

RMSD100 protein-size-normalized root-mean-square-distance. 17, 19, 29, 37–41, 43, 45–47, 56, 60, 61, 65, 83, 87–89, 91–97, 99, 105–110, 130, 160, 161, 163–165, 171, 172, 175, 176, 183

SAS solvent accessible surface. 50, 56

SASA solvent accessible surface area. 31, 33, 128

SDS sodium dodecyl sulfate. 113

SDSL site-directed spin labeling. 4, 5, 9–11, 29, 33, 83–87, 89, 90, 92–98, 100–109, 163, 169, 183–186, 188, 189, 191, 198

SMR small multidrug resistance. 1, 111, 131

SpcU specific *Pseudomonas* chaperone for ExoU. xviii, 99, 100, 104, 109, 131

SSE secondary structure element. 7, 12, 13, 15–17, 19, 25, 27, 29–33, 35–37, 39–42, 44, 45, 52, 54, 55, 58, 62, 64, 69, 70, 72, 73, 77, 84–87, 90, 92, 93, 95–97, 100, 102, 103, 113, 114, 126, 127, 131, 159, 164, 165, 167–170, 176, 179, 181, 184, 185, 188–190, 194

TMH transmembrane helix. 30, 37, 44, 45, 111, 112, 115–119, 121–123, 125

TPP tetraphenylphosphonium. xviii, xix, 111–118, 120–125, 131, 202

WT wild type. 121

XL cross-linking. xvii, 11–14, 17, 19, 24, 25, 48, 49, 52, 53, 60, 62–66, 126–130, 158, 171, 176

SUMMARY

The overall focus of this dissertation was the development and enhancement of computational methods for the *de novo* prediction of protein structures and ensembles. The developed algorithms were integrated into the BioChemical Library (BCL),[a] which is a software suite for molecular modeling and drug design developed in the Meiler Laboratory at Vanderbilt University. A major focus of my work was put on the integration of limited experimental data from electron paramagnetic resonance (EPR) spectroscopy and chemical cross-linking (XL) into the prediction algorithms to compensate for necessary simplifications in the computational approaches.

Chapter I on page 1 provides a summary of the physical laws that govern protein ensemble formation, the relevance of knowledge about a protein's equilibrium constitution, and experimental approaches to probe protein ensembles. An additional focus was put on computational approaches for protein ensemble prediction — in particular in conjunction with limited experimental data. This chapter was written for this dissertation.

Chapter II on page 12 details an evaluation of a modular protein structure prediction pipeline based on the *de novo* prediction method BCL::Fold in the eleventh double-blind Critical Assessment of Protein Structure Prediction (CASP) experiment in 2014 (CASP11). Over the course of the experiment, the tertiary structures of twenty topologically dissimilar proteins were predicted either from their primary structure alone or using additional experimental data from nuclear magnetic resonance (NMR) spectroscopy, chemical XL in conjunction with mass spectrometry (MS), and contact prediction. This chapter is based on a manuscript that was published as “CASP11 – An Evaluation of a Modular BCL::Fold-Based Protein Structure Prediction Pipeline”.[1]

Chapter III on page 29 describes a model for the integration of data from EPR distance and accessibility measurements into computational protein structure prediction methods to predict the tertiary structure of membrane proteins. A major goal of this study was the establishment of a protocol for membrane protein structure prediction from EPR distance and accessibility data. The proposed protocol was evaluated on twenty-nine membrane proteins. This chapter is based on a manuscript that was published as “BCL::MP-Fold: Membrane protein structure prediction guided by EPR restraints”.[2]

Chapter IV on page 48 details a model and potential function to incorporate data from chemical XL in conjunction with MS into computational protein structure prediction methods. A major focus of this work was the evaluation of the influence of the cross-linker spacer length and cross-linker reactivity on protein structure prediction accuracy. This chapter is based on a manuscript that was published as “Protein structure prediction guided by crosslinking restraints – A systematic evaluation of the impact of the crosslinking spacer length”.[3]

Chapter V on page 67 details an algorithm for efficient sampling of loop conformations. The algorithm was developed with the goal of sampling structurally diverse ensembles of loop conformations while minimizing the required amount of central processing unit (CPU) time. To facilitate this goal, the algorithm uses a combination of conformational hashing and cyclic coordinate descent (CCD). This chapter is based on a manuscript that was published as “Efficient sampling of loop conformations using conformation hashing in conjunction with cyclic coordinate descent”.[4]

Chapter VI on page 83 describes the application of the protein structure prediction protocols developed in previous studies to predict the soluble monomeric and membrane-associated homodimeric states of the Bcl-2-associated X protein (BAX). A major focus of this study was the evaluation if in-

[a] http://www.meilerlab.org/bclcommons

trinsic limitations of computational protein structure prediction methods can be overcome through integration of limited experimental data from EPR spectroscopy. This chapter is based on a manuscript that was published as "Pushing the size limit of de novo structure ensemble prediction guided by sparse SDSL-EPR restraints to 200 residues: The monomeric and homodimeric forms of BAX".[5]

Chapter VII on page 99 details the application of the protein structure prediction protocols developed in previous studies to predict the structure and dynamics of the type III secretion effector protein exotoxin U (ExoU). The predictions were performed in conjunction with EPR distance restraints and provide evidence that the unbound structure of ExoU matches the topology observed in complex with the specific *Pseudomonas* chaperone for ExoU (SpcU). This chapter is based on a manuscript that was published as "Structure and Dynamics of Type III Secretion Effector Protein ExoU As determined by SDSL-EPR Spectroscopy in Conjunction with De Novo Protein Folding".[6]

Chapter VIII on page 111 describes the determination of the protonation-dependent dynamics of the efflux-multidrug resistance protein (EmrE). The focus of this work was employing a combined approach of EPR distance and accessibility measurements in conjunction with computational methods to predict the tertiary structures of EmrE in the protonated and tetraphenylphosphonium (TPP)-bound states to derive a model of the protein's transportation mechanism. This chapter is based on a manuscript that was published as "Protonation-dependent conformational dynamics of the multidrug transporter EmrE".[7]

Chapter IX on page 126 discusses the studies presented in this dissertation. The main focus points of the discussion are the evaluation of hierarchical protein structure and ensembles prediction pipelines that were employed throughout most experiments, the influence of limited experimental data on the prediction accuracy, and the general applicability of computational modeling to the field of structural biology. Additionally, a discussion is provided that details future directions regarding method development for *in silico* prediction of protein ensembles building on this work. This chapter was written for this dissertation.

The appendices are structured according to the dissertation chapters, with one appendix per chapter (with the exception of the chapters regarding introduction and discussion). They provide comprehensive protocol captures for their respective dissertation chapters as well as supplementary data that was not provided in the chapter.

Appendix A on page 155 provides supplementary information and a protocol capture for the study detailed in chapter II on page 12. In particular, an evaluation of potential correlations between protein properties on one hand and prediction accuracy and model discrimination is shown. All prediction results from the CASP experiment are provided in tabulated form. Additionally, a protocol capture of the computation procedured employed duing the CASP experiment is listed.

Appendix B on page 163 lists all prediction results for study detailed in chapter III on page 29 in tabulated form. The prediction results are quantified to structure dissimilarity and model discrimation metrics and listed for different combinations of EPR distance and accessibility data. Additional experiments are shown that study the influence of the number of restraints on the prediction accuracy and model discrimination. This appendix also lists the protocol captures for predicting membrane protein structure from EPR data and for simulation of EPR distance restraints.

Appendix C on page 171 provides the tabulated prediction results for chapter IV on page 48. It lists the cross-link yield per spacer length for the different benchmark proteins, prediction accuracies and model discrimination for different cross-linker reactivities, and visualizations of the translation from cross-linking data into structural restraints. This appendix also provides a protocol capture for protein structure prediction from cross-linking data using the BCL.

Appendix D on page 178 details the protocol capture for chapter V on page 67. It lists the computational procedures for generating a loop template library from a set of protein models and using the generated templated library to sample loop conformations using the BCL. This appendix also provides the configuration files to sample loop conformations using a combination of conformation hashing and CCD.

Appendix E on page 183 lists the agreements of the X-ray- and NMR-derived models of monomoeric and homodimeric BAX with the EPR distance data in tabulated form. This appendix also visualizes the dependence of the -value, which was reported in chapter VI on page 83, on the number of models and provides a detailed protocol capture of the computational procedures employed to predict the tertiary structures of monomeric and homodimeric BAX from EPR distance data.

Appendix F on page 191 provides supplementary data and computational procedures for chapter VII on page 99. This appendix provides the agreement of the X-ray-derived reference structure with the glsepr data in tabulated form. Additionally, the distance distributions derived from the double electron-electron resonance (DEER) experiment are shown and compared to explicitly simulated distance distributions for the structure that was predicted from the EPR data. This appendix is concluded by a protocol capture of the computational methods employed to predict the tertiary structure of ExoU from EPR data.

Appendix G on page 198 lists supplementary data for chapter VIII on page 111. This appendix provides descriptions of the ligand-dependent conformational dynamics of EmrE, the equilibrium of EmrE, and the effect of protonation-mimetic mutations of acidic residues on the distance distributions in β-dodecyl maltoside (DDM) micelles. This appendix is concluded by a comparison of the experimentally obtained distances for the TPP-bound state with the distances simulated for the predicted models.

CHAPTER I
CURRENT METHODS FOR *IN SILICO* PREDICTION OF PROTEIN ENSEMBLES

Determination of a protein's structural equilibrium constitution remains an unsolved problem. In the equilibrium, proteins exist in multiple conformations with the population sizes defined by the free energy differences between them. Although techniques like X-ray crystallography or NMR spectroscopy can frequently determine a structural model of the protein's conformation at the Gibbs free energy minimum, alternative populations remain elusive. In this manuscript, we review the physical concepts that govern the formation of protein ensembles, we evaluate the abilities of the different spectroscopic techniques in the context of protein ensemble determination, and we summarize the current state of in silico *protein ensemble prediction. A particular focus of this manuscript was put on the combination of experimental and computational techniques to compensate for information gaps in the experimental data as well as for ambiguities arising from simplifications in the computational methods.*

I.1. Introduction

Proteins are the main actors in biological processes. Examples of functions facilitated by proteins include but are not limited to the transport and storage of other molecules,[8–11] carrying out chemical reactions,[12–14] transduction of signals,[15–18] and providing structural support for the cell.[19–22] Typically, carrying out its function requires a protein to react to specific stimuli by performing specific actions that materialize as changes of its conformation. One example for this is provided by the small multidrug resistance (SMR) transporter EmrE. EmrE is a homodimeric membrane protein from *Escherichia coli* that extrudes cytotoxic molecules, which are predominately hydrophobic cations.[23] This energetically unfavorable extrusion, if seen as an isolated process, is facilitated by coupling the export of the cation to the import of two protons alongside their electrochemical gradient.[23–26] This coupled transport requires complex conformational changes of the protein to enable the binding of protons and substrates as well as the release of them on the opposite sides of the membrane.[7] Consequently, a comprehensive characterization of this protein requires substantiated knowledge of its multiple conformations. Additionally, when attempting to structurally characterize a dynamic protein like this, it might seem tempting to view it as existing within a certain state given a certain environment and changing into a different state given a specific environmental change. However, this formulation as a triggered conformational change does not comprehensively capture the underlying physical laws. A closer examination of the physical foundations that determine a protein's equilibrium constitution will provide a better understanding of how to develop computational methods for the prediction of protein ensembles and how to interpret and use structural information from spectroscopic experiments.

In the following sections of this introduction, we provide an overview about the physical laws that govern the formation of protein ensembles in the equilibrium (section I.1.1 on the next page) and a discussion of the importance of knowing alternative conformations for *in silico* drug design (section I.1.2 on page 3). The introduction is followed by a discussion of the experimental approaches to obtain information about a protein's equilibrium constitution (section I.2 on page 4) and a discussion of computational methods to determine protein ensembles completely *in silico* or in conjunction with experimental data, which is discussed in section I.3 on page 6 and in section I.3.3 on page 9. This manuscript is concluded by a discussion of the provided material in section I.4 on page 10.

I.1.1. The physical laws governing the formation of protein ensembles

Like all matter, proteins adhere to the laws of thermodynamics. Most relevant for the structural characterization of a protein is the second law of thermodynamics, which states that for any spontaneous process, the Gibbs free energy of the system will be at a minimum for the resulting equilibrium.[27,28] Another conclusion from this statement was Anfinsen's thermodynamics hypothesis of protein folding that was postulated in 1973.[29] Anfinsen states that the three-dimensional structure of a protein in its native state is the conformation for which the Gibbs free energy of the system is at the minimum.[28,29] Using Anfinsen's definition, the native state of a protein is defined through one conformation. However, this one conformation must not necessarily be the only conformation of the protein that exists with a significant population size in the protein's equilibrium.

The population sizes of the different conformations can be derived from their Gibbs free energy $G = H - TS$, where H is the enthalpy, T is the temperature, and S is the entropy of the system. For determining the population sizes, the Gibbs free energy G does not need to be known on an absolute scale, but it is sufficient to know the Gibbs free energy differences, ΔG, between the different conformations. The ratios of the Gibbs free energy differences can then be translated into ratios of relative population sizes. A starting point for the derivation of the formula defining the population sizes is the equilibrium constant of the system. The equilibrium constant K_{AB}, which is shown in equation (I.1), for a system with the two possible states A and B is defined as the quotient of the respective probabilities for observing either conformation in the equilibrium.

$$K_{\mathrm{AB}} = \frac{[A]}{[B]} = \frac{P_{\mathrm{A}}}{P_{\mathrm{B}}} \tag{I.1}$$

where:

K_{AB} = equilibrium constant for the states A and B
$[X]$ = concentration of state X in the equilibrium
P_{X} = unnormalized probability of observing state X in the protein's equilibrium

The equilibrium constant — and therefore the ratio of the population sizes for the different conformations — is determined by the temperature T of the system and the free energy differences ΔG between the different conformations of the system, which is formalized by the Boltzmann relation (equation (I.2)). By combining the formula for the equilibrium constant (equation (I.1)) with the Boltzmann relation, we arrive at equation (I.3) that puts the free energy difference between two conformation into relation with their equilibrium constant — hence, the relative probabilities of the protein assuming each state in the equilibrium. For a system that consists of more than two states, the equation can be extended and rearranged to compute the relative population size of a specific state of the protein (equation (I.4) on the next page).

$$\Delta G_{\mathrm{AB}} = -RT \ln K_{\mathrm{AB}} \tag{I.2}$$

$$K_{\mathrm{AB}} = \frac{P_{\mathrm{A}}}{P_{\mathrm{B}}} = e^{-\Delta G_{\mathrm{AB}}/RT} \tag{I.3}$$

$$\begin{aligned} P_{\mathrm{A,abs}} &= \frac{P_{\mathrm{A}}}{\sum_i^N P_i} = \frac{e^{-G_{\mathrm{A}}/RT}}{\sum_i^N e^{-G_i/RT}} \\ &= \frac{1}{\sum_i^N e^{(G_{\mathrm{A}}-G_i)/RT}} \\ &= \frac{1}{\sum_i^N e^{-\Delta G_{\mathrm{A}i}/RT}} \end{aligned} \tag{I.4}$$

where:

ΔG_{AB} = $G_{\mathrm{B}} - G_{\mathrm{A}}$, free energy difference between the states A and B
R = gas constant
T = temperature of the system
K_{AB} = equilibrium constant for the states A and B
P_{X} = unnormalized probability of observing state X in the protein's equilibrium
$P_{\mathrm{X,abs}}$ = normalized probability of observing state X in the protein's equilibrium

I.1.2. Knowledge about alternative conformations is important for drug design

The value of knowing a protein's equilibrium constitution becomes immediately clear when designing drugs *in silico*. A computational approach frequently used by drug design methods is the "key-lock" approach. In this context, the protein's conformation is the "lock" and the "key" — the drug — is designed to fit into the lock.[30] However, this approach will not account for the drug binding to alternative conformations of the protein and therefore discards potentially effective drug candidates. Alternative methods include the "induced fit"[30] or "conformational selection"[31,32] approaches. For those approaches, the protein target is either treated with some conformational flexibility or an ensemble of conformations is provided. For the induced fit approach, the tertiary structure of the protein target is perturbed during the design process to simulate the existence of alternative conformations. Although this improves the probability of designing an effectively docking drug by accounting for the flexibility of the protein, this approach is also limited with regards to the conformational space that can be covered. The conformational selection approach on the other hand designs the drug against an ensemble of conformations, therefore also accounting for drugs binding to the protein in alternative conformations. This approach is more comprehensive but also more difficult to perform since the conformations of the protein's major populations have to been known. In the following section we discuss the currently available structural information about a protein's major populations and how computational methods can be able to provide additional information about alternative conformations.

The conformational selection method has been used successfully on multiple occasions for various targets.[32] The most prominent examples include the discovery of novel binding trenches in HIV integrase[33,34] and the *in silico* discovery of enzyme inhibitors.[35] The method of choice in those studies were molecular dynamics (MD) simulations that were employed to structurally diversify an initial structural model obtained through an orthogonal structure determination technique. Although the diversification approach using MD simulations provided promising results in those studies, the conformational space that can be covered by MD simulations might still be a limiting factor.[32]

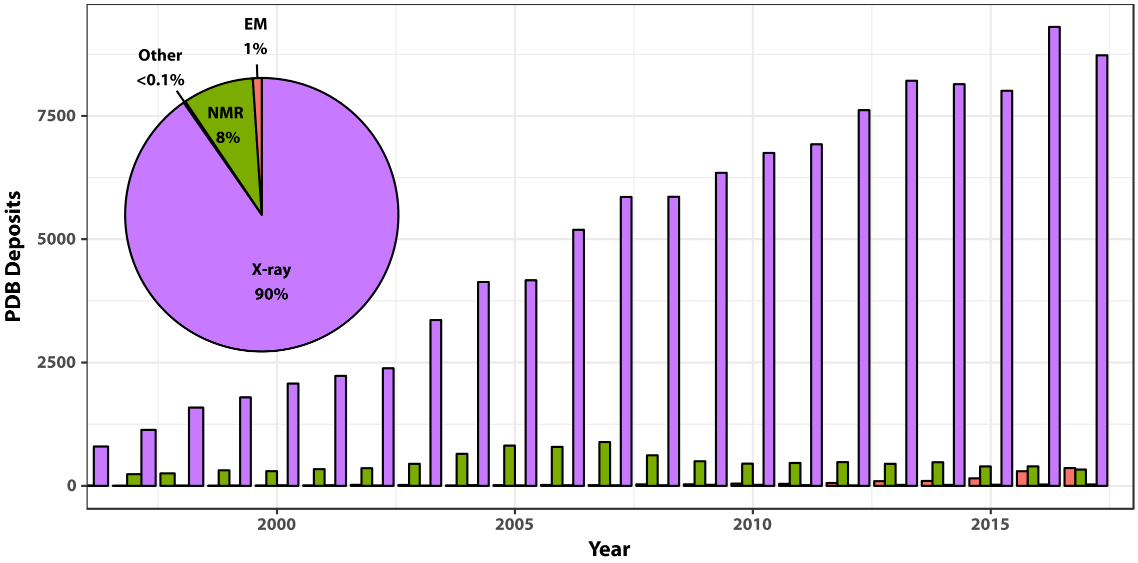

Figure I.1.: Techniques used for protein structure determination. *The pie chart shows the percentage of structure deposits in the PDB categorized by technique. The bar graph lists the PDB deposits per technique over the last twenty years. The years prior to 1997 were omitted in the bar graph due to low deposit numbers.*

I.2. Experimental techniques can obtain information about a protein's equilibrium constitution

In this section, we provide a summary of the structural information that is obtainable by different techniques to probe the structure and dynamics of proteins. The techniques discussed in detail are X-ray crystallography, NMR spectroscopy, EPR spectroscopy in conjunction with site-directed spin labeling (SDSL), electron microscopy (EM), and NMR spectroscopy. The Protein Data Bank (PDB)[36,37] has seen a rapid increase in the number of deposited structural models of proteins, mainly derived from X-ray crystallography, NMR spectroscopy, and EM. As of December 2017,[a] 125 799 structural models of proteins have been deposited in the PDB. Of those, 113 609 were derived from X-ray crystallography, 10 563 from NMR spectroscopy, 1322 from EM, and 305 were determined through other techniques or hybrid approaches (see figure I.1). Especially EM has experienced an increasing usage over the recent three years.

X-ray crystallography determines the molecular structure of a protein through an incident beam of X-rays. The crystalline atoms diffract the X-rays and through measuring the intensities and angles of the diffracted beams, the three-dimensional structure of the protein is computed. X-ray crystallography is frequently able to achieve sub-angstrom resolutions, which has been show in the recent past when structural models of the NADH-cytochrome b_5 reductase and of a hydrogenase were determined at resolutions of 0.78 Å[38] and 0.95 Å,[39] respectively. However, its major limitation is requirement of the protein entities to crystallize into regular crystal lattices.[40] Consequently, the studied protein frequently needs to be subjected to stabilizing mutations, resulting in a structural model that may or may not accurately represent one of the protein's major populations under *in vivo* conditions. Any information about alternative populations is lost due to the stabilization and crystallization process.

[a] The statistics were obtained from `https://www.rcsb.org/pdb/` on December 3, 2017.

EM has seen increasing use for the determination of structural models of proteins. In its newer variant cryo-EM, a flash-frozen sample of the protein is subjected to electron rays and an electron detector located behind the sample collects the passing through or moderately deflected electrons.[41] The rapid freezing of the specimen prevents the destruction of the sample through crystal formation and ensures that the structural equilibrium constitution is preserved. The result of the electron detection is a two-dimensional projection of the protein's electron density. Due to the protein being present and observed from multiple angles, a three-dimensional scalar field representing the electron densities of the protein can be reconstructed using computational methods. For several years, it has only been possible to obtain low- to medium-resolution density maps, e.g. ranging from 9 Å to 20 Å. Examples for this are the density maps of an adenovirus, which was determined at a resolution of 9 Å.[42] However, recent progress in the electron detectors and the computational methods enabled the determination of structural models at higher resolutions. This has been shown for density maps of a glutamate dehydrogenase and of a β-galactosidase that were determined at resolutions of 1.8 Å[43] and 2.2 Å,[44] respectively. Generally, EM is able to capture multiple states of the protein but to recompute the three-dimensional electron density map, the different states have to be identified and clustered. This is not always possible for states that only contribute a smaller population to the equilibrium and can result in a low resolution of the density map or discarding these states at general.

NMR spectroscopy can also provide information about the dynamics of the studied protein.[45] In NMR spectroscopy, the protein is subjected to a static magnetic field and the reaction of the nuclear spins to an electromagnetic pulse is measured. The spectroscopic data yields atom distance and bond angle restraints, which are used to derive either a single structural model or an ensemble of structural models. One of the major problems is the refinement of the structural models for agreement with the restraints derived from the spectroscopic data. The obtained restraints are observed on the equilibrium of the protein, therefore representing multiple states. However, frequently each structural model in the NMR ensemble is fitted against the whole set of restraints although it might not be possible for one model to satisfy all restraints at the same time. Additional difficulties arise from the size limit of NMR spectroscopy, which makes this technique frequently unsuitable for studying membrane proteins or large soluble proteins.[40]

EPR spectroscopy in conjunction with SDSL is typically performed on a cys-less variant of the protein by introducing two cysteine residues at the desired spin labeling sites. These are coupled with the spin label MTSL that carries an unpaired electron. Through the DEER experiment,[46] the dipolar interaction of the two unpaired electrons is measured. Because the interaction intensity of the two unpaired electrons is inversely proportional to their cubed distance, the Euclidean distance between the unpaired electrons can be computed[47,48] from the measured spectra. This measurement is typically performed on a flash-frozen sample that captures the protein's equilibrium constitution and therefore provides a distribution of distances that represents all present conformations in the equilibrium. However, due to the indirectness of the measurement — the experiment measures the distance between the free electrons that are close to the tip of a flexible spin label — and the sparseness of the data, it is typically not possible to unambiguously determine structural models for the protein's major populations from the SDSL-EPR data alone.[5]

In summary, X-ray crystallography, NMR spectroscopy, and EM account for more than 99.9 % of the structural models deposited in the PDB (figure I.1 on the previous page). However, all these techniques have specific limitations when it comes to the prediction of protein dynamics and alternative populations of the protein in question. SDSL-EPR spectroscopy on the other hand does not have any limits regarding the protein's size or dynamics but the yielded data is typically too sparse to unambiguously determine an ensemble of conformations. Pure computational approaches have their own limitations (see section I.3.1) but a combination of computational and experimental methods could compensate for the weaknesses in both approaches (see section I.3.2 on page 8).

I.3. Computational methods for the prediction of protein ensembles

In this section, we discuss the different computational approaches for the prediction of a protein's tertiary structure or an ensemble describing its structural constitution in the protein's equilibrium. In section I.3.1, we detail the general limitations of computational approaches in the context of structure and ensemble prediction. This is followed by a discussion of the advantages gained by incorporating limited experimental data into the computational prediction methods, which is provided in section I.3.2 on page 8. This section is concluded by a discussion of application examples for computational protein ensemble prediction in section I.3.3 on page 9.

I.3.1. Limitations of computational approaches

The free energy of a conformation depends on interactions of the protein with itself and interaction of protein with its environment. Interactions of the protein with itself include backbone hydrogen bonds, disulfide bonds, and ionic interactions.[49] However, in the crowded cell that is observed *in vivo*, there is also a wide variety of potential interactions between the protein and other molecules. These interactions among others include hydrogen bonding between the protein and cytosol, hydrophobic interactions between the protein and a cell membrane, and also interactions between the protein and other macromolecules in the cell.

Performing a full-atom simulation of the crowded cell is unfeasible due to the large number of potential interactions. Consequently, the biological system has to be simplified. This is usually achieved by discarding other macromolecules in the cell and only simulating the protein itself and its direct interaction partners, which include the molecules in the cytosol as well as lipids in the case of a membrane protein.

Many methods also simplify the representation of the solvent — cytosol and lipids. This is typically achieved by representing the solvent implicitly as volume units with certain assigned physical properties. Examples for this are the Rosetta software suite and BCL::Fold.[50] Both methods represent membranes implicitly as infinite plane. Accordingly, interactions between the protein and the lipids cannot be evaluated explicitly but need to be approximated using statistics. For both methods, this is achieved through the usage of knowledge-based potentials that quantify the likelihood of observing a certain residue type a specific depth in the membrane. For quantifying the interaction strength, different metrics are used. Most prevalent is the neighbor count that counts the number of residue neighbors weighted by their Euclidean distance for each residue. Similar approaches are used for evaluating interactions with the cytosol. A drawback of this approach is that the statistics used for generating the knowledge-based potentials are not fully reliable. For example, the statistics encompass different solutions or membrane compositions that differ from *in vivo* conditions. Additionally, the location of the membrane is frequently uncertain and its *in silico* representation is idealized and inflexible. In consequence, the interactions between the protein and its environment cannot be evaluated at atomic detail but a more coarse-grained approach is needed, introducing further inaccuracy into the approximation of a conformation's free energy.

High-resolution methods like MD simulations on the other hand attempt an approach with higher resolution. In many cases the environment of the protein is simulated explicitly, with the solvent being represented through actual atoms.[51,52] However, even these approaches subject the system to certain simplifications. Frequently the actual lipid composition of the membrane is not accurately simulated but represented through a small number of lipid types.[53,54] Alternative MD approaches completely avoid explicit simulation of the solvent and choose to use an implicit representation,[55,56] which can introduce additional inaccuracy.[57,58]

Further simplifications of the system can also affect the protein representation itself. Each residue of the protein contributes multiple degrees of freedom to the protein model. The backbone exhibits two rotational degrees of freedom around the (ϕ, ψ) angles. Additional rotational degrees of freedom are contributed by the side chain of the residue. Using a frequency-weighted average, each side chain exhibits an additional two rotational degrees of freedom on average. Consequently, exhaustive sampling of all possible conformations is frequently unfeasible and the number of sampled conformations needs to be restricted. This can be directly achieved using a simplified representation of the protein, e.g. where side chains are not modeled explicitly. Examples for these simplified model representations are especially prevalent in *de novo* protein structure prediction methods.

For example, the Monte Carlo Metropolis (MCM) method BCL::Fold assembles the protein model from idealized secondary structure elements (SSEs)[59] with only allowing limited deviations from the idealized dihedral angles. The side chains of the residues are also not modeled explicitly but represented by a single "superatom". The Rosetta software suite assembles the tertiary from fragments collected from the PDB, avoiding exhaustive sampling of the torsion angles. The Rosetta algorithm consists of multiple stages with the later stages sampling the side chain conformations using a rotamer library; again avoiding exhaustive sampling of the torsion angles. These simplifications result in problems when the conformations of the studied protein significantly differ from idealized structures or structures found in the PDB.

Using those approaches makes it unlikely to sample an accurate conformation for proteins that significantly deviate from structural models observed in the PDB. In consequence, computation of interactions at an atomic level are futile because the energetically most stable conformation might not even get sampled.

Aforementioned simplifications of system and model representation directly result in another problem — evaluation of a conformation's free energy. The system representations in conjunction with the non-exhaustive sampling hinder an accurate computation of the model's free energy since there is uncertainty about an atoms location or, due to the implicit system representation, the location of the atom is not known at all. Accordingly, the approximations made to the free energy evaluation have to match the resolution of the system representation. As an example, if the torsion angles of the system are sampled at steps of 45°, a scoring function approximating the free energy of the system may not be sensitive to torsion angle changes below the step size. Otherwise the system representation would not allow comprehensive exploration of the energy landscape defined by the approximation. Consequently, the resolution of the free energy evaluation needs to be reduced, which again results in additional ambiguity. An example of free energy approximations enabling adaptive reduction of the resolution is the knowledge-based potential. In this context, the inverse Boltzmann relation is used to derive the free energy of a conformation from statistics collected from structural models deposited in the PDB.[60] This approach is used by a variety of protein structure prediction methods like BCL::Fold,[60] Rosetta,[61] and I-TASSER.[62,63]

As a consequence of aforementioned simplifications made to the system representation, to the sampling of different conformations, and to the evaluation of a conformation's free energy, it is frequently not possible to distinguish low-energy conformations from high-energy conformations. This is best demonstrated by examining the results of the CASP experiment. The CASP experiment is community-wide double-blind protein structure prediction study. A variety of research groups participate with their protein structure prediction methods and the summarized prediction results are available on the CASP website.[b] The setup of the experiment as double-blind study ensures that the research group not only publish prediction cases for which their method works. Briefly, the tertiary structures of the benchmark proteins is determined by a group who does not know the groups participating in the CASP experiment. The participating groups send their predictions to the CASP organizers, who do not know the experimentally determined tertiary structures. The groups' predictions are anonymized and sent alongside the experimentally determined structures to another group that evaluates the prediction quality. The results are subsequently sent back to the CASP organizers, who create a ranking of the prediction groups based on the evaluation of the prediction qualities. The CASP experiment encompasses two protein target classes — free modeling targets and template-assisted targets. Whereas for template-assisted targets, a structural template with related primary structure is available, the free modeling targets typically have to predicted purely *de novo*. At the two most recent CASP experiments in 2014 and 2016,[64,65] the template-based modeling category continued to experience remarkable successes.[66] However, the free-modeling category — protein targets for which no suitable structural template was available — continued lagging behind.[67] This was particularly pronounce for proteins with a larger sequence length,[67] which indicates that *de novo* prediction of protein structures and therefore protein ensembles is still not feasible.

In summary, *de novo* protein structure prediction methods are not able to routinely predict the tertiary structure of a protein at its Gibbs free energy minimum within a reasonable accuracy limit. Consequently, it is doubtful that these methods would be able to predict a protein's equilibrium constitution because they are not able to routinely distinguish low-energy from high-energy conformations. Consequently, these methods need to be supplemented with additional data that simplifies model discrimination.

I.3.2. Combining experimental data with computational approaches

Structure and ensemble prediction for proteins is hindered by the necessary simplifications and approximations to the representation of the biological system and to the computation of a system's free energy (see section I.3.1 on page 6 for a detailed discussion). The currently available computational resources don't allow a comprehensive simulation of the biological system and, in consequence, an accurate energy evaluation. Although the limitation of computational resources is unlikely to change over the near future, the problems arising from the applied approximations can be mitigated through the incorporation of limited experimental data. As detailed in section I.2 on page 4, methods like EM or EPR can provide information about the structural constitution of a protein's equilibrium. The yielded data can be interpreted geometrically, therefore circumventing the need for a more accurate computation of the system's free energy. For the prediction of a singular tertiary structure, this has been demonstrated for both, EM and EPR.

EM has been successfully combined with computational methods in the past. In particular, when no high-resolution electron density maps are available, the computational methods can fill the information gaps in the EM data. This has been shown for medium-resolution density in conjunction with the BCL software suite and usage of the cross-correlation coefficient for quantifying the agreement of the

[b] http://www.predictioncenter.org

structural model with the EM data.[68–72] Alternative studies used the MultiFit method in conjunction with template-based modeling and complex docking to predict the macromolecular assembly from EM data.[73,74] However, also in those approaches only one structural model was predicted from the EM density maps.

EPR spectroscopy in conjunction with SDSL has been successfully combined with computational methods for the prediction of protein structures. Typically, this combinations requires two components: i) a method that explores the possible conformational space of the protein by sampling different conformations, and ii) a method that quantifies the agreement of the conformation or the ensemble of conformations with the measured SDSL-EPR data. A successful protocol published by Jeschke *et al.*[75] used the observed spin-spin distance as hard cut-off for the $C_\beta - C_\beta$ distance in conjunction with the MODELLER[76,77] software suite. Alternative approaches used the motion-on-a-cone (CONE) model[78] to translate the $C_\beta - C_\beta$ distance into a likelihood of an observed spin-spin distance.[2] However, those two examples only consisted of the prediction of a single conformation, which is presumably representing the protein's major population at the environmental conditions of the EPR experiment. Both approaches try to satisfy restraints derived from multiple states with a single state. Although the results of these studies demonstrate the suitability of this approach to improve the prediction accuracy when using an X-ray- or NMR-derived model as reference, new methods are needed for the prediction of protein ensembles.

In conclusion, experimental data from SDSL-EPR and EM experiments can successfully be combined with computational methods to fill information gaps in the experimental data and to compensate for the limitations intrinsic to the computational methods. However, the methods discussed so far have only been used to predict one structural model from the experimental data. First, this does not provide comprehensive information about the studied protein and, second, it might not even be feasible to fit one structural model against experimental data representing an ensemble of conformations.

I.3.3. Computational methods for protein ensemble prediction

Depending on the available experimental data, different computational approaches for the prediction of protein ensembles are viable. For many proteins it is possible to determine the tertiary structure of at least one state through experimental techniques like X-ray crystallography, NMR spectroscopy, or EM. This is demonstrated by the pace at which structural models get deposited in the PDB (see section I.2 on page 4 for details). For other proteins, it might be possible to determine a viable conformation using template-based modeling in conjunction with a structural model of a homolog. This initial conformation can then serve as starting point from where the conformational space is explored to identify alternative conformations with significant population sizes. One computational method to guide the exploration of the conformational space are MD simulations. This approach has been used in the past to structurally diversify drug receptors in support of computer-aided drug design.

One example is the *in silico* docking of the first inhibitor to the HIV integrase that was marketed as a drug.[33] Schames *et al.* simulated a 2 ns MD trajectory of the integrase, starting from a structural model derived from X-ray crystallography.[79] The structural information provided by the X-ray-derived model was deemed insufficient due to a flexible loop with uncertain conformation and location.[33,80] Consequently, the tertiary structure of the X-ray-derived model had to be diversified to find alternative conformations that might exist in the protein's equilibrium. MD trajectories were simulated and snapshots of the trajectories were taken as target for the ligand docking. Using this approach, Schames *et al.* found a novel binding trench in the HIV integrase that seemed inaccessible in the X-ray-derived model.[33]

Another example is provided by Durrant *et al.*, who identified an enzyme inhibitor using computational methods.[35] Also in this case, MD simulations were employed to structurally diversify the receptor and to gain a better understanding of the receptor dynamics. Unlike in the aforementioned example, five independent 20 ns MD trajectories were simulated to increase the structural diversity as opposed to one long simulation.[35] Clustering was subsequently used to identify twenty-four states that served as receptor ensemble for the ligand docking simulations. A conformational selection approach was then used to identify potential ligands.[31,35] Notably, none of these approaches used the Boltzmann relation (see section I.1.1 on page 2) to relate the *in silico* simulations and free energy approximations to population sizes. Instead, they performed dynamics simulations and clustering.

There are also cases when no suitable template is available. This happens when no conformation of the protein could be determined experimentally through X-ray crystallography, NMR spectroscopy, or EM and additionally there is no suitable homolog with known tertiary structure. This frequently is the case for membrane proteins, which are either too flexible for X-ray crystallography or too large for NMR spectroscopy.[40] For those cases, two approaches are possible: a) *de novo* prediction of one conformation and subsequent exploration of alternative conformations from this starting point using previously described approaches or b) *de novo* prediction of all conformations with a significant population size. If experimental data providing structural information about the protein's equilibrium constitution is available, the second approach seems more suitable because it allows the fitting of an ensemble against the experimental data.

I.4. Discussion

In order to acquire a comprehensive understanding of a protein's function and to effectively develop molecules able to affect a protein's function, its major populations in the equilibrium have to be determined. Knowing the tertiary structures associated with these populations as well as their population sizes enables the effective design of drugs against their receptor using a conformational selection approach. Most experimental techniques used to derive structural models of proteins are not able to provides this information in most cases. Orthogonal techniques like SDSL-EPR spectroscopy provide structural information about the protein's different conformations but the data is too sparse to unambiguously derive three-dimensional models (see section I.2 on page 4 for details).

Although structural models derived from X-ray crystallography or NMR spectroscopy have been used successfully in ligand docking and *in silico* drug design studies, there are documented cases where structural snapshots obtained from X-ray crystallography were insufficient for the development of ligands. To diversify the structure of the receptor and explore alternative conformations, research groups have successfully employed MD simulations using an X-ray-derived structural model as starting point (see section I.3.3 on the preceding page for details). However, those approaches were purely computational, which inevitably introduces additional uncertainty (see section I.3 on page 6 for details). This approach has also not been demonstrated to work if either there is no structural model that can be used as starting point or if there is a substantial structural dissimilarity between the starting model and the conformation binding the ligand. In addition sometimes experimental data about a protein's structural equilibrium constitution is available in the form of SDSL-EPR measurements or other data. This data could also be used to mitigate the limitations of computational methods.

In this dissertation, we discuss the different aspects for the development of a protein ensemble prediction pipeline. In order to achieve a computational ensemble prediction pipeline, several problems need to be overcome: i) A computationally efficient prediction pipeline for protein structures needs to be developed and tested on a preferably unbiased data set. In particular, the limitations in conformation sampling and in the approximation of the Gibbs free energy differences between the sampled conformations need to be identified. ii) Methods for the integration of experimental data for the purpose of model discrimination and structure validation have to be developed. In addition, it needs to be evaluated to what extent sparse experimental data is able to overcome problems in the approximation of the Gibbs free energy differences. iii) As opposed to the prediction of a singular model, an ensemble prediction pipeline must be able to cover all conformations with a significant population size within a limited number of CPU cycles. In consequence, methods need to be developed to speed up the conformation sampling. iv) Protocols for the prediction of protein structures, protein ensembles, and protein dynamics from limited experimental data need to be established and evaluated on biologically relevant systems.

This work, depicted in the following chapters, provides a discussion of methods and protocols relevant to evaluating aforementioned items. Chapter II on the following page presents and discusses a *de novo* protein structure prediction pipeline that was evaluated in the CASP experiment. This is followed by a description and evaluation of a protocol for using SDSL-EPR distance and accessibility measurements for membrane protein structure prediction in chapter III on page 29. Chapter IV on page 48 details a method to incorporate data from XL-MS experiments into protein structure prediction. Additionally, this chapter discusses what spacer length is optimal for maximizing the obtainable structural information. This is followed by a discussion and evaluation of a method for the rapid sampling of loop conformations using conformation hashing in chapter V on page 67. The subsequent chapter VI on page 83 evaluates to what extent limited experimental data from EPR experiments can overcome problems in the approximation of Gibbs free energy differences between the sampled conformations. This is followed by two chapters detailing protocols and results for the structure, ensemble, and dynamics predictions of the proteins ExoU (chapter VII on page 99) and EmrE (chapter VIII on page 111). The body of this work is concluded by a discussion of the presented results and the future directions in chapter IX on page 126.

CHAPTER II
EVALUATION OF BCL::FOLD IN THE CASP11 EXPERIMENT

This chapter is based on the publication "CASP11 – An Evaluation of a Modular BCL::Fold-Based Protein Structure Prediction Pipeline".[1] Axel W. Fischer contributed to the development of the protein structure prediction pipeline, performing the experiment, analyzing the data, and writing the article.

In silico *prediction of a protein's tertiary structure from its primary structure remains an unsolved problem. The community-wide CASP experiment provides a double-blind study to evaluate improvements in protein structure prediction algorithms. We developed a protein structure prediction pipeline employing a three-stage approach, consisting of low-resolution topology search, high-resolution refinement, and MD simulation to predict the tertiary structure of proteins from the primary structure alone or including distance restraints either from predicted residue-residue contacts, NMR-nuclear overhauser effect (NOE) experiments, or XL-MS data. The protein structure prediction pipeline was evaluated in the CASP experiment 2014 (CASP11) on twenty regular protein targets as well as thirty-three "assisted" protein targets, which also had distance restraints available. Although the low-resolution topology search module was able to sample models with a global distance test total score (GDT_TS) greater than* 30 % *for twelve out of twenty protein targets, it was frequently not possible to select the most accurate models for refinement, resulting in a general decay of model quality over the course of the prediction pipeline. In this study, we provide a detailed overall analysis, study one target protein in more detail as it travels through the protein structure prediction pipeline, and evaluate the influence of limited experimental data on the prediction accuracy.*

II.1. Introduction

In silico prediction of a protein's tertiary structure from its primary structure remains an unsolved problem. The vast size of the conformational space that needs to be sampled with a limited number of CPU cycles requires simplifications in sampling and scoring, often in conjunction with a simplified representation of the protein. Consequently, the depth of the native energy minimum is reduced, making it difficult to distinguish it from alternative energy minima.[81–85] The limited sampling density results in an intrinsic, minimal deviation of the conformations sampled from the lowest energy conformation that exists in each region of the conformational space further adding to the uncertainty.[81,83] In addition, the environment of the protein — the cytoplasm or the membrane — is represented in an implicit and static way, adding another layer of inaccuracy to free energy evaluation.

The *de novo* protein structure prediction algorithm BCL::Fold[59] was developed as part of the BCL to overcome aforementioned problems and efficiently predict the topology of larger proteins with up to 400 residues. The necessary complexity reduction of the sampling space is achieved by assembling predicted SSEs using a Monte Carlo (MC) algorithm and omitting more flexible loop regions. The energy evaluation of the sampled models is performed using knowledge-based scoring functions,[60] which provide a rapid way to approximate the free energy of the sampled conformations. In a previous study, it was demonstrated that BCL::Fold is able to efficiently sample the topologies of larger proteins.[86] Problems in model discrimination, which can arise from necessary simplifications made to sampling, scoring, and system representation, could be compensated for through incorporation of limited experimental data from EM,[68,70,87] NMR spectroscopy,[50] EPR spectroscopy,[2,7] XL-MS experiments,[3] small angle X-ray and neutron scattering,[88] and predicted residue-residue contacts.

To evaluate the accuracy of the described protein structure prediction pipeline, we participated in the community-wide CASP experiment in 2014 (CASP11), which takes place every two years.[89] Due to its setup as a double-blind study, the CASP experiment provides an impartial benchmark for protein structure prediction algorithms. The experimentally determined tertiary structures of the benchmark proteins are withheld from predictors, assessors, and organizers until conclusion of the experiment. After conclusion of the experiment, the experimentally determined structures are released to predictors and assessors and the predicted structures are released to the assessors, who determine the accuracy of the predictions. At the CASP experiment, the amino acid sequences of fifty-five proteins were released to human predictors as regular targets (T0), *i.e.*, without additional experimental restraints. Several regular targets were rereleased as "assisted" targets with additional structural information in terms of predicted residue-residue contacts (TP), only correct residue-residue contacts (TC), NMR-NOE (TS), and XL-MS restraints (TX). As of June 2015, experimentally determined structures have been released for thirty regular protein targets. Of those, we predicted the tertiary structure of twenty targets during the CASP experiment. Therefore, the analyses in this study are based on twenty T0, twelve TP, twelve TC, eight TS, and one TX protein target (see table II.1 on the following page for a list of all benchmark proteins).

In section II.2, we describe in detail the protein structure prediction pipeline employed in the CASP experiment. In addition, we define the different quality metrics used in this study and we introduce the subset of the CASP benchmark set analyzed in this study. Section II.3 on page 20 reports the accuracy results for the different pipeline modules and describes the protein structure prediction pipeline on hand for one regular target in detail. A major focus is put on the model selection problem and the resulting model accuracy decay over the course of the prediction pipeline. Section II.4 on page 27 discusses the successes and failures of our pipeline.

II.2. Materials and methods

This section details the different modules of the employed protein structure prediction pipeline — low-resolution topology search, high-resolution refinement and loop construction, and MD refinement. This description is followed by a subsection describing how clustering was used to aggregate and transfer models between the different pipeline modules. The subsequent subsections describe the quality metrics used to quantify the prediction results in terms of sampling accuracy and model discrimination. This section is concluded by a summary of the proteins used in this study.

II.2.1. Low-resolution topology search with BCL::Fold

BCL::Fold was specifically developed to predict the topologies of large proteins with a low-resolution approach. The BCL::Fold method was also specifically designed to complement Rosetta[94] by predicting SSE-only models with likely topologies of the protein and feeding them into Rosetta for loop and side chain construction as well as high-resolution refinement. The complexity of the conformational space grows exponentially with the number of residues in the protein, rendering exhaustive sampling of the conformational space impossible even for proteins with sequence lengths less than 100 residues. Protein structure prediction groups have come up with different approaches to address this problem. For example, Rosetta assembles the tertiary structure of proteins by assembling short fragments collected

Target	#aas	#α	#β	CO	PDB	Res. (Å)	Predicted as	Baker (%)	Zhang (%)	Kihara (%)
T0759	113	5	6	24	4Q28	2.6	—	38	40	32
T0761	252	5	13	55	4PW1	2.1	TP,TS,TC	14	15	16
T0763	134	3	11	52	4Q0Y	1.7	TP,TS,TC	15	19	18
T0765	98	3	4	59	4PWU	2.5	—	74	79	47
T0767	296	8	15	57	4QPV	1.8	TP,TS,TC,TX	13	16	15
T0769	112	2	4	44	2MQ8	4.3	—	75	81	74
T0771	186	3	10	56	4QE0	1.9	—	23	24	22
T0781	390	12	17	75	4QAN	2.1	—	17	17	11
T0783	411	14	20	52	4CVH	2.4	—	44	47	44
T0785	145	1	9	30	4D0V	1.7	TP,TS,TC	26	27	26
T0794	506	6	28	61	4CYF	2.3	TP,TS,TC	45	44	36
T0803	520	2	12	34	4OGM	2.2	—	52	40	38
T0814	403	3	38	78	4R7F	2.3	TP,TS,TC	11	15	15
T0818	138	4	12	34	4R1K	2.6	TP,TS,TC	35	42	43
T0831	420	15	2	116	4QN1	2.5	TP,TC	16	16	16
T0832	241	11	0	67	4RD8	1.7	TP,TS,TC	13	17	17
T0834	222	10	6	51	4R7Q	2.0	TP,TC	14	16	20
T0848	326	9	18	50	4R4G	2.6	TP,TC	30	29	28
T0853	152	3	10	38	2MQB	1.0	TP,TC	16	35	31
T0855	119	4	6	37	2MQD	1.3	—	40	45	45

Table II.1.: The proteins used in this study for the CASP benchmark. *Twenty regular protein targets from the CASP benchmark set were used in this study. The proteins covered a wide range of structural features, like the sequence length (#aas), the number of α-helices and β-strands (#α and #β), as well as the complexity of their fold quantified through the contact order (CO). Several regular targets were rereleased with limited experimental data in terms of predicted residue-residue contacts (TP), only correct residue-residue contacts (TC), NMR-NOE restraints (TS), and XL-MS restraints (TX). GDT_TS-values of the submitted models are shown for three other groups (Baker,[90] Zhang,[91,92] and Kihara[93]) for comparison.*

from experimentally determined structures deposited in the PDB. This approach substantially reduces the complexity of the sampling space because the dihedral angles are not exhaustively sampled. Using rotamer libraries provides a similar simplification for the side chain conformations. However, even with mentioned simplifications the size of the conformational space remains too large for many proteins with more than 100 residues. Additionally, previous studies found that *de novo* prediction using Rosetta exhibits a bias towards structures with low CO.[95]

Unlike Rosetta, BCL::Fold assembles disconnected fragments with limited internal flexibility to remove this bias. Secondary structure prediction methods are employed to assign the secondary structure to the primary structure. For the resulting SSEs, initial conformations are created from idealized dihedral angles (ϕ, ψ): (−60°, −40°) for α-helices and (−135°, 135°) for β-strands. BCL::Fold assembles the SSEs in the three-dimensional space using an MCM algorithm. Unlike in Rosetta, loop regions connecting the SSEs are not constructed explicitly, further reducing the complexity of the sampling and allowing for changing the overall topology in a small number of MC steps. Instead, the likelihood of the loop being able to close on the current conformation is predicted. Further complexity reduction is achieved by representing the side chains as "superatoms", avoiding explicit sampling of side chain conformations. BCL::Fold has this approach in common with Rosetta and other modeling approaches. Although these simplifications of the structural representation allow for an efficient enumeration of different topologies, a high-resolution scoring of the sampled models is no longer possible. Instead, BCL::Fold employs low-resolution scoring functions to evaluate geometrical parameters of the sampled models. These scoring functions include the likelihood of closing a loop given the number of amino acids and the Euclidean distance between two SSEs or if the twist angle between SSEs allows for side chain interaction among others. Most scoring functions used in BCL::Fold are knowledge-based, meaning they are derived from statistics over known protein structures deposited in the PDB and based on the inverse Boltzmann relation that is described in equation (II.1).

$$E = -RT \cdot \ln \frac{P_\mathrm{o}}{P_\mathrm{b}} \tag{II.1}$$

where:

E = free energy of the protein structure
P_o = probability of observing a specific feature
P_b = probability of observing a specific feature by chance
R = gas constant
T = temperature

The normalization of P_o by P_b is conducted to ensure that favorable states are assigned a negative score and unfavorable states are assigned a positive score. For example, the scoring function evaluating the burial of residues quantifies the degree of burial using the neighbor count metric.[96] For each amino acid type, the occurrences of its neighbor count-values were collected from structures deposited in the PDB. The values were binned and the probability of each bin was computed and used as P_o. The background probability, P_b, was in this context the normalized sum of all normalized amino acid exposure distributions.[60] The BCL scoring function is the weighted sum of all scoring terms.

II.2.2. Protein structure prediction pipeline

The protein structure prediction pipeline consisted of three modules (figure II.1 on the next page). The first module consisted of a low-resolution topology search, which applied large-scale structural

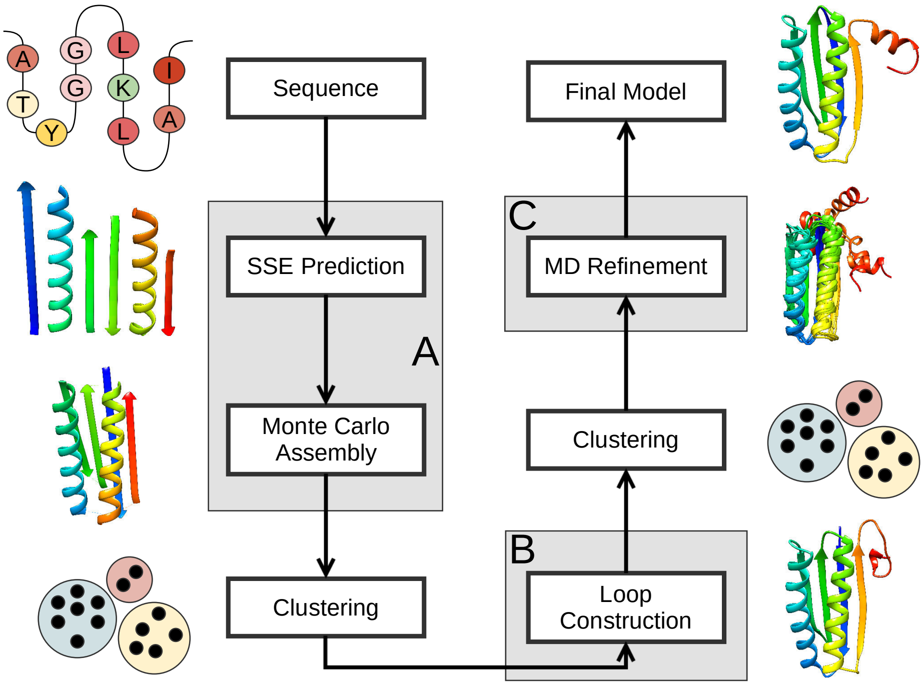

Figure II.1.: Protein structure prediction pipeline used in the CASP experiment. *The protein structure prediction pipeline consisted of three modules — low-resolution topology search from predicted SSEs using BCL::Fold (A), high-resolution refinement and loop construction using Rosetta (B), and MD refinement using the Amber package (C).*

perturbations to the model in conjunction with a rapid low-resolution scoring function (see section A.2.1 on page 159). The second module consisted of a high-resolution structural refinement, which only applied small-scale structural perturbations to the model in conjunction with a high-resolution scoring function while also constructing loop regions and placing the side chains (see section A.2.3 on page 161). The third module consisted of a MD simulation for further structural refinement and evaluation of model stability. The three modules were connected through filtering and clustering steps to transfer protein models from one module to the other (see section A.2.2 on page 160 for details).

The protocol for the first module was based on the previously published protein structure prediction protocol of BCL::Fold for soluble proteins.[59] In a first step, the secondary structure prediction methods Jufo9D,[97] PSI-blast based secondary structure PREDiction (PSIPRED),[98] and MASP[99] were employed to predict the protein's secondary structure. The protein's tertiary structure was subsequently assembled from the predicted SSEs through an MC sampling algorithm in conjunction with a Metropolis criterion. After each MC step, the model was evaluated using the weighted sum of multiple knowledge-based scoring functions including SSE packing, radius of gyration, residue exposure, residue-residue interactions, loop closure geometry as well as residue-residue and SSE-SSE clashes.[60] Depending on the score difference to the previous MC step and the simulated temperature, the new model was either accepted or rejected by the Metropolis criterion. The MCM optimization was broken down into six stages. The first five stages consisted of large-scale structural perturbations to search the energy landscape for minima. The employed perturbations included adding SSEs from the predicted SSE pool, removing SSEs from the model, large-scale translations and rotations of SSEs as well as the flipping and swapping of SSEs and SSE domains. Over the course of the first five stages, the weights of the scores evaluating clashes between residues and SSEs ramped up from 0 to 125, 250, 375 and 500. The five stages applying large-scale structural perturbations were followed by one stage of small-scale structural perturbations to transfer the model to the local energy minimum. If residue-residue contacts, NMR-NOE data, or XL-MS data were available, the scoring function was extended by the appropriate scoring terms.[3,50] For each protein target, the first module sampled 20 000 SSE-only models without explicitly modeled side chains or loop regions.

On conclusion of the first module, the models were ranked according to their completeness. The 25 % to 50 % of the models with the lowest completeness were filtered out. The filtering threshold was chosen in dependence of the maximum completeness achieved throughout the conformation sampling. For the different targets, 10 000 to 15 000 models remained. For the remaining models, clustering was used to detect limit points in the sampling space, which indicate energy minima. The clustering was performed using a k-means implementation[100] in the R[a] package. For the different targets, this resulted in 10 to 50 clusters. The cluster medoids were subsequently selected as start models for the second module.

The protocol for the second module was based on Rosetta,[94,101] added loop regions and side chains to the model, and conducted a high-resolution refinement. For each of the models resulting from the previous clustering step, 1000 models were sampled using Rosetta's CCD algorithm[101] followed by model relaxation using Rosetta's "relax" application.[94] Per target, this module resulted in 10 000 to 50 000 models.

On conclusion of the second module, the models were ranked according to their Rosetta score and the 80 % of the models with the worst score were discarded. The remaining 2000 to 10 000 models were clustered according to the same criteria as the first clustering step. After filtering out clusters with a population of less than 0.5 % of all models, this step resulted in 10 to 35 clusters. The cluster medoids were subsequently selected for high-resolution refinement and stability evaluation through MD simulations.

The third module consisted of MD simulations using the Amber package.[102] Hierarchical clustering was used to identify the sub-states for each model. Subsequently, a representative of each cluster was relaxed and scored using Rosetta. This module resulted in 10 to 35 models, which were visually inspected. The visual inspection was performed to filter out models with trivial implausibilities, like occlusion of known binding sites or mechanic frustrations of knows flexible regions. From the remaining models, five models were selected for submission to the CASP organizers. The selection criterion for this step were the Rosetta scores of the models.

II.2.3. Using clustering for model selection

The protein-size-normalized root-mean-square-distance (RMSD100) metric[103] was used to quantify the distance between models. The metric can be computed from the C_α-root-mean-square-distance (RMSD) and the protein's sequence length l as defined in equation (II.2).

$$RMSD100 = \frac{RMSD}{1 + \log \sqrt{l/100}} \tag{II.2}$$

where:

[a] https://www.r-project.org

$RMSD100$ = protein-size-normalized $RMSD$
$RMSD$ = root-mean-square distance of the C_α-coordinates
l = number of residues in the protein

The set of all models was sorted by their score and divided into the disjoint sets *high* and *low*. The set *low* contained the 20 % of the models with the most favorable score, whereas the set *high* contained the remaining models. Both sets were clustered independently. The number of clusters was optimized to minimize the cluster radii and to maximize the separation between clusters, with an allowed maximum radius of 5 Å. Clusters that contained less than 0.5 % of all models were also filtered out. The clustering after loop construction and side chain placement was conducted the same way as the clustering after the low-resolution topology search, but only the 20 % of the models with the most favorable Rosetta score were considered.

II.2.4. Molecular dynamics simulations

All simulations were prepared using Tleap[102] and simulations were performed with the Amber package[102] using the ffSB98ildn force field.[104] Each refinement target was solvated in a 10 Å TIP3P[105] water box with neutralizing Na^+ or Cl^- ions then equilibrated following a modified procedure.[106] First, the solvent was minimized for 500 steps using steepest descent, followed by 5000 steps of conjugate gradient minimization. Next, the systems were heated from 100 K to 300 K over 20 ps with 500 $\text{kcal mol}^{-1}\,\text{Å}^{-2}$ restraints on the protein followed by 30 ps of NPT at 300 K and 1 atm pressure. This process was repeated with restraints of 100, 50, 25, 12.5 and 1 $\text{kcal mol}^{-1}\,\text{Å}^{-2}$. After equilibration, each structure consisted of a 50 ns NPT production run at 300 K with periodic boundary conditions using Langevin dynamics[107] with a collision frequency of 5 ps^{-1}. The electrostatics were calculated using particle mesh-ewald[108] while a 10 Å cut-off was used to calculate long-range interactions. The SHAKE[109,110] algorithm constrained all covalent bonds with hydrogen atoms allowing a 2 fs time step. Each production trajectory was analyzed using Cpptraj.[111] Hierarchical clustering using complete-linkage was used to identify all sub-states for each model. Subsequently, one representative from each cluster was scored with the Rosetta[94] application.

II.2.5. Evaluation of the prediction accuracy

The prediction results were evaluated in terms of aampling accuracy and model discrimination. The sampling accuracy was quantified using the GDT_TS[112] metric. The GDT_TS of a model is the average percentage of C_α-coordinates in the model with a maximum deviation of 1 Å, 2 Å, 4 Å and 8 Å from the experimentally determined structure. The GDT_TS is computed as $GDT\_TS = (P_1 + P_2 + P_4 + P_8)/4$ with P_i being the percentage of residues in the model that can be superimposed with maximum deviation of i Å from the experimentally determined structure. The model discrimination is quantified through the enrichment metric, which equates to the percentage of the most accurate models that can be selected by the scoring function (see section II.2.6).

II.2.6. Computation of enrichments

The enrichment describes the correlation between model accuracy and score; thus, quantifying how well the scoring function is able to distinguish accurate models from inaccurate models. To compute the enrichment, the set of the sampled models S is divided into the disjoint subsets P (positive) and N

(negative). The positive set contains the 10 % of the models in S, which have the lowest RMSD100-value. The negative set contains the remaining models in S. In a second step, S is divided again into the disjoint subsets PS (positive score) and NS (negative score). The set PS contains the 10 % of the models in S, which have the best score, whereas the set NS contains the remaining models. By computing the intersection $TP = P \cap PS$, the set of the models, which can be identified by the scoring function, can be calculated. The enrichment is then calculated as described in equation (II.3).

$$e = \frac{|TP|}{|P|} \cdot 10 \tag{II.3}$$

where:

e = enrichment
P = 10 % most accurate models according to their GDT_TS-value
TP = 10 % most accurate models that are also among the 10 % best scoring models

Thus, the enrichment metric describes, which fraction of the most accurate models can be identified by the scoring function. Therefore, the enrichment can span a range from 0 to 10, with 1 indicating random selection, enrichments greater than 1 indicating that the scoring function has the ability to recognize native-like models, and an enrichment of less than 1 indicating that the scoring function is selecting against accurate models

II.2.7. The CASP benchmark subset used in this study

The analyses in this study are based on twenty soluble proteins released as targets during the CASP experiment. The twenty benchmark proteins covered a wide range of structural properties (table II.1 on page 14), making them an appropriate test case for protein structure prediction algorithms. The sequence length ranged from 109 to 470 residues and the secondary structure content ranged from 6 to 41 SSEs. SSE definitions were obtained through the algorithm Dictionary of Secondary Structure of Proteins (DSSP).[113] The α-helical content ranged from 1 to 15 SSEs, whereas the β-strand content ranged from 0 to 38 SSEs. The fold complexity quantified through the CO metric[114] ranged from 34 to 116. Twelve of the twenty regular targets were also studied using additional structural data such as residue-residue contacts, NMR-NOE restraints, or XL-MS restraints (table II.1 on page 14).

II.2.8. The available experimental data

For twelve protein targets, limited experimental data was provided by the CASP organizers. The experimental data included predicted residue-residue contacts (TP and TC), NMR-NOE restraints (TS), and XL-MS restraints (TX). The residue-residue contacts were predicted by research groups participating in the CASP contact prediction experiment and included correct and incorrect residue-residue contacts for the TP targets. After completion of the TP predictions, a subset of the contacts only containing correct residue-residue contacts was released. The NMR-NOE restraints were simulated by Gaetano Montelione's group and incorrect restraints were added purposefully. The XL-MS restraints were determined experimentally by Juri Rappsilber's group.

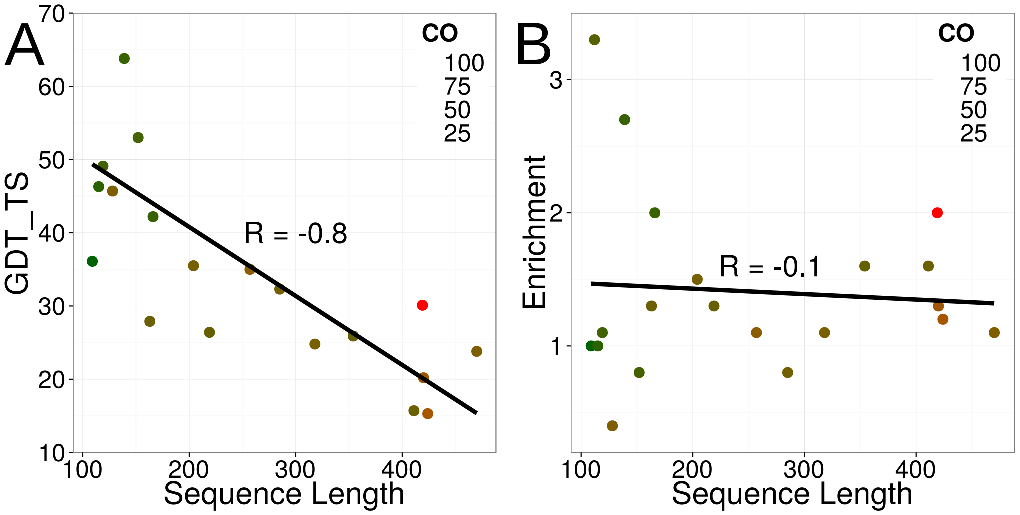

Figure II.2.: Sampling accuracy and model discrimination for the CASP protein targets. *(A) GDT_TS-value of the most accurate model for each regular target sampled by the low-resolution topology search in dependence of the sequence length. (B) Model discrimination for each regular targets as quantified through the enrichment metric in dependence of the proteins' sequence lengths. The coloring is according to the proteins' CO-values.*

II.3. Results

This section is divided into subsections discussing the sampling accuracy and model discrimination of the low-resolution topology search module, followed by a subsection discussing the general decay of model accuracy over the course of the protein structure prediction pipeline. Subsequently, a case study for target T0769 describes in detail the processing of the data through the protein structure prediction pipeline. This section is concluded by a subsection describing the impact of different types of limited experimental data on protein structure prediction accuracy.

II.3.1. BCL::Fold sampled models with a GDT_TS greater than 30 % *for twelve out of twenty regular targets*

To quantify the ability of BCL::Fold to sample the topology of the target proteins, the GDT_TS metric was used. The GDT_TS metric computes the average percentage of C_α-coordinates in the model that deviate maximally 1 Å, 2 Å, 4 Å and 8 Å from the experimentally determined structure (see section II.2.5 on page 18 for details). For twelve out of twenty regular targets, BCL::Fold sampled models with a GDT_TS-value greater than 30 % (table A.1 on page 157 and figure II.2). The average GDT_TS-value over all twenty regular targets was 36 % (table A.1 on page 157). The success in sampling accurate models strongly depended on the length of the protein's sequence (R-value of −0.8, table A.1 on page 157 and figure II.2). Notably, there was no dependence of the sampling accuracy on the complexity of the protein's topology as quantified through the CO metric (R-value of 0.0).

II.3.2. The BCL::Fold scoring function was frequently unable to select accurate models

After conclusion of the first pipeline module — the low-resolution topology search — models were selected for high-resolution refinement and loop construction with Rosetta. Although the model

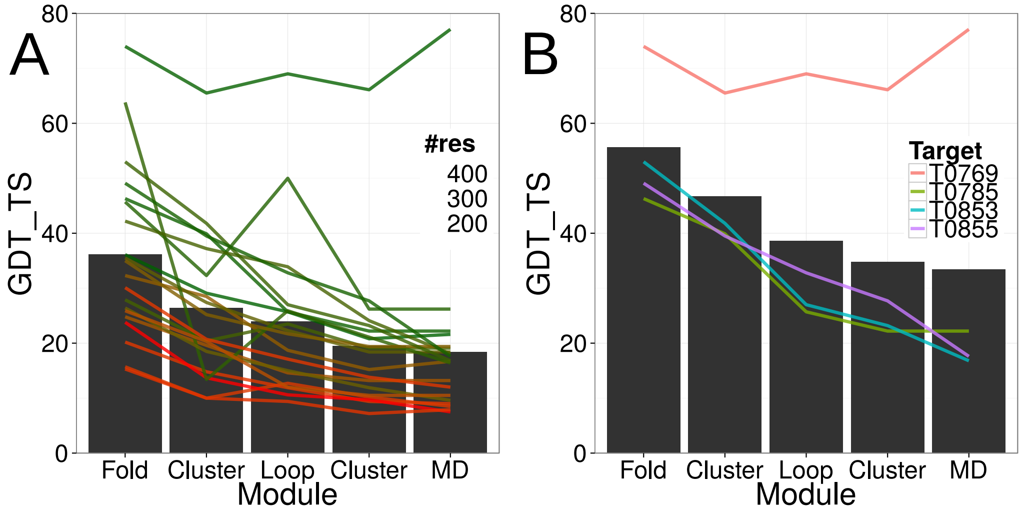

Figure II.3.: Model accuracy decay over the course of the CASP protein structure prediction pipeline. *(A) The model accuracy decayed over the course of the protein structure prediction pipeline. The black bars show the average GDT_TS-value of the most accurate model over all twenty regular targets after each pipeline module. The lines show the development of model accuracy for each target over the course of the pipeline. The coloring is according to the number of residues in the protein target. (B) Same as in (A) for four selected targets with a GDT_TS-value of greater than* 40 % *after the first clustering step.*

selection was conducted using a clustering approach, how well the BCL::Fold scoring function identifies accurate models remains an interesting question. The ability of the scoring function to select the accurate models among the sampled models was quantified using the enrichment metric. The enrichment metric computes the percentage of the most accurate models that can be selected by the scoring function (see section II.2 on page 13 for a definition).

Over all twenty regular targets the average enrichment was 1.4 (table A.1 on page 157), meaning that 14 % of the most accurate 10 % models could be selected by the BCL::Fold scoring function, which is only slightly better than random selection. There was no clear correlation between the enrichment and the sequence length, the complexity of the protein's fold, or the number of α-helices and β-strands in the protein (figure A.1 on page 155). However, the model selection in our pipeline was not conducted through direct usage of the BCL score, but through clustering to identify limit points, which indicate score minima. To evaluate the success of this approach, we computed for each protein target the percentage of models that had a GDT_TS-value greater or equals 40 %, assuming with a high enough percentage, those models can be detected through clustering. A density in this context could be seen as significant if it surpassed the population cutoff of 0.5 % during the first clustering step. For the regular targets T0769, T0785, T0803, T0853, and T0855 significant densities accounting for 54 %, 1 %, 47 %, 4 % and 1 % of all models could be detected. For the remaining targets, the percentages of models with a correct topology were below 0.5 %. Notably, for four out of the five of the aforementioned protein targets, models with a GDT_TS-value greater or equals 40 % could be detected through clustering (table A.1 on page 157).

II.3.3. Model accuracy decayed over the course of the pipeline

The three different modules of our protein structure prediction pipeline were connected through filtering and clustering. In an optimal scenario, the most accurate models would be detected through clustering and transferred to the subsequent module. However, ambiguities in the employed scoring function and the consequently biased sampling lead to difficulties in detecting the most accurate models. In clustering, native-like conformations become detectable if a sufficiently high density of models exists around it. For the four targets T0769, T0785, T0853, and T0855, models with a GDT_TS-value greater or equals 40 % could be detected through clustering after the low-resolution topology search and transferred to the second module for loop construction and side chain placement (table A.1 on page 157 and figure II.3 on the previous page). The average GDT_TS-value of the most accurate models for the four regular targets developed from 56 % to 47 %, and to 39 % over the course of the low-resolution topology search, the first clustering, and the loop construction and side chain placement steps (figure II.3 on the preceding page). At general, a decay of model accuracy was observable over the course of the protein structure prediction pipeline (table A.1 on page 157 and figure II.3 on the preceding page). The average GDT_TS-values over all twenty regular targets dropped from 36 % (low-resolution topology search) to 26 % (first clustering), to 24 % (loop construction and side chain placement), to 20 % (second clustering), and to 18 % (MD refinement). Expectedly, the most significant loss in model accuracy happened during the transition for the low-resolution topology search to loop construction and side chain placement where the average GDT_TS-value over all twenty regular targets dropped from 36 % to 24 %. A significant improvement through MD refinement could only be observed for regular target T0769 for which the GDT_TS-value of the most accurate model improved from 66 % to 77 %. For the other regular targets, the GDT_TS-value of the most accurate start model was 27 % or less and MD refinement consequently was not able to improve the accuracy of the model. For the previously mentioned regular target T0765, the most accurate models sampled by the loop construction and side chain placement module could not be detected through the clustering and filtering steps before the MD refinement. Consequently, the accuracy of the starting model for the MD refinement was low and the resulting models also exhibited low structural similarity to the experimentally determined reference structure.

II.3.4. A case study of regular target T0769

The regular target T0769 was a 112-residue-long soluble protein consisting of two α-helices and four β-strands, resembling a ferredoxin fold. The first module of our protein structure prediction pipeline — the low-resolution topology search — sampled models with GDT_TS-values of up to 74 % (table A.1 on page 157 and figure II.4 on the following page). An enrichment of 3.3 was observed indicating that 33 % of the 10 % most accurate models could be selected by the scoring function. About 69 % of all models had the correct topology. Through clustering, a model with a GDT_TS-value of 65 % could be detected (figure II.4 on the next page). In the second module of the pipeline, the loop regions were constructed and the side chains were placed. The most accurate model resulting from this pipeline module arrived at a GDT_TS-value of 69 % (figure II.4 on the following page). The models resulting from the second module were clustered again and the cluster medoids selected for MD refinement. The most accurate medoid had a GDT_TS-value of 66 %. Upon conclusion of the MD simulations, the refined models were rescored using Rosetta and the model with the most favorable Rosetta score was designated as final model. The final model arrived at a GDT_TS-value of 77 % (figure II.4 on the next page).

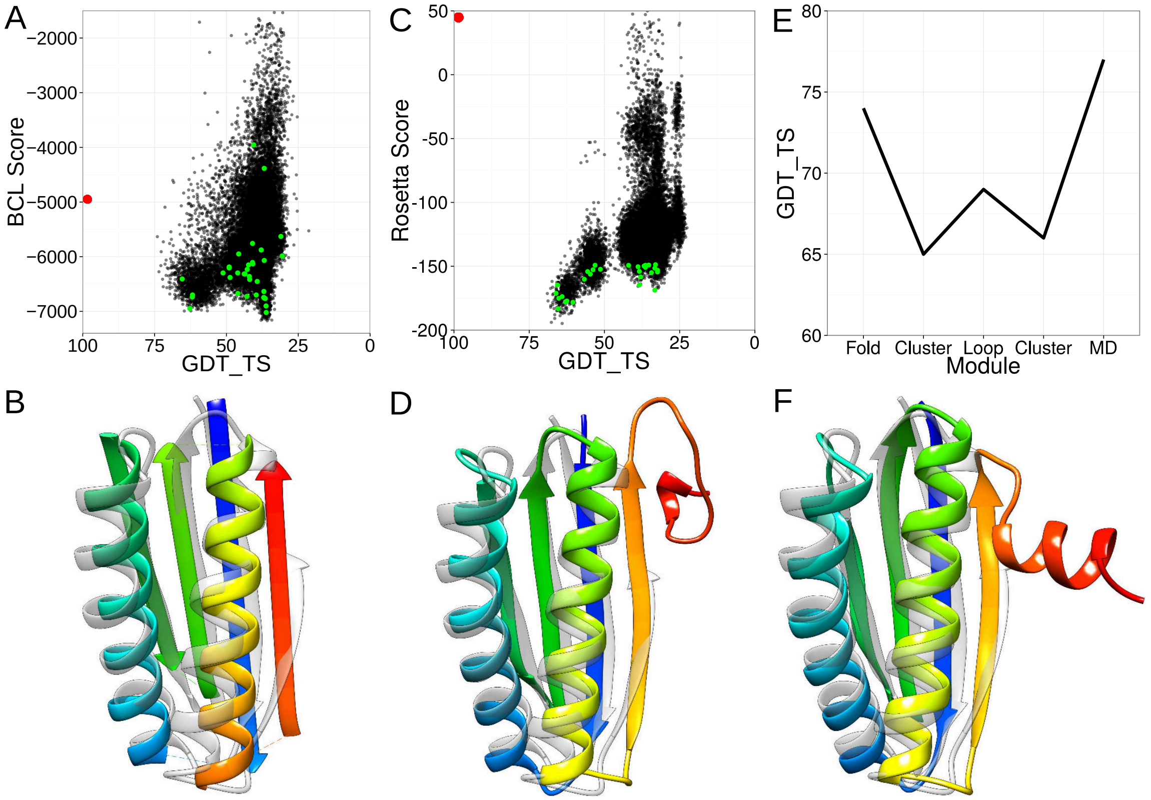

Figure II.4.: Case study of regular target T0769. *(A) Results for the low-resolution topology search. Each black dot represents one sampled model. The NMR structure is shown in red. The green dots are the cluster medoids selected after the topology search. (B) Most accurate model after the topology search (rainbow) superimposed with the NMR structure (grey). (C) Results for high-resolution refinement and loop construction. Each black dot stands for one sampled model. The NMR structure is shown in red. The green dots are the cluster medoids selected after the high-resolution refinement. (D) Most accurate model after the high-resolution refinement (rainbow) superimposed with the NMR structure (grey). (E) Development of the GDT_TS of the most accurate model over the course of the pipeline. (F) Most accurate model after the molecular dynamics refinement (rainbow) superimposed with the NMR structure (grey).*

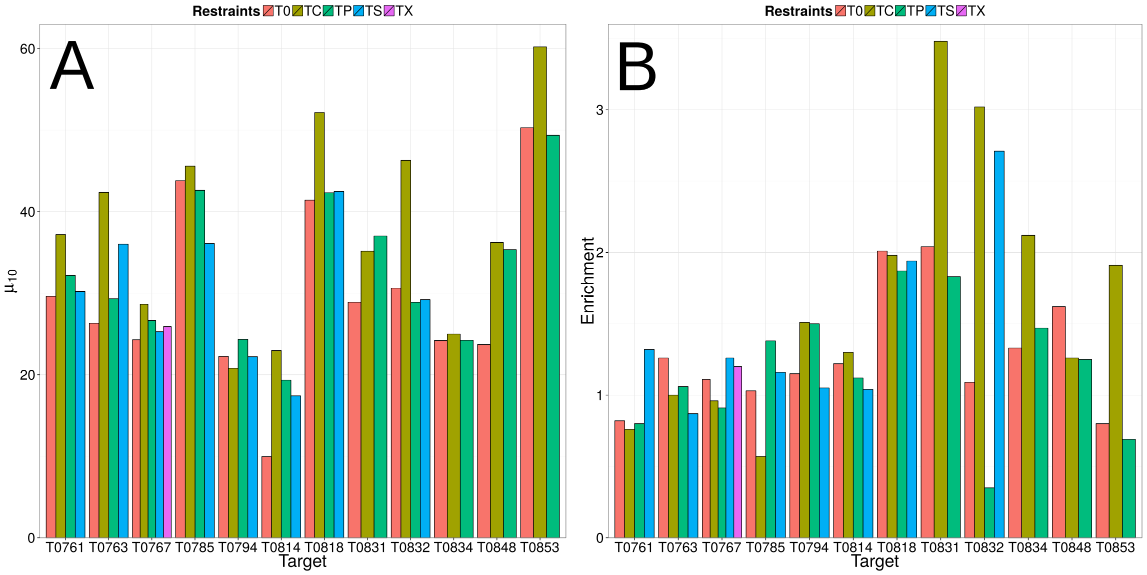

Figure II.5.: Sampling accuracy and model discrimination for "assisted" targets. *(A,B) The average GDT_-TS-values of the most accurate models (μ_{10}) and the enrichments are compared for protein structure prediction without restraints (T0), with predicted residue-residue contacts (TP), only correct residue-residue contacts (TC), NMR-NOE restraints (TS), and XL-MS restraints (TX).*

II.3.5. The impact of limited experimental data on protein structure prediction accuracy

If none of the participating groups in the CASP experiment was able to accurately predict the tertiary structure of a regular target, this target was rereleased as "assisted" target and additional limited experimental data was provided. Of the twenty regular targets analyzed in this study, twelve targets were rereleased as "assisted" targets (table II.1 on page 14). Of those, predicted residue-residue contacts (TP) and only correct residue-residue contacts (TC) were provided for all twelve assisted targets. NMR-NOE data (TS) was provided for eight assisted targets, and XL-MS data was provided for one assisted target (TX). To evaluate the impact of different kinds of experimental data on the sampling accuracy of the low-resolution topology search module, we compared the average GDT_TS-value of the ten most accurate models (μ_{10}) for each restraint type and protein target. The comparison is based on ten models instead of one model to account for the randomness of the sampling. The impact of limited experimental data on model discrimination was evaluated by comparing the achieved enrichments (see section II.2 on page 13).

For the predicted residue-residue contacts (TP), a data set that also includes incorrect residue-residue contacts, only minor improvements in sampling accuracy could be observed. Whereas the average μ_{10}-value over the twelve TP targets was 30 % when predicting without residue-residue contacts, incorporating residue-residue contacts improved the average μ_{10}-value to 33 % (table A.2 on page 158 and figure II.5). There was also no beneficial impact on model discrimination. Actually, the average enrichment-value dropped from 1.3 to 1.2 when using predicted residue-residue contacts. Incorporation of only correct residue-residue contacts (TC), had a more significant impact on the sampling accuracy, which is demonstrated by an improved average μ_{10}-value of 38 %. A similar beneficial impact could be observed on model discrimination, which is demonstrated by an improved enrichment-value of

1.7 (table A.2 on page 158 and figure II.5 on the previous page). NMR-NOE restraints (TS) were only available for eight protein targets. For those eight protein targets, only minor improvements in sampling accuracy and model discrimination could be observed. The average μ_{10}- and enrichment-values improved from 29 % to 30 % and from 1.2 to 1.4, when compared to the prediction results without using additional structural information (see table A.2 on page 158 and figure II.5 on the previous page for details). XL-MS restraints (TX) were only available for one regular target (T0767) analyzed in this study. For this protein target, incorporation of XL-MS data also only had a minor impact on the sampling accuracy and model discrimination. The μ_{10}- and enrichment-values improved from 24 % to 26 % and from 1.1 to 1.2 (see table A.2 on page 158 and figure II.5 on the previous page).

II.3.6. The low-resolution topology search fails in some instances to sample the correct topology

In an MCM algorithm, the sampling depends on the scoring because the probability with which an MC step is accepted depends on the score difference to the previous MC step.[59] To further investigate limitations in sampling and scoring, we relaxed the experimentally determined structures in the BCL::Fold force field. In this process, small structural perturbations are applied to the experimentally determined structure in order to find a structurally similar conformation with a more favorable BCL score. For sixteen out of the twenty benchmark proteins (80 % of all targets), the relaxation resulted in structurally similar conformations (GDT_TS greater than 70 %), which had a favorable BCL score (among the top 20 % of the sampled models). We conclude that these topologies should therefore be selectable through the BCL scoring function and within the sampling range of BCL::Fold (figure II.6 on the following page and figure A.2 on page 156). For T0781, conformations with a GDT_TS-value greater than 80 % exist that score as favorably as the best scoring *de novo* sampled conformations during the CASP experiment (see figure II.6 on the following page). To further investigate, why none of the well scoring conformations were sampled, we folded an additional 500 000 conformations for T0781 with additional correct residue-residue contact restraints to further limit the size of the sampling space. Despite that, it was not possible to sample a conformation with a GDT_TS greater than 25 %, which indicates that the sampling algorithm needs to be revisited. Visual inspection of a clustered representation of the sampled models revealed that the SSEs in all cluster medoids exhibited a strong bias towards Rossmann-like[115,116] α-β-α-sandwich topologies (figure II.6 on the next page), whereas the experimentally determined structure (PDB entry 4QAN) is categorized as α-β-roll, according to a search of the CATH[117] database. In a future step, the sampling of β-strand containing topologies needs to be thoroughly revisited.

For four benchmark targets (T0759, T0771, T0818, and T0831, see figure A.2 on page 156), the relaxation of the experimentally determined structure did not result in conformations with a score as favorable as the score of the *de novo* folded models. Whereas this did not pose any problem for target T0818 because conformations with favorable score and GDT_TS-value greater than 40 % exist (figure A.2 on page 156), this could have had detrimental effect on the structure prediction for the other three targets. The remaining targets are outliers to the statistics the BCL::Fold scoring function is based on (see section II.2 on page 13 for detail). The scores of the targets T0759 and T0831 (PDB entries 4Q28 and 4QN1) are heavily penalized for their large radius of gyration — the spatial extent of the proteins' tertiary structures with respect to their sequence lengths.[60] The radius of gyration score introduces a bias towards globular folds and it will have to be evaluated on a large benchmark set if turning off this scoring term will have a negative impact on structure prediction at general. For the remaining target T0771 (PDB entry 4QE0), multiple properties of the experimentally determined structure — burial of

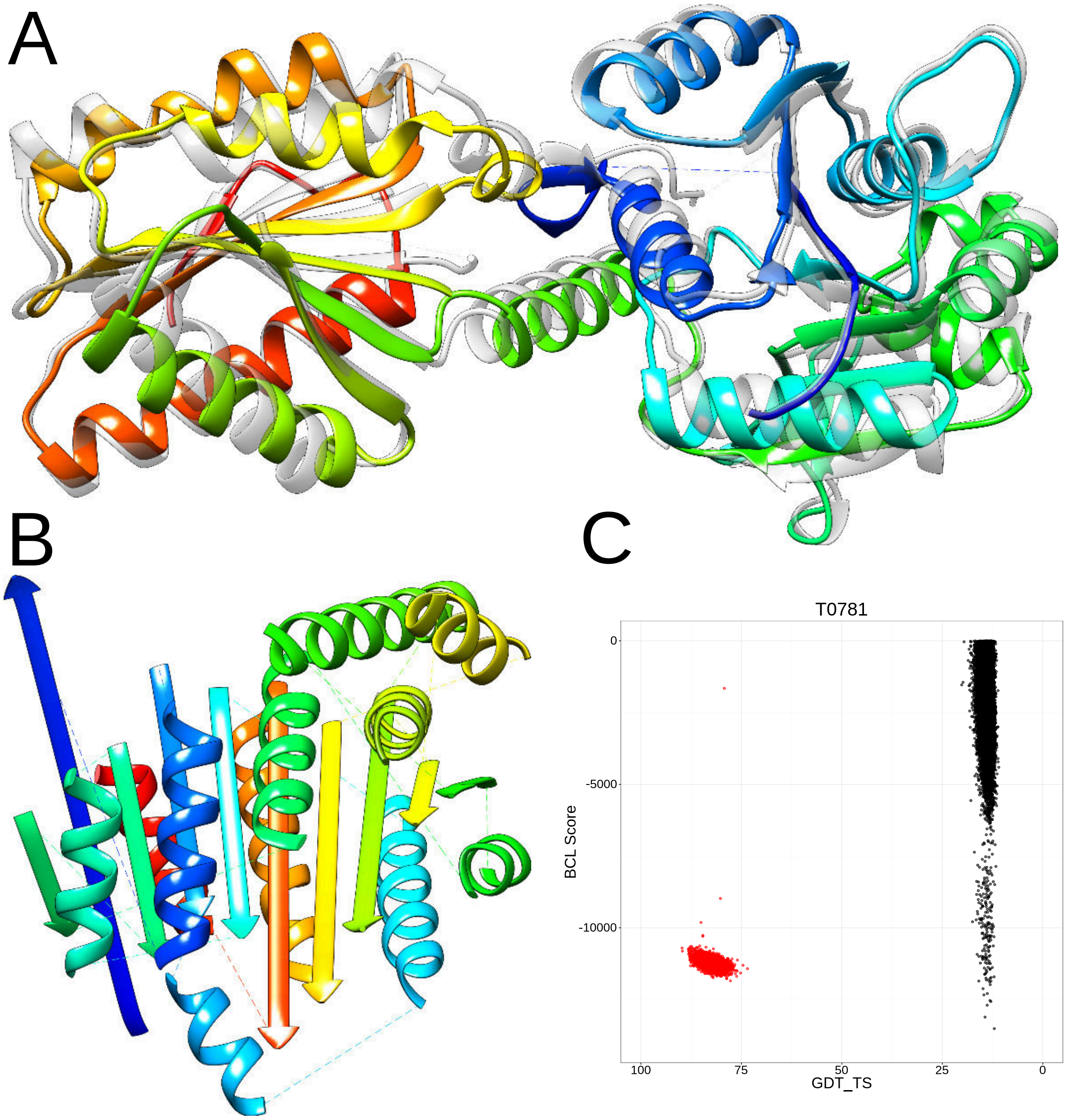

Figure II.6.: Limitations in the conformational sampling hinder structure prediction for regular target T0781. *(A) Experimentally determined structure of T0781 (PDB entry 4QAN, grey) superimposed with the same structure after relaxation with the BCL scoring function (rainbow). (B) Best scoring* de novo *model predicted by BCL::Fold. (C) Shown are the BCL score of the models (y-axis) and the GDT_TS of the model relative to the experimentally determined structure (x-axis). Relaxing the experimentally determined structures in the BCL::Fold scoring function reveals native-like conformations with a favorable score (red dots). In comparison, the* de novo *folded conformations observed during the CASP experiment (black dots) achieve comparable scores but don't include conformations, which are structurally similar to the experimentally determined structure.*

residues, residue-residue interactions, SSE packing — scored worse than the *de novo* predicted models and the scoring function was not able to identify a native-like conformation. This target represents an outlier to our statistics over protein structure properties and would have to be complemented with experimental restraints.

II.4. Discussion

II.4.1. Necessary simplifications in the topology search hinder protein structure prediction

The vast size of the conformational space does not allow for exhaustive sampling of all possible conformations of a protein's chain. BCL::Fold reduces the complexity of the search space by assembling the protein's tertiary structure from idealized SSEs and only allowing for limited deviations from the idealized dihedral angles. Although this approach reportedly worked well for α-helical proteins[59] and, in particular, membrane proteins,[50] the protein targets in the CASP benchmark set contained many proteins with a large percentage of β-strand content (table II.1 on page 14). Many of those proteins contained strongly bent β-strands, making it impossible for the low-resolution topology search module to sample and select models having the correct topology (table A.1 on page 157 and figure II.2 on page 20). Although BCL::Fold was able to sample models with a GDT_TS-value of at least 40 % for seven out of twenty regular targets, only four of those targets had accurate models in a sufficient density to be detectable through clustering (see table A.1 on page 157 and figure II.3 on page 21). Consequently, future work needs to be focused on the development of efficient algorithms to assemble the topologies of β-sheet domains and domains significantly deviating from idealized dihedral angles at general.

II.4.2. The high-resolution refinement protocol requires additional optimization

Over the course of the protein structure prediction pipeline, a general decay of model accuracy was observed (table A.1 on page 157 and figure II.3 on page 21). During the loop construction and side chain placement step using Rosetta, the average GDT_TS-value of the most accurate models over all twenty regular targets dropped from 27 % to 24 % (figure II.3 on page 21). Only for one regular target (T0765), a significant improvement in model accuracy could be observed. Those findings are less surprising since the Rosetta loop construction and refinement step, only applies small-scale perturbations to the start model, and therefore did not further explore the conformational space to transform a topologically incorrect model into an accurate conformation. Consequently, future work needs to be focused on the development of more accurate scoring functions to increase the sampling density of accurate models. A similar observation was made for the atomic-detail MD refinement step. The average GDT_TS-value of the most accurate models over all twenty regular targets dropped from 20 % to 18 %. A significant improvement in model accuracy was only observed for one regular target (T0769), for which the GDT_TS-value of the most accurate model improved from 66 % to 77 % (figure II.3 on page 21 and figure II.4 on page 23). However, we cannot necessarily conclude that MD refinement is unable to recover from inaccurate starting models. Previous work by the groups of David E. Shaw, Chaok Seok, and J. Andrew McCammon demonstrated that MD refinement is able to improve the accuracy of a model.[72,118–121] An evaluation of the CASP refinements through MD also reports some success.[122] Whereas Shaw describes a successful approach using simulations at least 100 µs long, we employed 50 ns simulations. In conjunction with the low accuracy of our start models, this could explain why our MD refinement was in most cases unable to significantly improve the accuracy of the model. In upcoming studies, we will therefore employ longer simulations to allow for sufficient coverage of the conformational space. Additional influence factors originate in the employed force field, which will have to be investigated in future studies.

II.4.3. Sampling problems could not be overcome through limited experimental data

Incorporation correct residue-residue contacts (TC) into the scoring function improved the average μ_{10}-values for the twelve "assisted" targets from 32 % to 40 % (table A.2 on page 158 and figure II.5 on page 24). Statistically significant improvements in sampling accuracy were only observed for the six targets T0763, T0814, T0818, T0832, T0848, and T0853, for which an average improvement of 13 % was observed. For the remaining targets, only minor improvements in sampling accuracy were observed, indicating that a conformation with high structural similarity to the experimentally determined structure is not part of the sampling space. The remaining twelve targets, for which no significant improvement could be observed, were either large or had contained a large number of β-strands. Expectedly, improvements in sampling accuracy and model discrimination by using NMR-NOE restraints and predicted residue-residue contact restraints were less pronounced, because those restraint sets also contained incorrect distance restraints. The NMR-NOE restraints were simulated and incorrect restraints were added purposefully by the CASP organizers (see section II.2 on page 13). Exemplary are the targets T0818 and T0832 for which incorporation of correct residue-residue contacts resulted in an improvement of the μ_{10}-values from 41 % and 31 % to 52 % and 46 %, whereas incorporation of NMR-NOE and predicted residue-residue contact restraints did not result in any improvement (table A.2 on page 158 and figure II.5 on page 24). Consequently, future work needs to be also focused on developing methods to properly handle incorrect experimental data.

II.5. Acknowledgments

The authors want to thank Jeffrey Mendenhall for computational advise as well as Amanda Duran and Jordan Willis for advise regarding the Rosetta software suite. The staff of the *Advanced Computing Center for Research & Education* and Roy Hoffman of the *Center for Structural Biology* at Vanderbilt University provided computational support. Parts of the data analysis were performed using the R[b] package with ggplot2[123] and cluster.[100] The renderings of the models were created using Chimera.[124] This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

[b] https://www.r-project.org

CHAPTER III
MEMBRANE PROTEIN STRUCTURE PREDICTION FROM EPR DATA

This chapter is based on the publication "BCL::MP-Fold: Membrane protein structure prediction guided by EPR restraints".[2] Axel W. Fischer contributed to the development of the prediction pipeline, performing the experiment, analyzing the data, and writing the article.

For many membrane proteins, the determination of their topology remains a challenge for methods like X-ray crystallography and NMR spectroscopy. EPR spectroscopy has evolved as an alternative technique to study structure and dynamics of membrane proteins. The present study demonstrates the feasibility of membrane protein topology determination using limited EPR distance and accessibility measurements. The BCL::MP-Fold algorithm assembles SSEs in the membrane using a MCM approach. Sampled models are evaluated using knowledge-based potential functions and agreement with the EPR data and a knowledge-based energy function. Twenty-nine membrane proteins of up to 696 *residues are used to test the algorithm. The RMSD100-value of the most accurate model is better than* 8 Å *for twenty-seven, better than* 6 Å *for twenty-two, and better than* 4 Å *for fifteen out of twenty-nine proteins, demonstrating the algorithm's ability to sample the native topology. The average enrichment could be improved from* 1.3 *to* 2.5*, showing the improved discrimination power by using EPR data.*

III.1. Introduction

Membrane protein structure determination continues to be a challenge. About 22 % of all proteins are membrane proteins and an estimated 60 % of pharmaceutical therapies target membrane proteins.[125] However, only 2.5 % of the proteins deposited in the PDB are classified as membrane proteins.[36,126] Protein structures are typically determined to atomic detail using X-ray crystallography or NMR spectroscopy. However, membrane proteins provide challenges for both techniques.[40] It is difficult to obtain quantities of purified membrane proteins sufficient for both X-ray crystallography and NMR spectroscopy. The two-dimensional nature of the membrane complicates crystallization in a three-dimensional crystal lattice. In order to obtain crystals, the target protein is often subjected to non-native-like environments and/or modifications such as stabilizing sequence mutations.[127,128] Additional problems may evolve from post-translational modification such as phosphorylation.[129] Many membrane proteins continue to be too large for structure determination by NMR spectroscopy.[130] Even if the target itself is not too large, the membrane mimic adds significant additional mass to the system.[131] Despite wonderful successes in determining the structure of high-profile targets, it is critical that the structural features observed with one technique are confirmed with an orthogonal technique.[132]

EPR spectroscopy in conjunction with SDSL provides such an orthogonal technique for probing structural aspects of membrane proteins.[133–135] Advantages of EPR spectroscopy include that the protein can be studied in a native-like environment and that only a relatively small sample amount is required. In addition, EPR spectroscopy can be used to study large proteins. Although EPR is a versatile tool for probing membrane protein structure, it has its own challenges: at least one unpaired electron (spin label) needs to be introduced into the protein. Typically, this requires mutation of all cysteine residues to either alanine or serine, introduction of one or two cysteines at the desired labeling sites, coupling to the thiol-specific nitroxide spin label *S*-(1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate (MTSL), and functional characterization of the protein. As a result, data sets from EPR spectroscopy are sparse containing only a fraction of measurements per residue in the target protein. EPR is not a high-throughput technique.

EPR provides two categories of structural information important to membrane protein topology: a) EPR can provide information about the local environment of the spin label.[136–138] The accessibility of the spin label to oxygen probe molecules indicates the degree of burial of the spin label within the protein in the transmembrane region. Accessibility measurements are typically performed in a sequence scanning fashion. This provides an accessibility profile over a large portion of the sequence.[139,140] The accessibility profile tracks the periodicity of SSEs as individual measurements rise and fall according to the periodic exposure and burial of residues. The exposed face of an SSE can be determined,[141] a task that is difficult within the hydrophobic environment of the membrane. b) When two spin labels are introduced, EPR can measure inter-spin label distances, routinely of up to 60 Å through the DEER experiment.[142,143] EPR distance measurements have been demonstrated on several large membrane proteins including MsbA,[144] rhodopsin,[145] and LeuT.[146] Given the sparseness of data, EPR has been frequently used to probe different structural states of proteins.[147,148] Changes in distances and accessibilities track regions of the protein that move when converting from one state into another. Such investigations rely upon an already determined experimental structure to define the protein topology and provide a scaffold to map changes observed via EPR spectroscopy.

One critical limitation for *de novo* protein structure prediction from EPR data is that measurements relate to the tip of the spin label side chain where the unpaired electron is located whereas information of the placement of backbone atoms is needed to define the protein fold. For distance measurements, this introduces an uncertainty in relating the distance measured between the two spin labels to a distance between points in the backbone of the protein. This uncertainty, defined as the difference between the distance between the spin labels and the distance between the corresponding C_β-atoms is up to 12 Å.[78,149] To address this uncertainty we previously introduced a CONE model, which provides a knowledge-based probability distribution for the C_β-atom distance given an EPR-measured spin label distance.[78,150] Using the CONE model, just twenty-five or even eight EPR measured distances for T4-lysozyme, enabled Rosetta to provide models matching the experimentally determined structure to atomic detail including backbone and side chain placement.[78] Further success was reported by Yang *et al.*,[151] who successfully determined the tertiary structure of a homodimer by using inter-chain restraints determined from NMR and EPR experiments. These studies demonstrate that *de novo* prediction methods can supplement EPR data sufficiently to allow structure elucidation of a protein.

De novo membrane protein structure prediction was demonstrated with Rosetta using twelve proteins with multiple transmembrane helices (TMHs).[152] The method was generally successful in determining the membrane topology of small proteins with up to 278 residues. However, the results of the study suggest that sampling of large membrane topologies requires methods that directly sample structural contacts between sequence-distant regions of the protein.[153]

For this purpose, we developed an algorithm as part of the BCL[a] that assembles protein topologies from predicted SSEs termed BCL::Fold.[59] The omission of loop regions in the initial protein folding simulation allows sampling of structural contacts between regions distant in sequence and thereby rapidly enumerates all likely protein topologies. A knowledge-based potential guides the algorithm towards physically realistic topologies. The algorithm is particularly applicable for the determination of

[a] http://www.meilerlab.org/bclcommons

membrane protein topologies as transmembrane spans are dominated by regularly ordered SSEs.[50] Loop regions and amino acid side chains can be added in later stages of modeling structure. The algorithm was tested in conjunction with medium-resolution density maps[69] achieving models accurate at atomic detail in favorable cases.[87] The algorithm was also evaluated in conjunction with sparse NMR data.[154]

The present study combines EPR distance and accessibility restraints with the BCL::Fold SSE assembly methodology for the prediction of membrane protein topologies. In the following sections we first introduce scores specific to EPR distances and accessibilities and demonstrate their ability to enrich for accurate models. Following that, we describe the approach and results for assembling twenty-three monomeric and six multimeric membrane proteins guided by EPR distance and accessibility restraints. The results demonstrate that the inclusion of protein specific structural information improves the frequency with which accurate models are sampled and greatly improves the discrimination of incorrect models.

III.2. Materials and methods

In this section, we first describe the set of proteins that was used to evaluate the performance of the algorithm. This is followed by sections describing the simulation procedures for EPR distance and accessibility restraints and how observed EPR was translated into structural restraints that could be used by the BCL::Fold algorithm. This section is concluded by a detailed description of the BCL::Fold algorithm and the conducted benchmark.

III.2.1. Compilation of the benchmark set

Twenty-nine membrane proteins of known structure were used to demonstrate the ability of EPR specific scores to improve sampling during protein structure prediction as well as selecting the most accurate models. The proteins for the benchmark were chosen to cover a wide range of sequence length, number of SSEs, and percentage of residues within SSEs (table III.1 on the following page). Twenty-three of the proteins were monomers ranging in size from 91 to 568 residues. One protein (2L35) has two chains, with the second chain being a single transmembrane span. The remaining five proteins were symmetric multimeric proteins of two or three subunits containing up to 696 residues. 5000 independent structure prediction trajectories were conducted for each protein without restraints, with distance restraints only, with accessibility restraints only, and with distance and accessibility restraints. In order to achieve results that are independent of one specific spin labelling pattern, ten different restraint sets were used for each protein. Those trajectories were conducted with SSEs predicted from sequence and, to test the influence of incorrectly predicted secondary structure, with the SSEs obtained from the experimentally determined structure. In addition, rhodopsin (PDB entry 1GZM) was added to the benchmark set to demonstrate the algorithm's ability to work with experimentally determined restraints.

III.2.2. Simulation of EPR restraints

For 1GZM, experimentally determined EPR distance restraints were available,[145] whereas for the other proteins EPR distance and accessibility restraints were simulated to obtain data sets for each of the twenty-nine proteins. Accessibility restraints were simulated by calculating the neighbor vector-value[96] for residues within SSEs of each protein. Unlike the neighbor count approximation of the solvent accessible surface area (SASA), the neighbor vector approach takes the relative placement of the neighbors with respect to the vector from the C_α-atom to the C_β-atom into account. It thereby becomes

Protein	#aas	#SSE	res$_{\text{SSE}}$ (%)	Source	res. (Å)
1IWG	68	5	90	X-ray	3.5
1GZM	349	7	62	X-ray	2.7
1J4N	116	4	80	X-ray	2.2
1KPL	203	8	76	X-ray	3.0
1OCC	191	5	74	X-ray	2.8
1OKC	297	9	71	X-ray	2.2
1PV6	189	8	87	X-ray	3.5
1PY6	227	9	75	X-ray	1.8
1RHZ	166	5	65	X-ray	3.5
1U19	278	7	66	X-ray	2.2
1XME	568	18	79	X-ray	2.3
2BG9	91	3	87	EM	—
2BL2	145	4	88	X-ray	2.1
2BS2	217	8	80	X-ray	1.8
2IC8	182	7	68	X-ray	2.1
2K73	164	5	62	NMR	—
2KSF	107	4	64	NMR	—
2KSY	223	7	78	NMR	—
2NR9	196	8	75	X-ray	2.2
2XUT	524	16	72	X-ray	3.6
3GIA	433	15	81	X-ray	2.2
3KCU	285	10	67	X-ray	2.2
3KJ6	366	8	47	X-ray	3.4
3P5N	189	6	70	X-ray	3.6
2BHW	669	12	45	X-ray	2.5
2H8A	363	12	79	EM	3.2
2HAC	66	2	79	NMR	—
2L35	95	3	81	NMR	—
2ZY9	344	16	90	X-ray	2.9
3CAP	696	18	68	NMR	2.9

Table III.1.: Proteins used for benchmarking the structure prediction algorithm. *The twenty-nine proteins for the benchmark were chosen to cover a wide range of sequence length, number of SSEs as well as number and percentage of residues within SSEs while having a mutual sequence identity of less than* 20 %. *The columns denote the sequence length, the number of SSEs, the number of residues within SSEs, and the percentage of the residues is within SSEs. The proteins above the separating line are monomeric proteins; below the separating line are multimeric proteins. 2HAC, 2ZY9, and 3CAP are homodimers, 2BHW and 2H8A are homotrimers, and 2L35 is a heterodimer. 1GZM was additionally included to evaluate the protocol on experimentally determined data.*

a more accurate predictor of SASA.[96] The resulting exposure-value for each residue was considered an oxygen accessibility measurement. One restraint per two residues within the transmembrane segment of each SSE was simulated.

Distance restraints were simulated using a restraint selection algorithm,[155] which distributes measurements across all SSEs (see section B.2.1 on page 169 for details). It also favors measurements between residues that are far apart in sequence. One restraint was generated per five residues within the transmembrane segment of an SSE, if not indicated otherwise. Distances are calculated between the C_β-atoms; for glycine, the $H_{\alpha 2}$-atom is used. To simulate a likely distance observed in an actual EPR experiment, the distance is adjusted by an amount selected randomly from the probability distribution of observing a given difference between the spin-spin distance (D_{SL}) and the back bone distance (D_{BB}).[149] In order to reduce the possibility of bias arising from restraint selection and spin labelling patterns, ten independent restraint sets were generated. For the five symmetric multimeric proteins, the same protocol was used, but only distance restraints between the same residues in the different subunits were considered.

III.2.3. Translating EPR accessibilities into structural restraints

EPR accessibility measurements are typically made in a sequence scanning fashion over a portion of the target protein. Although each individual accessibility measurement is difficult to interpret, the pattern of accessibilities over a stretch of amino acids within an SSE indicates reliably, which phase of the SSE is exposed to solvent/membrane versus buried in the protein core. We found accessibility restraints to have a limited impact on structure prediction for soluble proteins.[78] We concluded that this is the case as knowledge-based potentials on their own can distinguish the polar phase of an SSE that is exposed to an aqueous solvent from a hydrophobic phase buried in the protein core. However, we also hypothesized that the situation will be different for membrane proteins where it would be harder to distinguish the membrane-exposed from the buried phase of an α-helix as both of these tend to be apolar.

Our approach for developing an EPR accessibility score takes advantage of the regular geometry within the SSE: the exposure moment $\boldsymbol{E}_\mathrm{w}$ of a window of amino acids N is defined as shown in equation (III.1).

$$\boldsymbol{E}_\mathrm{w} = \sum\nolimits_{n=1}^{N} e_n \cdot \boldsymbol{s}_n \tag{III.1}$$

where:

$\boldsymbol{E}_\mathrm{w}$ = exposure moment of the residue window
N = number of residues in the window
e_n = exposure-value of residue n
$\boldsymbol{s}_n$ = normalized vector from the C_α-atom to the C_β-atom of residue n

This equation was inspired by the hydrophobic moment as previously defined.[156] Calculating the exposure moment $\boldsymbol{E}_\mathrm{w}$ from the SASA has been previously demonstrated to approximate the exposure moment calculated from SDSL-EPR accessibility measurements.[141]

However, during *de novo* protein structure prediction, the protein is represented only by its backbone atoms, which hampers calculation of the SASA. Furthermore, calculation of the SASA from an atomic detail model would be computationally prohibitive for a rapid scoring function for usage in *de novo* protein structure prediction. Therefore, the neighbor vector approximation for the SASA is used.[96] In this context, the exposure moment is calculated for overlapping windows of length seven for

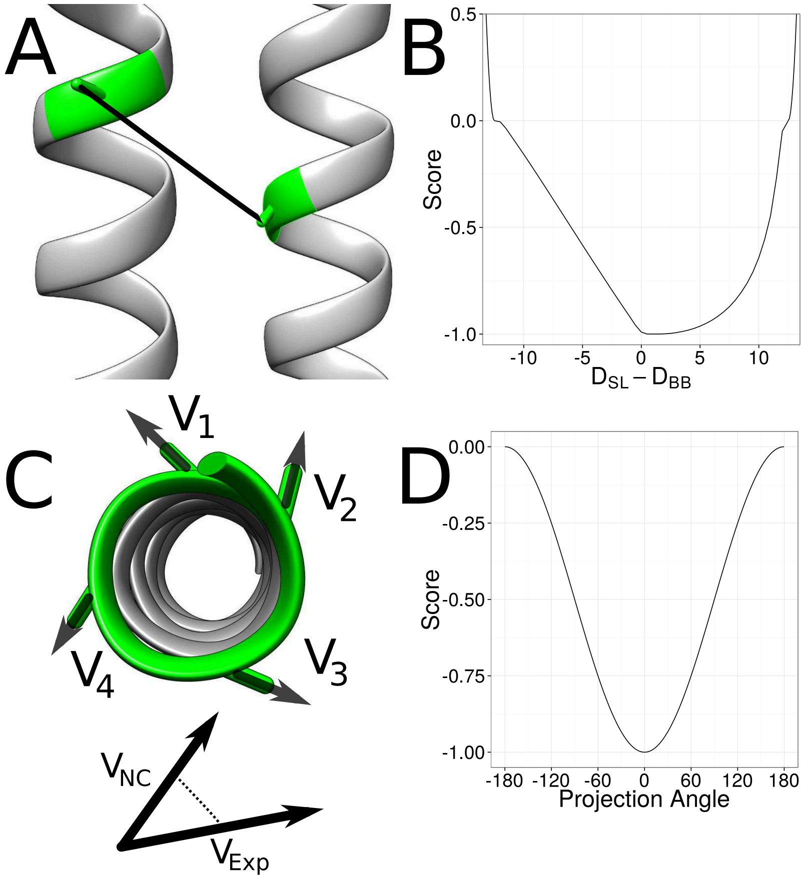

Figure III.1.: Translation from EPR data into structural restraints. *EPR distance measurements measure distances between residues in a protein indirectly. Whereas the experiment determined the spin-spin distance* (D_{SL})*, a distance between the backbone atoms* (D_{BB}) *is needed during the* de novo *protein structure prediction process. Therefore a translation from* D_{SL} *to* D_{BB} *is necessary. BCL::Fold uses a knowledge-based potential to evaluate the agreement of the distance between the* C_β*-atoms in the model with the experimentally determined spin-spin distances (B). EPR accessibility data is translated into structural restraints by summing up the hydrophobic moment vectors* (C_α*-atom to* C_β*-atom) of four consecutive residues (C). This is done twice: first the normalized* $C_\alpha - C_\beta$ *vectors are multiplied with the accessibility determined in the EPR experiment, the second time they are multiplied with the neighbor count of the residue in the model. The vectors are summed up for each approach and the projection angle between the two resulting vectors is scored, with an angle of* 0° *being the best and* 180° *being the worst agreement (D).*

α-helices and length four for β-strands. The pseudo-energy score is computed accordingly as shown in equation (III.2).

$$S_{\text{orient}} = -0.5 \cdot cos(\theta) \tag{III.2}$$

where:

S_{orient} = pseudo-energy score for the exposure moment
θ = torsion angle between exposure moments

This method evaluates to a pseudo-energy score of −1 if $\theta = 0°$ (the vectors parallel) and to a pseudo-energy score of 0 if $\theta = 180°$ (the vectors are anti-parallel; see figure III.1 on the preceding page for a plot of the scoring function).

It has previously been demonstrated that the burial of sequence segments relative to other sequence segments can be determined from the average accessibility-values measured for that stretch of sequence.[157] To capture this information, the magnitude of the exposure moment for overlapping residue windows is determined from the model structure and from the measured accessibility. The Pearson correlation is then calculated between the rank order magnitudes of the structural versus experimental moments. This gives a value between −1, which indicates the structural and exposure magnitudes are oppositely ordered, and 1, which means the structural and exposure magnitudes are ordered equivalently. The score S_{magn} is obtained by negating the resulting Pearson correlation-value so that matching ordering will get a negative score and be considered favorable.

III.2.4. Translating EPR distances into structural restraints

The CONE model[78] yields a predicted distribution for the difference between D_{SL} and D_{BB}. This distribution was converted into a knowledge-based potential function, which is used to score the agreement of models with experimentally determined EPR distance restraints.[149] This score spans a range of $D_{SL} - D_{BB}$ between −12 Å and 12 Å. D_{SL} is the EPR measured distance between the two spin labels; D_{BB} is the distance between the corresponding C_β- or $H_{\alpha 2}$-atoms on the residues of interest; $D_{SL} - D_{BB}$ is the difference between these two distances (figure III.1 on the previous page).

In addition, we found it beneficial to add an attractive potential on either side of the range spanned by the scoring function to provide an incentive for the MCM minimization to bring structures within the defined range of the scoring function. These attractive potentials use a cosine function to transition between a most unfavorable score of 0 and a most favorable score of −1. The attractive potential is positive for $30\,\text{Å} \geq | D_{SL} - D_{BB} | \geq 12\,\text{Å}$. It levels to 0 when the difference between D_{BB} and D_{SL} approaches 12 Å (figure III.1 on the preceding page).

III.2.5. Summary of the folding protocol

The protein structure prediction protocol (figure III.2 on the next page) is based on the protocol of BCL::Fold for soluble proteins.[59] The method assembles SSEs in the three-dimensional space, drawing from a pool of predicted SSEs. A MC energy minimization with the Metropolis criteria is used to search for models with favorable energies. Models are scored after each MC step using knowledge-based potentials describing optimal SSE packing, radius of gyration, amino acid exposure, and amino acid pairing, loop closure geometry, secondary structure length and content, and penalties for clashes.[60]

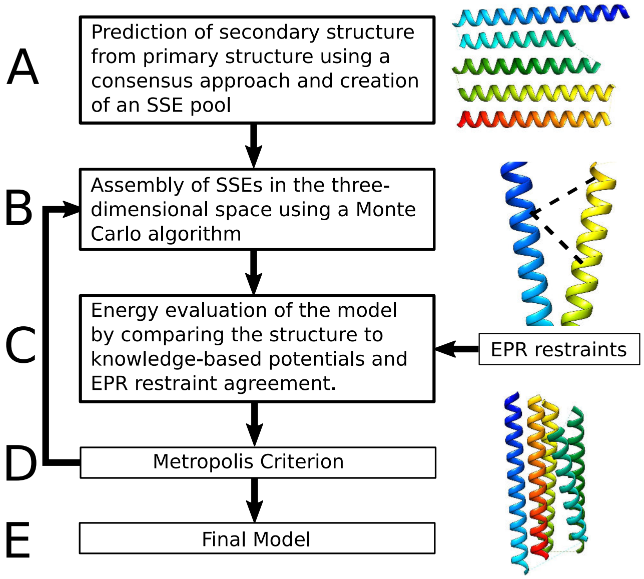
A
Prediction of secondary structure from primary structure using a consensus approach and creation of an SSE pool
B
Assembly of SSEs in the three-dimensional space using a Monte Carlo algorithm
C
Energy evaluation of the model by comparing the structure to knowledge-based potentials and EPR restraint agreement.
EPR restraints
D
Metropolis Criterion
E
Final Model

Figure III.2.: Structure prediction protocol for using EPR data. *(A) SSEs are predicted using machine learning. (B) BCL::Fold arranges predicted glsplsse using an MC algorithm in conjunction with knowledge-based potentials (C) and Metropolis criterion (D).*

The algorithm was adapted for membrane protein folding by altering the amino acid exposure potential according to an implicit membrane environment.[50] Additional scores are used, which favor orthogonal placement of SSEs relative to the membrane and penalizing models with loops going through the membrane. All moves introduced for soluble proteins are used.[59] In addition, we include perturbations that optimize the placement of the protein in the membrane such as translation of individual SSEs in the membrane as well as rigid body translation and rotation of the entire protein.

The assembly of the protein structure is broken down into five stages of sampling with large structural perturbation moves that can alter the topology of the protein. Each of the five stages lasts for a maximum of 2000 MC steps. If an energetically improved structure has not been generated within the previous 400 MC steps, the minimization for that stage will cease. Over the course of the five assembly stages, the weight of clashing penalties in the total score is ramped as 0, 125, 250, 375 and 500.

Following the five stages of protein assembly, a structural refinement stage takes place. This stage lasts for a maximum of 2000 MC steps and will terminate sooner if an energetically improved model is not sampled within the previous 400 steps. The refinement stage consists of small structural perturbations, which will not drastically alter the topology of the protein model.

After 5000 models have been generated for each protein, the models are filtered according to EPR distance score. The top 10 % or 500 models resulting from the structure prediction protocol are selected for a second round of energy minimization. The second round occurs as described above, the only difference being that the minimization uses the SSE placements of a given protein as a starting point. For each starting structure, ten models are sampled, resulting in 5000 models. This boot strapping approach, which re-optimizes structures that are in good agreement with the EPR restraints and with the knowledge-based potential was beneficial when combining BCL::MP-Fold with limited NMR data and is not applied when no experimental data are used.[154]

III.2.6. Summary of the benchmark setup

To test the influence of EPR restraints, each protein besides 1GZM was folded in the absence of restraints, with just distance restraints, with just accessibility restraints, and with distance and accessibility restraints. To test the influence of secondary structure prediction accuracy (see section III.2.7), the experiment was repeated with optimal SSEs derived from the experimentally determined structure. 1GZM was only folded without restraints and with the experimentally determined distance restraints. 5000 models were created for each of the benchmark proteins in independent MCM folding trajectories. EPR distance and accessibility scores are used during the five assembly and one refinement stages of structure prediction protocol. The EPR distance scores have a weight of 40 during all assembly and refinement stages using either pool.

III.2.7. Structure prediction protocol

For each protein, two sets of SSE pools are generated for use during structure assembly. The first SSE pool consists of the TMHs as predicted by obtainer of correct topologies for uncharacterized sequences (OCTOPUS). The second SSE pool contains elements predicted by OCTOPUS as well as SSEs predicted from sequence by Jufo9D (see section B.2.2 on page 169 for details). Using these two SSE pools, the structure prediction protocol is independently conducted twice: a) once using the SSE pool containing predictions from OCTOPUS and Jufo9D ("full pool") and b) once emphasizing the predictions by OCTOPUS ("OCTOPUS pool"). Emphasis is placed on OCTOPUS predictions by using only the OCTOPUS generated SSE pool during the first two stages of assembly. During last three stages of structure assembly, the SSEs predicted from Jufo9D are added to the pool. This allows for better coverage of SSEs within the structure, since OCTOPUS only predicts transmembrane spanning helices.

EPR specific scores are used during the five assembly and one refinement stages of structure prediction (see section B.2.2 on page 169 for details). The EPR distance scores have a weight of 40 over the course of the assembly and refinement stages.

III.2.8. Calculating EPR score enrichments

The enrichment-value is used to evaluate how well a scoring function is able to select the most accurate models from a given set of models. The models of a given set are sorted by their RMSD100-values. The 10 % of the models with the lowest RMSD100-values are put into the set P (positive) the rest of the models will be put into the set N (negative). The models of S are then also sorted by their assigned scoring-value and the 10 % of the models with the lowest (most favorable) score are put into the set T. The models, which are in P and in T are the models, which are correctly selected by the scoring function and their number will be referred to as TP (true positives). The number of models, which are in P but not in T are the models, which are not selected by scoring function despite being among the most accurate ones. They will be referred to as FN (false negative). The enrichment will then be calculated according to equation (III.3).

$$e = \frac{\#TP}{\#P} \cdot \underbrace{\frac{\#P + \#N}{\#P}}_{=10.0} \tag{III.3}$$

where:

e = enrichment
P = 10 % most accurate models
N = remaining models
TP = 10 % most accurate models that are also among the 10 % best scoring models

The positive models are in this context the 10 % of the models with the lowest RMSD100-values. Therefore, $(\#P + \#N)/\#P$ evaluates to a constant of 10.0. Consequently, no enrichment would be result in a value of 1.0 and an enrichment-value between 0.0 and 1.0 indicates that the score selects against accurate models.

III.3. Results

In this section, we report the results of the benchmark. The results are analyzed under two aspects: the RMSD100 metric, which quantifies the structural dissimilarity between the sampled models and the experimentally determined reference structure and the enrichment metric (see section III.2.8 on the preceding page for details), which quantifies how well the scoring function is able to distinguish between accurate and inaccurate models.

III.3.1. Using EPR-specific scores during membrane protein structure prediction improves sampling accuracy

For each protein, the ten models sampled with the best RMSD100-values[103] are used to determine ability to sample accurate models by taking their RMSD100-value average, μ_{10}. Using the best ten models by RMSD100 provides a more consistent measure of sampling accuracy compared to looking at the single best because of the random nature of the structure prediction protocol. Additionally, the percentages of models with an RMSD100-value less than 4 Å and less than 8 Å, τ_4 and τ_8, were calculated.

By using EPR distance and accessibility scores, not only is the frequency increased with which higher accuracy models are sampled, but the best models achieve an accuracy not sampled in the absence of EPR data (table B.1 on page 165). Across all proteins, μ_{10} is, on average, 6.0 Å when EPR distance and accessibility scores are not used. When adding restraints for distances and then both distances and accessibilities, the average μ_{10}-value drops to 5.1 Å and 5.0 Å, respectively (table B.1 on page 165). By only adding EPR accessibility restraints, the average μ_{10}-value over all proteins improves only slightly to 5.8 Å. This demonstrates that the accuracy of the models is primarily improved by using EPR distance restraints in the structure prediction process. With the exception of 1KPL and 2XUT, all proteins achieve a μ_{10}-value of less than 8.0 Å. This indicates the placement of the transmembrane spanning regions follow the experimentally determined structures and the correct fold could be predicted. figure III.3 on the next page compares the RMSD100-values of the average of the 1 % most accurate models with and without the usage of EPR distance restraints — an average improvement of 0.8 Å over the benchmark set is observed. The shift to lower RMSD100-values in distributions for selected benchmark proteins is shown in figure III.3 on the following page. The average τ_4- and τ_8-values improve from 3 % and 13 %, when folding without EPR restraints, and to 6 % and 19 % when using EPR restraints, respectively.

The six multimeric proteins achieve an average μ_{10}-value of 5.0 Å when the structure prediction was conducted without using EPR restraints. By using EPR distance and accessibility restraints μ_{10} could be improved to 2.9 Å. The τ_4- and τ_8-values could be improved from 13 % and 24 % to 21 % and 41 % when using EPR distance and accessibility restraints in the structure prediction process.

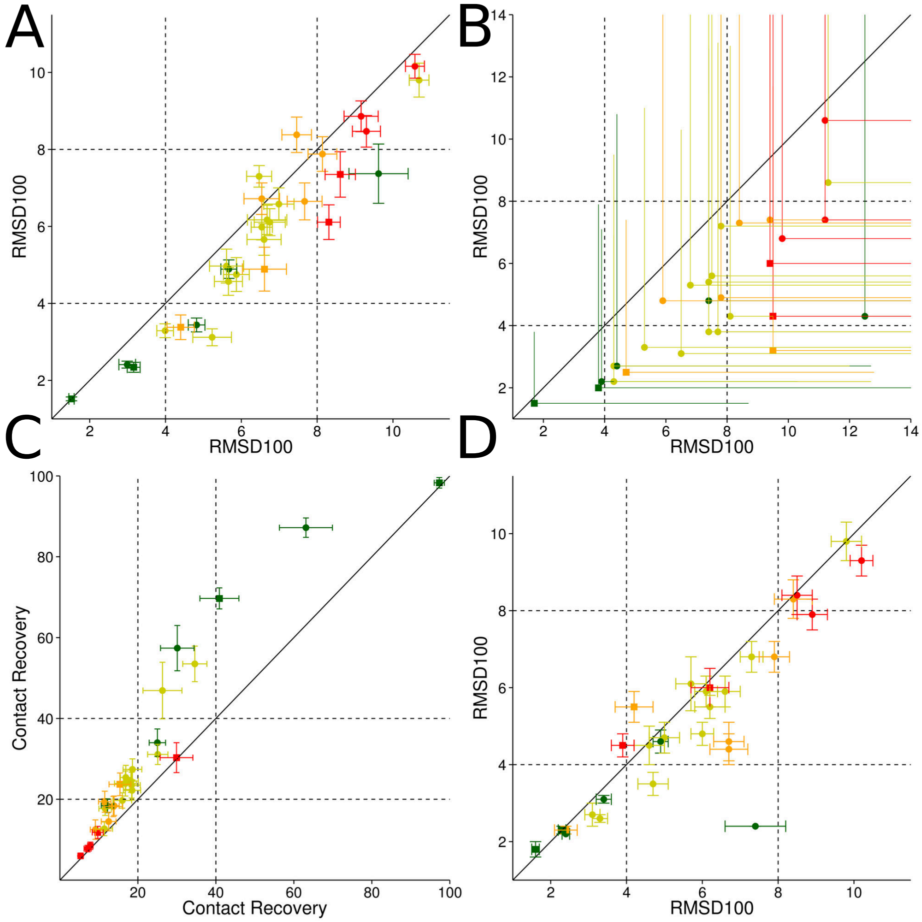

Figure III.3.: Sampling accuracy, contact recovery, and enrichment results when using EPR data. *By using EPR distance and accessibility data in the structure prediction process the sampling accuracy can be improved significantly for monomeric (circles) as well as oligomeric (squares) proteins (A). The sampling accuracy could be improved in twenty-five out of twenty-nine cases by using EPR distance and accessibility data, which is demonstrated by comparing the average RMSD100-values of the* 1 % *most accurate models predicted without (x-axis) and with EPR data (y-axis) in (A). Adding protein specific structural information in the form of EPR distance and accessibility restraints also improves our ability to select the most accurate models among the sample ones. In each of the twenty-nine cases EPR distance and accessibility restraints enable us to select more accurate models when compared to structure prediction without EPR data available. Shown are the average (line) and best (dot/square) RMSD100-values of the best* 1 % *models by BCL score with (y-axis) and without (x-axis) EPR restraints (B). By using EPR accessibility data only (y-axis) the contact recovery could be improved in twenty-two out of twenty-nine cases (C) when compared to structure prediction without EPR accessibility restraints (x-axis). Improvements in SSE prediction methods would also lead to improved sampling accuracies (D, see also table B.4 on page 168). In twenty-one out of twenty-nine cases the average RMSD100 of the ten most accurate models could be improved by using SSE definitions obtained from the experimentally determined structure (y-axis) compared to using predicted SSEs.*

III.3.2. EPR accessibility scores are important for improving contact recovery

Data from EPR accessibility measurements were previously used in conjunction with the Rosetta protein structure prediction algorithm.[78] The derived scores were applied in a benchmark to predict the structures of the small soluble proteins T4-lysozyme and αA-crystallin. The improvement in sampling models that are more accurate was compared between prediction trajectories using an EPR distance score and trajectories using an EPR distance score coupled with an accessibility score. For T4-lysozyme and αA-crystallin, using the accessibility score did not result in a significant improvement in the accuracy of models sampled. This was attributed to the simple rule of exposure that is well captured by the knowledge-based potentials: polar residues tend to be exposed to solvent; apolar residues tend to be buried in the core of the protein.

Membrane proteins are subjected to a more complex set of possible environments. Any given residue can reside buried in the core of the protein or exposed to different environments ranging from the membrane center to a transition region to an aqueous solvent. If the protein fold contains a pore, a residue can be solvent-exposed deep in the membrane.[158] Such a complex interplay of environments will not be as easily distinguished by knowledge-based potentials. Here it has been demonstrated that using EPR accessibility information consistently improves the contact recovery for highest accurate models.

Although improvements regarding sampling accuracy and selection of the most accurate models by RMSD100 is mainly achieved by using EPR distance restraints, EPR accessibility restraints help determining the correct rotation state of SSEs and therefore improves the number of recovered contacts (figure III.3 on the previous page). A contact is defined as being between amino acids, which are separated by at least six residues and have a maximum Euclidean distance of 8 Å. We are measuring the percentage of the contacts in the experimentally determined protein structure, which could be recovered in the models. In order to be independent of huge deviations occurring when only looking at the best model sampled, we quantify the average contact recovery of the ten models with the highest contact recovery (ϕ_{10}) and the percentage of models, which have more than 20 % and 40 % of the contacts recovered (γ_{20} and γ_{40}).

For folding without EPR restraints, the average ϕ_{10}-value over all twenty-three monomeric proteins was 23 % whereas with accessibility restraints it was 31 % (table B.2 on page 166). Using distance restraints additionally to the accessibility restraints, ϕ_{10} remains at 31 %. This is demonstrating that improvements in contact recovery are mainly achieved by using EPR accessibility restraints in the structure prediction process. The average γ_{20}- and γ_{40}-values over all twenty-nine proteins for structure prediction without EPR restraints were 5 % and 3 %. By using EPR accessibility restraints, the values could be improved to 12 % and 16 %, respectively.

For the six multimeric proteins, improvements in contact recovery by the usage of EPR accessibility restraints are observed as ϕ_{10}-, γ_{20}-, and γ_{40}-values could be increased to 46 %, 25 % and 16 % from the previous values of 38 %, 17 % and 14 % when performing protein structure prediction without EPR data. By complementing the accessibility with distance restraints, ϕ_{10}-, γ_{20}-, and γ_{40}-values can be improved to 50 %, 30 % and 16 %.

III.3.3. EPR-specific scores select for accurate models of membrane proteins

The ability of EPR specific scores to select for accurate models is tested by calculating enrichment-values for structure prediction trials of twenty-nine membrane proteins (table B.3 on page 167). The enrichment of a scoring function indicates how well the score identifies a protein model that is accurate

	1/10			1/3			1/2		
Protein	μ_{10} (Å)	τ_4 (%)	τ_8 (%)	μ_{10} (Å)	τ_4 (%)	τ_8 (%)	μ_{10} (Å)	τ_4 (%)	τ_8 (%)
1OCC	3.3	2.0	42.4	1.9	5.6	52.6	2.0	5.4	51.0
1PV6	5.3	0.0	8.3	4.3	0.0	35.9	4.2	0.0	34.6
1PY6	4.2	0.0	19.8	3.5	0.0	27.7	3.3	0.6	32.7
1RHZ	4.7	0.0	5.5	3.3	0.7	22.2	3.5	0.4	24.0

Table III.2.: Sampling accuracy is improving with an increasing number of EPR restraints. *The percentages of models sampled with RMSD100-values less than 4 Å and 8 Å (τ_4 and τ_8) are increasing with the number of restraints increase from one distance restraint per ten residues within SSEs to one restraints per three residues within SSEs to one restraint per two residues within SSEs. An upper limit is met at one restraint per three residues for 1OCC, 1PV6, and 1RHZ since the further accuracy improvements would require a more effective sampling of possible dihedral angle conformations.*

by a good score. It computed as the cardinality of the intersection $I = HS \cap P$ with P being the set of the accurate models and HS being the set of the 10 % of the models with the most favorable score (see section III.2.5 on page 35).[60] Accurate is defined as the 10 % of the models with the lowest RMSD100 when compared to the experimentally determined structure. Therefore, if a score correctly identifies all accurate models as being accurate, a perfect enrichment would result in a value of 10.0.

Enrichment-values were computed for protein models sampled without experimental restraints. For protein structure prediction without EPR data, the average enrichment-value for using just the knowledge-based potentials over all twenty-nine proteins is 1.3. By using EPR distance and accessibility data, the average enrichment is improved to 2.5. The enrichment for using EPR distance and accessibility restraints ranges from 1.1 to 6.2. In seventeen out of twenty-nine cases, the enrichment is greater than 2.0. In twenty-three out of twenty-nine cases the enrichment could be improved by at least 0.5 (table B.3 on page 167). By using EPR accessibility data only the average enrichment over all proteins is 1.6, demonstrating that improvements regarding the selection of the most accurate models are mainly caused by EPR distance restraints.

III.3.4. The number of restraints determines the significance of improvements in sampling accuracy

For four proteins, the influence of varying numbers of restraints was examined. In addition to the one restraint per five residues within SSEs setup used for all benchmark cases, the tertiary structure of 1OCC, 1PV6, 1PY6, and 1RHZ was predicted using one restraint per ten residues, one restraint per three residues, and one restraint per two residues within SSEs. For 1PY6, the sampling accuracy could be steadily improved with an increasing number of restraints demonstrated by τ_8-values increasing from 15 % to 20 % to 24 % to 28 % to 33 % and μ_{10}-values improving from 4.4 Å to 4.2 Å to 3.6 Å to 3.5 Å to 3.3 Å for structure prediction without restraints, one restraint per ten residues, one restraint per five residues, one restraint per three residues and one restraint per two residues (see table III.2 and figure B.1 on page 164). For 1OCC, 1PV6, and 1RHZ, a significant improvement in sampling accuracy is observed for using one restraint per three residues instead of one restraint per ten residues within SSEs, which is demonstrated by improvements in τ_8-values from 42 % to 53 %, from 8 % to 36 %, and from 6 % to 22 % and by improvements in μ_{10}-values from 3.2 Å to 1.9 Å, from 5.3 Å to 4.3 Å, and from 4.7 Å to 3.3 Å, respectively. Increasing the number of restraints to one restraint per two residues within

SSEs fails to further improve the sampling accuracy. We attribute this observation to significant bends in some of the SSEs that are currently not sampled sufficiently dense by BCL::MP-Fold.

III.3.5. Using experimentally obtained EPR distance restraints for rhodopsin

The benchmark was extended to also contain rhodopsin (PDB entry 1GZM) for which EPR distance measurements were available.[145] Although only sixteen EPR distance restraints were available, which amounts to less than one restraint per ten residues within SSEs, the sampling accuracy as well as the enrichment improve significantly. The μ_{10}-values improved from 4.9 Å for folding without restraints to 4.4 Å when using restraints. The enrichment-values could be improved from 0.6 to 1.2 demonstrating that even a small number of restraints improves discrimination of incorrect models.

III.4. Discussion

EPR distance and accessibility restraints can aid the prediction of membrane protein structure. For this purpose, EPR-specific scores were combined with the protein structure prediction algorithm BCL::MP-Fold. BCL::MP-Fold assembles predicted SSEs in space without explicitly modeling the SSE connecting loop regions. This allows for rapid sampling of complex topology that is not easily achieved when an intact protein backbone must be maintained. By adding EPR specific scores to the knowledge-based scoring function, sampling of accurate structures is increased. Additionally the selection of the most accurate models could be improved significantly.

However, it has to be clearly stated that — with the exception of bovine rhodopsin (PDB entry 1GZM) — all EPR restraints used in this study were simulated using the CONE model. Therefore, the relevance of our findings depends on how well the CONE model describes the nature of experimental DEER measurements and in particular the mobility of the spin label.

III.4.1. EPR distance scores improve the accuracy of topologies predicted for membrane proteins

EPR distance measurements are associated with large uncertainties with regards to the translation of the measured spin label – spin label distances into backbone distances. In spite of this, EPR distance measurements provide important data on membrane protein structures.[145,146,159] In the present study, it has been demonstrated that EPR distance data can significantly increase the frequency with which the correct topology of a membrane protein is sampled (figure III.3 on page 39 and figure III.4 on the next page). This is important because as the correct topologies are sampled with higher accuracy, models start to reach the point where they can be subjected to atomic detail refinement to further increase their accuracy.[160]

It is crucial to distinguish between the two major challenges in *de novo* structure prediction — sampling and scoring: The average improvement in sampling accuracy — *i.e.*, the best model built among 5000 independent folding trajectories — of 0.8 Å is moderate but significant. However, inclusion of the EPR data does not only allow folding of models that are more accurate, it greatly improves discrimination of incorrect models with a scoring function that combines BCL knowledge-based potentials and EPR restraints. Without using EPR restraints the average enrichment is 1.3, *i.e.*, 13 % of the most accurate models are in a sample of the 10 % best scoring models, which is close to chance. By using EPR data in addition to the knowledge-based score enrichment increases to 2.5, *i.e.*, one out of four models in the 10 % best scoring models also has the correct fold. This is important as it greatly improves the chance to identify correctly folded models, e.g. through clustering of good-scoring models.

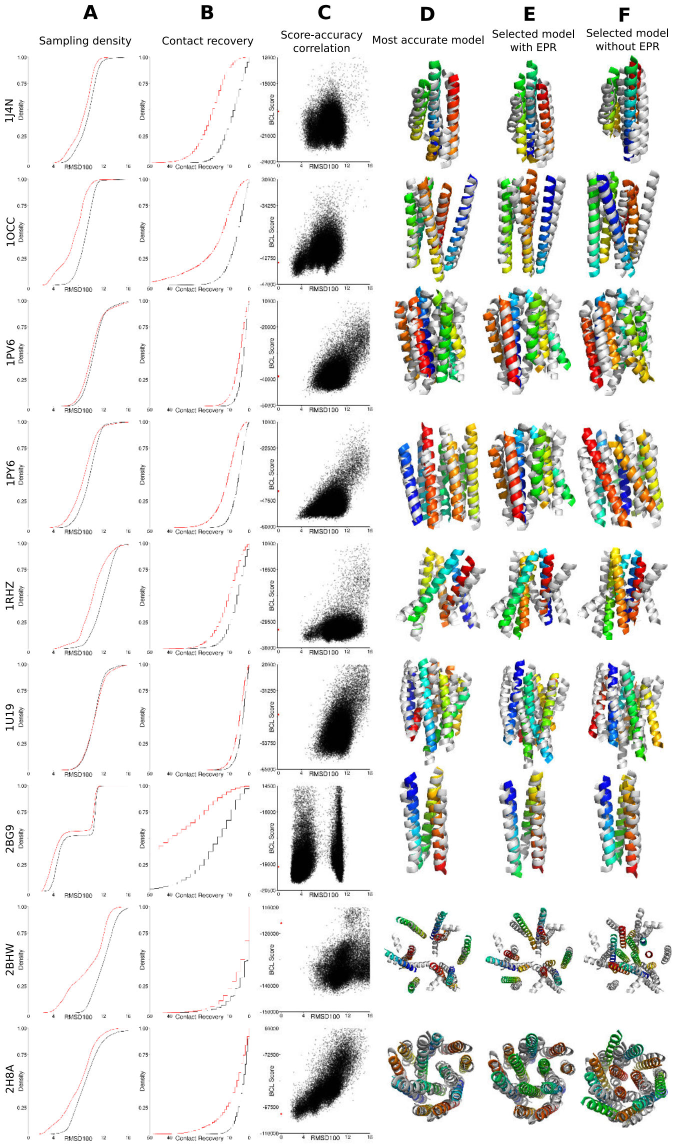

Figure III.4.: Gallery of the structure prediction results when using EPR data. *By using EPR distance and accessibility restraints, the sampling accuracy is significantly improved as the selection ability regarding accurate models. For selected proteins, a comparison of the RMSD100 (column A) and contact recovery (column B) distributions for sampling with (red) and without (black) EPR restraints is shown. The y-axis of column A shows the cumulative density of models with respect to the RMSD100. The y-axis of column B shows the cumulative density of models with respect to their contact recovery. Column C shows the correlation between the BCL score and the RMSD100 for the models sampled with EPR restraints (black dots) and the experimentally determined structure (red dot). The y-axis is the pseudo-energy score the algorithm assigned to the structure; the x-axis is the RMSD100 relative to the experimentally determined structure. The superimpositions show the best models by RMSD100 for folding with EPR restraints (column D), the best model by pseudo-energy score for folding with EPR restraints (column E), and the best model by pseudo-energy score for folding without EPR restraints (column F) superimposed with the experimentally determined structure (grey).*

The combination of improved sampling and discrimination thereby significantly improves the reliability with which were able to predict the tertiary structure of a protein.

The EPR distance data used for the present study is simulated from known experimental structures. It will be interesting to repeat this benchmark once sufficiently dense experimental data sets for several membrane proteins become available. For now, considerable effort was put forth to ensure that the simulated data mimics what would be obtained from a true EPR experiment, so that any results are unbiased by the simulated data. The previously published method for selecting distance restraints was used to create ten different data sets per protein.[155] This ensures results are not biased by a particularly selected data set. Previously, the uncertainty in the difference between spin label distances and the corresponding C_β distance ($D_{SL} - D_{BB}$) was accounted for in simulated distance restraints by adding a random value between 12.5 Å and −2.5 Å.[155] Here, the probability of observing a given $D_{SL} - D_{BB}$ is used to determine the amount that should be added to the $C_\beta - C_\beta$ distance measured from the experimental structure.

Using a method developed for soluble proteins to select restraints for membrane proteins is not necessarily ideal. The constraints already imposed upon membrane proteins by the membrane geometry suggest that optimized methods for selecting restraints for membrane proteins should be developed. One such strategy could be to measure distances between transmembrane segments on the same side of the membrane, with the assumption that TMHs are mostly rigid, parallel structures. Further, additional work is needed to account for topologically important SSEs that do not span the membrane, as well take into account the deviations of transmembrane segments from ideal geometries.

The improved sampling accuracy in the protein structure prediction process is primarily caused by the distance restraints. Whereas by using EPR accessibility restraints the average μ_{10}-value over all twenty-nine proteins drops from 6.0 Å to 5.8 Å, by using EPR distance restraints the average μ_{10}-value could be improved to 5.1 Å.

III.4.2. Why not use the membrane depth parameter as additional restraint?

Of note is that EPR-derived accessibility measurements have also previously been used to the determine membrane depth parameter Φ.[161–163] For this purpose, the accessibility Π of a single residue to two paramagnetic reagents are compared: the water-soluble (nickel-(II)-ethylenediaminediacetate — NiEDDA) and the membrane-soluble (molecular oxygen — O_2). The ratio of both values is used to compute the membrane depth parameter according to equation (III.4).

$$\Phi_n = \ln \frac{\Pi_{O_2}}{\Pi_{\text{NiEDDA}}} \tag{III.4}$$

where:

Φ_n = membrane depth parameter
Π_k = accessibility to paramagnetic relaxation agent k

The present approach does not test effectiveness of a score that relies on the membrane depth parameter for membrane protein structure prediction for several reasons: a) we hypothesize that knowledge-based potentials will be capable of placing transmembrane SSEs at the right depth for this placement should again be dominated by polarity which is well captured in such potentials (read above), and b) the membrane depth parameter Φ_n is affiliated with a larger error margin for NiEDDA accessibilities become very small in the core of the membrane and they omit averaging over multiple residues. Nevertheless, testing if a membrane depth related score can improve BCL::MP-Fold could be a goal in a future experiment.

III.4.3. Improved secondary structure predictions will improve the accuracy of predicted structures

The SSE pools are created in order to reduce the possibility of missing a SSE, which is generally a successful approach as demonstrated previously for soluble proteins.[59] The helical transmembrane span prediction software OCTOPUS[164] is used in conjunction with Jufo9D.[97] Jufo9D provides predictions for SSEs that do not necessarily span the membrane and therefore will not be predicted by OCTOPUS. Improved secondary structure prediction methods will benefit membrane protein structure prediction. In addition, it has been demonstrated that the pattern of accessibility-values for measurements along a sequence follow the periodicity of the SSE on which they are measured.[139,144,159] Measured accessibility profiles could therefore be used to inform the pool of SSEs used for structure prediction.

The pool of SSEs used to assemble the membrane protein topologies is the most important determinant in successfully predicting the membrane proteins' structure. This is seen for 1U19 and 2BL2. With predicted SSEs, the structure of the two proteins can be sampled to μ_{10}-values of 5.9 Å and 6.2 Å, respectively (table B.1 on page 165). By using SSE definitions extracted from the experimentally determined structure, the proteins can be sampled at μ_{10}-values of 4.4 Å and 2.6 Å, respectively. This is caused by secondary structure prediction methods breaking up TMHs into several short helices making it harder to assemble the tertiary structure that does not have loop going through the membrane. The experiment was repeated with SSE definitions obtained from the experimentally determined structures of the proteins. Whereas with predicted SSEs average μ_{10}-, τ_4-, and τ_8-values of 5.0 Å, 6 %, and 19 % are achieved over all twenty-nine proteins, by using the SSE definitions from the experimentally determined structure we could improve them to 4.5 Å, 8 %, and 25 %. In twenty-one out of twenty-nine cases the average accuracy of the ten best models by RMSD100 could be improved by using SSE definitions obtained from the experimentally determined structure (figure III.3 on page 39). This demonstrates that further improvements of the secondary structure prediction will also lead to an improved sampling accuracy of BCL::Fold.

III.4.4. Limitations of the CONE model knowledge-based potential

The unknown label conformation is taken into account by the CONE model, which yields a $D_{SL} - D_{BB}$ distribution. This wide probability distribution accounts for two inherently different aspects — a structural and a dynamical aspect: the structural effect looks at the relative position of the unpaired electron with respect to the backbone coordinates of the protein. This positioning is dependent on the protein structure, specifically the direction in which the $C_\alpha - C_\beta$ vector project into space with respect to the $C_\alpha - C_\alpha$ vector that links the two labeling site. As the CONE model is applied in a model-independent fashion, it does not consider these geometric features but expresses the resulting ambiguity as part of the probability distribution. Second, chemical environment and exposure cause variable levels of spin label dynamics. These result in distance distributions of variable tightness in EPR experiments. This information is currently not considered as parameter in the CONE model but absorbed by using a very wide $D_{SL} - D_{BB}$ probability distribution. This approach has the advantage that it is very robust with respect to uncertainties within the EPR experimental parameters and very fast to compute. At the same time, the CONE model knowledge-based potential neglects important geometric parameters. Developing and testing approaches that take these parameters into account and

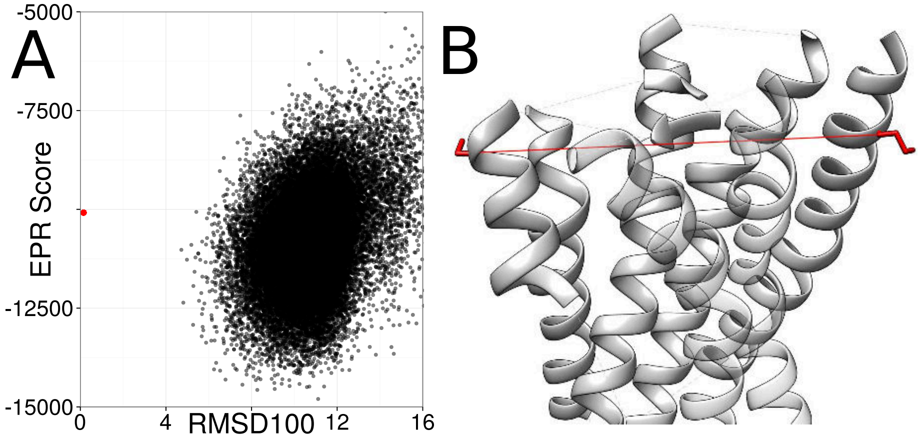

Figure III.5.: Limitations of the CONE model. *For 1U19, the most accurate model cannot be reliably selected (A). One reason for that is, that the translation from the observed spin-spin distance to the backbone distance is inaccurate resulting in models which deviate topologically from the experimentally determined structure achieving a better agreement with the EPR distance restraints than the experimentally determined structure (B). This is demonstrated by the plot showing the correlation between the agreement with the EPR distance restraints (y-axis) and the RMSD100 relative to the experimentally determined structure (x-axis). The EPR potential does not take the exposure of the spin labeling site and the orientation of the* $C_\alpha - C_\beta$ *vectors into account leading to inaccuracies when translating* D_{SL} *into* D_{BB} *for the residues* 7 *and* 170 *of 1U19. Both spin labels are at the outside of the protein and on different sides of the structure leading to greater difference between* D_{SL} *and* D_{BB}.

lead to tighter distance distributions without losing the advantages of speed and robustness is an active area of our research.

Not considering geometrical features hinders the selection of accurate models for 1U19. EPR distance restraints improved the sampling accuracy, but it is still not possible to reliably select accurate models (figure III.5). Although the distances observed in EPR experiments are typically long and therefore allow a broad range of topologically different models to fulfill them, inaccuracies in the translation from D_{SL} to D_{BB} also contribute to the selection problem. In the case of 1U19 the experimentally determined structure, which served as the template for the simulation of the EPR distance restraints, shows a worse agreement with the restraints than the best scoring models. The spin-spin distance between residue 7 and residue 170 is 43.6 Å, whereas the distance between the C_β-atoms is 35.7 Å resulting in an agreement score of 0.3 on a scale from 0 to 1. Following the EPR potential, a $C_\beta - C_\beta$ distance of 41.1 Å is favorable, which is accomplished by the sampled models with the best score leading to the selection of models, which deviate significantly from the experimentally determined structure. Both spin labeling sites are exposed, indicating they are at the outside of the protein. The projection angle between the $C_\alpha - C_\beta$ vectors is greater than 160°, making it more likely that the spin labels are pointing away from each other. Those two properties allow the inference that we would expect a larger difference between D_{SL} and D_{BB} than 2.5 Å. By using a knowledge-based potential, which also takes the exposure of the spin labeling sites and additional geometrical information into account a better ranking of the sampled models would be possible.

III.4.5. Ambiguities in the ranking of models remain

Although the usage of restraints obtained from EPR experiments significantly improves the discrimination of incorrect models, ambiguities in the ranking of the models remain for multiple proteins in the benchmark set. This observation was especially pronounced for the proteins 1J4N, 1PV6, 1PY6, and 1U19 (figure III.4 on page 43). In those cases, the best 10 % of the models by BCL score cover a wide range of topologies. For 1PV6, the best 10 % of the models by BCL score cover an RMSD100 range of 8 Å when compared to the experimentally determined structure. Multiple factors are contributing to this observation. First, the BCL::Fold scoring function is an inaccurate approximation of free energy, which limits its discriminative power.[60] Although adding a term that measures agreement with experimental data will improve its discriminative power, it appears that sparse restraints from EPR data are sometimes insufficient to remove all ambiguities. This is also because, second, the translation of spin label distance distributions into a backbone structural restraint introduces a substantial uncertainty and therefore allows sometimes multiple topologies to fulfill the restraint. One side effect of these approximations is that — as shown in figure III.4 on page 43 — the native structure is not always in the global minimum of the BCL scoring function. Relaxing the experimentally determined protein structures in the BCL force field indicate that the closest minimum in the scoring function is between 1.5 Å and 4.1 Å in RMSD100 separate relative to the experimentally determined structures.

III.5. Conclusion

The determination of membrane protein folds from EPR distance and accessibility data is within reach if these restraints aid protein folding protocols such as BCL::MP-Fold. The ability of EPR data to improve the sampling of native-like topologies and the importance of EPR accessibility data for obtaining highest contact recovery-values was demonstrated. Further, the EPR specific scores allow the selection of close-to-native models, thereby overcoming a major obstacle in *de novo* protein structure prediction. Refining EPR distance potentials to also take the exposure of the spin labeling sites as well as relative orientation of the $C_\alpha - C_\beta$ vector might provide a more accurate translation from spin-spin distance into backbone distance, thereby further increasing model quality.

III.6. Acknowledgments

We thank Cristian Altenbach and Wayne Hubbell for sharing their EPR data for rhodopsin (PDB entry 1GZM) with us and therefore enabling us to evaluate our algorithm based on experimentally determined data. Parts of the data analysis were performed using the R[b] package with ggplot2.[123] The renderings of the models were created using Chimera.[124] The composite figures were created using Inkscape.[c] This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

[b] https://www.r-project.org

[c] https://inkscape.org

CHAPTER IV
PROTEIN STRUCTURE PREDICTION FROM CROSS-LINKING DATA

This chapter is based on the publication "Protein structure prediction guided by crosslinking restraints – A systematic evaluation of the impact of the crosslinking spacer length".[3] Axel W. Fischer contributed to the development of the potential function, performing the experiment, analyzing the data, and writing the article.

> *Recent development of high-resolution MS instruments enables chemical XL to become a high-throughput method for obtaining structural information about proteins. Restraints derived from XL-MS experiments have been used successfully for structure refinement and protein-protein docking. However, one formidable question is under which circumstances XL-MS data might be sufficient to determine a protein's tertiary structure* de novo*? Answering this question will not only include understanding the impact of XL-MS data on sampling and scoring within a* de novo *protein structure prediction algorithm, it must also determine an optimal cross-linker type and length for protein structure determination. Whereas a longer cross-linker will yield more restraints, the value of each restraint for protein structure prediction decreases as the restraint is consistent with a larger conformational space.*
>
> *In this study, the number of cross-links and their discriminative power was systematically analyzed* in silico *on a set of* 2055 *non-redundant protein folds considering Lys-Lys, Lys-Asp, Lys-Glu, Cys-Cys, and Arg-Arg reactive cross-linkers between* 1 Å *and* 60 Å*. Depending on the protein size a heuristic was developed that determines the optimal cross-linker length. Next, simulated restraints of variable length were used to* de novo *predict the tertiary structure of fifteen proteins using the BCL::Fold algorithm, which is part of the BCL.*[a] *The results demonstrate that a distinct cross-linker length exists for which information content for* de novo *protein structure prediction is maximized. The sampling accuracy improves on average by* 1.0 Å *and up to* 2.2 Å *in the most prominent example. XL-MS restraints enable consistently an improved selection of native-like models with an average enrichment of* 2.1.

IV.1. Introduction

"Structural Genomics" — the determination of the structure of all human proteins — would have profound impact on biochemical and biomedical research with direct implication to functional annotation, interpretation of mutations, development of small molecule binders, enzyme design, or prediction of protein-protein interaction.[165] Although significant progress towards this goal has been made through X-ray crystallography and NMR spectroscopy, tertiary structure determination continues to be a challenge for many important human proteins. At present, high-resolution structures exist for about 5 % of all human proteins in the PDB.[36] For many uncharacterized human proteins, construction of a comparative model is possible starting from the experimental structure of a related protein. Nevertheless, for about 60 % (about 7800) of known protein families in the Pfam database[166] not a single structure is deposited.[167] Many of these proteins will continue to evade high-resolution protein structure determination.

[a] http://www.meilerlab.org/bclcommons

Accordingly, researchers strive to develop alternative approaches. The most extreme approach includes computational methods that predict the tertiary structure of proteins from their sequence alone. Although computational methods are sometimes successful at the predicting the tertiary structure of small proteins with up to one hundred residues,[168] for larger proteins the size of the conformational space to be searched as well as the discrimination of incorrectly folded models hinder structure prediction.[95,169,170]

However, recent studies demonstrate that combining *de novo* protein structure prediction with limited experimental data,[2,68,70,78,149,154,171] *i.e.*, experimental data that alone is insufficient to unambiguously determine the fold of the protein, can yield accurate models for larger proteins. The structural restraints in those studies were acquired using EPR spectroscopy,[2,78,149] EM,[68,70] or NMR spectroscopy.[154]

As an alternative technique, XL in combination with MS can be applied to obtain distance restraints, which can be used to guide protein structure prediction.[172–175] Using bifunctional reagents with a defined length, functional groups within the protein can be covalently bridged in a native-like environment. Thus, it is possible to determine an upper limit for the distance between those residues after enzymatic proteolysis and identification of cross-linked peptides.

This method allows for a fast analysis of protein structures in a native-like environment at a low concentration and can even be applied to high molecular weight proteins,[176] membrane proteins,[177] or highly flexible proteins.[178] If combined with affinity purification it becomes possible to study proteins inside the cell.[179] Currently, the XL-MS technology is rapidly gaining importance driven by the liquid chromatography (LC)-MS instrument development, the generation of advanced analysis software,[180] and the direct integration in protein structure prediction workflows.[181–183] Furthermore, hundreds of different cross-linking reagents with different spacer lengths, reactivities, and features for specific enrichment and improved detectability are now commercially available.[184]

However, whereas the potential to combine XL-MS and computational modeling has been frequently demonstrated and many technical problems of XL-MS have been solved, several central questions have not yet been evaluated systematically.

(i) Cross-linking reagents are available with a spacer length ranging from 0 Å to more than 35 Å. Whereas longer reagents are likely to provide more distance restraints, shorter cross-links have higher information content in *de novo* structure prediction as the conformational search space is more restricted. Thus, the question arises, which cross-linker spacer length supports structure prediction best?

(ii) Cross-linking results are often used to confirm already existing structures. However, what is the average gain in model accuracy and selection of correct models when using cross-linking data in conjunction with *de novo* protein structure prediction?

(iii) Cross-linking reagents vary in reactivity towards different functional groups present in different amino acids. For *de novo* protein structure prediction, what is the gain of using additionally cross-linkers with different reactivities?

In this study, we simulated cross-linking experiments on more than 2000 non-redundant protein structures to determine the number of possible and structurally relevant cross-links depending on the size of the protein as well as on the length and reactivity of the applied cross-linking reagents. We then tested the impact of cross-linking restraints on *de novo* protein structure prediction for fifteen selected proteins.

IV.2. Materials and methods

IV.2.1. *Software and databases*

A subset of the PDB containing 2055 non-redundant protein structures was downloaded from the protein sequence culling server (PISCES) server (version 08.2012).[185] This PDB subset was created by filtering all available structures with a resolution of at least 1.6 Å, a maximum sequence identity of 20 %, and an R-factor cutoff of 0.25. Euclidean distances and shortest solvent accessible surface (SAS) path lengths between $C_\beta - C_\beta$, $N_z - N_z$ (Lys-Lys), $N_z - C_\gamma$ (Lys-Asp), and $N_z - C_\delta$ (Lys-Glu), as well as $N_{H2} - N_{H2}$ (Arg-Arg) and $S_G - S_G$ (Cys-Cys) atom pairs with a maximum intramolecular distance of 60 Å were determined through the command line version of Xwalk.[186]

IV.2.2. *Generation of sequence dependent distance functions*

Tables containing the Euclidean distances and the sequence separation between cross-linking target amino acids (i) Lys-Lys, (ii) Lys-Asp, (iii) Lys-Glu, (iv) Arg-Arg, and (v) Cys-Cys were generated. Amino acid pair distances were sorted into 2.5 Å bins. The total number of observed pairs for each sequence and Euclidean distance was counted. Based on the result an approximation of the distance distribution for every sequence distance was created. The median of the distribution was determined. A logarithmic function, described in equation (IV.1), was calculated as a regression curve to correlate the sequence separation S to the median Euclidean distances E_{med}.

$$E_{\text{med}} = a \cdot \ln(S) + b \tag{IV.1}$$

where:

E_{med} = median Euclidean distance between residues
S = sequence separation between residues

IV.2.3. *Calculation of the amino acid side chain length*

Based on the structure of calmodulin (PDB entry 2KSZ) the average $C_\beta - Nz$, $C_\beta - C_\gamma$, $C_\beta - C_\delta$, $C_\beta - N_{H2}$, and $C_\beta - S_G$ distances of the side chains of lysine, aspartic acid, glutamic acid, arginine, and cysteine were determined to be 4.5 Å, 2.3 Å, 3.6 Å, 5.1 Å, and 1.8 Å, respectively.

IV.2.4. *Distinguishing between impossible, possible and structurally valuable cross-links*

Cross-linker spacer lengths between 1 Å and 60 Å distances were evaluated and classified in either (i) impossible cross-links, meaning that the distance between the C_β-atoms of the cross-linked amino acids exceeds the sum of the spacer lengths and the side chain lengths, or (ii) possible cross-links, meaning that the $C_\beta - C_\beta$ distance is below the sum of the spacer lengths and side chain lengths. The latter group was subdivided into cross-links potentially useful for structure determination (valuable cross-links) and those that are unlikely to contribute much information (non-valuable cross-links). We defined cross-links as valuable if the spacer length was shorter than the median distance expected for the given sequence separation by the equations derived in section IV.2.2. For these calculations, all proteins were grouped into 2.5 kDa bins. The calculations were performed for cross-linker lengths from 1 Å to 60 Å with a step size of 1 Å.

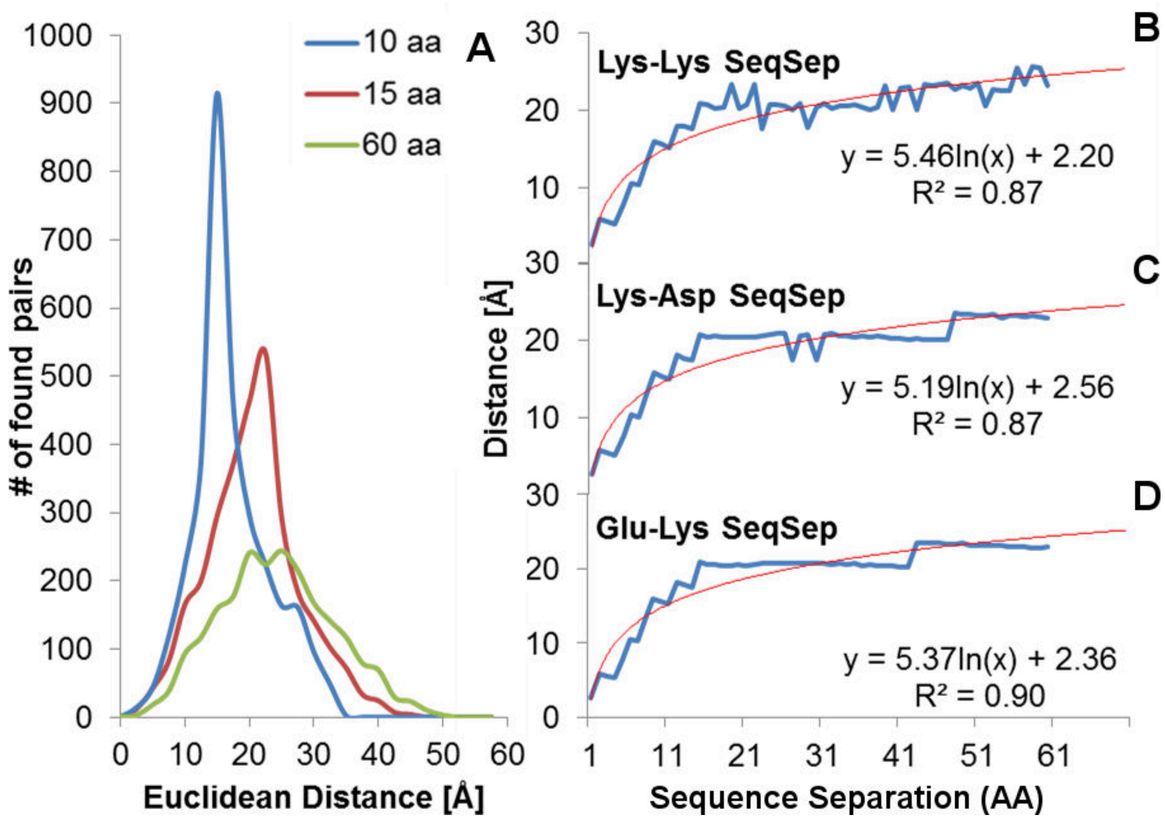

Figure IV.1.: Residue pair distance distributions. *(A) Distribution of the number of Lys-Lys pairs in respect to their Euclidean distance and (B-D) functions representing the relationship between sequence and spatial distance approximated by method of least squares to a logarithmic equation (see equation* (IV.1) *on the preceding page) for (B) Lys-Lys, (C) Lys-Glu, and (D) Lys-Asp.*

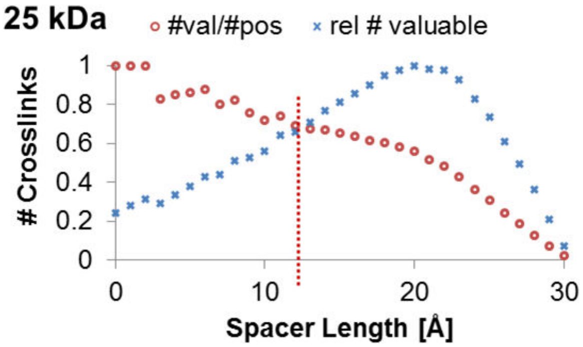

Figure IV.2.: Cross-link yield depending on the spacer length. *Behavior of valuable and possible cross-links in the MW bin* 25 kDa *and localization of the optimal spacer length. Shown is the number of valuable cross-links for every tested spacer length in red. These values are normalized to a dimension spanning* 1*. Blue points show the share of valuable cross-links among the physical possible ones. The dotted line meets the intersection of both curves and represents the optimal spacer length where the best ratio between valuable and possible cross-links is attained and the number of valuable cross-links is maximized in respect to this ratio.*

IV.2.5. Estimation of the optimal spacer lengths for a given protein molecular weight

Over all proteins in each of the molecular weight (MW) bin, the total number of possible distance pairs (P) as well as the number of distance pairs useful for structure determination (V) were computed for each cross-linker spacer length. Furthermore, the maximum number of valuable cross-links observed for all spacer lengths (V_{max}) was determined. For each MW bin the ratios (V/P) and (V/V_{max}) were plotted as a function of the cross-linker spacer length. The optimal cross-linker length for each MW bin was approximated as intersection points of the two functions using a local regression (figure IV.2). The estimated values for the optimal cross-linker spacer length were plotted as a function of the MW and were fitted using a cubic regression curve. The script used for the calculation is available at `http://www.ufz.de/index.php?en=19910`.

IV.2.6. Simulation of cross-linking restraints

Seventeen proteins with known tertiary structure determined via X-ray crystallography (resolution of less than 1.9 Å) were selected from the data set of structures as test cases to evaluate the influence of cross-linking restraints on *de novo* protein structure prediction. To thoroughly benchmark the algorithm, the benchmark set covers a wide range of protein topologies and structural features. The sequence lengths of the proteins range from 105 to 303 residues, the number of SSEs ranges from 5 to 19 with varying α-helical and β-strand content (see table IV.1 on the next page). For these proteins, all solvent accessible surface $C_\beta - C_\beta$ distances between target amino acids in the structure which were within the range of either homobifunctional Lys-reactive cross-linkers or heterobifunctional Lys-Asp/Glu reactive cross-linkers were determined through Xwalk. For the predicted optimal cross-linker length (read above) and spacer lengths of 2.5 Å, 7.5 Å, 17.5 Å and 30.0 Å lists of structurally possible cross-links were generated.

For the two proteins horse heart cytochrome *c* (PDB entry 1HRC) and oxymyoglobin (PDB entry 1MBO) restraints were also derived from published cross-linking MS experiments deposited in the XL database.[181] Experimental cross-linking data of FGF2 (PDB entry 1FGA) and P11 (PDB entry 4HRE) were derived from Young *et al.*[175] and Schulz *et al.*,[187] respectively.

Protein	Uniprot	resolution (Å)	weight (Da)	length	Lysine (%)	α-helix (%)	β-strand (%)
1HRC	P00004	1.9	12 368	105	18	40	1
3IV4	Q7A6S3	1.5	13 235	112	6	49	25
1BGF	P42228	1.5	14 504	124	5	79	1
1T3Y	Q14019	1.2	15 835	141	9	35	29
3M1X	C4LXT9	1.2	15 882	138	7	25	28
1X91	Q9LNF2	1.5	16 419	153	7	76	0
1JL1	P0A7Y4	1.3	17 483	155	7	34	30
1MBO	P02185	1.6	17 980	153	12	77	0
2QNL	Q11XA0	1.5	19 218	162	5	70	2
2AP3	Q8NX77	1.6	23 190	199	23	81	0
1J77	Q9RGD9	1.5	24 226	209	8	62	1
1ES9	Q29460	1.3	25 876	232	3	41	11
3B5O	D0VWS1	1.4	27 506	244	3	71	0
1QX0	P0A2Y6	2.3	32 821	293	7	38	20
2IXM	Q15257	1.5	34 798	303	7	60	3
FGF2	P09038	1.5	17 859	145	10	9	34
P11	P60903	2.0	11 071	95	13	63	3

Table IV.1.: Proteins used for the cross-link spacer length benchmark. *The fifteen proteins for the benchmark set were selected from high-resolution structures deposited in the PDB with varying content of lysines. The structures were selected to cover a wide range of the structural features sequence length as well as the percentage of residues within α-helices and β-strands.*

IV.2.7. Translating cross-linking data into structural restraints

Explicitly rebuilding coordinates for a cross-link is comparable to solving the loop closure problem.[101] During *de novo*, protein structure prediction the cross-link would have to be reconstructed each time the conformation of the protein changes. In a typical MC simulation with a maximum of 12 000 MC steps per model and 5000 models for each protein this would result in a maximum number of 60 million attempts to build the cross-link, which is too resource demanding for usage in *de novo* protein structure prediction. Therefore, we developed a fast approach to estimate the chance that a particular model fulfills a XL-MS restraint. The surface path of a cross-link is approximated by laying a sphere around the protein structure and computing the arc length between the cross-linked residues (figure C.1 on page 173). The geometrical center of the protein structure is used as the center of the sphere. If takeoff and landing point have different distances to the center of the sphere, the longer distance is used as the radius. During the protein structure prediction process, the side chains of the residues are not modeled explicitly but represented on a simplified way through a 'super atom'. While this simplification vastly reduces the computational demand of the algorithm, it also adds additional uncertainty due to the unknown side chain conformations. The agreement of the model with the cross-linking data is quantified by comparing the distance between the cross-linker lengths ($l_{\mathrm{XS}} + l_{\mathrm{SS1}} + l_{\mathrm{SS2}}$) with the computed arc lengths (d_{arc}), with −1 being the best agreement and 0 being the worst agreement. To account for the uncertainty of side chain conformations a cosine-transition region of 7 Å was introduced (figure C.1 on page 173).

IV.2.8. Structure prediction protocol for the benchmark set

The protein structure prediction protocol was based on the previously published BCL::Fold protocol for soluble proteins.[59] In a preparatory step, the SSEs are predicted using the SSE prediction methods PSIPRED[98] and Jufo9D[97] and an SSE pool is created (see section C.2.1 on page 176). Subsequently a MCM energy minimization algorithm draws random glsplsse from the predicted SSE pool and places them in the three-dimensional space (figure IV.3 on the next page; see section C.2.2 on page 176 for details). Random transformations like translation, rotation or shuffling of glsplsse are applied. After each MC step the energy of the resulting model is evaluated using knowledge-based potentials which, among others, evaluate the packing of glsplsse, exposure of residues, radius of gyration, pairwise amino acid interactions, loop closure geometry and amino acid clashes.[60] Based on the energy difference to the previous step and the simulated temperature a Metropolis criterion decides whether to accept or reject the most recent change.

The protein structure prediction protocol is broken into multiple stages, which differ regarding the granularity of the transformations applied, and the emphasis of different scoring terms. The first five stages apply large structural perturbations, which can alter the topology of the protein. Each of the five stages lasts for a maximum of 2000 MC steps. If an energetically improved structure has not been generated within the previous 400 MC steps, the stage terminates. Over the course of the five assembly stages, the weight of clashing penalties in the total score is ramped up as 0, 125, 250, 375 and 500.

The five protein assembly stages are followed by a stage of structural refinement. This stage lasts for a maximum number of 2000 MC steps and terminates if no energetically improved model is sampled for 400 MC steps in a row. Unlike the assembly stages, the refinement stage only consists of small structural perturbations, which will not drastically alter the topology of the protein model.

Through multiple prediction runs with different score weights, the optimal contribution of the cross-linking score to the total score was determined to be 40 % to 50 %. Consequently, the weight for the scoring term evaluating the agreement of the model with the cross-linking data was set to 300 over all six stages, which ensures that the cross-linking score contributes between 40 % and 50 % to the total score.

IV.2.9. De novo *folding simulations without and with cross-linking restraints*

To evaluate the influence of cross-linking restraints on protein structure prediction accuracy, each protein was folded in the absence and in the presence of Lys-Lys, Lys-Glu, and Lys-Asp cross-linking restraints. Independent structure prediction experiments were performed for the predicted optimal as well as two shorter and two longer cross-linker spacer lengths each of the five spacer lengths (table C.1 on page 171). Additionally, predictions were performed using combination of all spacer lengths as well as using restraints obtained by the optimal spacer length of all three cross-linker reactivities. For the two proteins of which experimentally determined cross-linking data were available, protein structure prediction was additionally performed for the experimentally determined restraints. For each protein and cross-linker length used, 5000 models were sampled in independent MCM trajectories. Due to the randomness of the employed MC algorithm, ten sets of 5000 models were sampled for each protein without restraints. Improvements in prediction accuracy can be compared to the standard deviations to identify statistically significant improvements (see table C.2 on page 172 for details).

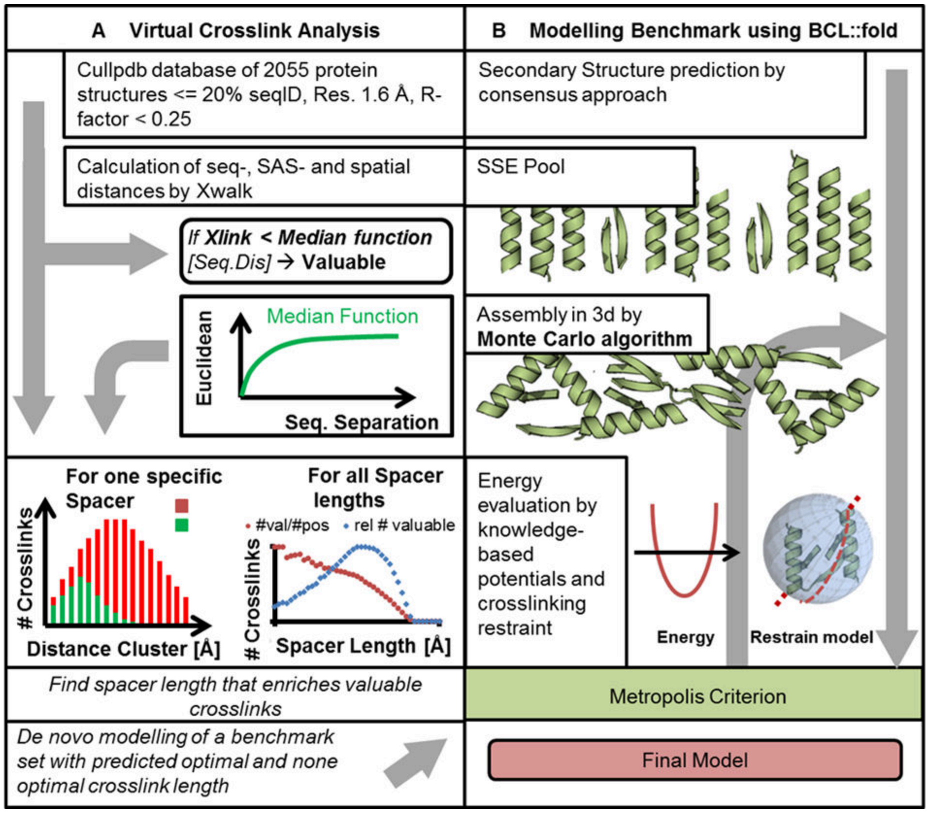
A Virtual Crosslink Analysis
B Modelling Benchmark using BCL::fold
Cullpdb database of 2055 protein structures <= 20% seqID, Res. 1.6 Å, R-factor < 0.25
Calculation of seq-, SAS- and spatial distances by Xwalk
If Xlink < Median function [Seq.Dis] → Valuable
Median Function
Euclidean
Seq. Separation
For one specific Spacer
# Crosslinks
Distance Cluster [Å]
For all Spacer lengths
• #val/#pos
• rel # valuable
# Crosslinks
Spacer Length [Å]
Find spacer length that enriches valuable crosslinks
De novo modelling of a benchmark set with predicted optimal and none optimal crosslink length
Secondary Structure prediction by consensus approach
SSE Pool
Assembly in 3d by Monte Carlo algorithm
Energy evaluation by knowledge-based potentials and crosslinking restraint
Energy
Restrain model
Metropolis Criterion
Final Model

Figure IV.3.: Spacer length and protein structure prediction workflow. *Workflow for (A) the prediction of optimal cross-linker spacer length and (B) for* de novo *protein structure prediction using BCL::Fold. (A) Workflow for the prediction of the optimal spacer length depending on the MW of the protein of interest. (B) Workflow for* de novo *protein structure prediction using BCL::Fold. The SSEs are predicted using the SSE prediction methods PSIPRED and Jufo9D. Subsequently, an MCM algorithm is employed to search the conformational space for the conformation with most favorable score.*

IV.2.10. Metrics for comparing calculating model accuracy and enrichment

The quality of the prediction results was quantified using the RMSD100[103] and enrichment[2,60] metrics. The RMSD100 metric was used to quantify the sampling accuracy by computing the normalized root-mean-square-deviation between the backbone atoms of the superimposed model and native structure. The enrichment metric was used to quantify the discrimination power of the scoring function by computing which percentage of the most accurate models can be selected by the scoring function. The enrichment metric is used to assess the influence of the cross-linking restraints to discriminate among the sampled models. First, the models of a given set S are sorted by their RMSD100 relative to the native structure. The 10 % of the models in S with the lowest RMSD100 are assigned to subset P (positives) and the remaining 90 % of the models are assigned to subset N (negatives). Second, the models in S are sorted by their BCL score. The 10 % of the models in S with the best score are assigned to subset FS (favorable score). The intersection $TP = FS \cap P$ contains the most accurate models which the scoring function can select (true positives). The enrichment $e = \frac{\#TP}{\#P} \cdot \frac{\#P+\#N}{\#P}$ of the most accurate models the scoring function can select. In order to reduce the influence of the sampling accuracy on the enrichment-values, the positive models are considered the 10 % of the models with the lowest RMSD100 and $\frac{\#P+\#N}{\#P}$ is fixed at a value of 10.0. Therefore, the enrichment ranges from 0.0 to 10.0, with a score of 1.0 indicating random selection and a value above 1.0 indicating that the scoring function enriches for native-like models.

IV.3. Results

IV.3.1. Creation of an in silico *cross-linking database*

We performed *in silico* cross-linking experiments on 2055 non-redundant proteins. Covering an MW range from 1.4 kDa to 139 kDa, 59 % of the proteins have an MW below 25 kDa. For each of those proteins all Lys-Lys, Lys-Asp, and Lys-Glu sequence and Euclidean distances as well as the SAS distance between the C_β-atoms were determined. Thus, the resulting database contained information on 391 902 Lys-Lys, 395 815 Lys-Glu, and 360 101 Lys-Asp pairs which built the basis for the determination of the number of possible cross-links, cross-links useful for structure prediction, and finally for the prediction of the optimal cross-linker length for studying a selected protein (figure IV.3 on the previous page).

IV.3.2. Estimation of the possible cross-links per protein

Next we estimated how many and which of the distances could be cross-linked with a cross-linker of a given length and specificity. We considered cross-links possible if the sum of the spacer length and the length of the two connected side chains ($C_\beta - C_\beta$, $N_z - N_z$ for Lys-Lys, $N_z - C_\gamma$ for Lys-Asp, or $N_z - C_\delta$ for Lys-Glu) is longer than the $C_\beta - C_\beta$-SAS-distance between the amino acids. As the lengths of the side chains of Lys ($C_\beta - N_z$), Asp ($C_\beta - O_z$), and Glu ($C_\beta - O_z$) 4.5 Å, 2.4 Å, and 3.6 Å were used, which were determined as average values from the crystal structure of calmodulin (PDB entry 1CLL). *In silico* cross-linking experiments were conducted for all of the 2055 proteins using homobifunctional Lys-Lys-reactive, as well as heterobifunctional (Lys-Asp- and Lys-Glu-reactive) cross-linking reagents with lengths from 1 Å to 60 Å (step size 1 Å).

To draw conclusions from the correlation of this *in silico* cross-linking experiments to the MW of the studied proteins the proteins were grouped into 45 bins with a step size of 2.5 kDa. For example, a protein with a MW in the range of 25 kDa to 27.5 kDa contains on average 15.1 Lys, 14.4 Asp, and 16.7

Glu. On average, 182 Lys-Lys, 173 Lys-Glu, and 144 Lys-Asp cross-links exist per protein within this specific MW bin. Theoretically, all of those could be cross-linked with a cross-linker of 60 Å. In contrast by utilization of cross-linkers of 13 Å (as e.g. BS3) only about 33 % of the cross-links are formed *in silico*. When going to a cross-linker of length of 1 Å (e.g. close to EDC), only 10 % of all possible amino acid pairs are linked.

IV.3.3. Estimation of structurally relevant cross-links

In protein structure prediction approaches, the enrichment of low RMSD structures among the thousands of generated models is crucial. Therefore, we hypothesized that the restraints, which are valuable for structure prediction will reduce the conformational search space substantially. For the present study, we classify a cross-linking restraint as useful for structure prediction if it discriminate at least 50 % of all possible conformations. Thus, in a second step each of the possible cross-links was evaluated in terms of its potential to discriminate at least 50 % of incorrect structures (useful for structure determination) or whether the cross-linked amino acids are so close in sequence that it can be derived from sequence separation that the distance can be bridged by the cross-linker independently of the protein's structure (not useful for structure determination).

In order to develop a stringent measure for usefulness we did not simply assume the maximum distance that can be bridged by an amino acid chain of a certain length. Instead, the Euclidean distance distributions for Lys-Lys, Lys-Glu, and Lys-Asp were computed for the sequence separations ranging from 1 to 60 amino acids within our database of protein structures. For example, in the more than 2000 analyzed structures there are 3132 Lys-Lys pairs, which are separated by ten amino acids. For this sequence distance Euclidean distances bins of 2.5 Å were defined in which the occurrences of residue pairs were counted. The pairs were present in bins ranging from 2.5 Å to 35.0 Å. As the median distance, we found 15.5 Å. For the same sequence distance the distribution of Lys-Glu (3336 pairs) and Lys-Asp (3010 pairs) are quite similar and the median values were 15.6 Å and 15.3 Å.

Similarly, for sequence separations of 15 amino acids, we observed 3024 Lys-Lys pairs, 3200 Lys-Glu pairs, and 2835 Lys-Asp pairs. The median values were 20.8 Å, 20.9 Å, and 20.7 Å, respectively. For sequence separations of 60 amino acids, we observed 2167 Lys-Lys pairs, 2212 Lys-Glu pairs, and 2167 Lys-Asp pairs. The median values are 23.0 Å, 23.0 Å, and 23.0 Å, respectively (see figure IV.1 on page 51 for details).

Approximating the proteins structures as spheres, we applied a logarithmic model to fit the relationship between the sequence separation S and the median Euclidean distance E_{med}. We find

$$E_{\text{med}} = \begin{cases} 5.46 \cdot \ln(S_{\text{Lys-Lys}}) + 2.2 & \text{for Lys-Lys,} \\ 5.37 \cdot \ln(S_{\text{Lys-Glu}}) + 2.36 & \text{for Lys-Glu,} \\ 5.19 \cdot \ln(S_{\text{Lys-Asp}}) + 2.36 & \text{for Lys-Asp.} \end{cases} \tag{IV.2}$$

for Lys-Lys, Lys-Glu, and Lys-Asp distances, respectively.

Secondly, using our derived functions constituting the S/E relationships, we considered every cross-link as of reasonable discriminative power, *i.e.*, which fulfills the criterion that the sum of the cross-linker spacer length and the average length of both contributing side chains is shorter than the median of the sequence/Euclidean-distance distribution. If we examine the 25 kDa MW bins of Lys-Lys targets with a 1 Å spacer cross-link 1167 of the possible 22 398 target pairs fulfilled this criterion and were considered as of sufficient discriminative power (figure C.2 on page 174). These cross-links, which represent 4 % of all Lys-Lys distances we defined therefore as useful for protein structure prediction. Application of a 13 Å spacer length results in 2935 valuable target pairs (12 % of all Lys-Lys distances, see figure C.2 on page 174). In contrast, a cross-linker with a spacer length of 60 Å would allow to cross-link all distances. However, none of the cross-links would have discriminative power for native-like models (figure C.2 on page 174). For the proteins of the 25 kDa MW bins the number of valuable cross-links as a function of the cross-linker length has a log-normal character never exceeding a roughly 25 Å spacer. The intermediate length of 13 Å resulted in an almost equal contribution of valuable and structurally invaluable cross-linking pairs. Whereas 29 % of all possible reactive amino acid pairs are linked, 12 % are considered valuable for structure prediction.

IV.3.4. Prediction of molecular weight dependent optimal cross-linker spacer lengths

Whereas usage of a short cross-linker will result in only a few but mostly structurally valuable restraints, a longer cross-linker will yield more restraints but a lower ratio of valuable restraints. Furthermore, the ratio of valuable restraints as well as the number of possible restraints depends on the size of the protein. In agreement with prior studies regarding structural modeling driven by sparse distance restraints,[188] we hypothesize that a compromise between maximizing the portion of valuable cross-links compared to all cross-links which can be formed with a given cross-linker length (V/P) and maximizing the relative number of achievable cross-links with any spacer length (V/V_{max}) might yield the best results.

Following our hypothesis, for each MW bin we derived the optimal spacer length as the intersection point of the two functions as it is shown exemplarily for an MW of 25 kDa in figure IV.2 on page 52.

The derived optimal spacer lengths for Lys-Lys, Lys-Asp, and Lys-Glu were plotted as function of the MW (figure IV.4 on the next page). The relationship was fitted using a cube root function. For our observable MW sample space for Lys-Lys cross-links, all spacer lengths reached dimensions between 5.0 Å and 18.6 Å. No optimal spacer length was further than 2.5 Å separated from the regression curve. The average distance from the modeled spacer lengths was 0.7 Å. The MW term as well as the side chain term has been modeled as an exponential fraction with respect to the relation between volume and distances in spherical objects.

Additionally, the optimal spacer lengths were also predicted for homobifunctional arginine and for homobifunctional cysteine cross-linking reagents analogously to the procedure being described for the homo- and heterobifunctional lysine-containing cross-links. Consistently, the optimal spacer lengths depend on the MW as well as the lengths of the cross-linked side chains $SS1$ and $SS2$ and could be calculated by $l_{\text{opt}}[\text{Å}] = k \cdot \sqrt[3]{MW} + \sqrt[3]{SS1 + SS2}$. The factor k was determined to be 0.32, 0.31, 0.34, 0.34 and 0.35 for Lys-Lys, Lys-Asp, Lys-Glu, Arg-Arg, and Cys-Cys pairs, respectively.

IV.3.5. Generation of in silico *and experimental cross-linking data for testing the effect of different spacer length for* de novo *modeling*

To evaluate the effect of cross-linking data derived from experiments with different spacer length we folded seventeen proteins *de novo* with BCL::Fold (figure IV.3 on page 55). Thirteen proteins were part of our data set while for four proteins experimental cross-linking data were available (1MBO, 1HRC, 1FGA, and 4HRE) (table IV.1 on page 53). All proteins have a MW in the range from 13 kDa to 27 kDa. Most structures were mainly α-helical with fewer β-strand SSEs. The β-strand content ranged from 0 % to 51 %. The α-helical content ranges from 2 % to 81 %. The highest β-strand content showed 1LMI with also the fewest α-helices. The portion of lysines was between 3 % and 23 %, which resulted in minimal 4 and maximal 46 lysine residues per structure. For the two structures 1MBO and 1HRC, which were studied experimentally, we used the published experimental data, which were obtained using DSG,

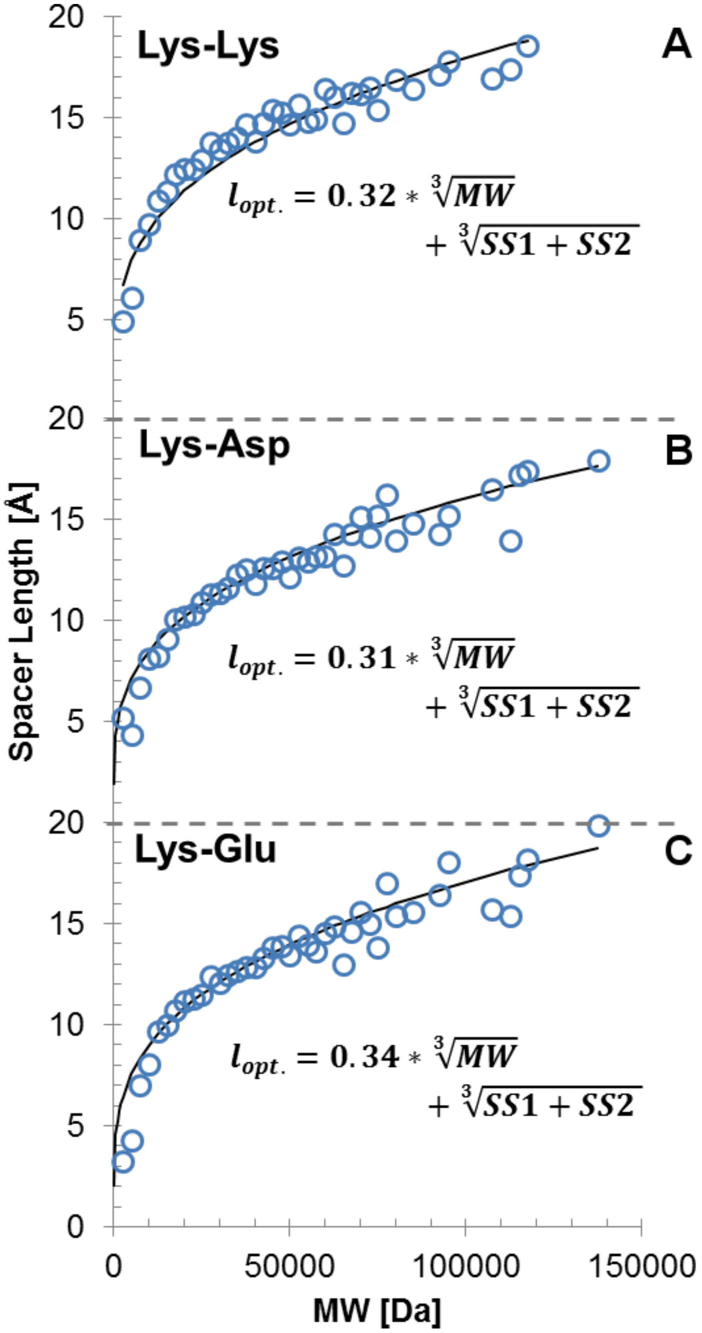

Figure IV.4.: Relationship between sequence distance and Euclidean distance. *Functions representing the relationship between sequence (S) and spatial distance (E). The equations approximated by method of least squares to a logarithmic equation for (A) Lys-Lys, (B) Lys-Glu, and (C) Lys-Asp.*

DSS/BS3, and DEST.[181] For 1MBO, there were 8 cross-links in total with the 11.4 Å reagent BS3 four of them confirmed with the 7.7 Å reagent DSG. For 1HRC, 48 cross-links were reported. 9 DSS, 31 BS3, 6 DSG, and 9 with DEST (11 Å). Six cross-links had been identified with different cross-linking reagents. 18 BS3 cross-links were published for 1FGA,[175] whereas 3 intramolecular BS3 cross-links were available for 1HRE.[187] For the thirteen proteins as well as for 1MBO and 1HRC, we predicted all cross-links, which are possible with the predicted optimal cross-linker length as well as with two shorter and two longer cross-linking reagents (table C.1 on page 171) and used these data as restraints during modeling (figure C.1 on page 173).

IV.3.6. Cross-linking restraints improve the sampling accuracy of de novo *protein structure prediction*

XL-MS data provides structural restraints that reduce the sampling space in *de novo* structure determination. Thereby a fraction of incorrect conformations is excluded and the sampling density in all other areas of the conformational space is increased. To evaluate the power of cross-linking restraints to guide *de novo* protein structure determination we computed the RMSD100-values[103] of the most accurate models for each protein for structure prediction with and without cross-linking restraints. Using cross-linking restraints not only increases the frequency with which accurate models are sampled, but the best models achieve an accuracy not sampled in the absence of cross-linking data (table C.2 on page 172). Across all benchmark proteins, the accuracy of the best models was, on average, 6.6 Å when no cross-linking data was used. By using cross-linking, data for the spacer length deemed optimal the average RMSD100-value was improved to 5.6 Å, which corresponds to two standard deviations. By using restraints obtained for all five spacer lengths, the average accuracy of the best model improved to 5.2 Å. For the proteins 1XQ0, 2IXM, and 3B50, even with cross-linking data, it was not possible to sample a native-like conformation. We attribute this to limitations in the sampling algorithm resulting in the native conformation not being part of the sampling space. For other proteins, significant improvements could be observed. While the accuracy of the best models for 1ES9 and 1J77 was 7.3 Å and 6.8 Å, cross-linking restraints improved the accuracy to 5.7 Å and 4.5 Å, respectively. For 1MBO, the accuracy could be improved from 7.1 Å to 4.2 Å by using a combination of Lys-Glu/Asp reactive cross-linkers (figure C.3 on page 175).

IV.3.7. Cross-linking restraints improve the discriminative power of the scoring function

The ability of the scoring function to identify the most accurate models among the sampled ones was quantified using the enrichment metric (see section IV.2 on page 50). Enrichments were computed for proteins predicted without cross-linking data, for each spacer length and for all spacer lengths combined. For protein structure prediction without cross-linking restraints an average enrichment of 1.1 was measured, which is barely better than random selection. The scoring function has almost no discriminative power. Using cross-linking restraints yielded by the optimal spacer length improved the enrichment to 2.1 (table C.2 on page 172), which corresponds to three standard deviations. Using all five spacer lengths to obtain additional restraints, further improves the enrichment to 2.4. The most significant improvement could be observed for the Gram-negative bacterial oxygenase (PDB entry 1J77), for which the enrichment could be improved from 0.5 to 2.4.

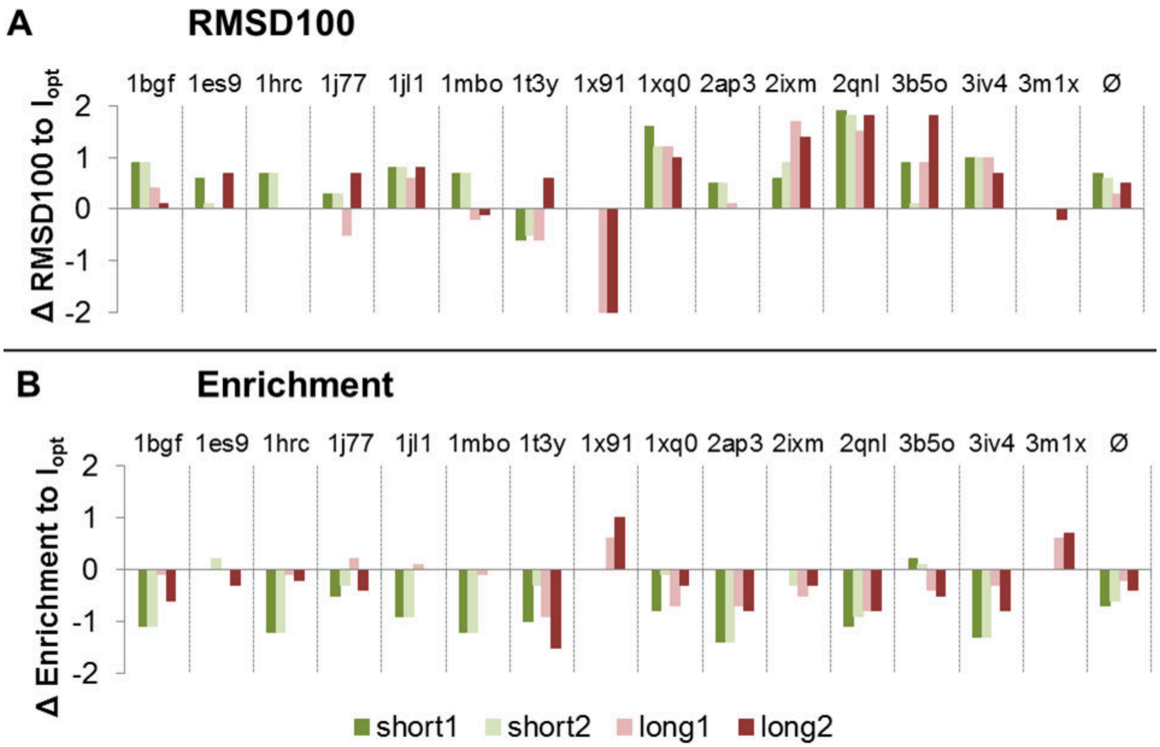

Figure IV.5.: Protein structure prediction results for various spacer lengths. *Cross-linking data improve the prediction accuracy and discrimination power. Using geometrical restraints derived from cross-linking experiments reduces the size of the conformational space, which needs to be searched for the conformation with the lowest free energy. This results in a higher likelihood of sampling accurate models and an improved discrimination power of the scoring function. Panel A compares the RMSD100-values of the most accurate model for structure prediction from different spacer lengths to the results for the optimal spacer length (horizontal line). Panel B compares the enrichments for different spacer lengths likewise.*

IV.3.8. The cross-linker length determines improvements in sampling accuracy and discrimination power

The length of the cross-linker determines the number of obtainable restraints as well as their information content.[78] Whereas a longer cross-linker is able to form more cross-links and therefore yields a larger number of restraints, the longer cross-linker length can be fulfilled by a larger number of conformations, reducing the discriminative power of the restraint. In order to assess the influence of the cross-linker length, and therefore the number of restraints and restraint distances, on the sampling accuracy and discrimination power, the protein structure prediction protocol was conducted with restraints derived from different cross-linker lengths.

The cross-linker lengths were separated into five groups: *optimal*, which is the predicted optimal cross-linker length, *short1* and *short2*, which are shorter cross-linker lengths, and *long1* and *long2*, which are longer cross-linker lengths. The cross-linker length predicted to be optimal yielded the most useful restraints for protein structure prediction in terms of sampling accuracy and discriminative power. Across all proteins the average RMSD100-values of the most accurate models for the optimal cross-linker length were 5.6 Å — an improvement by 15 % — while they were 6.3 Å, 6.2 Å, 5.9 Å and 6.1 Å — improvements by 5 %, 6 %, 11 % and 8 % — for the shorter and longer cross-linker lengths, respectively (figure IV.5). The longest cross-linkers have a less significant impact on the sampling accuracy due to their ambiguity, whereas the shortest cross-linkers failed to yield a sufficient number of distance restraints to impact prediction accuracy. The method's discriminative power, as quantified through the enrichment metric, for the optimal cross-linker length was 2.1, whereas it was 1.4, 1.5, 1.9 and 1.7 for the shorter and longer cross-linkers, respectively (see figure IV.5 on the previous page for details). For the proteins 1X91 and 3M1X, the optimal cross-linker length did not yield any cross-links with a sequence separation of at least ten and therefore did not provide relevant structural information. In those cases protein structure prediction with longer cross-linker lengths provided better results. By combining restraints obtained for all five cross-linker lengths, the average enrichment-value could be improved to 2.4.

IV.3.9. Combination of cross-linkers with different reactivities results in improvements larger than seen when varying the spacer lengths

In order to obtain valuable restraints for *de novo* protein structure prediction a maximum number of SSE pairs needs to be cross-linked. The availability of Lys-Asp/Glu reactive cross-linkers allows for a better sequence coverage and therefore a wider coverage of SSE pairs. Cross-links with different spacer lengths were simulated for the proteins in the benchmark set using Xwalk. To assess the influence of Lys-Asp/Glu reactive cross-linkers on protein structure prediction the same protocol was applied as for the Lys-Lys reactive cross-linkers. For the Lys-Glu reactive cross-linkers a prediction accuracy of 5.7 Å and enrichment of 2.2 on average could be achieved, which is comparable to the results for the Lys-Lys reactive cross-linkers.

While Lys-Asp reactive cross-linkers also achieve improvements in prediction accuracy and enrichment when compared to protein structure prediction without restraints, the results are slightly worse than for Lys-Lys reactive cross-linkers with an average prediction accuracy of 6.0 Å versus 5.6 Å and an average enrichment of 1.7 versus 2.1 (table C.2 on page 172). To a large part, the difference in the overall results is caused by the results for the proteins 1ES9, 1T3Y, and 3M1X for which Lys-Asp reactive cross-linkers failed to yield restraints between SSE pairs and therefore failed to reduce the conformational space significantly. Besides deviations regarding the average improvements over all proteins, the spacer length deemed optimal also provides the best results for Lys-Asp/Glu reactive cross-linking. Combining the restraints yielded for the optimal spacer lengths with Lys-Lys/Asp/Glu reactive cross-links improves the sampling average sampling accuracy to 5.1 Å and the average enrichment to 2.6. Combining the restraints yielded by all spacer lengths and cross-linker reactivities failed to further improve prediction results.

IV.4. Discussion

IV.4.1. Prediction of the optimal cross-linker spacer length

It has been demonstrated frequently that chemical cross-linking data can be used to guide *de novo* structure prediction and selection of native-like models. Certainly, the sensitivity, the broad applicability to almost all proteins, the almost physiological experimental condition during the chemical cross-linking reaction, and the potential of automation are the main advantages for using XL-MS to generate such restraints. However, the small number and high uncertainty of restraints from chemical cross-links limit impact on *de novo* proteins structure prediction, in particular when compared to more data rich methods such as NMR spectroscopy.[154]

One major limitation arises from the fact that distances between the functional groups in long and flexible amino acid side chains are measured. Therefore, a significant uncertainty is added to the cross-linker length when converting XL-MS data into $C_\beta - C_\beta$ restraints. A second challenge of chemical cross-links is that only the maximum distance is restricted, but no information is obtained on the minimum distance or the favored distance distribution. In result, even a "zero length" cross-linker restricts the $C_\beta - C_\beta$ distance to the sum of the lengths of the two connected side chains (e.g. for Lys-Lys cross-links 9.1 Å).

In most of the cross-linking experiments, lysine residues are targeted. Lysines are excellent targets because of their overrepresentation on protein surfaces and the clean chemistry of amine modification. Nevertheless, their frequency is on average only about 7 %. As a consequence the number of cross-links, which can be formed for example in a 25 kDa protein with a standard homobifunctional Lys-Lys-reactive cross-linking reagents with a spacer length of 6.4 Å (length of DST) are in the range of about 20. Only a small fraction of these restraints will substantially limit the conformational space for the protein. This number is usually too small to restrict the conformational space to an unambiguous single protein fold. To increase the number of restrains it is possible to use cross-linkers with longer spacer length or target amino acids such as Asp, Glu, Tyr, Ser, Thr, Arg, or Cys.

Restraints obtained with longer cross-linking reagents are less restrictive to the conformational space. To evaluate the value of cross-links for protein structure prediction we determined for each sequence distance (0 to 60 amino acids) how long a cross-linker has to be to link the target amino acids. For example two lysines, which are separated by eight amino acids in sequence were found to be linkable in all 3488 cases by a homobifunctional cross-linker with a length greater than 30 Å (as it is in BS(PEG)9). In our study, we stated the hypothesis that it would be desirable if two target amino acids can only be linked in 50 % of all models created meaning that 50 % of all structures could be discarded based on a single cross-link. For example, for two lysines separated by 10 amino acids this would be the case for cross-linker lengths of 14.8 Å (distance distributions for other amino acids distances are shown in figure IV.1 on page 51). Cross-links, which could only be formed in less than 50 % for the corresponding sequence distance, were considered as being valuable. Based on this definition for all 2055 structures of the applied non-redundant protein structure database the optimal spacer length was calculated. With this optimal spacer length, the number of structural valuable cross-links has been maximized taking into account that in general for modeling approaches few distance restraints of highly discriminative character are less favorable than a higher number with a smaller discriminative power.[183,188]

Since the optimal cross-linker length should depend on the protein size in a cubic root fashion to convert volume into distance, it is not unexpected that this was also observed for the dependency on the MW (figure IV.4 on page 59). However, one has to keep in mind that the formula might not be applicable to non-globular proteins and multi-domain proteins. However, in case of multidomain proteins this formula should be applicable to the separated domains. Remarkably, based on our simulation for proteins with a MWs of 10 kDa, 25 kDa, 50 kDa and 100 kDa the recommended spacer lengths are 9.0 Å, 11.5 Å, 13.9 Å and 17.0 Å, respectively, which is quite close to the homobifunctional amine-reactive commercially available cross-linkers DSG (7.7 Å), BS3 (11.4 Å), and EGS (16.1 Å), which are currently the preferred choice to study small (less than 20 kDa), medium (20 kDa to 50 kDa), and large proteins (greater than 50 kDa), respectively.

Addressing different functional groups is a second approach to increase the total number of distance restraints. The consequence is that the cross-linking reaction is either less effective or specific (Asp, Glu, Tyr, Ser, Thr) creating challenges in data interpretation or the target amino acids are less frequent (Arg and Cys) limiting the number of restraints observed. However, using the same theoretical approach revealed that optimal spacer length for heterobifunctional Lys-Asp and Lys-Glu cross-linker (figure C.3 on page 175) as well as homobifunctional Cys-Cys and Arg-Arg cross-linker can be predicted with the same equation: $l_{\text{opt}}[\text{Å}] = k \cdot \sqrt[3]{MW} + \sqrt[3]{SS1 + SS2}$ with $k \approx \frac{1}{3}$ where $SS1$ and $SS2$ are the average lengths of the cross-linked side chains.

When comparing the two approaches to increase the number of valuable cross-links, it should be pointed out that using several cross-linking reagents with different reactivities results in significantly higher improvement of the model quality than using only lysine reactive cross-linking reagent but with different spacer length.

IV.4.2. Challenges in using cross-linking data to guide de novo *modeling*

To test whether the cross-linker with the predicted optimal spacer length indeed perform best in modeling we have chosen a *de novo* structure prediction approach BCL::Fold for testing. Even though comparative modeling using known protein structures as a template usually performs better then *de novo* modeling, our goal was to maximize impact of XL-MS restraints.

A major limiting factor for *de novo* protein structure prediction methods is the vast size of the conformational space. Cross-linking restraints can aid the computational prediction of a protein's tertiary structure by drastically reducing the size of the sampling space. Cross-linking experiments yield a maximum Euclidean distance between the cross-linked residues, which increases the sampling density in the relevant part of the conformational space.

A major limitation of using cross-linking restraints to guide protein structure prediction when compared to restraints obtained from EPR and NMR spectroscopy is that the cross-linker length cannot be directly translated into a Euclidean distance between the cross-linked residues. While cross-link prediction and evaluation methods like Xwalk[178] are successful at predicting if a certain cross-link can be formed in a given structure, explicit modeling approaches are computationally too expensive for usage in a rapid scoring function required for protein structure prediction. Approximations, such as the great circle on a sphere presented here, are fast to compute but associated with increased uncertainties. Most of the cross-linkers used can cover a long Euclidean distance and therefore the yielded restraints can be fulfilled by a wide variety of conformations. In spite of this, cross-linking restraints display some potential in limiting the size of the sampling space, resulting in a higher density of accurate models. Further, the geometrical restraints derived from XL-MS allow for the discrimination of a significant fraction of models representing incorrect topologies and therefore improve the discriminative power of the scoring function.

IV.4.3. Abilities and limitations of protein structure prediction from limited experimental data

We showed that incorporation of cross-linking data into a *de novo* protein structure prediction method improves the accuracy of the structure prediction. The two major challenges of *de novo* predictions are the sampling of structures as well as the discrimination of inaccurate structures. In this study reduction of the conformational space was achieved through the assembly of predicted SSEs with limited flexibility and the incorporation of geometrical restraints derived from cross-linking data. The discrimination of inaccurate models is performed through a scoring function which approximates the free energy. Assuming that the native structure is in the global energy minimum, complete sampling and an accurate methods to measure free energy would lead to the correct identification of the native conformation. However, the conformational space is too large to be extensively sampled and the free energy needs to be approximated, which results in ambiguity regarding the model which is most similar to the native structure. Incorporating cross-linking data provides geometrical restraints, which can be used as additional criteria to discriminate inaccurate models. While an average sampling accuracy of 5.1 Å, when using restraints yielded by XL-MS, is a significant improvement over the 6.6 Å, when not using cross-linking data at all, only for four proteins it was possible to sample models with an RMSD100 of less than 4 Å when compared to the crystal structure. Cross-linking data yields an upper boundary for the Euclidean distance of the cross-linked residues, which allows for the placement of the second residue within a sphere of volume $\frac{4}{3}\pi r^3$ around the first residue. Depending on the cross-link distribution, topologically different models can fulfill the same restraint set. Discrimination among those models is impossible with XL-MS restraints.

IV.4.4. Comparison of experimental and in silico *cross-links*

In order to draw general conclusion based on the analysis of hundreds of different structures this study relies mainly on virtual cross-linking experiments. Unfortunately, although extensive XL-MS data sets have been published for several proteins, it proved difficult to obtain suitable experimental data sets for the present benchmark due to additional requirements: (i) the protein must be monomeric and small enough for *de novo* protein folding with BCL::Fold, (ii) an experimental atomic detail structure for comparison, and (iii) a large data set of intramolecular cross-links must be available. Results for the four cases P11, FGF2, cytochrome *c*, and oxymyoglobin that came closest are reported to demonstrate our efforts to work not only with simulated data. However, for P11 and FGF2 using experimentally determined restraints did not improve the prediction results in a statistically significant way. For P11, only three restraints were available with a maximum sequence separation of nine residues. Because of the small sequence separation, these restraints contain very limited structural information and no improvement in *de novo* folding can be expected. The tertiary structure of FGF2 contains twelve β-strands with several β-strands that are strongly bent. This protein is too large for *de novo* structure determination with BCL::Fold. As BCL::Fold is unable to sample the conformation of the protein in the first place, no significant improvement was expected or observed when XL-MS data were added. Nevertheless, the value of the predicted cross-links in comparison to experimental cross-links could be validated with the two proteins cytochrome *c* and oxymyoglobin for which experimental cross-links had been published in the XL database.[172] For cytochrome *c* (PDB entry 1HRC), we indeed found that the cross-linker with predicted optimal spacer length of 10.2 Å performed best. However, for oxymyoglobin (PDB entry 1MBO) the longer spacers improved the accuracy slightly more than the cross-linker with the optimal spacer length. Interestingly, on the one hand for both proteins several cross-links, which should be possible, could not be detected, which might be due to experimental or analytical reasons. On the other hand, also several cross-links, which were experimentally, identified which were not predicted. An examination of these data revealed that most of these cross-links are not present in the virtual data set because their $C_\beta - C_\beta$ distances exceed the expected maximum length. This finding is in agreement with Merkley *et al.*,[189] who evaluated protein structures by molecular dynamics and reported that usually a high number of experimental approved cross-links exceed the theoretical maximal spatial distance due to structure flexibility. It was concluded for the investigation of Lys-Lys distances using a BS3/DSS cross-linking reagent an upper bound of 26 Å to 30 Å for C_α-atoms.[189]

On the other hand, spacer conformations usually adapt lengths that are somehow distributed between their minimal and maximal lengths. In line it was also reported that many spacers in commercially available cross-link agents preferable adopt conformations, which are significantly below the cited maximal spacer length.[190] Thus, ideally cross-linking results should be evaluated based on experimentally derived or simulated ensembles of in-solution structures instead of using X-ray structures as reference. However, to address all degrees of flexibility during the *de novo* structure prediction is currently too resource intensive. Furthermore, there are many additional practical challenges, which may prevent the formation or identification of cross-links, and thus may result in more meaningful results using a cross-linker with a non-optimal length. Nevertheless, for both structures the sampling accuracies could also be improved by 0.7 Å based on the experimentally determined restraints, which is only slightly worse than the improvement of 1.0 Å observed based on *in silico* cross-links.

IV.5. Conclusion

Recent development of high-resolution MS instruments enables the analysis of proteins not accessible to NMR spectroscopy and X-ray crystallography. Data obtained from those experiments can be translated into structural restraints to guide protein structure prediction. The information content of a geometrical restraint obtained from XL-MS experiments is directly dependent on the used spacer length. Thus, the choice of the spacer length is an important step.

Firstly, for amino acids pairs close in sequence only minimum structural information is obtained if the spacer is too long. Here we determine the optimal spacer length to gain structural information on lysines with a sequence separation of S, we estimated a length as $E = 5.5 \cdot \ln(S) + 2.2$. Secondly, we demonstrate that for *de novo* protein structure prediction the optimal spacer length depends on the MW of the protein of interest and the length of the cross-linked side chains ($SS1$ and $SS2$) and can be predicted as $l_{\text{opt}} = k \cdot \sqrt[3]{MW} + \sqrt[3]{SS1 + SS2}$, with $k \approx \frac{1}{3}$.

We also demonstrate that restraints obtained from cross-linking experiments contribute moderately to solving the major challenges of *de novo* protein structure prediction — the vast size of the conformational space and discrimination of inaccurate models. Using restraints from cross-linking experiments significantly increases the sampling density of native-like models and contribute to the discrimination of incorrect models. By combining cross-linking restraints with knowledge-based scoring functions, the average accuracy of the sampled models could be improved by up to 2.2 Å and the average enrichment of accurate models could be improved from 11 % to 24 %.

Conclusively, we believe this study can help in the planing of XL-MS experiments as well as to evaluate how much information can be gained by XL-MS experiments and the ambiguity that remains.

IV.6. Acknowledgments

This study was supported by grants from Deutsche Forschungsgemeinschaft Transregio 67 (subproject Z4) and ESF Investigator group GPCR 2. Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 GM099842, R01 DK097376) and NSF (CHE 1305874). This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. Parts of the data analysis were performed using the R package.[b] The renderings of the models were created using Chimera.[124]

[b] https://www.r-project.org

CHAPTER V
EFFICIENT SAMPLING OF LOOP CONFORMATIONS

This chapter is based on the publication "Efficient sampling of loop conformations using conformation hashing in conjunction with cyclic coordinate descent".[4] Axel W. Fischer contributed to the development of the loop construction algorithm, performing the experiment, analyzing the data, and writing the article.

De novo *construction of loop regions is a critical and resource-consuming task. In the absence of periodic backbone hydrogen bonds that define secondary structure, loop regions are more likely to exhibit significant conformational flexibility. Accordingly, an ensemble of physically realistic conformations could provide a more relevant representation of the protein in its native state. We developed a loop construction algorithm using conformational hashing complemented with cyclic coordinate descent. It achieves a closure rate of* 100 % *on a benchmark set consisting of twenty-nine proteins with* 296 *non-terminal loops, while requiring only* 161 ms *on average to close one loop. The efficiency of the algorithm enables it to be used for protein ensemble prediction and simulation of spectroscopic data observed on the equilibrium constitutions of proteins. In this manuscript, we investigate the bottlenecks and limitations of conformational hashing and provide a detailed technical description of the algorithm to enable implementation by other researchers.*

V.1. Introduction

The construction of loop regions in proteins remains a challenge. The absence of periodic backbone hydrogen bonds that define regular secondary structure allows for a large conformational space that needs to be searched. Often, multiple conformations with low differences in their free energy might exist, making it difficult to identify the single conformation in the absolute free energy minimum or to determine an ensemble of conformations that accurately mimics the loop flexibility at room temperature. Benchmarking of loop construction algorithms is complicated, as loop conformations can be perturbed in experimental structures from the global free energy minimum, for example through contacts in the crystal lattice.

Multiple research groups have developed loop modeling approaches for usage within their respective software suites. Rohl *et al.* developed a method based on Rosetta[94] to predict the conformations of structurally diverse regions in comparative models.[191] Their approach constructs the initial conformations of short variable segments from structural templates selected from the PDB. The conformations for the longer segments are constructed through the Rosetta software suite from fragments with a sequence length of three and nine residues. The resulting conformations are refined using gradient minimization, MC minimization, and rapid repacking of the side chains. Their method was evaluated in the CASP experiment and compared favorably to alternative approaches.

Canutescu *et al.* developed the CCD algorithm,[101] which was inspired by the random tweak algorithm used in robotics. The CCD algorithm starts from an extended loop conformation and employs random rotations around the loop's rotatable bonds to optimize the superimposition of the loop's virtual terminus and the loop's anchor point. This approach was shown to achieve a high closure rate even for longer loops.[101] However, limitations of this approach are its time complexity that depends on the sequence length of the loop and potential distortions in the loop's dihedral angles.

Tyka *et al.* published an orthogonal approach using conformational hashing.[192] In their approach, a template library of known loop conformations is generated from protein models deposited in the PDB. To construct a missing loop region, their algorithm performs a look-up in the template library and fits the loop's sequence against the selected template. This approach has the advantage of being fast even for long loops while using dihedral angles that have been experimentally observed before.[192] However, a severe limitation is the underrepresentation of long loops in the PDB, which results in a lack of templates for longer loops. Consequently, this approach is only suitable for shorter loops.

Mandell *et al.* detailed the method "kinematic closure (KIC)" that was inspired by robotics and achieved sub-angstrom accuracy.[193] For the reconstruction of a loop region of sequence length N, KIC designates three of the loop's C_α-atoms as pivot points, whereas the remaining $N-3$ C_α-atoms are non-pivot. The method samples the non-pivot torsion angles using a Ramachandran map and the torsion angles of the pivot points are subsequently sampled to close the loop. This method was evaluated on a benchmark set consisting of twenty-five loops and improved the median accuracy from 2.0 Å to 0.8 Å.[193]

An optimal loop construction algorithm will find middle ground between low time complexity, high closure rate, and physically realistic dihedral angles of the sampled loop conformations. To achieve this goal, we developed a conformational hashing algorithm that was extended to include functionality for fragment recombination to counteract the lack of templates for long loops. In this approach, long loops can be constructed from shorter fragments within a MCM framework. The algorithm was developed to work in conjunction with a CCD implementation, which is only applied to loops that even with fragment recombination cannot be closed. The resulting compound algorithm combines the advantages of conformational hashing and CCD while mitigating their limitations.

In this study, we describe the implementation details of the conformational hashing algorithm with fragment combination and its performance with and without its CCD complement. This algorithm was implemented as part of the BCL.[a] The materials and methods section starts with a description of the general methodology — how conformational hashing and CCD are combined within an MCM framework. This is followed by an analysis of the generated loop template library and a mathematical description how loop conformations can be parameterized. The subsequent sections provide a mathematical description how templates and fragments can be combined into longer templates to compensate for the lack of long loops in the PDB, a technical description of the employed CCD algorithm, and a summary of the benchmark used to quantify the performance of the algorithm. The materials and methods section is concluded by a description how the performance of our algorithm was compared to Rosetta's "loophash" algorithm. In the results section, we evaluate the performance of the approach in terms of loop closure rate and CPU time consumption in dependence of the algorithmic parameters as well as discussing the effective loop length limit of the conformational hashing algorithm. The reported results are also compared to Rosetta's conformational hashing algorithm.

V.2. Materials and methods

In this section, we provide a detailed technical description of the algorithm to enable other researchers to readily implement and enhance this algorithm. The following subsections describe the implementation

[a] http://www.meilerlab.org/bclcommons

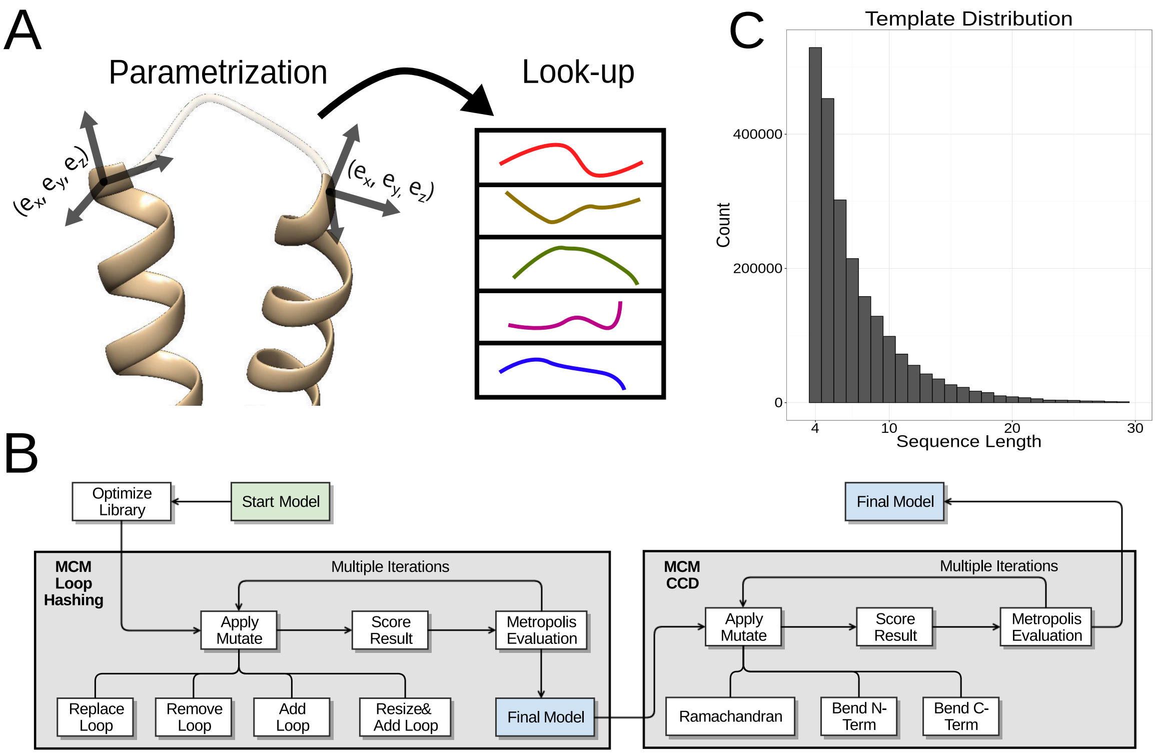

Figure V.1.: Algorithmic approach for loop construction using conformational hashing in conjunction with CCD. *(A) Loops are parameterized through a hash key that is computed from the sequence length of the loop and the relative orientation of its anchor points. The selection of suitable templates is performed through a hash look-up of suitable conformations. (B) The loop construction algorithm consists of a conformational hashing and a CCD stage. Both stages are embedded into MCM frameworks featuring mutates for adding, replacing, and removing loops as well as bending the termini of the anchor SSEs. (C) The initial template library consisted of about* 3.7 *million loop conformations with different sequence lengths that were collected from a set of about* 87 000 *protein structures deposited in the PDB.*

of the conformational hashing and CCD algorithms as well as how they are embedded into an MCM algorithm to facilitate efficient construction of missing loop regions. This is followed by a subsection describing the benchmark set used to evaluate the performance of the algorithm and a subsection describing how additional loop templates can be constructed *in silico* from already known loop templates. This section is concluded by a description of the comparison to Rosetta's algorithm.

V.2.1. General methodology and obtaining the loop template library

The loop construction method described in this manuscript uses a sampling approach consisting of two stages, which is followed by a post-processing step. In the first stage, the loop regions are constructed from precomputed templates in a sequence-independent manner using conformational hashing. Since suitable templates are not available for all loop regions, the second stage uses a CCD algorithm to close the loops that could not be closed by the first stage. The two stages are followed by a post-processing step to construct the terminal loop regions.

The general methodology of the first stage, the conformational hashing, can be broken down into two steps: i) creation of a library containing loop templates and ii) construction of missing loop regions in protein models using the gathered templates (figure V.1 on the preceding page). In the first step, the initial template library was compiled from experimentally determined protein structures deposited in the PDB.[194] The Dunbrack lab's PISCES[185,195] server[b] was used to select a subset from the PDB consisting of structures with a minimum resolution of 3.0 Å while excluding structures that completely or partially consisted of C_α-traces. Those selection criteria resulted in a set of about 87 000 protein structures. The loop regions in the protein structures were determined using the SSE definition program DSSP.[113] The SSE definitions provided in the PDB files were ignored to ensure that the same definition criteria were applied to all structures and our results are reproducible. After filtering out loops containing unresolved backbone coordinates, a set consisting of about 3.7 million loop conformations was collected (figure V.1 on the previous page). Subsequently, each of the loop conformations was parametrized according to geometrical aspects and stored alongside the loop's conformation in the template library (see the following sections for details). In the second step, the parametrization of the templates was translated into a hash key and the loop conformations associated with this key were stored in a hash table. For the construction of loop regions, an MCM algorithm was employed (figure V.1 on the preceding page), which used mutates to add and replace loops for given protein models (see the following sections for details). To add a loop, the mutate computes the parameterization of the missing loop, computes the hash key for the parametrization, and randomly selects a suitable template from the hash table. After selection of a suitable template, the loop's sequence is fitted against the template and inserted into the protein model. The advantage of this approach is that the template look-up can be performed in $O(1)$ and generally requires CPU time in the order of microseconds. To counteract the lack of templates for long loops, an additional mutate constructs missing loop regions from shorter fragments (see the following sections for details).

The general methodology of the second stage, the CCD algorithm, was previously described by Canutescu *et al.*[101] and functions by calculating the rotation that must occur around a given axis (ϕ, ψ) in order to minimize the distance between a moving (loop end) and target (anchor) set of coordinates over many iterations in order to close a chain break. In order to prevent getting stuck in a non-closable conformation based on the starting loop conformation, in this implementation, the residue and ϕ orψ are randomly chosen at every step. In addition, only a random fraction of the rotation is applied. This protocol can be performed in the presence or absence of scoring functions. If used in the presence of scoring potentials, rotations that lead to severe clashes between amino acids or SSEs for example would be rejected. The original protocol was extended to allow for bending of the terminal regions of the preceding and succeeding SSEs (figure V.1 on the previous page).

Following the second stage, the terminal loop regions were constructed starting from an extended conformation exhibiting backbone dihedral angles that were randomly drawn from a (ϕ, ψ) distribution derived from experimentally determined protein structures (see the following sections for details). The CCD algorithm in conjunction with knowledge-based potentials was subsequently applied to the constructed terminal loop regions to resolve steric interferences and optimize other energetically unfavorable or forbidden conformations.

V.2.2. Parametrization of loop conformations and selection of suitable conformations

Loop construction finds a peptide conformation that is able to bridge a gap between the C-terminal residue of the N-terminal SSE and the N-terminal residue of the C-terminal SSE — henceforth called the

[b] http://dunbrack.fccc.edu/PISCES.php

two anchor points of the loop. Consequently, each loop can be defined through the relative translational and rotational orientation of the anchor points and the sequence length of the loop. To formalize the relative translational and rotational orientation of the anchor points, we defined local orthonormal coordinate systems for the anchor points of each loop as $(e_\mathrm{x}, e_\mathrm{y}, e_\mathrm{z})$ based on the backbone coordinates of the anchor points (figure V.1 on page 69), with $\boldsymbol{e}_\mathrm{x}$ being the normalized $\mathrm{C}_\alpha - \mathrm{C}$ vector, $\boldsymbol{e}_\mathrm{y}$ being the normalized component of the $\mathrm{C}_\alpha - \mathrm{O}$ vector that is orthogonal to $\boldsymbol{e}_\mathrm{x}$, and $\boldsymbol{e}_\mathrm{z}$ being computed from $\boldsymbol{e}_\mathrm{x}$ and $\boldsymbol{e}_\mathrm{y}$ as $\boldsymbol{e}_\mathrm{z} = \boldsymbol{e}_\mathrm{x} \times \boldsymbol{e}_\mathrm{y}$. The origins of the coordinate systems reside in the C_α-atoms of the anchor points and the translation vector between the coordinate systems is defined accordingly as $\boldsymbol{t} = (t_\mathrm{x}, t_\mathrm{y}, t_\mathrm{z}) = (\alpha_{c,x} - \alpha_{n,x}, \alpha_{c,y} - \alpha_{n,y}, \alpha_{c,z} - \alpha_{n,z})$, with $\alpha_{n,x}$ being the x-coordinate of the C_α-atom of the N-terminal anchor and $\alpha_{c,x}$ being the x-coordinate of the C_α-atom of the C-terminal anchor. The relative rotational orientation of the two anchor points was quantified using Euler angles (α, β, γ) following the extrinsic x-y-z convention.[196] The Euler angles can be readily extracted from the matrix $\boldsymbol{M}_\mathrm{r}$[197] describing the rotation between both coordinate systems that can be computed as $\boldsymbol{M}_\mathrm{r} = \boldsymbol{M}_\mathrm{n}^{-1} \cdot \boldsymbol{M}_\mathrm{c}$, with $\boldsymbol{M}_\mathrm{n}$ and $\boldsymbol{M}_\mathrm{c}$ being the transformation matrices of the local coordinate systems at the N- and C-terminal anchor points.

The resulting parameterization of each loop consists of seven parameters: the sequence length d of the loop, the three components of the translation vector $(t_\mathrm{x}, t_\mathrm{y}, t_\mathrm{z})$, and the three Euler angles (α, β, γ). Those seven parameters are discretized through binning and translated into a one-dimensional hash key k using the hash function f defined in equation (V.1).

$$f : d \times (t_\mathrm{x}, t_\mathrm{y}, t_\mathrm{z}) \times (\alpha, \beta, \gamma) \rightarrow k \tag{V.1}$$

where:

d	= sequence length of the loop
$(t_\mathrm{x}, t_\mathrm{y}, t_\mathrm{z})$	= translation vector between the anchor points
(α, β, γ)	= Euler angles between the anchor points

The discretization aims at grouping structurally similar loops and is necessary to prevent sparse population of the hash map. For the discretization, we evaluated different bin sizes in this study (see the results section for details) and found bin sizes of 1 Å for the components of the translation vector and 60° for the Euler angles to provide the best closure rates while maintaining reasonable accuracy for the given use case. Additionally, for each loop, the conformation c is stored as a sequence of dihedral angles (ϕ, ψ), resulting in a sequence of length $2d + 2$: $c = (\psi_\mathrm{n}, \phi_1, \psi_1, \dots, \phi_\mathrm{d}, \psi_\mathrm{d}, \phi_\mathrm{c})$, where ψ_n is the ψ-angle of the N-terminal anchor point, $(\phi_\mathrm{i}, \psi_\mathrm{i})$ are the dihedral angles of the i-th residue of the loop, and ϕ_c is the ϕ-angle of the C-terminal anchor point. Consequently, each loop results in a key-value pair (k, c) that, can be stored in a hash map.

The look-up of suitable conformations for a given loop is performed accordingly. For each missing loop, the local coordinate systems of its anchor points are computed and the resulting hash key (see equation (V.1)) is determined. Suitable conformations that would close this loop can subsequently be looked up in the hash map within constant time complexity. The sequence of the loop is then fitted against the dihedral angles of the selected template and inserted into the protein model. It needs to be noted that although the look-up happens in $O(1)$ and therefore is independent of the length of the loop, the fitting can only be performed in $O(n)$, which results in linear time complexity for the overall algorithm.

V.2.3. Construction of loop regions using conformational hashing within a Monte Carlo Metropolis framework

Insertion of different loop conformations can be mutually exclusive. For example, in a protein model containing multiple loop regions, closing a certain loop with a certain conformation might prevent the other loops from being closed. Consequently, an algorithm is needed that can efficiently sample different combinations of loop conformations while scaling well with the available computational resources. Previous studies have demonstrated the successful application of MCM algorithms on protein structure prediction problems,[1,59,94] which convinced us to embed the loop construction algorithm into an MCM framework. Effectively, the loop construction algorithm consists of two MCM algorithms executed in sequence. The first MCM algorithm embeds the conformational hashing algorithm and the second MCM algorithm embeds the CCD algorithm, with the latter only being applied to the loops that could not be constructed by the conformational hashing algorithm.

In the MCM implementation of the conformational hashing algorithm (figure V.1 on page 69), a mutate is randomly selected and applied to the protein model; the resulting protein model is subsequently evaluated using a scoring function, and, depending of the score difference to the previous protein model, the new model is either accepted or rejected.[59] Our implementation consists of four mutates for entire loops — *Add*, *Remove*, *Replace*, and *AddResize* — and three mutates for the fragment-based loop construction — *FragAdd*, *FragDel*, and *FragReplace* (figure V.1 on page 69). The mutate *Add* randomly selects a missing loop in the protein model, looks up the set of suitable conformations in the hash map, and inserts a randomly selected suitable conformation. The mutate *Remove* removes a randomly selected existing loop from the protein model. The mutate *Replace* replaces a randomly selected existing loop in the protein model with another randomly selected suitable conformation. The mutate *AddResize* cuts back one or both anchor SSEs by one to three residues before applying the mutate *Add* to the resized loop. The result of each mutate is evaluated through the weighted sum of previously published knowledge-based scoring terms evaluating steric interference between residues, agreement of the dihedral angles with Ramachandran's distribution, and residue-residue interactions.[60] To counteract uncontrolled cutting back of SSEs by the *AddResize* mutate, scoring terms were added to evaluate the agreement of the protein model with the secondary structure prediction methods PSIPRED,[98,198] Jufo9D,[97] and MASP.[99] To support loop closure, an additional scoring term penalizing protein models with missing loops was also introduced. This scoring term adds a penalty score that is linearly based on the number of missing loop residues and is weighted to contribute about 30 % to the total score. The fragment-based construction of loops is detailed in the following sections.

In each MCM cycle, one randomly selected mutate is applied to the protein model, the result is scored, and the changed protein model is either accepted or rejected. We evaluated different cycle lengths to find the optimal balance between closure rate and time requirement and found that the following termination criteria provided the best results (see the results section for details): for the conformational hashing algorithm, the maximum number of iterations per protein model was set to 500 and the algorithm terminated early if no score improvement was found for 50 iterations in a row. For the CCD algorithm, setting the maximum number of iterations per protein to 2000 and the early termination criterion to 500 iterations provided the best results.

V.2.4. Construction of missing loop regions using cyclic coordinate descent

The CCD algorithm was used to construct loop regions or parts of loop regions that could not be constructed using the conformational hashing algorithm — *i.e.*, if the conformational hashing algorithm only constructed part of loop, the CCD algorithm was employed to close the remaining gap. Its implementation was inspired by an algorithm published previously by Canutescu *et al.*,[101] which is based on the random tweak algorithm.[199,200] The CCD implementation employed in this study used a pre-stage and a main-stage approach for the conformation sampling.

During the pre-stage, an algorithm dynamically adds missing residues and samples the (ϕ, ψ) backbone angles. Residues are added with initial (ϕ, ψ) values that are derived from a probability distribution of experimentally observed backbone dihedral angles. The (ϕ, ψ) angles are subsequently perturbed and evaluated using a knowledge-based potential. The residues are added to both the C-terminus of the N-terminal anchor SSE and the N-terminus of the C-terminal anchor SSE. To account for potential inaccuracies in the secondary structure prediction, residues can be added or removed from the anchor SSEs. The sampling of this pre-stage is guided by scoring terms evaluating the completeness of the amino acid sequence, steric interference between residues, residue-residue interactions, and the loop trajectory towards its anchor point. This module was also employed to construct the terminal loop regions of the protein models.

The main-stage is based on the previously published CCD algorithm by Canutescu *et al.*.[101] CCD calculates the rotation that must occur around a given axis (ϕ or ψ) to minimize the distance between a moving (loop end) and target set of coordinates over many iterations to close a chain break. In order to help prevent getting stuck in a non-closable conformation based on the starting loop conformation, the algorithm was extended to select the residue and rotation axis (ϕ or ψ) randomly at every step. In addition, only a random fraction of the rotation is applied. This protocol can be performed in the presence or absence of scoring functions. We used this approach in conjunction with scoring functions to evaluate residue-residue interactions and steric interferences between residues.

V.2.5. The benchmark set used to evaluate the algorithm

To evaluate the performance of the algorithm, we selected a benchmark set consisting of twenty-nine soluble and membrane proteins (table V.1 on the next page). This benchmark set consisted of a set of soluble proteins used previously by Tyka *et al.* to benchmark the Rosetta conformational hashing algorithm[192] and was extended by eleven membrane proteins that were used to benchmark the protein structure prediction algorithm BCL::MP-Fold.[2] The proteins in the benchmark set ranged in size from 57 to 1560 residues with varying α-helical and β-strand secondary structure content. The proteins contained 296 non-terminal loop regions with lengths up to 41 residues. For the benchmark proteins, secondary structure definitions were obtained using DSSP[113] and loop regions were removed accordingly from the experimentally determined structures collected from the PDB. Almost two thirds of the proteins in the benchmark set either had missing coordinates in terminal loop regions or very short terminal loop regions. Due to the lack of reference points for those loops, we opted to exclude them from the evaluation except for checking whether an energetically unforbidden conformation has been constructed.

The benchmark was not performed using the original template library; instead all proteins with a sequence identity greater or equals 25 % to any protein in the benchmark set were removed and a new template library was generated. For heterooligomers, only the chains fulfilling this criterion were removed from the initial template set. Using this criterion, about 3000 protein structures were completely or partially removed from the initial set consisting of about 87 000 protein structures. Subsequently, a new template library was created from the remaining proteins and used in the benchmark that is detailed in the following sections.

Soluble proteins

Protein	Length	#Loops
1A32	88	3
1ACF	125	10
1BK2	57	4
1BKR	109	4
1ELW	252	12
1OPD	85	6
1PGX	83	4
1R69	69	4
1TTZ	87	6
1UBI	76	5
1VCC	77	5
1VKK	154	11
1WDV	304	24
2ACY	98	6
2CHF	128	9
2H28	260	12
2HE4	90	7
2ICP	94	4

Membrane proteins

Protein	Length	#Loops
1J4N	271	8
1OKC	297	10
1PY6	498	20
2BL2	1560	56
2IC8	182	8
2K73	183	7
2KSF	107	3
2KSY	247	10
2NR9	196	7
3GIA	444	17
3P5N	378	14

Table V.1.: The protein structures used to benchmark the loop construction algorithm. *Twenty-nine protein structures with different sequence lengths and a total number of* 296 *non-terminal loops were used to evaluate the performance of the algorithm.*

V.2.6. Compensating for the lack of templates for long loops through fragment combination

From the initial set of about 87 000 protein structures about 3.7 million loop templates were collected. Of those, about 2.2 million loop templates had a sequence length of at least four residues (figure V.1 on page 69) and therefore their conformation was not overdetermined. However, only 12 % of the templates had a sequence length of ten or more residues, which posed a substantial problem, since longer loops can cover a larger conformational space and are therefore less likely to be closed by the conformational hashing algorithm using a template library that only has a limited number of conformations for such loops. To compensate for the lack of templates that are suitable for long loops, we developed a two-pronged approach.

First, we developed an algorithm to combine short loop templates to longer loop templates. This is performed by superimposing the backbone coordinates of the C-terminal anchor point of the first template with the backbone coordinates of the N-terminal anchor point of the second template. Accordingly, the resulting template has a sequence length of $d = d_1 + d_2 + 1$, with d_1 and d_2 being the sequence lengths of the original templates. The remaining parameters, the translation vector $\boldsymbol{t}$ and the Euler angles (α, β, γ) could be computed by transforming the local coordinate system of the N-terminal anchor point of the second template into the local coordinate system of the N-terminal anchor point of the first template, which can be achieved by multiplying the rotation matrix of the first coordinate system with the inverse rotation matrix of the second coordinate system $\boldsymbol{M} = \boldsymbol{M}_2^{-1} \cdot \boldsymbol{M}_1$. Multiplying the translation vector $\boldsymbol{t}_2$ of the second template with this matrix will allow for simple vector addition to compute the resulting translation vector $\boldsymbol{t} = \boldsymbol{t}_1 + (\boldsymbol{t}_2 \cdot \boldsymbol{M})$. Although this approach works in theory, we experienced numerical inaccuracies that became significant when constructed templates were used to construct additional templates. Instead, we opted for a slower but more accurate approach. In this approach, we combined the stored dihedral angles for both templates into a sequence consisting of $2 \cdot (d_1 + d_2 + 2)$ dihedral angles: $(\psi_{n,1}, \phi_{1,1}, \psi_{1,1}, \dots, \phi_{1,d}, \psi_{1,d}, \phi_{1,c}, \psi_{n,2}, \phi_{2,1}, \psi_{2,1}, \dots, \phi_{2,d}, \psi_{2,d}, \phi_{2,c})$. An artificial amino acid sequence of length $d + 2$ was generated within the algorithm and fitted against the combined sequence of dihedral angles, which allowed for accurate computation of the parameterization for this sequence.

Second, we added mutates to enable a step-wise construction of loops within the MCM framework. If a suitable loop with sequence length d is missing in the template library, the *AddFrag* mutate randomly selects a loop conformation with a sequence length less than $d - 2$ from the template library and fits the terminal part of the loop against the selected template. This approach required additional scoring terms to prevent the algorithm from sampling incomplete loop conformations that cannot possibly be closed. To address this problem, we added two previously described scoring terms, `loop` and `loop_closure`.[60] Briefly, the scoring term `loop` quantifies the likelihood of closing a gap with a certain Euclidean distance given a certain sequence distance and the scoring term `loop_closure` evaluates if a fully extended loop with the given sequence length would be able to bridge the gap. To achieve a sufficient coverage of the conformational space, we also added mutates that are inverse to the mutate *AddFrag*. The mutate *DelFrag* removes a randomly chosen loop fragment from the model and the mutate *ReplaceFrag* replaces a randomly selected fragment in the protein model with a randomly selected fragment from the template library.

V.2.7. Comparison with Rosetta Loophash

The loophash protocol of Tyka *et al.*[192] was employed as a comparison method, though slight changes to the protocol were needed due to the change in focus from loop diversification to loop construction. Most notably, the "relax" stage of the procedure was omitted, and only the loop resampling plus minimization-based closure stages were performed. As loophash assumes "ideal" bond length and angles, all input structures were passed through the idealize application of Rosetta prior to use. The same set of 85 808 homology-culled PISCES structures were used to create a database for loophash, though all portions of the protein were included in the loophash database, not just the DSSP-defined loops. The 296 non-terminal loops present in the benchmark set were each tested individually, in the context of the full experimentally determined structure of the remaining loops, with parameters set to skip RMSD-based filtering. A total of 100 output structures were sampled for each loop in the benchmark set, and each output structure represents the result of 100 randomly selected database loops which match the required geometry. Loophash was always able to produce samples, but due to the substitution and minimization approach used, loophash may significantly perturb the protein structure, even in regions outside the loop. For this reason, we counted as "closure failure" those loops for which none of the 100 output structures were within 1 Å whole structure C_α-RMSD from the input structure. Runtimes for loophash are reported as an average for a single output structure, and include only the time spent actively sampling the loop.

V.3. Results

In this section, we present the distribution of templates collected from structures in the PDB and rationalize the need for template recombination. Following this, we report the performance of the conformational hashing algorithm in terms of closure rate and time requirement. Additionally, the

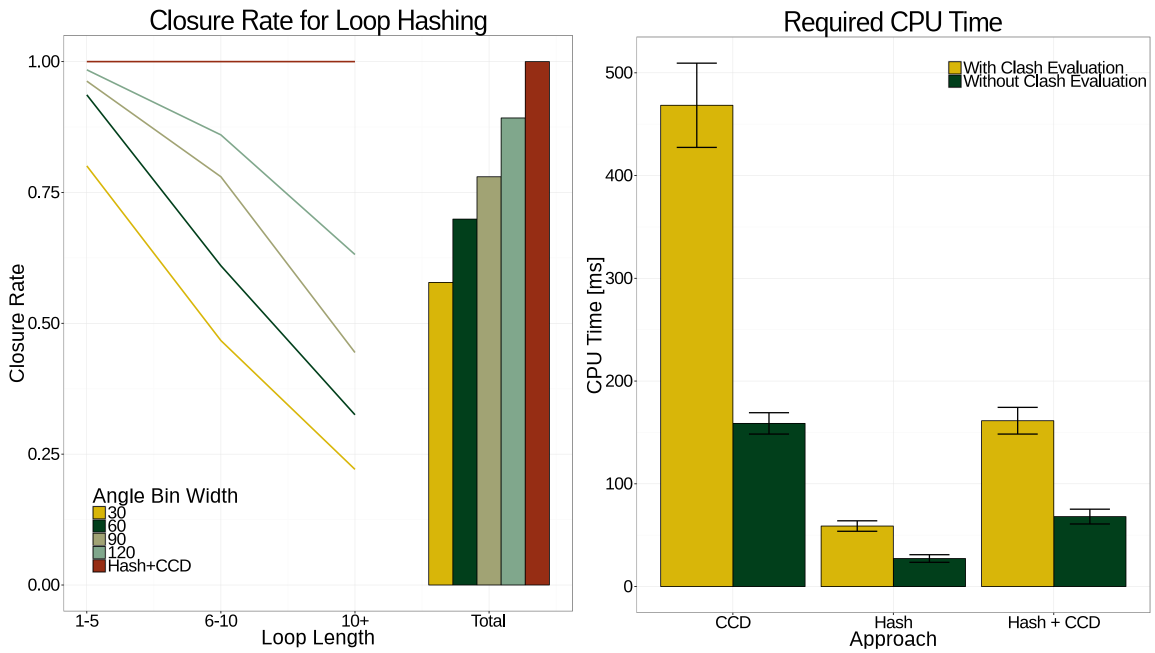

Figure V.2.: Performance of the conformation hashing and CCD algorithms for loop construction. *(A) The closure rate of the conformational hashing algorithm depended on the rotation angle bin width and the length of the loops. High closure rates were achieved for short loops. Combining conformational hashing with CCD improved the closure rate to 100 % for the given benchmark set. (B) Evaluation of steric interferences dominated the CPU time requirement per loop. A combination of conformational hashing and CCD maintained the high closure rate of CCD while improving CPU time efficiency.*

performance in dependence of different optimization parameters are reported. This section is concluded by the results for combining conformational hashing with CCD. Where applicable, a comparison to Rosetta's loophash is provided.

V.3.1. The closure rate of conformational hashing strongly depends on the bin size for the parameterization and the loop length

Before computing the hash key for a loop template, the loop's parametrization needs to be discretized, which is achieved through binning (see section V.2.2 on page 70 details). The bin width directly determines how the hash map is populated and therefore is expected to have significant influence on the closure rate and the physical reasonability of the constructed loop regions. Whereas larger bin widths result in a denser population of the hash map, the constructed loop regions can be physically unreasonable due to unstable bond lengths or bond angles. Smaller bin widths on the other hand result in a sparser population of the hash map leading to physically stable bond lengths and bond angles at the expense of a lower closure rate. To evaluate the influence of the bin width on the closure rate, we repeated the loop construction for the benchmark set with different bin widths. The rotation bin width was increased in 30° steps from 30° to 120°. The translation bin width was kept at 1 Å because larger translation bin widths would require post-processing of the fitted loop to avoid overextension of the peptide bonds. Although the post-processing could be achieved using gradient minimization or similar structure optimization methods, we opted for exclusion of post-processing to avoid additional computational burden on the algorithm.

Overall, for all four angle bin widths, the closure rate decreased approximately linearly with the length of the loop. For the selected bin width of 60° and the extended loop library, the closure rate for loops up to five residues was 94 %, for loops with a length between six and ten residues, it was 61 %, and for loops with a length of more than ten residues, the closure rate dropped to 33 %, resulting in a total closure rate of 70 % (figure V.2 on the preceding page). A similar, almost linear relation between closure rate and loop length was also observed for the other three evaluated angle bin widths.

The total closure rate was strongly dependent on the angle bin width. For the evaluated angle bin widths of 30°, 60°, 90° and 120°, the total closure rate of the conformational hashing algorithm arrived at 58 %, 70 %, 78 % and 89 % (figure V.2 on the previous page). Although these results might suggest using a larger angle bin width, this approach would also be prone to producing unnatural angles of the peptide bonds connecting the anchor SSEs and the loop. This problem could again be mitigated by applying gradient minimization or similar structure optimization methods as post-processing, but would increase the required computation time. Instead, we opted to use an angle bin width of 60°, which provided a compromise between closure rate and reasonable bond angles.

V.3.2. Conformational hashing achieves a high closure rate for short loops but CCD is required for long loops

To evaluate the closure rate of the algorithm — what percentage of the missing loop regions can be successfully constructed — we removed the non-terminal loop regions from each protein structure in the benchmark set (see table V.1 on page 74 for a list of the benchmark proteins). The resulting structural models were used as input for the described algorithm (see materials and methods for details). For each input protein structure, the algorithm constructed the missing loop regions and the closure rate per benchmark protein was computed. We performed the experiment twice: once with the pruned template library collected from the about 84 000 PDB structures and once with the template library that was extended using template recombination and fragment-based loop construction (see materials and methods for details).

Using the original template library with an angle bin width of 60° and without additional templates through fragment combination, the algorithm achieved a closure rate of 54 % over all twenty-nine benchmark proteins. The closure rate strongly depended on the sequence length of the loops, which was demonstrated by a closure rate of 87 % for loops with a maximum sequence length of five residues and a closure rate of 21 % for loops with a sequence length greater than ten residues. Using the extended template library in conjunction with fragment-based loop construction, the overall closure rate could be improved to 70 % (figure V.2 on the previous page). Further increase of the template library through fragment combination or increasing the number of MC steps did not improve the closure rate in any significant manner. However, by combining the conformational hashing algorithm with CCD, as described in figure V.1 on page 69, the closure rate could be improved to 100 % on the given benchmark set (figure V.2 on the previous page). While Rosetta loophash is able to achieve a 99 % closure rate on the loops of length ten residues or less, it is only able to successfully close 39 % of the loops eleven residues or longer. The high closure rate of Rosetta for loops up to a length of ten residues is achieved in part by using a wider bin width of the translation vector as compared to the implementation described in this manuscript, 2 Å versus 1 Å, which consequently results in the need of structure optimization and refinement following the fitting. This additional computational effort is reflected in a substantially higher CPU time requirement as described in section V.3.3 on the following page.

V.3.3. CPU time requirement is dominated by the evaluation of steric interferences

The time complexity of the template look-up is within $O(1)$, whereas the fitting of the target sequence against the template is within $O(n)$, resulting in linear time complexity for the overall loop construction algorithm. The absolute time requirement to construct loop regions was evaluated by measuring the time between entering the MCM algorithm and leaving the MCM algorithm (see figure V.1 on page 69 for details) divided by the number of successfully constructed loop regions. To evaluate the contribution of the scoring term evaluating steric interference between residues relative to the overall CPU time requirement, the computation time needed by this term was evaluated separately. This experiment was repeated for the CCD algorithm and the combination of conformational hashing and CCD.

Excluding the time required to evaluate steric interferences, conformational hashing required on average (27 ± 4) ms CPU time per loop, whereas CCD required (159 ± 11) ms and a combination of conformational hashing and CCD required (68 ± 7) ms (figure V.2 on page 76). Including the time required to evaluate steric interferences, the requiredglscpu time increased to (59 ± 5) ms for conformational hashing, (468 ± 41) ms for CCD, and (161 ± 13) ms for the combination of conformational hashing and CCD (figure V.2 on page 76). For the given benchmark set and the three different algorithmic approaches, the evaluation of steric interferences dominated the computational burden by accounting for 54 %, 66 % and 58 % of the total CPU time requirement, respectively.

In contrast, the Rosetta loophash procedure requires on average a CPU time of 160 s to sample each loop. This runtime is highly correlated with total protein size ($R^2 = 0.8$). The longer runtime of loophash is driven primarily by the minimization-based approach used, with 95 % of the runtime being devoted to minimization.

V.3.4. The algorithm samples the conformation of a protein's major population in most cases

The algorithm was developed to achieve efficient sampling of a protein's major and minor populations in the equilibrium. Whereas we cannot validate minor populations, in this study we assume that the experimentally determined structures deposited in the PDB correctly and accurately represent one of the proteins' major populations. To test whether the proposed loop construction algorithm correctly samples those conformations, we used the protocol described in the materials and methods section to sample 100 models with constructed loop regions for each protein in the benchmark set. The loop conformations of the sampled models were subsequently compared to the conformations of the experimentally determined structures using the RMSD of the C_α-atoms. For the membrane protein structure 3P5N, two non-terminal loops were not resolved in the X-ray-derived model and consequently excluded from this comparison. Of the remaining 294 non-terminal loops in the benchmark set, 94 % could be sampled with a C_α-RMSD less than 2 Å relative to the experimentally determined reference structure (examples and distribution shown in figure V.3 on the following page). The remaining eleven loops all had a sequence length greater than ten residues (example shown in figure V.3 on the next page). In contrast, Rosetta loophash was only able to sample a native-like loop conformation in 76 % of the cases, with forty failures in loops of sequence length ten or less, and thirty-two failures (89 %) in loops greater than ten residues.

V.4. Discussion

In this section, we discuss the applicability of conformation hashing to the loop construction problem and why this approach needs to be complemented with a template-independent approach (for details,

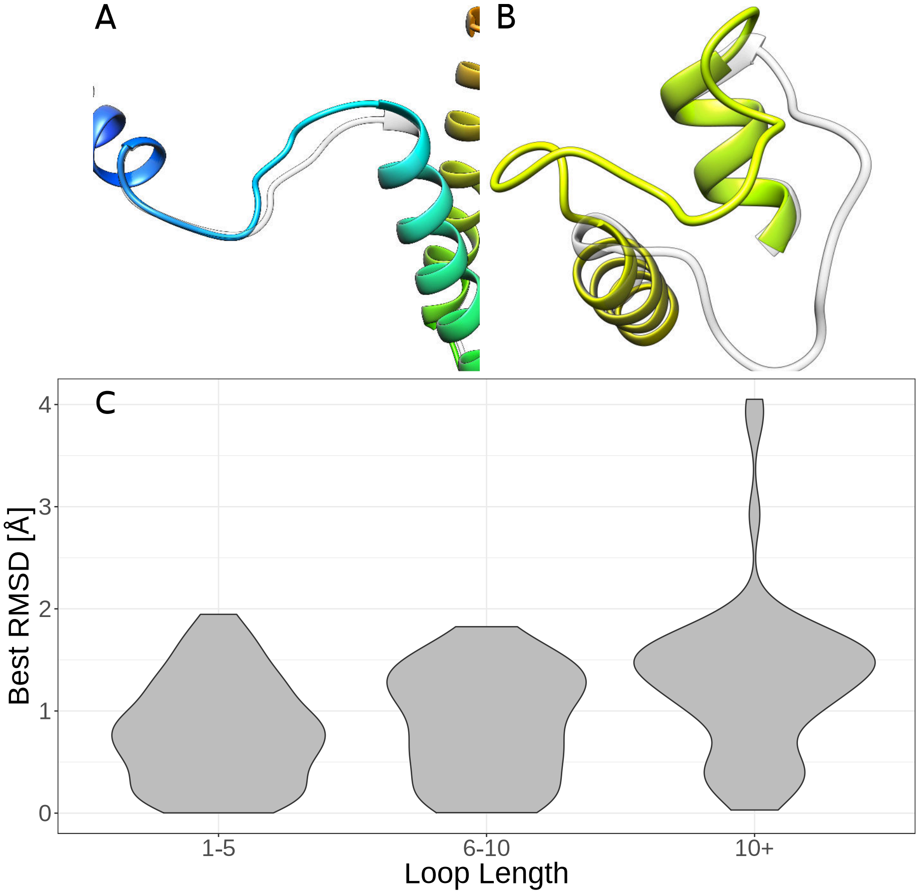

Figure V.3.: Examples for structural similarity of constructed loops to experimentally determined major populations. *For more than 96 % of all non-terminal loops, a conformation structurally similar to the major population could be sampled (the example shown in panel A depicts parts of 1A32). For some long loops, the major population could not be sampled (the example shown in panel B depicts parts of 3GIA). The sampled models are rainbow-colored, whereas the experimentally determined structures are shown in transparent gray). The violin plot (C) details the accuracy distribution of the most accurate models depending on the loop length.*

see section V.4.1). Additionally, we discuss extensions to the algorithm to compensate for the lack of templates for long loops (section V.4.2).

V.4.1. Conformational hashing needs to be complemented with a template-independent algorithm

Employing conformational hashing to construct loop regions faces the major problem that most loop regions in structures deposited in the PDB have a sequence lengths less than ten residues. In our initial template library set gathered from about 87 000 protein structures, 88 % of all loops lay below the ten-residue threshold (figure V.1 on page 69). Additionally, longer loops can cover a larger conformational space, furtherly impeding construction of long loops using conformational hashing. The problem is demonstrated by substantial discrepancies between the closure rates for different loop lengths: whereas for loops with up to five residues a closure rate of 96 % could be achieved, the closure rate dropped to 61 % for loops with a sequence length between six and ten residues, and finally to 33 % for loops with more than ten residues (figure V.2 on page 76). This problem could only be mitigated but not solved by using template combination and fragment-based loop construction (see materials and methods for details), which was demonstrated by an improvement of the total closure rate from 54 % to 70 %. These results suggest that conformational hashing needs to be complemented with a template-independent loop construction algorithm. In this study, we applied a CCD implementation to the loop regions or parts of loop regions that could not be constructed by the conformational hashing algorithm, which compensated for the latter's inability to close long loops and was demonstrated by an improvement of the closure rate to 100 %. Consequently, a stand-alone conformational hashing approach can be sufficient to construct very short loop regions but for an average protein, it needs to be complemented with template-independent loop construction algorithms that are more suitable for long loops.

V.4.2. Combining conformational hashing with CCD provides an efficient way to sample structurally diverse loop ensembles

Prediction of structural ensembles for proteins requires algorithms to facilitate efficient sampling of diverse conformations to capture major and minor populations of the protein in the equilibrium. Although CCD achieves a high closure rate, its relatively high demand for CPU time requires a more efficient approach. The limited closure rate of the conformational hashing algorithm for long loops demonstrated that CCD and other template-independent loop construction programs cannot be entirely replaced by the more CPU-efficient conformational hashing approaches. However, constructing as many loops as possible using conformational hashing and only construct the remaining, missing loops or parts of the loops with CCD resulted in a significant reduction of the required CPU time. This was demonstrated by a drop of required CPU time from 468 ms, when using CCD only, to 161 ms, when using a combination of conformational hashing and CCD (figure V.2 on page 76). This demonstrates that a combination of conformational hashing and CCD can result in a reasonable consensus between the efficiency of conformational hashing and the high closure rate of CCD. The reduction of the required CPU time will allow for sampling more conformations with the same amount of computational resources available, therefore allowing to sample a wider range of possible loop conformations (see figure V.4 on the next page for examples). However, it needs to be noted that for about 6 % of the loops in the benchmark set, the conformation of the experimentally determined structure was not sampled (see figure V.3 on the preceding page for an example). All these loops had a sequence length greater than ten residues resulting in a larger conformational space. This problem might be solved through incorporation of limited experimental data from technique like electron paramagnetic resonance spectroscopy or

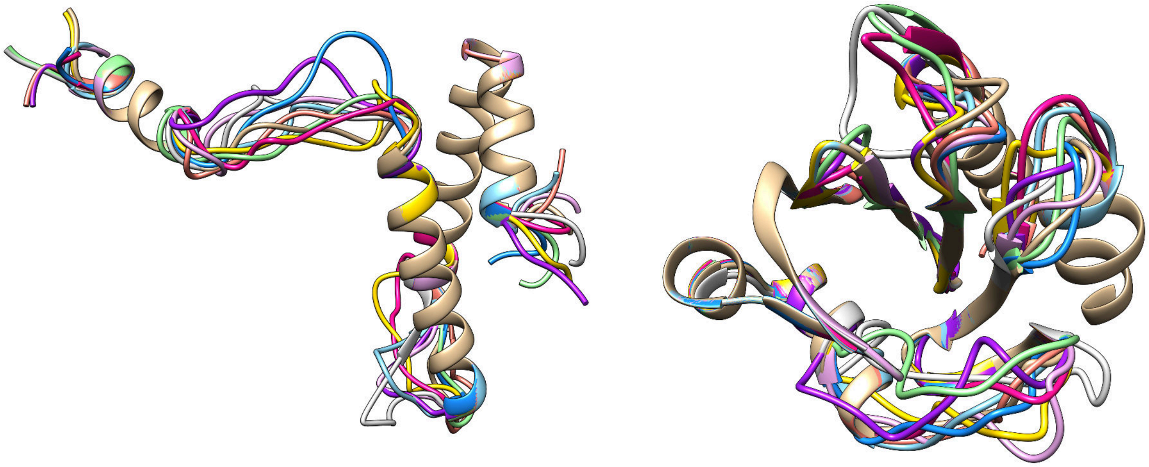

Figure V.4.: Examples for predicted loop ensembles using conformational hashing in conjunction with CCD. *Loop ensembles for the protein structures 1A32 (left) and 1ACF (right) were predicted using a combination of conformational hashing and CCD.*

other sources that can be used to limit the size of the allowed conformational space and should be evaluated in future studies.

V.5. Conclusion

The proposed loop construction algorithm consisting of conformational hashing complemented by a template-independent approach like CCD provides an efficient way to sample a structurally diverse ensemble of loop conformations that is significantly faster than using the template-independent approach alone. Under the assumption that the loop conformations in the experimentally determined structures deposited in the PDB represent a major population in the respective protein's equilibrium, we conclude that the proposed algorithm is able to sample the major populations in the vast majority of all cases. These results indicate, that the algorithmic approach described in this study could be used for the prediction of protein ensembles, for which a structurally diverse set of conformations will be fitted against experimental data to determine a weighted ensemble that represents the equilibrium constitution of the protein in question.

As additional benefit, the constructed loop regions largely exhibit naturally occurring dihedral angles due to the construction from conformations observed in the PDB. Due to a smaller number of elucidated conformations for long loops in the PDB, the conformational hashing approach cannot be used on its own for the construction of long loops but needs to be complemented with a template-independent approach.

V.6. Acknowledgments

We want to thank Marion F. Sauer for thorough proofreading of this manuscript. Work in the Meiler laboratory is supported through NIH (R01 GM080403 and R01 GM073151). Parts of the data analysis were performed using R[c] in conjunction with the ggplot2 package.[123] Figures depicting protein models

[c] https://www.r-project.org

were created using Chimera[124] and composite figures were created using Inkscape.[d] To determine sequence identities between the template set and the benchmark set, the sequences were aligned using Clustal Omega.[201]

[d] https://inkscape.org

CHAPTER VI
PREDICTING THE MONOMERIC AND HOMODIMERIC FORMS OF BAX

This chapter is based on the publication "Pushing the size limit of de novo structure ensemble prediction guided by sparse SDSL-EPR restraints to 200 residues: The monomeric and homodimeric forms of BAX".[5] Axel W. Fischer contributed to the development of the prediction pipeline, performing the experiment, analyzing the data, and writing the article.

Structure determination remains a challenge for many biologically important proteins. In particular, proteins that adopt multiple conformations often evade crystallization in all biologically relevant states. Although computational de novo *protein folding approaches often sample biologically relevant conformations, the selection of the most accurate model for different functional states remains a formidable challenge, in particular, for proteins with more than about* 150 *residues. EPR spectroscopy can obtain limited structural information for proteins in well-defined biological states and thereby assist in selecting biologically relevant conformations. The present study demonstrates that* de novo *folding methods are able to accurately sample the folds of* 192*-residue long soluble monomeric BAX. The tertiary structures of the monomeric and homodimeric forms of BAX were predicted using the primary structure as well as twenty-five and eleven EPR distance restraints, respectively. The predicted models were subsequently compared to respective NMR/X-ray structures of BAX. EPR restraints improve the RMSD100 of the most accurate models with respect to the NMR/crystal structure from* 5.9 Å *to* 3.9 Å *and from* 5.7 Å *to* 3.3 Å, *respectively. Additionally, the model discrimination is improved, which is demonstrated by an improvement of the enrichment from* 5 % *to* 15 % *and from* 13 % *to* 21 %, *respectively.*

VI.1. Introduction

Proteins undergo conformational changes while performing their biological function. Although X-ray crystallography provides snapshots of important conformations, often not all biologically relevant conformations can be crystallized. NMR spectroscopy, the premier method to study protein dynamics at atomic detail, suffers from a size limit that complicates a detailed analysis of larger proteins. EPR spectroscopy in conjunction with SDSL offers an alternative approach to study protein structure and dynamics. Briefly, typically two cysteine residues are introduced into a cys-less variant of the protein and coupled with *S*-(1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate (MTSL), which carries an unpaired electron. The dipolar interaction of the two unpaired electrons is inversely proportional to the cubed distance and can be measured with high sensitivity with a pulsed dipolar spectroscopy technique called DEER or pulsed electron paramagnetic resonance (PELDOR).[75,150] As for every distance measurement a dedicated protein double-mutant needs to be created and tested for functional viability, data obtained from SDSL-EPR measurements are sparse. Thus, such data typically fail to unambiguously determine the structure of a protein at atomic detail. However, it has been demonstrated that in conjunction with *de novo* protein structure prediction algorithms determination of a protein's fold might be within reach.[2,78,149] Whereas previous studies were performed on smaller proteins[78] or mainly based on simulated SDSL-EPR restraints,[2] this study evaluates the impact of experimental SDSL-EPR distance restraints on *de novo* protein structure prediction for larger proteins that adopt multiple biologically relevant conformations.

The major challenges of *de novo* protein structure prediction are the vast size of the conformational space that needs to be sampled as well as the discrimination of inaccurate models, *i.e.*, the identification of low-energy, biologically relevant states of a protein with a simplified energy function. The simplified macromolecule representations used in *de novo* folding simulations prohibit computation of accurate free energy differences between different conformations. Instead, the approach employed in this study uses knowledge-based energy functions to determine the likelihood of proposed protein models.[60] In parallel, SDSL-EPR distance restraints restrict the sampling space to conformations that are in agreement with the SDSL-EPR data,[202] thus increasing the frequency with which models in agreement with the SDSL-EPR data are sampled. Through incorporation into the scoring function, SDSL-EPR distance restraints also improve the discrimination of inaccurate models. Studying soluble monomeric and homodimeric BAX in this context is especially intriguing due to the large size of the protein and the availability of high-quality experimental SDSL-EPR data sets.

BAX plays a central role in the apoptotic cell death, which is fundamental to the survival of mammals and related to various diseases. Whereas unwanted apoptosis is seen as cause for ischemia and Alzheimer's disease,[203] failure of apoptosis is a key step in developing tumors and autoimmune diseases.[204–207] As many different signals for cell death converge on mitochondrial outer membrane (MOM) permeabilization, a better understanding of this mechanism is pivotal for the treatment of diseases related to the apoptotic process.[208] MOM permeabilization is controlled by members of the Bcl-2 family, and the pro-apoptotic protein BAX is described to execute it.[209] In a healthy cell, BAX is a monomeric, cytosolic protein, whose structure was determined by NMR spectroscopy.[210] Upon pro-apoptotic stimuli, BAX inserts into the MOM, oligomerizes, and creates pores.[208,209] Through the pores, cytochrome *c* and other pro-apoptotic proteins are released into the cytosol, initiating a proteolytic cascade leading to cell death. The structure of the membrane-embedded active BAX remains elusive. However, three recent publications have provided valuable new structural insights.[202,211,212]

Here we apply the BCL::Fold[59] algorithm, which is part of the BCL, to predict the tertiary structure of soluble monomeric BAX and of the dimerization domain of membrane-embedded BAX oligomers. For the solution structure of BAX (PDB entry 1F16) and the BAX BH3-in-groove dimer (PDB entry 4BDU), high-resolution structures are published[210,211] and a number of SDSL-EPR measurements exist.[202] Therefore, this study represents a benchmark test if SDSL-EPR data are sufficient to determine the structure of biologically important states of large, membrane-associated proteins. BCL::Fold is tailored towards assembly of large protein structures from predicted SSEs.[59,86] In a first step, the tertiary structure of soluble monomeric BAX was predicted from twenty-five SDSL-EPR distance restraints,[202] demonstrating the feasibility of the protocol as well as the influence of the limited SDSL-EPR data on *de novo* protein structure prediction. In a second step, the tertiary structure of the dimerization domain of homodimeric BAX (α-helices 2 to 5) was predicted from eleven SDSL-EPR distance restraints,[202] demonstrating the applicability of the protocol to oligomeric proteins. In both cases, usage of SDSL-EPR distance restraints significantly improved the accuracy of the sampled models as well as the accuracy with which the models in best agreement with the NMR- and X-ray-derived models could be selected.

VI.2. Materials and methods

The tertiary structures of soluble monomeric and homodimeric BAX were predicted using the previously published BCL::Fold[59] algorithm, which is part of the BCL.[a] A summary of the structure prediction protocol is given in the following section, followed by a section describing how SDSL-EPR distances were translated into structural restraints. The accuracy of the predictions was evaluated by computing a protein-size normalized root-mean-square-deviation of the backbone coordinates.[103] Further, we compute the enrichment metric,[2] which quantifies how well the employed scoring function is able to distinguish accurate models from inaccurate models.

VI.2.1. Structure prediction protocol

The protocol used to predict the tertiary structure of soluble monomeric BAX and homodimeric BAX is based on the BCL::Fold protocol for soluble proteins.[59] As in the original protocol, a pool containing the SSEs was predicted from the primary structure using the secondary structure prediction algorithms PSIPRED[98] and Jufo9D[97] (see section E.2.3 on page 188). BCL::Fold subsequently uses a MC sampling algorithm to assemble the predicted SSEs in the three-dimensional space. BCL::Fold uses the MC sampling algorithm in conjunction with the Metropolis criterion for energy minimization to search the conformational space for models with a likely overall fold (see section E.2.4 on page 189).[59] After each MC step, models are scored using knowledge-based potentials evaluating different scoring terms like SSE packing, radius of gyration, amino acid exposure, amino acid interactions, loop closure geometry, secondary structure length and content, as well as penalizing potentials for SSE and amino acid clashes.[60] The potential functions for each scoring term were derived from statistics over protein structures deposited in the PDB using the inverse Boltzmann relation[60] as described in equation (VI.1).

$$E = -RT \cdot \ln \frac{P_\mathrm{o}}{P_\mathrm{b}} \tag{VI.1}$$

where:

E = free energy of the protein structure
P_o = probability of observing a specific feature
P_b = probability of observing a specific feature by chance
R = gas constant
T = temperature

For each scoring term, the probability of observing a specific feature (P_o) was computed from statistics derived from structures deposited in the PDB. This probability is normalized by the probability of observing this feature by chance (P_b). This normalization ensures that favorable features are assigned negative scores. The term RT is set to 1 for convenience.[60] For example, one scoring term (S_NC) evaluates the burial of residues. The degree of burial was quantified using the neighbor count metric,[96] which assigns a non-negative number — the neighbor count value — to each residue. For each amino acid type, statistics over the neighbor count distributions were collected from structures deposited in the PDB. The distributions were binned and the probability of each bin (P_o) was computed.[60] After normalization by P_b, the inverse Boltzmann relation can be used to compute S_NC for each residue in the sampled models. The total score of a protein model — the BCL score — is the weighted sum of

[a] http://www.meilerlab.org/bclcommons

the different scoring terms.[60] Additional scoring terms based on the CONE model[78,149] were used to quantify the agreement of the sampled models with the available SDSL-EPR data.

The folding simulation is broken down into five assembly stages. Each stage lasts for a maximum of 2000 MC steps but is terminated early if a maximum of 400 MC steps without score improvement in a row is reached. The assembly stages consist of large-scale sampling moves like adding or removing SSEs, flipping and swapping SSEs, as well as large-scale translations and rotations. Over the course of the five assembly stages, the weights for the potentials penalizing SSE and amino acid clashes ramp up to 0, 125, 250, 375 and 500. The weight for scoring the agreement of the model with the SDSL-EPR data remains constant at 50 over all stages. As a result, agreement with SDSL-EPR distance restraints contributes about 45 % to the total score, if provided. For previous benchmark studies, various weights for the SDSL-EPR agreement score were evaluated and a weight of 50, which equates to a contribution of 40 % to 50 % to the total score, provided the best prediction results.[2]

After the assembly stages the model is refined. This process is encapsulated in one stage that consists of small structural perturbations like low-amplitude translations and rotations of SSEs. This stage does not change the overall topology drastically. This stage lasts for a maximum of 2000 MC steps but is terminated early if a maximum of 400 MC steps without score improvement in a row are reached. During the refinement stage, the weight for the SDSL-EPR score remains at 50. For homodimeric BAX, the protein structure prediction protocol was slightly altered to assemble and refine the models in C2-symmetry mode.[50]

VI.2.2. Translating SDSL-EPR distances into structural restraints

Through the DEER/PELDOR experiment, SDSL-EPR spectroscopy measures the distance between two unpaired electrons located in the N−O group of spin labels (D_{SL}) that are covalently attached to cysteines in the protein. The DEER experiment consists of microwave pulses at two different frequencies used to measure the dipolar coupling between two electron spins. The pulse sequence at the observer frequency produces an echo. The pulse at the pump frequency flips the coupled spin, thus changing the local field at the observer spin by the dipole-dipole coupling. Variation of the pump pulse delay leads to modulation of the intensity of the refocused echo. The periodicity is a function of the distance-dependent coupling between the spin labels.[213]

For effective usage of the SDSL-EPR data in a *de novo* structure prediction algorithm that relies on a backbone-only protein model, those distances need to be translated into possible distance restraints for the closest atoms represented in the model, which in our case are the distances between the C_β-atoms of the spin labeling sites (D_{BB}). In the case of glycine, which lacks a C_β-atom, the $H_{\alpha2}$-atom is used instead. The side chain flexibility of the spin label prevents an unambiguous translation from D_{SL} into D_{BB} due to its unknown conformation on the protein. Additionally, the SDSL-EPR experiment is conducted on a double-cysteine mutant protein to which spin labels have been covalently bounded — a species that is distinct from the wild type protein and might have a different structure and dynamics. Lastly, the SDSL-EPR experiment itself and the fitting procedures used to translate the primary DEER data into a distance distribution are accompanied by uncertainties. To quantify the agreement of D_{SL} with D_{BB} a knowledge-based potential based on the CONE model was introduced.[78,149] The scoring function scores $D_{SL} - D_{BB}$ ranges of −12.5 Å to 12.5 Å, which covers the minimum and maximum difference between D_{SL} and D_{BB}.[78,149] It assigns a score ranging from 0 (no agreement) to −1 (optimal agreement) to each $D_{SL} - D_{BB}$ pair in a protein model. An additional scoring function is used to penalize conformations with $D_{\mathrm{SL}} - D_{\mathrm{BB}}$ differences less than −12.5 Å or greater than 12.5 Å with the purpose of drawing restraints into the −12.5 Å to 12.5 Å range.[2]

VI.2.3. Benchmark setup

To evaluate the influence of SDSL-EPR-derived structural restraints on *de novo* protein structure prediction, multiple folding simulations were performed. In a first experiment, the conformational space of soluble monomeric BAX was sampled in the absence of SDSL-EPR restraints. Therefore, the above-mentioned structure prediction protocol was altered so that the SDSL-EPR potential was turned off. Additional folding simulations with the experimentally determined SDSL-EPR distance restraints were performed for soluble monomeric BAX as well as with multiple sets of simulated SDSL-EPR restraints. For each setup, 7500 models were sampled in independent folding trajectories. The sampling accuracy was quantified by computing the RMSD100[103] with respect to the soluble monomeric BAX structure determined by NMR spectroscopy (PDB entry 1F16, model 8). The discrimination power of the scoring functions was computed using the enrichment metric (see section VI.2.5).[60] For homodimeric BAX, the same approach was used for the dimerization domain (α-helices 2 to 5). RMSD100 computation (see section VI.2.5) was performed with respect to the crystal structure (PDB entry 4BDU).

VI.2.4. Simulation of additional SDSL-EPR distance restraints for soluble monomeric BAX

It seems reasonable to assume that a larger number of SDSL-EPR distance restraints would result in improvements regarding the accuracy of the sampled models as well as the reliability with which accurate models can be selected. To evaluate the influence of the number of restraints on sampling accuracy and model selection, we simulated additional SDSL-EPR distance restraints based on the NMR structure for soluble monomeric BAX (PDB entry 1F16, model 8). The simulation of the additional SDSL-EPR distance restraints consisted of two steps: the selection of pairs of spin labeling sites and the simulation of the spin-spin distance between the two spin labeling sites (see section E.2.5 on page 189). The selection of suitable spin labeling sites was performed using a location selection algorithm that relies on the protein's sequence and predicted secondary structure.[155] The algorithm employs MC sampling to distribute spin labeling pairs over all SSEs. To avoid buried spin labeling sites, only residues that are predicted to be solvent-exposed were considered. For the resulting set of spin labeling pairs, the spin-spin distance was simulated using the CONE model.[78,149] Briefly, the CONE model implicitly models the structure and dynamics of MTSL as a motion-on-a-cone. It yields a probability distribution for the difference between the spin-spin distance (D_{SL}) and the $\mathrm{C}_{\beta} - \mathrm{C}_{\beta}$ distance (D_{BB}) of the spin labeling sites. This model has been successfully evaluated on experimentally determined SDSL-EPR distances for T4-lysozyme and αA-crystallin.[78,149] By adding the predicted distribution to D_{BB} in the NMR structure of soluble monomeric BAX, the spin-spin distance for a pair of spin labeling sites can be simulated. Using this protocol, three additional sets consisting of 30, 40 and 50 SDSL-EPR distance restraints were simulated for soluble monomeric BAX.

VI.2.5. Calculating SDSL-EPR score enrichments

The RMSD100 metric[103] was used to quantify the structural dissimilarity between different models. The RMSD100 is the protein-size normalized root-mean-square-deviation of the backbone coordinates computed as shown in equation (VI.2).

$$RMSD100 = \frac{RMSD}{1 + \log \sqrt{L/100}} \tag{VI.2}$$

where:

$RMSD100$ = protein-size-normalized $RMSD$
$RMSD$ = root-mean-square distance of the C_α-coordinates
L = length of the protein chain

The enrichment is used to evaluate how well a scoring function is able to select the most accurate models from a given set of models. The models of a given set S are sorted by their RMSD100-values and the 10 % of the models with the lowest RMSD100-values put into the set P (positive) the rest of the models will be put into the set N (negative). The models of S are then also sorted by their assigned scoring-value and the 10 % of the models with the lowest (most favorable) score are put into the set T. The models, which are in P and in T, are the models, which are correctly selected by the scoring function, and their number will be referred to as TP (true positive). The number of models, which are in P but not in T, are the models, which are not selected by the scoring function despite being among the most accurate ones. They will be referred to as FN (false negative). The enrichment is then calculated as shown in equation (VI.3).

$$e = \frac{\#TP}{\#P} \cdot \underbrace{\frac{\#P + \#N}{\#P}}_{=10.0} \tag{VI.3}$$

where:

e = enrichment
P = 10 % most accurate models
N = remaining models
TP = 10 % most accurate models that are also among the 10 % best scoring models

The positive models are in this context the 10 % of the models with the lowest RMSD100-values. Therefore, $\frac{\#P+\#N}{\#P}$ is fixed at a value of 10.0. Consequently, the enrichment can range from 0.0 to 10.0. An enrichment-value of 1.0 indicates that the scoring function is unable to discriminate between accurate and inaccurate models and the probability of selecting and accurate model corresponds to random chance. Enrichment-values greater than 1.0 indicate that the scoring function is able to select accurate models with a probability that is greater than random chance. Enrichment-values smaller than 1.0 indicate that the scoring function selects against accurate models and the probability of selecting accurate models is less than random chance.

VI.2.6. Using clustering for model selection

Clustering of the backbone coordinates by RMSD was used for additional model selection trials. A partitioning-based clustering approach was used, which is based on *k*-means and implemented in the cluster package[100] in R. Clustering was performed using a maximum average dissimilarity between cluster members of 3 Å. Clusters were only considered if their population size was at least 1 % of all models sampled. The reported RMSD100-values are between the cluster centers (medoids) and the experimentally determined structure.

VI.3. Results

In this section, the effect of SDSL-EPR distance restraints on *de novo* protein structure prediction is evaluated under the aspects of sampling accuracy and discrimination power. The features of BAX that complicate *de novo* protein structure prediction in the absence of experimental data are discussed. Subsequently, the effect of SDSL-EPR distance restraints on sampling accuracy and discrimination power are evaluated. Reported results are the accuracies of the models with the lowest RMSD100-values (henceforth labeled as most accurate models) as well as the percentage of models with an RMSD100-value (see section VI.2.5 on page 87) of less than 8 Å with respect to the corresponding NMR or X-ray crystal structure available. Additionally, the enrichment is reported, which is the percentage of the accurate models that can be selected by the scoring function (see section VI.2.5 on page 87).

VI.3.1. Summary of the available SDSL-EPR data for soluble monomeric and homodimeric BAX

The benchmark was performed on the soluble monomeric and the homodimeric states of BAX. Here, we give a summary about the SDSL-EPR data available for both states and how well the respective experimentally determined reference structures (PDB entry 1F16 for soluble monomeric BAX and PDB entry 4BDU for homodimeric BAX) agree with SDSL-EPR data. The latter is important because we evaluate the accuracy of the predicted models based on their structural similarity to the respective experimentally determined structure.

Data was taken from the literature where Bleicken *et al.* measured twenty-five distances for soluble monomeric BAX by Q-band DEER (table E.1 on page 184).[202] In their study, the spin labeling sites were selected based on several criteria: While the spin labels should reveal relevant information about the protein structure, their introduction should not change the protein's fold or affect the stability or function of the protein. The spin labeled proteins used in Bleicken's study were shown to retain their fold and the ability to permeabilize large unilamellar vesicles with a composition mimicking the MOM.[202] The structure of soluble monomeric BAX was determined by Suzuki *et al.* through NMR spectroscopy (PDB entry 1F16)[210] and was used here as a baseline for comparison. To evaluate the suitability of the available SDSL-EPR distance data for protein structure prediction, all models from the NMR ensemble were scored for agreement with the SDSL-EPR restraints using the CONE model.[78,149] The average difference between the observed D_{SL} and D_{BB} was 6.3 Å with an average score of −0.84 (table E.1 on page 184, perfect agreement score is −1.00 whereas the worst possible agreement score is 0.00).

Data was taken from the literature where SDSL-EPR distance measurements were performed on membrane embedded, active and homooligomeric BAX by Bleicken *et al.*.[202] Of the forty-one measured distances, seventeen are within the dimerization domain whereas the remaining twenty-four are within the piercing domain or between dimerization and piercing domain.[202] A crystal structure of a truncated BAX variant covering only the dimerization domain was published by Czabotar *et al.* (PDB entry 4BDU).[211] In order to benchmark our algorithm, we consequently opted for predicting the dimerization domain only, for which a reference structure was available. Although the reference structure (PDB entry 4BDU) was crystalized in the absence of the membrane, Bleicken *et al.*[202] showed that 4BDU well represents the fold of the dimerization domain as its present in the full length active protein embedded in liposomes and consequently is suitable as a baseline for comparison. This is in agreement with our evaluation, in which we used the CONE model[78,149] to evaluate the agreement of the X-ray crystal structure with the SDSL-EPR data measured by Bleicken *et al.*: the average difference between D_{SL} and D_{BB} was 3.1 Å with an SDSL-EPR agreement score of −0.94 (table E.2 on page 185), indicating that the crystal structure is in good agreement with the SDSL-EPR data. In this study, we folded residues 54 to 122, which is identical to the region determined in the crystal structure (PDB entry 4BDU). Of the seventeen published SDSL-EPR distance restraints within the dimerization domain, we only used eleven restraints. The six discarded restraints are between the dimerization domain and residue 126, which is not included in 4BDU.

Additional analysis was conducted to evaluate if bendings of SSEs are required to satisfy the SDSL-EPR restraints. This is important because the complexity of structural sampling does not allow for exhaustive sampling of all possible conformations. On these grounds, BCL::Fold reduces the complexity of the sampling space by assembling the tertiary structure from idealized, straight SSEs, only allowing small deviations from idealized parameters. Therefore, in a second test, α-helices in the NMR models and the X-ray crystal structure were straightened before scoring in order to quantify the influence of bent SSEs on the agreement with the SDSL-EPR distance restraints. In this context, idealization means setting the dihedral angles (ϕ, ψ) to (−60°, −40°) for α-helices and to (−135°, 135°) for β-strands. To evaluate the influence of deviations from idealized dihedral angles (bendings or kinks) on the agreement with the SDSL-EPR distance restraints, the experimentally determined structures for soluble monomeric BAX (PDB entry 1F16, model 8) and homodimeric BAX (PDB entry 4BDU) were idealized using the BCL software suite (see section E.2.6 on page 190), which sets the dihedral angles of the SSEs to aforementioned idealized values. The agreement of the idealized structures with the SDSL-EPR data was subsequently quantified, showing an average agreement score of −0.88 for soluble monomeric BAX. The resulting agreement is no diminishment from the agreement score for the non-idealized structure of −0.88. This indicates that a structure with idealized SSEs can achieve agreement with the SDSL-EPR distance data and focusing the sampling on SSEs with idealized dihedral angles won't negatively influence the prediction of the protein's tertiary structure. Based on this analysis, the eighth model of the NMR ensemble for soluble monomeric BAX was selected as reference structure for the benchmark because it had the best agreement with the SDSL-EPR data. Notably, the same model was selected based on the RMSD by Bleicken *et al.*[202] between the experimental time domain DEER traces and those simulated with the software Multiscale Modeling of Macromolecular systems (MMM) 2013.2,[214] based on a rotamer library approach. For homodimeric BAX and straightened SSEs, the average difference between D_{SL} and D_{BB} was 3.8 Å with an SDSL-EPR agreement score of −0.90 (table E.2 on page 185), which again does not constitute a significant diminishment of the SDSL-EPR agreement score for idealized SSEs; indicating that structure assembly from idealized SSEs won't hinder the prediction for homodimeric BAX.

VI.3.2. *The properties of BAX complicate* de novo *protein structure prediction in the absence of experimental data*

BCL::Fold scores protein structures using knowledge-based potentials derived from statistics over properties of protein structures deposited in the PDB (see section VI.2.1 on page 85 for details). Therefore, if a protein structure significantly deviates from the statistics, an unfavorable score is assigned as compared to alternative conformations, hindering prediction of the protein's tertiary structure with BCL::Fold.

BAX monomers consist of 192 residues, forming nine α-helices. Due to its ability to interact with membranes, some portions of the soluble monomeric BAX structure are outliers to statistics collected from experimentally determined structures of soluble proteins. Specifically, the exposure of the residues in α-helix 9 as well as the relative orientation of α-helix 9 with respect to other α-helices feature poor

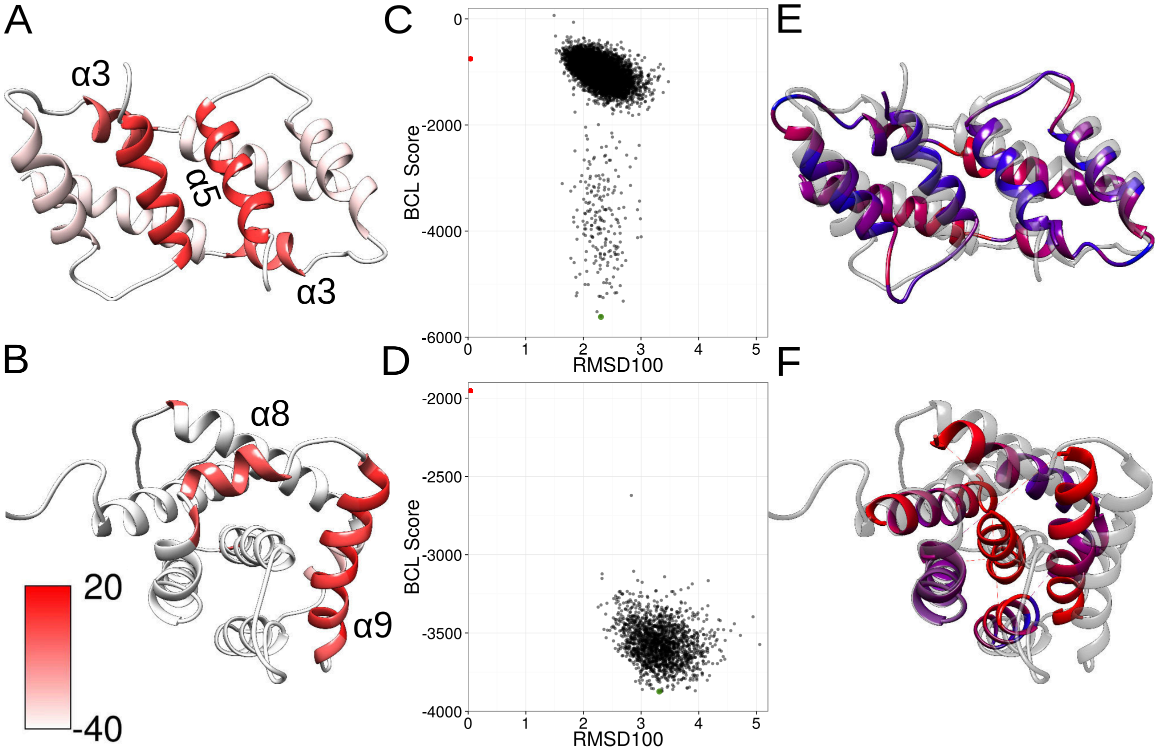

Figure VI.1.: The properties of soluble monomeric BAX and homodimeric BAX hinder* de novo *protein structure prediction. *(A,B) Due to their properties, parts of homodimeric BAX (A) and soluble monomeric BAX (B) score poorly when evaluated in the BCL::Fold knowledge-based scoring function, hindering prediction of the tertiary structure. Color code: white-red scale with white being good score and red being poor score. (C,D) Relaxing the NMR and X-ray crystal structures in the BCL::Fold force field shows score minima for alternative conformations. Black dots represent alternative conformations with their BCL score (y-axis) and the RMSD100 relative to the NMR/X-ray crystal structure (x-axis). The NMR/crystal structure is shown as red dot. Green dots are the best scoring structures, which are shown in (E,F). (E) Relaxing the X-ray crystal structure of homodimeric BAX (PDB entry 4BDU) in the BCL::Fold force field results in tighter packing of α-helices 3 and 5 and a slightly reduced radius of gyration. The relaxed model is shown on a blue-red scale, with blue being structural similarity to the crystal structure (grey) and red being structural dissimilarity. (F) Relaxing the NMR structure of soluble monomeric BAX (PDB entry 1F16, model 8) in the BCL::Fold force field results in α-helix 9 moving closer into a pocket formed by α-helices 2 to 5. The relaxed model is shown on a blue-red scale, with blue being structural similarity to the NMR structure (grey) and red being structural dissimilarity.*

agreement with statistics (see section VI.2.1 on page 85) collected from experimentally determined structures in the PDB (figure VI.1 on the preceding page). Notably, α-helix 9 is proposedly transmembrane after membrane insertion.[202,212] In consequence, a knowledge-based potential function, as used by many *de novo* folding algorithms, ranks the experimentally determined structure of soluble monomeric BAX poorly compared to alternative arrangements (see figure VI.1 on the previous page for details). BCL::Fold,[59] which uses knowledge-based potentials to evaluate the accuracy of a model,[60] is no exception.

This can be demonstrated though relaxing the experimentally determined structure of soluble monomeric BAX (PDB entry 1F16) in the BCL::Fold force field. During the relaxation, small structural perturbations are applied to the NMR structure. After each perturbation, the resulting structure is scored using the BCL score. This results in a set of models, which structurally deviate from the NMR structure but have a more favorable BCL score. The structures with the lowest score are most likely to be predicted by BCL::Fold as the native structure of soluble monomeric BAX. Figure VI.1 on the preceding page shows the BCL scores and dissimilarities to the NMR structure for a set of relaxed models. Soluble monomeric BAX has a local score minimum for conformations with an RMSD100-value of 3 Å to 4 Å (figure VI.1 on the previous page). The model with the most favorable score shows α-helix 9 moving closer into a pocket formed by α-helices 2 to 5 (figure VI.1 on the preceding page), which reduces the exposure of the residues in α-helix 9 and results in a more favorable score.

These difficulties in scoring/ranking the sampled models make soluble monomeric BAX an appropriate test case to evaluate if scoring problems can be overcome by incorporating limited structural data from SDSL-EPR experiments. Further, as SDSL-EPR data recently became available for the dimerization domain of homooligomeric BAX, BAX is also a test case for determining a protein's structure in different, biologically relevant conformations. For homooligomeric BAX, similar challenges in the ranking of models in the absence of experimental data can be observed. The radius of gyration of the crystal structure of the dimerization domain (PDB entry 4BDU) significantly deviates from statistics collected from known structures in the PDB. Additional SSE- and residue-based deviations are observed for the exposure of residues and SSE orientations. These deviations are particularly pronounced for the α-helices 3 and 5 (figure VI.1 on the previous page). Repeating the relaxation experiment as described above for homodimeric BAX, shows a local score minimum for structures with an RMSD100-value between 2 Å and 3 Å relative to the crystal structure (figure VI.1 on the preceding page). The model with the most favorable BCL score shows a change in the packing of α-helices 2 to 5 and a slight reduction of the radius of gyration (figure VI.1 on the previous page). Comparably to soluble monomeric BAX, these scoring problems make the tertiary structure of homodimeric BAX hard to predict with BCL::Fold because the scoring function does not detect the crystal structure as native-like.

VI.3.3. SDSL-EPR distance restraints can overcome de novo *sampling and scoring problems*

By using SDSL-EPR distance restraints, it is possible to overcome scoring and sampling problems, which hinder *de novo* protein structure prediction. As demonstrated in the previous section, the NMR ensemble of soluble monomeric BAX and the X-ray crystal structure of homodimeric BAX score poorly in the BCL::Fold knowledge-based scoring function, which hinders prediction of a model that is in good agreement with the NMR- or X-ray-derived models.

The usage of SDSL-EPR distance restraints for soluble monomeric BAX results in a shift of the RMSD100 distributions by around 1.5 Å to models in better agreement with the NMR-derived model (figure VI.2 on the following page). Whereas without SDSL-EPR data, the most accurate model sampled

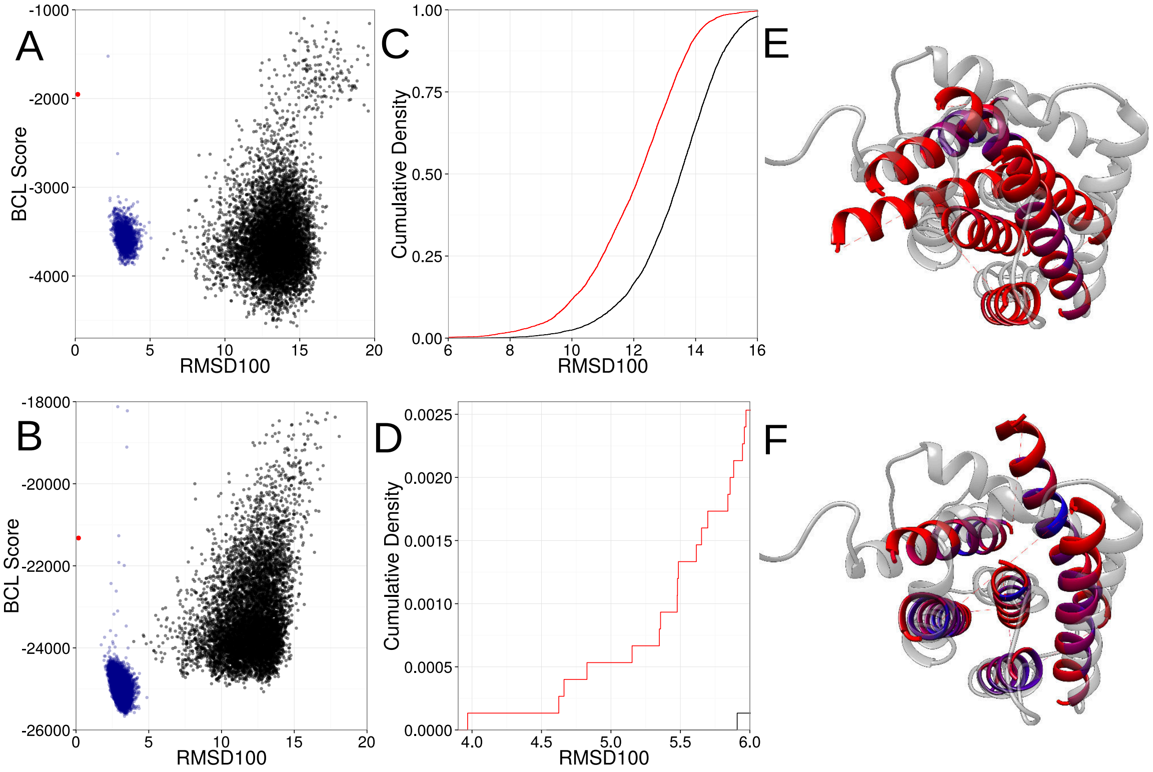

Figure VI.2.: Structure prediction results for soluble monomeric BAX. *(A) Protein structure prediction without SDSL-EPR distance restraints results in a poor correlation between the score of the* de novo *predicted models (black dots) and their accuracy (quantified as RMSD100 relative to the experimentally determined structure). The experimentally determined structure (red dot) and the experimentally determined structure relaxed in the BCL::Fold force field (blue dots) score worse than the* de novo *predicted models. (B) By using SDSL-EPR distance restraints, the score gap between the experimentally determined structure (red dot) and the* de novo *sampled models (black dots) could be reduced. The experimentally determined structure relaxed in the BCL::Fold force field (blue dots) scores better than the* de novo *sampled models. The BCL score of the experimentally determined structure and the relaxed structures includes the EPR agreement score, resulting in lower scores than in (A). (C,D) Using SDSL-EPR distance restraints increased the sampling density of models in agreement with the NMR-derived model (red — with SDSL-EPR distance restraints, black — without). (E,F) In the most accurate model sampled with SDSL-EPR distance restraints (F, blue-red scale, RMSD100 =* 3.9 Å*), the placement of the SSEs is more similar to the experimentally determined structure (PDB entry 1F16, model* 8*, grey), than for sampling without SDSL-EPR distance restraints (E, blue-red scale, RMSD100 =* 5.9 Å*). Color coding: blue-red scale with blue being structural similarity to the experimentally determined structure and red being dissimilarity.*

Protein	Restraints	best (Å)	μ_{10} (Å)	τ_8 (%)	e
Monomer	Without	5.9	7.0	0.2	0.4
Monomer	25 experimental	3.9	5.0	2.5	1.5
Monomer	30 simulated	4.2	4.9	7.5	4.1
Monomer	40 simulated	4.1	4.4	11.4	4.2
Monomer	50 simulated	3.9	4.2	11.5	4.5
Dimer	Without	5.7	6.8	0.1	1.3
Dimer	11 experimental	3.3	3.4	13.7	2.1

Table VI.1.: Sampling accuracy and enrichment are improved by SDSL-EPR distance restraints. *By using SDSL-EPR distance restraints in protein structure prediction the sampling accuracy can be improved as it is seen for the RMSD100-values of the most accurate (by RMSD100) model sampled (best), the average of the ten best models sampled (μ_{10}) and the percentage of the models with an RMSD100-value less than* 8 Å *(τ_8). A larger number of SDSL-EPR restraints leads to more substantial improvements, which was demonstrated by simulating additional SDSL-EPR restraint sets consisting of* 30, 40 *and* 50 *restraints. More restraints constantly result in more pronounced improvements in the sampling accuracy. SDSL-EPR distance restraints also improve the ability to select the accurate models among the sampled models, which is shown by improved enrichment-values (e).*

had an RMSD100-value of 5.9 Å; by using SDSL-EPR data, the RMSD100-value of the most accurate model could be improved to 3.9 Å (see figure VI.2 on the previous page for details). For further evaluation of the sampling accuracy, the ten best models by RMSD100 were selected and their average RMSD100-value, μ_{10}, was calculated (the average RMSD100-values for different numbers of models are shown in figure E.1 on page 183). In the absence of SDSL-EPR data, the μ_{10}-value was 7.0 Å, whereas with SDSL-EPR data the μ_{10}-value improved to 5.0 Å. Additionally, the percentage of models with an RMSD100-value of less than 8 Å, τ_8, was calculated. For folding without SDSL-EPR data, the τ_8-value was 0.3 %, whereas when folding with SDSL-EPR distance restraints the τ_8-value improved to 1.9 %. Using SDSL-EPR restraints for the dimerization domain of homooligomeric BAX improved the RMSD100-value of the most accurate model from 5.7 Å to 3.3 Å. The μ_{10}- and τ_8-values improved from 6.8 Å to 3.4 Å and from 0.1 % to 16.7 %, respectively (table VI.1 and figure VI.3 on the following page). Additional model selection trials were performed using clustering. For soluble monomeric and homooligomeric BAX — in the absence of SDSL-EPR distance restraints, the clusters closest to the experimentally determined structure had an RMSD100 value of 9.2 Å and 11.4 Å, respectively. By using SDSL-EPR distance restraints, clusters with an RMSD100 of 7.1 Å and 4.8 Å relative to the respective NMR/X-ray-derived model could be detected for soluble monomeric and homooligomeric BAX.

Besides the sampling of conformations, a protein structure prediction method must be able to select the most accurate models among the sampled models. To evaluate the ability of the scoring function to select the most accurate models sampled during *de novo* folding, score enrichments were calculated. The enrichment indicates how well the scoring function is able to distinguish between accurate and inaccurate models (see section VI.2.5 on page 87 for details). The term accurate is hereby defined as being among the 10 % of the models with the lowest RMSD100-value relative to the experimentally determined structure. For the models generated in the absence of SDSL-EPR data, the enrichment for soluble monomeric BAX was 0.4 (table VI.1). The enrichment of less than 1.0 for the BCL::Fold energy function indicates that it actually selects against topologies in agreement with the X-ray-derived model, presumably due to the poor score of the α-helix 9 as discussed in section VI.3.2 on page 90. With SDSL-EPR distance restraints, the enrichment improved to 1.5. The improvement in enrichment

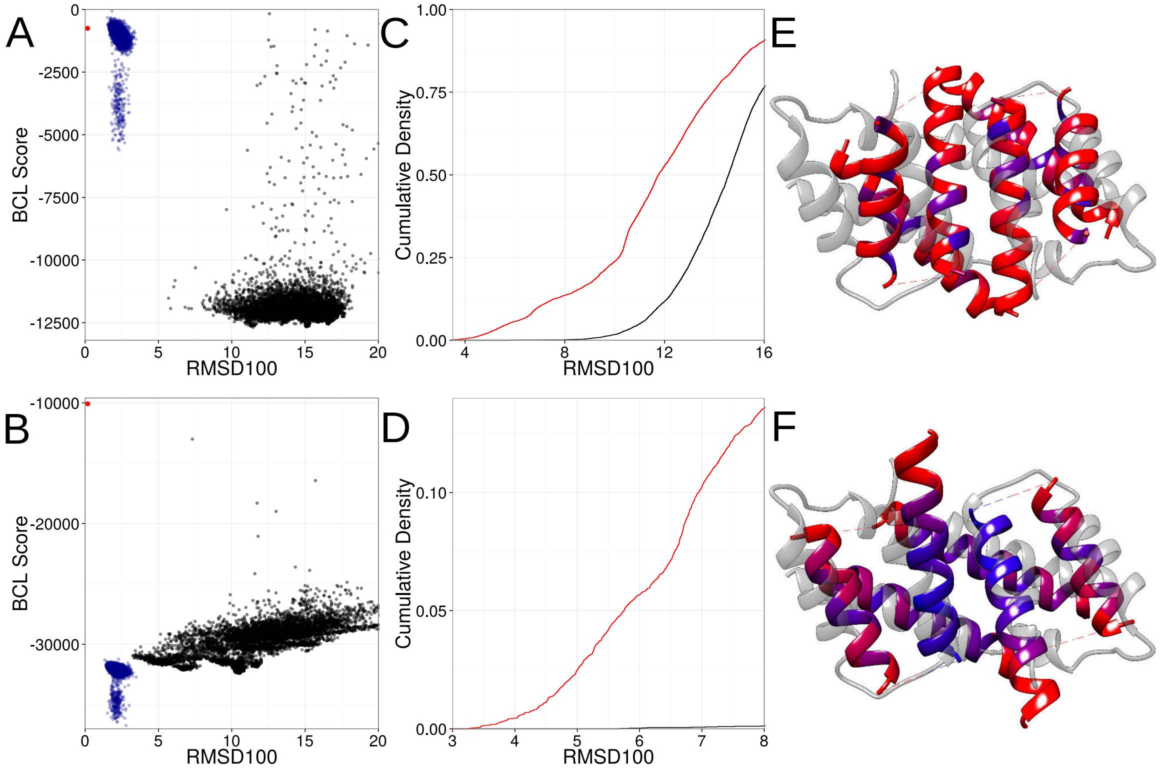

Figure VI.3.: Structure prediction results for homodimeric BAX. *(A) Protein structure prediction without SDSL-EPR distance restraints results in a poor correlation between the score of the* de novo *sampled models (black dots) and their accuracy (quantified as RMSD100 relative to the experimentally determined structure). The experimentally determined structure (red dot) and the experimentally determined structure relaxed in the BCL::Fold force field (blue dots) score significantly worse than the* de novo *sampled models. (B) Protein structure prediction with SDSL-EPR distance restraints results in an improved correlation between the score of the sampled models (black dots) and their accuracy. Whereas the experimentally determined structure (red dot) scores worse than the sampled models, the relaxed experimentally determined structure (blue dots) scores better than the sampled models. The BCL score of the experimentally determined structure and the relaxed structures includes the EPR agreement score, resulting in lower scores than in (A). (C,D) Using SDSL-EPR distance restraints significantly improves the sampling density of models in agreement with the NMR- and X-ray-derived models, result in a shift of the distribution of about* 6 Å *(red — with SDSL-EPR distance restraints, black — without). (E) Without SDSL-EPR distance restraints the placement of the SSEs of the most accurate model sampled (blue-red scale, RMSD100 =* 5.7 Å*) is dissimilar to the experimentally determined structure (grey). (F) By using SDSL-EPR distance restraints the SSE placement of the most accurate mode sampled (blue-red scale, RMSD100 =* 3.3 Å*) resembles the experimentally determined structure (grey). Color coding: blue-red scale with blue being structurally similar to the experimentally determined structure and red being dissimilarity.*

demonstrates that by using SDSL-EPR distance restraints, protein structure prediction methods can overcome model discrimination challenges. For homooligomeric BAX, usage of SDSL-EPR restraints improved the enrichment from 1.3 to 2.1 (table VI.1 on page 94).

VI.3.4. A larger number of restraints improves sampling accuracy and selection of accurate models

To evaluate the influence of the number of restraints on the sampling accuracy as well as the algorithm's ability to select accurate models, three additional restraint sets with different numbers of restraints were simulated based on the NMR-derived model of BAX (PDB entry 1F16, model 8). The spin labeling sites were chosen in order to distribute measurements across all SSEs (see section E.2.5 on page 189 for details). The experimentally determined restraint set consisted of twenty-five restraints, whereas the simulated restraint sets for the NMR structure of soluble, monomeric BAX (PDB entry 1F16, model 8) consisted of thirty, forty, and fifty restraints, respectively. To fold soluble BAX with the simulated restraints the same protocol was used as for the experimentally determined restraint set. The number of restraints had a significant effect on the sampling accuracy as well as on the algorithm's ability to select accurate models (table VI.1 on page 94). Whereas the μ_{10}-value for the twenty-five experimentally determined restraints was 5.0 Å, it was 4.9 Å for thirty restraints, 4.4 Å for forty restraints, and 4.2 Å for fifty restraints. For folding with twenty-five restraints, the τ_8-value was 1.9 %, with thirty restraints 7.5 %, with forty restraints 11.4 %, and with fifty restraints 11.5 %. The enrichment of 1.5 for folding with twenty-five restraints improves to 4.1 for thirty restraints, 4.2 for forty restraints, and 4.5 for fifty restraints (table VI.1 on page 94).

VI.4. Discussion

VI.4.1. Interpretation of the reported sampling accuracies and enrichments

It should be noted that comparison to 1F16 and 4BDU is somewhat limited: The RMSD100-values between the twenty individual models in 1F16 ranges from 1.7 Å to 4.7 Å with an average of 3.0 Å. The relatively low precision of the NMR-derived models represents an upper limit for the accuracy of 1F16. In result, any model that approaches this accuracy limit is in agreement with 1F16 within its accuracy limits. Additionally, in the case of the homodimeric structure, deviations may be caused by 4BDU being derived from a protein crystal with a reported resolution of 3.0 Å and in absence of membranes or membrane mimics, whereas the SDSL-EPR measurements were completed on full-length BAX variants inserted into large unilamellar vesicles mimicking the mitochondrial outer membrane lipid composition (MOM-LUVs), *i.e.*, in a more native-like environment.[202] Arguably, a comparison to the SDSL-EPR relaxed version of 4BDU and 1F16 could provide a more accurate measure of success of the folding simulation. As such models are however biased by the BCL::Fold scoring function we opted for comparison with the original PDB entries.

VI.4.2. Energy function and sampling limitations hinder in silico *protein structure prediction*

The major obstacle and challenge of *in silico* determination of a protein's tertiary structure is the vast conformational search space combined with the complicated models needed to compute an accurate estimate of a proteins free energy. These obstacles are overcome by simplifications in the scoring function and sampling space that are often coupled to a simplified representation of the protein. In concrete terms, simultaneous and exhaustive sampling of the ϕ- and ψ-angles in the protein backbone and -angles in the protein side chains is prohibitive.

BCL::Fold drastically reduced the search space by eliminating all χ-angles — side chains are represented as “superatoms”, eliminating ϕ- and ψ-angles in flexible loop regions by not explicitly modeling loop regions, and assembling predicted SSEs starting from idealized ϕ- and ψ-angles allowing only for limited deviations. Additionally, explicit simulation of the protein’s environment, like the membrane or the solvation water molecules, is circumvented by implicit models. Still, enumeration of all possible folds within an acceptable time frame remains prohibitive for larger proteins. As shown in figure VI.2 on page 93 and figure VI.3 on page 95, in the absence of any experimental data neither are models in agreement with the NMR- and X-ray-derived models sampled in a frequent manner, nor is it possible to distinguish more accurate models from less accurate models. For soluble monomeric BAX and the dimerization domain of membrane-embedded homooligomeric BAX, the experimentally determined structures both score poorly in the BCL scoring function. Even after relaxing the experimentally determined structures in the BCL::Fold force field to find a conformation in agreement with the NMR- and X-ray-derived models in a score minimum, the relaxed structures score worse than models that are not in agreement with the NMR- and X-ray-derived models (figure VI.2 on page 93 and figure VI.3 on page 95).

VI.4.3. SDSL-EPR measurements can overcome the limitations of de novo *protein structure*

SDSL-EPR distance measurements can be performed in a native-like environment and provide experimental data that can be interpreted as structural restraints, thus compensating for the algorithm’s limitation in sampling the large conformational space and estimating the free energy of these conformations accurately. Direct incorporation of the SDSL-EPR distance data into the BCL::Fold scoring function reduces the complexity of the energy function by removing local minima in the scoring function that are inconsistent with the experimental SDSL-EPR distance data, reinforcing conformations that are. Therefore, incorporation of SDSL-EPR distance restraints can overcome limitations in sampling and scoring. This was demonstrated by relaxing the experimentally determined structures in the BCL::Fold force field using SDSL-EPR restraints (figure VI.2 on page 93 and figure VI.3 on page 95). The relaxed structures are similar to the NMR- and X-ray-derived models and have a more favorable score than most of the sampled models. As a direct result of the improved pseudo-energy landscape, the MCM algorithm favors conformations that are in agreement with the SDSL-EPR data, leading to the sampling of models that are in better agreement with the NMR- and X-ray-derived models. Significant shifts of the accuracy distributions are observed for soluble monomeric BAX as well as the dimerization domain of homooligomeric BAX (see figure VI.2 on page 93 and figure VI.3 on page 95). For soluble monomeric BAX, the accuracy distribution improves by about 1.5 Å, whereas for homooligomeric BAX the improvement is about 4 Å. Additionally, usage of SDSL-EPR distance data mitigates the previously described problem of distinguishing accurate models from inaccurate models (figure VI.2 on page 93, figure VI.3 on page 95, and table VI.1 on page 94). This effect is more pronounced for homooligomeric BAX, for which the enrichment improves from 1.3 to 2.1. The score-accuracy plots in figure VI.3 on page 95 show an improved correlation between score and RMSD100. Although the best scoring model is still not in perfect agreement with the X-ray-derived model, a high model density exists in the 3 Å to 5 Å range, which could be detected through clustering. The results of this study demonstrate that SDSL-EPR distance restraints can mitigate the limitations of *de novo* protein structure prediction algorithms, by increasing the sampling frequency of the models that are in agreement with the SDSL-EPR data and by complementing the energy evaluation with structural restraints.

VI.5. Conclusion

This study demonstrates that even a limited number of SDSL-EPR distance restraints are able to introduce score minima for conformations, which have better agreement with the structural models derived from NMR or X-ray crystallography. Therefore, challenges in conformational sampling and model discrimination in *de novo* protein structure prediction can be overcome through incorporation of sparse SDSL-EPR distance restraints. This was demonstrated by the improved accuracy of the models as well as the improved enrichment of accurate models. In conclusion, a combined approach of *de novo* protein structure predictions methods and SDSL-EPR distance restraints is able to predict the fold of larger proteins that adopt multiple conformations.

VI.6. Acknowledgments

We want to thank Benjamin Mueller for thorough proofreading of the manuscript. Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 GM099842, R01 DK097376) and NSF (CHE 1305874). This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. A.J.G-S. and S.B. were supported by the Max Planck Society and the German Ministry for Education and Research (BMBF, grant N.0312040). Parts of the data analysis were performed using the R[b] package with ggplot2[123] and cluster.[100] The renderings of the models were created using Chimera.[124] The composite figures were created using Inkscape.[c]

[b] https://www.r-project.org

[c] https://inkscape.org

CHAPTER VII
STRUCTURE AND DYNAMICS OF TYPE III SECRETION EFFECTOR PROTEIN EXOU

This chapter is based on the publication "Structure and Dynamics of Type III Secretion Effector Protein ExoU As determined by SDSL-EPR Spectroscopy in Conjunction with De Novo Protein Folding".[6] Axel W. Fischer contributed to the development of the prediction pipeline, performing the experiment, analyzing the data, and writing the article.

ExoU is a 74-kDa cytotoxin that undergoes substantial conformational changes as part of its function, i.e.*, it has multiple thermodynamically stable conformations that interchange depending on its environment. Such flexible proteins pose unique challenges to structural biology: not only is it i) often difficult to determine structures by X-ray crystallography for all biologically relevant conformations, because of the flat energy landscape ii) experimental conditions can also easily perturb the biologically relevant conformation. The first challenge can be overcome by applying orthogonal structural biology techniques that are capable of observing alternative, biologically relevant conformations. The second challenge can be addressed by determining the structure in the same biological state with two independent techniques under different experimental conditions. If both techniques converge to the same structural model, the confidence that an unperturbed biologically relevant conformation is observed increases. To this end, we determine the structure of the C-terminal domain of the effector protein ExoU from data obtained by electron paramagnetic resonance spectroscopy in conjunction with site-directed spin labeling and* in silico de novo *structure determination. Our protocol encompasses a multi-module approach, consisting of low-resolution topology sampling, clustering, and high-resolution refinement. The resulting model was compared with an ExoU model in complex with its chaperone SpcU obtained previously by X-ray crystallography. The two models converged to a minimal RMSD100 of* 3.2 Å*, providing evidence that the unbound structure of ExoU matches the fold observed in complex with SpcU.*

VII.1. Introduction

ExoU is a cytotoxin with a molecular weight of 74 kDa that is encoded by the Gram-negative bacterium *Pseudomonas aeruginosa*.[215–218] Using the type III secretion system, ExoU is injected directly into eukaryotic cells, significantly increasing the severity of the infection.[219–221] Because of its function, ExoU needs to undergo substantial conformational changes; *i.e.*, depending on interaction partners and environment, different conformations of the protein will be thermodynamically most stable. One conformation of ExoU, in complex with its chaperone SpcU, has previously been elucidated through X-ray crystallography (PDB entry 3TU3).[216] This X-ray-derived model depicts ExoU as consisting of four domains. The C-terminal domain is of particular interest because it mediates association of ExoU with the membrane,[216,222,223] *i.e.*, is expected to undergo major conformational changes. However, all three structural models obtained through X-ray crystallography (PDB entries 3TU3,[216] 4AKX,[215] and 4QMK[222]) depict ExoU's C-terminal domain as exhibiting the same conformation — a four-helical bundle. Experiments performed by Gendrin *et al.* showed that even the presence of the chaperone SpcU does not occlude the residues involved in lipid binding.[215] Through EPR spectroscopy, Benson *et al.* provided evidence that the presence of the substrate induces conformational changes in ExoU's C-terminal domain.[224] Given the expected intrinsic flexibility of this domain, we set out to confirm a) that the conformation of the C-terminal domain observed in the X-ray crystallography-derived model in complex with its chaperone SpcU is consistent with structural data observed for ExoU in solution, and b) probe the structural dynamics of this domain. We chose EPR spectroscopy in conjunction with SDSL in combination with computational *de novo* protein folding to approach these questions.

EPR spectroscopy in conjunction with SDSL provides an alternative approach to probe the structure and dynamics of a protein. Briefly, SDSL-EPR is typically employed to measure the distance between two residues. To facilitate that, two cysteine residues are introduced at the sites of interest into a cys-less variant of the protein and coupled with *S*-(1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate (MTSL), which carries an unpaired electron. Through the DEER experiment,[75,150] the distance-dependent dipolar interaction of the two unpaired electrons can be me asured and translated into a distance distribution. Since every measurement requires a distinct protein double-mutant, the structural information gained from SDSL-EPR experiments is typically too sparse to unambiguously determine the protein's tertiary structure. However, in conjunction with *de novo* protein structure prediction methods, SDSL-EPR data could focus the sampling on conformations that are in agreement with the experimental data.

The computational protein structure prediction pipeline employed in this manuscript is based on the *de novo* method BCL::Fold[59] in the BCL,[a] which was specifically developed to predict the tertiary structure of large proteins. To facilitate this objective, the SSEs of the protein are predicted using machine learning methods. Conformations of the predicted SSEs exhibiting idealized dihedral angles are subsequently arranged in the three-dimensional space by a MCM algorithm. The intermediary and final models are evaluated using knowledge-based potentials that assign a pseudo-energy score to each model.[60] Although this method has been successful at *de novo* sampling the tertiary structure of large proteins, distinguishing between accurate and inaccurate models based on their pseudo-energy score alone remains a challenge.[1] However, it was demonstrated that incorporation of limited experimental data significantly mitigates problems in model discrimination.[2,3,5,88] The Rosetta method[94,149] was used to add atomic detail and energy-optimize the final models.

In this manuscript, we discuss the structure and dynamics of ExoU as determined by SDSL-EPR spectroscopy in conjunction with *de novo* protein structure prediction and provide a benchmark evaluating the influence of SDSL-EPR data on *de novo* protein structure prediction. In section VII.2, we detail the computational protein structure prediction pipeline, describe the available SDSL-EPR data, and compare the prediction results to the X-ray-derived model of ExoU and evaluate the influence of SDSL-EPR data on *de novo* protein structure prediction. In section VII.4 on page 109, we describe the experimental approach used to obtain the SDSL-EPR data.

VII.2. Results and discussion

Here, we report the results of the *de novo* protein structure prediction with and without the inclusion of SDSL-EPR data. The results were evaluated in the terms of sampling accuracy and discrimination of inaccurate models as described above. We begin with an analysis of improvements in sampling accuracy when SDSL-EPR data is incorporated into the protein structure prediction algorithm. The influence of SDSL-EPR data on the discrimination of inaccurate models is then considered. This section starts with

[a] http://www.meilerlab.org/bclcommons

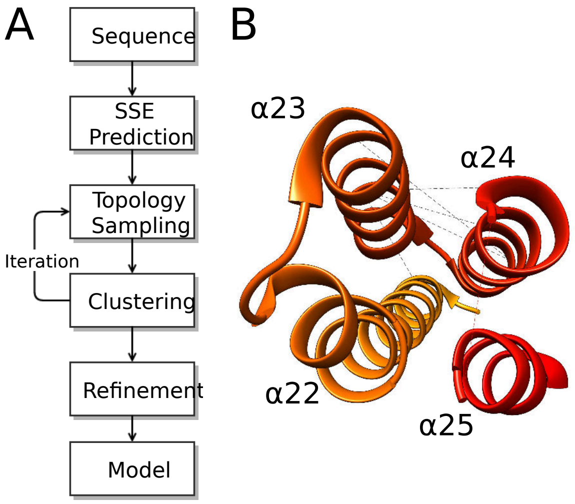

Figure VII.1.: Protein structure prediction pipeline and SDSL-EPR data for the C-terminal domain of ExoU. *(A) The* de novo *protein structure prediction pipeline for the C-terminal domain of ExoU employed a hierarchical approach consisting of modules for secondary structure prediction, low-resolution topology sampling, and high-resolution refinement. (B) Seven intra-domain SDSL-EPR measurements were available (shown as dashed lines) for the C-terminal domain of ExoU.*

an outline of the benchmarking procedure that was used to evaluate the influence of SDSL-EPR data on *de novo* protein structure prediction accuracy. This outline is followed by sections providing a detailed description of the protein structure prediction protocol, an analysis of the available SDSL-EPR data for the C-terminal domain of ExoU, and a description of the algorithm used to translate the SDSL-EPR data into structural restraints that are usable by the prediction algorithm. This section is concluded by an evaluation of the predicted tertiary structure — its agreement with the X-ray-derived model and a discussion of its consistency with the SDSL-EPR data.

VII.2.1. Summary of the available SDSL-EPR data for the C-terminal domain of ExoU

For the C-terminal domain of ExoU, seven intra-domain SDSL-EPR distance measurements were available (see the agreement with the SDSL-EPR data in table F.1 on page 191, the simulated distance distributions in figure F.1 on page 192, and the summary of the approach in figure VII.1 for details). The X-ray-derived model is in good agreement with the restraints derived from the SDSL-EPR measurements, as indicated by an average agreement score of −0.88 (table F.1 on page 191, see section VII.2.4 on page 104 for details regarding quantifying the agreement of models with the SDSL-EPR data). Of the seven restraints, four are between α-helices 23 and 24, one is between α-helices 22 and 23, one is between α-helix 23 and the loop region connecting α-helices 24 and 25, and one is between α-helix 24 and the loop connecting α-helices 24 and 25. As shown in figure VII.1, the SDSL-EPR restraints well describe the relative positions of α-helices 23 and 24.

VII.2.2. Summary of the benchmark setup to evaluate the influence of SDSL-EPR data on de novo *structure prediction*

The influence of SDSL-EPR data on *de novo* protein structure prediction was evaluated by performing two independent structure predictions runs, one with incorporated SDSL-EPR data and one in the absence of SDSL-EPR data, for the C-terminal domain of the effector protein ExoU. The protocols for both prediction runs were predominately identical, only differing in the scoring function that was extended by a scoring term quantifying the agreement of the model with the SDSL-EPR data for one prediction run (see section VII.2.4 on page 104 for details). For each prediction run, about 100 000 low-resolution models and about 50 000 high-resolution full-atom models were sampled and subsequently analyzed under the aspects of sampling accuracy and discrimination of inaccurate models (see the following sections for details). A previously published X-ray-derived model of ExoU (PDB entry 3TU3)[216] was used as reference structure for evaluating sampling accuracy and model discrimination — the reported RMSD-values are between the sampled models and the X-ray-derived model of ExoU. However, no information about the X-ray-derived model was used in the protein structure prediction protocol.

VII.2.3. Protein structure prediction protocol

The protein structure prediction protocol (see figure VII.1 on the previous page) consisted of two modules: a module for low-resolution sampling of possible topologies and a module for the construction of loop regions and high-resolution refinement of the resulting model. The two modules were connected through a data aggregation step using filtering and clustering. The low-resolution topology sampling was performed iteratively: upon conclusion of the first iteration of the low-resolution topology sampling, the most favorable models by pseudo-energy score and agreement with the SDSL-EPR data (if applicable) were selected as start models for a second round of optimization using the topology sampling module.

In the first module (see section F.2.1 on page 194 and section F.2.2 on page 194), low-resolution topology sampling, the secondary structure of the protein was predicted using PSIPRED[98,225] and Jufo9D.[97] The resulting SSEs were subsequently arranged in the three-dimensional space using the *de novo* protein structure prediction algorithm BCL::Fold.[59] BCL::Fold employs an MCM algorithm to sample possible topologies arising from the predicted SSEs. The BCL::Fold prediction consists of six stages: five assembly stages and one refinement stage. In each MC step, a randomly chosen perturbation (mutate) is applied to the current protein model. The assembly and refinement stages differed in the mutates applied by the MCM algorithm. Whereas the mutates during the assemble stages apply topology changing perturbations like large-scale translations of SSEs or swapping of SSEs, the mutates during the refinement stage only apply small-scale perturbations like rotating helices around their main axes. After each application of a mutate, the resulting protein model m is evaluated using a scoring function $E(m)$.[60] The scoring function $E(m)$ is the weighted sum of various knowledge-based scoring terms E_i and assigns a pseudo-energy score E_m to a protein model m by computing equation (VII.1).

$$E_\mathrm{m} = E(m) = \sum_i w_i \cdot E_i(m) \tag{VII.1}$$

where:

E_m = pseudo-energy score of protein model m
m = protein model
E_i = scoring term evaluating a specific property i of protein model m
w_i = weighting factor of scoring term i

The scoring terms E_i each evaluate different properties of the protein model like steric interferences, residue-residue interactions, SSE-SSE packing, or residue exposure.[60] Depending on the score difference between the current model and last accepted model, a Metropolis criterion either accepts or rejects the new model.[59,60] The Metropolis criterion in conjunction with simulated annealing is used to prevent sampling trajectories from getting trapped in local pseudo-energy minima. Subsequently, the MCM algorithm resumes with the latest accepted model. The MCM optimization of each assembly and refinement stages lasted for a maximum of 4000 MC steps with the optimization terminating early if no improvement in the pseudo-energy score were achieved for 800 MC steps in a row. In total, this module resulted in 50 000 models.

Upon conclusion of the first topology sampling module, the models were ranked according to their pseudo-energy score. The best 10 % of the sampled models (about 10 000 models) were selected for clustering using a k-means implementation in R,[100] with the RMSD between the backbone C_α-coordinates being the metric for quantifying the dissimilarity between models (see section F.2.3 on page 195 for details). The number of clusters was dynamically adjusted to maximize the average Silhouette width,[226,227] which quantifies how tight the grouping of the data points in each cluster is. Briefly, the Silhouette $s(n)$ of a data point n is computed according to equation (VII.2).

$$s(n) = \frac{b(n) - a(n)}{\max\{a(n), b(n)\}} \tag{VII.2}$$

where:

$s(n)$ = silhouette of data point n
$a(n)$ = average dissimilarity between n and all other data points in the same cluster
$b(n)$ = lowest average dissimilarity between n and a data point in any other cluster

The average Silhouette width of the clustering is computed accordingly as $\sum_{n=1}^{k} s(n)/k$, where k is the number of data points. The Silhouette width ranges from −1 to 1 with a higher value indicating a good matching of the data elements to their respective clusters and poor matching to other clusters — therefore indicating that clustering represents the underlying data well. For structure prediction with and without SDSL-EPR data, the clustering resulted in four and seven clusters, respectively. The cluster medoids and the most favorable model by pseudo-energy score were chosen for another round of optimization using the topology sampling module. The protocol of this optimization matched the protocol described above but used the selected models as start models. Upon conclusion of the second round of optimization, the same clustering protocol was applied and resulted in three and seven clusters, respectively. The cluster medoids and the most favorable model by pseudo-energy score were selected as start models for the second module — the construction of loop regions and high-resolution refinement of the models using Rosetta.

In the second module (see section F.2.4 on page 195 for details), the Rosetta software suite[94,228] was used to construct loop regions, add side chain coordinates, and perform a high-resolution refinement of the provided protein models. The CCD algorithm[101] was employed for construction of loop regions and SDSL-EPR data was incorporated into CCD and the subsequent refinement using the CONE model[2,78] (see the following sections for details) following previously published protocols.[149] The weight of the score quantifying the agreement of the model with the SDSL-EPR data was set to 40 to ensure that the score accounts for approximately 40 % of the total pseudo-energy score. For each of the provided start models, 500 full-atom models were sampled using this protocol, resulting in about 20 000 models.

VII.2.4. Incorporating SDSL-EPR data into computational protein structure prediction

To use SDSL-EPR spectroscopy for distance measurements in a protein, a spin label carrying a free electron needs to be introduced at the two sites of interest. The distance between the two spin labeling sites is then determined indirectly by measuring the dipolar interaction between the two free electrons, which is inversely proportional to their cubed distance.[75,150] The indirect nature of this measurement poses challenges for using the observed data in a protein structure prediction algorithm. First, even if the backbone of the protein is inflexible, the proteins in the sample for the measurement will exhibit different conformations of the spin label resulting in a distribution of distances rather than one observed distance. Second, depending on the type of spin label and its conformation, the distance between the free electron and the backbone of the spin labeled residue can be rather large adding uncertainty to the measurement. For example, for the spin label MTSL, the Euclidean distance between the C_β-atom of the spin labeled residue and the spin label's free electron can be up to 8.5 Å.[78]

To use the distances measured in the SDSL-EPR experiment within a protein structure prediction algorithm, a function to quantify the agreement between the experimental data and a protein model needs to be defined. This function needs to capture both aforementioned properties of the SDSL-EPR measurement — the flexibility of the spin label and the indirectness of the measurement. Previously, different approaches to define such a function have been published. The CONE model,[2,78,149] uses a knowledge-based approach to account for these factors. This implicit approach provides a rapid way to estimate the probability of observing a certain $C_\beta - C_\beta$ distance (D_{BB}) given a measured spin-spin distance (D_{SL}). The agreement score based on the CONE model is defined based on the difference between D_{BB} and D_{SL}, which can range from −12 Å to 12 Å. The value of the scoring function ranges from 0.0 which means no agreement, to −1.0, which means best possible agreement. This approach has been successfully used for *de novo* prediction of membrane proteins[2] and soluble proteins that exist in multiple relevant states.[5]

Due to its significantly faster computation time, we employed the CONE model[78] to translate the measured spin-spin distances into structural restraints for the *de novo* protein structure prediction algorithm. For the structure prediction algorithm, the weight of the CONE-based score quantifying the agreement of the protein model with the SDSL-EPR data was set to 40, which ensured that this score accounted for about 40 % of the total score — a contribution percentage for limited experimental data that provided the best prediction results in previous studies.[1] Additionally, we added a quadratic potential function to penalize models with $D_{SL} - D_{BB}$ values outside of the range of the CONE model.

VII.2.5. De novo *prediction results confirm the correctness of the X-ray-derived model*

The X-ray-derived model of the ExoU/SpcU (PDB entry 3TU3)[216] structure is of high quality (Resol. = 1.9 Å, R_{free} = 0.225, R_{work} = 0.191). The C-terminal domain makes few crystal lattice contacts that are overall unlikely to perturb its confirmation: V57, L55, G82 of SpcU appear to form a hydrophobic pocket for α-helix 23 and SpcU S51 and R83 potentially hydrogen bond to ExoU residues N657 and E636, respectively. Otherwise, SpcU does not appear to influence the structure of the C-terminal four-helix-bundle. Hence, we started with the hypothesis that *de novo* structure prediction in conjunction

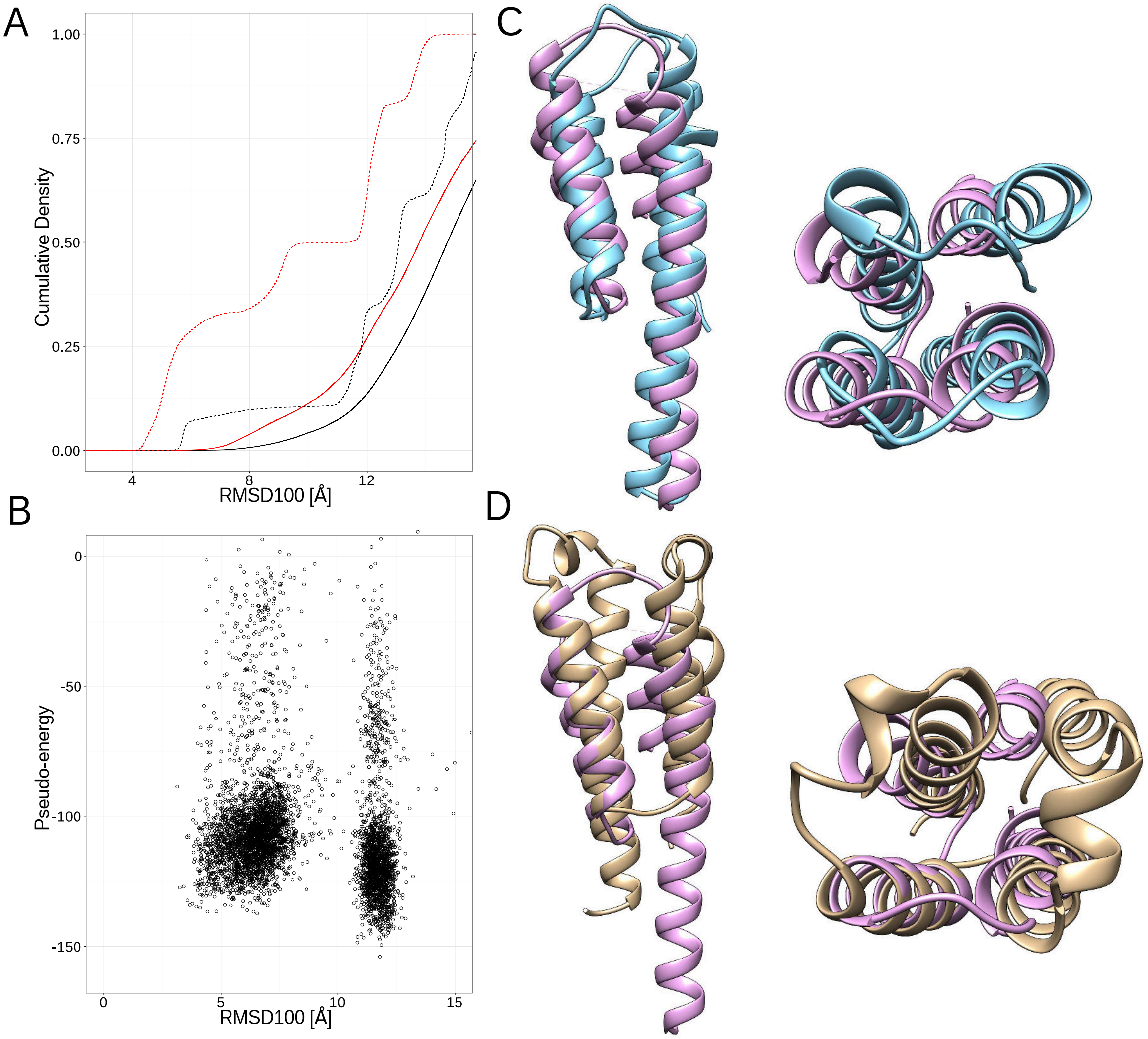

Figure VII.2.: Prediction results for the C-terminal domain of ExoU. *(A) Comparison of the sampling densities between prediction with (red) and without (black) SDSL-EPR data. Results are shown for the first (solid line) and second (dashed line) iterations of the low-resolution topology sampling. (B) Sampled models are shown as black dots with their pseudo-energy score and RMSD100 relative to the X-ray-derived model. (C) The most accurate model predicted (blue) superimposed with the X-ray-derived model (purple, PDB entry 3TU3) from top and side views. (D) Alternative model (beige) predicted by the prediction pipeline superimposed with the X-ray-derived model (purple, PDB entry 3TU3).*

	Low-resolution I			Low-resolution II			High-resolution		
Setup	μ_{10} (Å)	τ_5 (%)	e	μ_{10} (Å)	τ_5 (%)	e	μ_{10} (Å)	τ_5 (%)	e
No data	6.0	0.0	0.6	4.9	<0.1	0.1	4.6	<0.1	0.1
SDSL-EPR data	5.1	<0.1	2.8	3.9	1.2	2.5	3.2	0.7	1.2

Table VII.1.: Prediction results for the C-terminal domain of ExoU with and without SDSL-EPR data. *Incorporation of SDSL-EPR data results in improved sampling accuracy and model discrimination, as shown by improvements in the average RMSD100 over the ten most accurate models sampled (μ_{10}), in the percentage of models with an RMSD100 less than* 5 Å *relative to the X-ray-derived model (τ_5), an in the enrichment (e).*

with SDSL-EPR data will ultimately be consistent with this conformation. The *de novo* prediction of the C-terminal domain of ExoU resulted in two dissimilar topologies (see figure VII.2 on the preceding page for details). Whereas one topology is represented by models exhibiting a structural dissimilarity to the X-ray-derived model as low as 3.2 Å, the other topology is structurally very dissimilar with an RMSD100 of about 12 Å relative to the X-ray-derived model. Notably, both topologies have comparable agreements with the SDSL-EPR data. The approach described in this study is orthogonal to the procedures used for obtaining the X-ray-derived model. Although there is not enough experimental data to rule out either of the two topologies, the partial convergence of the *de novo* method on the topology of the X-ray-derived model reassures its correctness. The topology that is structurally dissimilar to the X-ray-derived model arrives at a more favorable pseudo-energy score than the structurally similar topology (figure VII.2 on the previous page). However, this does not necessarily mean that the alternative topology is energetically more stable but could also be an artifact caused by inaccuracies of the free energy approximations. Artifacts like this have been observed in previous studies and might be eliminated by obtaining additional distance measurements.[2,5]

We were also interested in examining the experimental bimodal distance distributions observed for A629-A645 and Q649-S672 (see figure F.1 on page 192 for details). To evaluate the agreement of the X-ray-derived model of ExoU with the determined distance distributions, we performed explicit simulation of the distance distribution for the double-mutant A629C-A645C, as described in materials and methods. The double-mutant Q649C-S672C was not evaluated because the residue S672 was not resolved in the X-ray-derived model and modeling the missing coordinates would introduce additional bias. For the double-mutant A629-A645, explicit simulation of the spin labels did not result in a bimodal distribution but in a distinct peak around 25.5 Å (figure F.2 on page 193). For comparison, EPR spectroscopy determined two peaks: (19.3 ± 1.6) Å and (24.1 ± 1.9) Å (figure F.1 on page 192). Taking the accuracy limit of the X-ray-derived model and the fixed backbone during the explicit simulation into account, we conclude that the X-ray-derived model is in agreement with the measured mean distance of 24.1 Å. Additional simulations were performed for the remaining double mutants that had both spin labeling sites resolved in the X-ray-derived model. The simulated peaks matched the experimentally determined peaks well (figure F.2 on page 193) given the accuracy limit of the X-ray-derived model and the fixed backbone during the simulation. For one double mutant, E636-N657, the simulation resulted in two peaks: one peak around 18.5 Å, which agrees with the experimentally determined peak at (20.0 ± 3.6) Å and one peak around 13.5 Å, which would be too short to detect through the DEER experiment.

VII.2.6. Incorporating SDSL-EPR data increases the probability of sampling accurate models

De novo sampling of conformations through an MCM algorithm is a statistical process. The similarity of the sampled models to the X-ray-derived model corresponds to a normal distribution. To evaluate if incorporation of SDSL-EPR data increases the probability of sampling accurate models, the shift between the distributions resulting from *de novo* sampling with and without SDSL-EPR data can be compared. However, the more important aspect is the accuracy of the most accurate models alongside the percentage of accurate models. To quantify improvements in sampling accuracy, we compared the average RMSD100-values of the ten most accurate models, μ_{10}, for the two prediction runs. We chose to compare the RMSD100 averages over ten models instead of one to mitigate the effect of statistical outliers. Moreover, the percentage of models with an RMSD100 of less than 5 Å relative to the X-ray-derived model, τ_5, was compared. In addition, we also investigated whether incorporation of SDSL-EPR data results in increased clustering of the sampled models, as would be expected since the incorporated restraints should exclude conformations that significantly violate the EPR-derived restraints.

The results of the iterative protocol clearly demonstrate that incorporation of even a small number of SDSL-EPR distance restraints significantly increases the probability of sampling accurate models. Additionally, the most accurate models sampled arrive at an accuracy not observed for *de novo* protein structure prediction in the absence of SDSL-EPR data. This is demonstrated by changes of the μ_{10}-values, that improve from 6.0 Å to 5.1 Å with inclusion of SDSL-EPR restraints for the first iteration of the low-resolution topology sampling (figure VII.2 on page 105 and table VII.1 on the previous page). The τ_5-values for the first iteration were too low to be compared in a meaningful way. Another notable effect of incorporating SDSL-EPR data was an increased clustering of the sampled models, which is likely caused by the exclusion of models that are not in agreement with the experimental data. The improved clustering is demonstrated by improvements of the average Silhouette width (see section VII.2.3 on page 102 for details), which was 0.21 for the first iteration of low-resolution topology sampling without SDSL-EPR data and improved to 0.57 when experimental data was included. Due to the improved clustering, a more accurate set of models could be selected for the second iteration of the low-resolution topology when including SDSL-EPR data. This is demonstrated by more favorable μ_{10}- and τ_5-values after the second iteration, 3.9 Å and 1.2 % as compared to 4.9 Å and less than 0.1 % when no experimental data was used. This pattern propagated to the high-resolution refinement step. For prediction with SDSL-EPR data, the μ_{10}- and τ_5-values arrived at 3.2 Å and 0.7 % whereas they were 4.6 Å and less than 0.1 % for the prediction without using experimental data (table VII.1 on the previous page).

In conclusion, incorporation of limited experimental data from SDSL-EPR spectroscopy into *de novo* protein structure prediction results in excluding models that violate the restraints. Although the experimental data in this test case are too sparse to unambiguously determine the tertiary structure of the C-terminal domain of ExoU, the probability of sampling accurate models is significantly improved. This was demonstrated by improvements of μ_{10}- and τ_5-values, as well as improvements of the average Silhouette width of the clusters, which indicates an increased clustering of the sampled models.

VII.2.7. Incorporation of EPR data improves the discrimination of inaccurate models

Distinguishing between accurate and inaccurate models resulting from *de novo* protein structure prediction is typically hindered by the reduced resolution of the sampled conformations that result from the relatively coarse-grained approaches used to approximate a model's free energy. This was demonstrated by the prediction results in the absence of SDSL-EPR data. Although moderately accurate models with an RMSD100 of 4.7 Å relative to the X-ray-derived model could be sampled during the first iteration of the low-resolution topology search, the employed scoring function was unable to correctly distinguish between accurate and inaccurate models, which is indicted by an enrichment-value of 0.6 (see table VII.1 on page 106 for a tabulation of the prediction results). Accurate models were sampled with low probability resulting in an accordingly low density. As a consequence, accurate models could not be detected through clustering. This was demonstrated by the Silhouette scores that were 0.21 when SDSL-EPR data were used and 0.57 otherwise, indicating a broader range of conformations considered favorable in the absence of SDSL-EPR data. In general, incorporation of SDSL-EPR data significantly improved the scoring function's ability to distinguish between accurate and inaccurate models, which can be shown by comparing the enrichment-values that arrived at 0.6, 0.1 and 0.1 for the two iterations of the low-resolution topology search and the high-resolution refinement in the absence of experimental data, respectively, but upon incorporation of the SDSL-EPR data improved to 2.8, 2.5 and 1.2 (table VII.1 on page 106).

VII.2.8. Limitations in the sampling of conformations and the model discrimination remain

The most accurate full-atom model sampled by the presented pipeline arrives at an RMSD100 of 3.2 Å (figure VII.2 on page 105 and table VII.1 on page 106) relative to the X-ray-derived model (PDB entry 3TU3), which was reported at a resolution of 1.9 Å. Therefore, the most accurate model sampled is not within the accuracy limit of the experimentally determined reference structure. Assuming that the X-ray-derived model correctly and accurately represents the protein's major population in the equilibrium, the most accurate model does not capture the protein's tertiary structure at atomic detail. This may be attributed in part to necessary simplifications when sampling conformations. Neither the low-resolution topology sampling nor the high-resolution refinement exhaustively search the conformational space, and the most accurate model sampled could indeed be the most accurate model possible when using these methods in a *de novo* approach. For future studies, it will be worth investigating if this pipeline should be augmented with MD simulations.

Although the discrimination of inaccurate models could be improved substantially through the incorporation of SDSL-EPR data, as it was demonstrated through the improvements of the enrichment-values (table VII.1 on page 106), it is still not possible to reliably select the most accurate models. The models with the most favorable score cluster around RMSD100-values between 7 Å and 13 Å (figure VII.2 on page 105). However, the difference in pseudo-energy between the models with the most favorable pseudo-energy and the models with the most favorable RMSD100 relative to the X-ray-derived model accounts for less than 15 % of the most favor able score. This indicates that the discrimination problem could be resolved through additional SDSL-EPR distance measurements. In this initial study, the C-terminal domain of ExoU was predicted using only seven EPR-derived restraints, which in conjunction with the low-resolution translation of experimental distances into structural restraints is not sufficient to remove ambiguity from the prediction. Nonetheless, significant improvements were made even with this modest set of distance measurements, providing a valuable benchmark for further studies evaluating the impact of a more comprehensive set of constraints on *de novo* structure prediction.

VII.3. Conclusions

Using EPR spectroscopy in conjunction with *de novo* protein structure prediction provided an orthogonal approach to probe the structure of ExoU. The prediction converged on a conformation that is topologically identical and st ructurally similar (RMSD100 of 3.2 Å) to the X-ray-derived model in complex with its chaperone SpcU (PDB entry 3TU3). This result confirms that the fold of the ExoU C-terminal domain in solution matches the fold when in complex with its chaperon SpcU. From a different perspective, we established a protocol to predict a model of a soluble protein from limited SDSL-EPR data using a combined approach consisting of BCL::Fold, R, and Rosetta. This approach can be applied to all soluble proteins.

VII.4. Materials and methods

In this section, we detail the experimental methods used to obtain the SDSL-EPR data and the computational methods to explicitly simulate EPR-derived distance distributions *in silico*. This section is concluded by a description of the quality metrics used to evaluate the protein structure prediction results.

VII.4.1. DEER spectroscopy and determination of distance distributions

Four-pulse DEER data were collected on a Bruker E4580 pulse EPR spectrometer (Bruker Biospin) operating at Q-band (34 GHz), equipped with an EN5107D2 resonator and a 10 W microwave amplifier. Selected MTSL-labeled double-cysteine mutants of ExoU were prepared in 20 mmol dm^{-3} MOPS (34[N-morpholino]propanesulfonic acid), 145 mmol dm^{-3} NaCl, pH 7.2 using perdeuterated water and containing 25 % (v/v) perdeuterated glycerol as cryoprotectant. Samples containing a final protein concentration of approximately 0.1 mmol dm^{-3} in a volume of 12 μL were flash frozen in liquid N_2 and immediately placed in the resonator where sample temperature was maintained at 80 K using an Oxford cryostat. Data were background corrected and analyzed by model-free Tikhonov regularization using DeerAnalysis2011.[229]

VII.4.2. Explicit simulation of EPR-derived distance distributions

To further evaluate the agreement of the X-ray-derived model with the SDSL-EPR data, explicit simulation of the spin label distance distribution was performed for double-mutants that had both spin labeling sites resolved in the X-ray-derived model. For the explicit simulation, the endogenous residues at the spin labeling sites were replaced with R1A, which is a cysteine residue spin labeled with MTSL, using Rosetta's application for fixed backbone design — "fixbb". The resulting model was subsequently energy-optimized using Rosetta's "relax" application. The relaxation was constrained to the start coordinates to avoid introducing bias through Rosetta's scoring function. Constraining to start coordinates limited backbone perturbations to less than 0.1 Å. Per double-mutant, 1000 independent trajectories were simulated and the spin-spin distances observed in each trajectory were extracted to determine the spin-spin distance distribution.

VII.4.3. Quantitative evaluation of the protein structure prediction results

The accuracy of the structure prediction results was evaluated under two aspects: the sampling accuracy, which is the structural similarity between the sampled models and the experimentally determined reference structure, and the discrimination of inaccurate models, which is how well the employed scoring function could distinguish between accurate and inaccurate models. For quantifying the sampling accuracy, the RMSD100 metric[103] was used, which can be computed according to equation (VII.3).

$$RMSD100 = \frac{RMSD}{1 + \log \sqrt{l/100}} \tag{VII.3}$$

where:

$RMSD100$ = protein-size-normalized $RMSD$
$RMSD$ = root-mean-square distance of the C_α-coordinates
l = number of residues in the superimposition

To quantify the model discrimination, the enrichment metric[2] was used, which can be computed as $e = \frac{\#TP}{\#P}$. The sets TP and P are both subsets of the set of all sampled models. The set P contains the 10 % of the models with the lowest RMSD100 relative to the experimentally determined reference structure. The set TP is computed from the sets P and PS, which contains the 10 % of the models with the most favorable pseudo-energy score, as $TP = P \cap PS$. Therefore, the set TP contains the 10 % most accurate models that are at the same time among the 10 % of the models with the most favorable pseudo-energy score. Accordingly, the enrichment ranges from 0 to 10 and an enrichment-value of 1.0 indicates that the selection by the employed scoring function is purely random and discrimination of inaccurate models does not take place. Enrichment-values greater than 1.0 indicate that the scoring function is able to distinguish between accurate and inaccurate models, whereas enrichment-values less than 1.0 indicate that the scoring function is selecting against accurate models. An enrichment-value of 1.0 indicates that 10 % of the most accurate models can be identified by the scoring function.

VII.5. Acknowledgments

This project was supported by NIH grants GM114234 (to J.B.F.) and AI104922 (to D.W.F.). Work in the Meiler laboratory is further supported through NIH (R01 GM080403, R01 GM099842) and NSF (CHE 1305874). We want to thank the staff of the *Advanced Computing Center for Research & Education* (ACCRE) as well as Roy Hoffman and William C. Riner of the *Center for Structural Biology* at Vanderbilt University for computational support. Parts of the data analysis were performed using R[b] in conjunction with the ggplot2[123] and cluster[100] packages. Figures depicting protein models were created using Chimera[124] and composite figures were created using Inkscape.[c]

[b] https://www.r-project.org
[c] https://inkscape.org

CHAPTER VIII
PROTONATION-DEPENDENT CONFORMATIONAL DYNAMICS OF EMRE

This chapter is based on the publication "Protonation-dependent conformational dynamics of the multidrug transporter EmrE".[7] Axel W. Fischer contributed to the development of the prediction pipeline for the TPP-bound conformation of EmrE and its prediction.

The SMR transporter from Escherichia coli*, EmrE, couples the energetically uphill extrusion of hydrophobic cations out of the cell to the transport of two protons down their electrochemical gradient. While principal mechanistic elements of proton/substrate antiport have been described, the structural record is limited to the conformation of the substrate-bound state, which has been shown to undergo isoenergetic alternating access. A central but missing link in the structure/mechanism relationship is a description of the proton-bound state, which is an obligatory intermediate in the transport cycle. Here we report a systematic spin labeling and DEER study that uncovers the conformational changes of EmrE subsequent to protonation of critical acidic residues in the context of a global description of ligand-induced structural rearrangements. We find that protonation of E14 leads to extensive rotation and tilt of TMHs* 1 *to* 3 *in conjunction with repacking of loops, conformational changes, which alter the coordination of the bound substrate and modulate its access to the binding site from the lipid bilayer. The transport model that emerges from our data posits a proton-bound, but occluded, resting state. Substrate binding from the inner leaflet of the bilayer releases the protons and triggers alternating access between inward- and outward-facing conformations of the substrate-loaded transporter thus enabling antiport without dissipation of the proton gradient.*

VIII.1. Introduction

Powered by the proton electrochemical gradient across the inner membrane of prokaryotes, SMR transporters extrude a spectrum of cytotoxic molecules that are primarily hydrophobic cations.[24] The functional unit is typically a homodimer wherein each protomer consists of four hydrophobic TMHs. TMHs 1 to 3 cradle a substrate binding pocket while TMH4 is involved in the contacts that stabilize the dimer. EmrE, the SMR transporter from *Escherichia coli*, has been a focal point of structural, spectroscopic and mechanistic investigations.[230–235] Seminal work from the Schuldiner lab over the last two decades has unlocked mechanistic principles of substrate-ion-antiport.[23–26] EmrE binds hydrophobic substrates in a membrane-embedded chamber coordinated by glutamate 14 (E14), an absolutely conserved, membrane-embedded, acidic residue in the middle of TMH1. Coupling between substrate and proton arises from the principle of mutual exclusion between the two ligands at the binding site.[23,26]

In contrast to the elaborate understanding of EmrE antiport mechanism, the conformational changes that enable binding and release of ligands have not been elucidated. While the general framework of antiport is presumed to follow the principles of alternating access, the only structure available is of EmrE bound to the substrate TPP.[231] EM analysis of two dimensional crystals established an antiparallel orientation of the protomers in the dimer.[230] This was later confirmed by the corrected crystal structure of TPP-bound EmrE trapped in an asymmetric state with an opening on one side of the transporter.[231] This asymmetry arises from distinct conformations of each protomer in the dimer. Fleishman *et al.*[236]

independently modeled EmrE on the basis of the EM structure and postulated the elegant notion that isomerization of the two protomers between the two conformations drives alternating access of the substrate-loaded transporter. Indeed solution NMR studies confirmed the isomerization of TPP-bound EmrE in bicelles and enabled the measurement of the time scale of the exchange between inward and outward facing states.[234] Structurally, the inward- and outward-facing states of the dimer are related by a 180° rotation around an axis parallel to the plane of the membrane, *i.e.*, they are identical except for their orientations in the membrane.

In addition to alternating access of the substrate-bound EmrE, transport requires the protonation/deprotonation of E14 in the context of proton translocation from the extracellular milieu.[23] It is typically assumed that the protonated intermediate has a similar structure to substrate-bound EmrE and undergoes isomerization by the isoenergetic exchange of the two protomers between two conformations.[237] However, this model has not been challenged experimentally and the structures of the protonated and apo states have not been determined. Thus critical steps in the transport cycle remain structurally and dynamically unexplored.

To illuminate the structural and dynamic aspects of protonation in the transport mechanism, we present a global perspective on the ligand-induced conformational changes of EmrE. A systematic analysis of distances and distance changes between spin labels, site-specifically introduced across the dimer, reveals distinct structural rearrangements associated with protonation and substrate binding. These rearrangements reconfigure the backbone and side chain orientations in the substrate binding chamber as well as modulate access to the bilayer. Protonation-induced movements are primarily dependent on E14 although residues E25 and D84 appear to influence the local conformation particularly in TMH3 suggesting a departure from rigid body movement of the protomer. When integrated into the current biochemical and structural framework, these results provide a novel model of coupled transport by EmrE.

VIII.2. Materials and methods

VIII.2.1. Mutagenesis, expression, purification, labeling, and reconstitution of EmrE

This study utilizes the previously generated constructs of single-cysteine mutants of EmrE.[233] Specific functional regions of EmrE including residues in GG7 dimerization motif, located in TMH4 were excluded from the analysis. Further functional mutants (E14Q, E25Q, D84N, E25Q/D84N) were introduced on the same constructs using site-directed mutagenesis. EmrE mutants were expressed, purified in 1.5 % β-DDM and spin labeled using the same protocol as previously described.[233] Purified EmrE were concentrated with Amicon Ultra-10 kDa centrifugal filter units (Millipore). Samples for DEER spectroscopy were prepared in the 100 mmol dm^{-3} to 200 mmol dm^{-3} protein concentration range. A final concentration of glycerol of 30 % (wt/vol) was used in all samples as a cryoprotectant. The TPP-bound state was obtained by addition of six fold molar excess of the substrate TPP. For pH 5 and apo pH 8 states, respectively, a calibrated volume of 1 N HCl or 1 N NaOH was added to samples in EmrE size exclusion chromatography buffer (50 mmol dm^{-3} sodium phosphate monobasic, 50 mmol dm^{-3} NaCl, 0.02 % β-DDM, and 0.02 % NaN_3, pH 7.2). For pH titration experiments, an Orion 9810BN micro pH electrode (Thermo Scientific) was used to adjust the pH-values. See section VIII.2.2 on the next page for reconstitution of EmrE in nanodiscs.[238]

VIII.2.2. Reconstitution of EmrE in nanodiscs

Escherichia coli polar lipid extract (Avanti Polar Lipids) was dissolved in chloroform, evaporated to dryness on a rotary evaporator and desiccated overnight under vacuum. Membrane scaffold protein (MSP1D1) was expressed and purified as described earlier.[238] The lipids were hydrated in MSP buffer (50 mmol dm^{-3} Tris, 0.1 mol dm^{-3} NaCl, pH 7.5) containing 0.5 % (w/v) β-DDM, filtered through 0.2 µm polycarbonate membrane (Whatman, Florham Park, N J) and stored in small aliquots at −80 °C. For reconstitution into nanodiscs, purified spin labeled proteins in β-DDM micelles were mixed with lipid, MSP1D1, and β-DDM in the following molar ratios: lipid:MSP1D1, 60:1; MSP1D1:EmrE, 8:1; β-DDM:lipid, 5:1. Mixtures were rocked at room temperature for two hours. Biobeads (0.8 g/mL) were then added to the solution and incubated overnight at 4 °C with rocking. The nanodiscs assembly solution was filtered using 0.4 5 mm filter to remove biobeads. Full nanodiscs were separated from empty nanodiscs by size-exclusion chromatography (50 mmol dm^{-3} sodium phosphate monobasic, 50 mmol dm^{-3} NaCl, 0.02 % $\mathrm{NaN_3}$, and 5 % glycerol, pH 7.2). Nanodiscs were concentrated using Amicon Ultra-50 kDa centrifugal filters (Millipore). Proteo-nanodiscs were then characterized using sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) to verify reconstitution. Concentration of spin labeled mutants in nanodiscs was determined by comparing the intensity of the integrated continuous wave (CW)-EPR spectrum to that of the same mutant in detergent Michelle's.

VIII.2.3. CW-EPR and DEER spectroscopy

For CW-EPR, spin labeled EmrE samples were loaded in capillaries and spectra were collected on a Bruker EMX spectrometer using 10 mW microwave power level and a modulation amplitude of 1.6 G. DEER spectroscopy was performed on an Elexsys E580 EPR spectrometer (Super QFT bridge with ELDOR; Bruker) operating at Q-band frequency (33.9 GHz) with the dead-time free four-pulse sequence at 83 K.[75] Primary DEER decays were analyzed using a home-written software operating in the Matlab environment.[238] Briefly, the software carries out global analysis of the DEER decays obtained under different conditions for the same spin labeled position. The distance distribution is assumed to consist of a sum of Gaussians, the number of which is determined based on a statistical criterion.

VIII.2.4. Refinement of the X-ray structure and modeling the EmrE structure at pH 5

The X-ray structure of EmrE in TPP-bound state (PDB entry 3B5D) was refined in several iterations using MODELLER version 9.10.[239] A previously built complete atomistic model of dimeric EmrE was used.[233] *In silico* spin labeling of the protein structure using rotamer library approach was performed using MMM 2013.2 software package.[214] See section VIII.2.5 on the following page for more details on refinement using MODELLER. The stereochemical quality of the generated models was evaluated using PROCHECK (table VIII.1 on the next page).[240]

The initial model of the EmrE structure at pH 5 was predicted by a two-step approach using BCL::Fold,[2] which is part of the BCL,[a] to assemble the SSEs in the three-dimensional space and subsequently Rosetta[94,149] to construct loop regions and predict side chain conformations. See section VIII.2.6 on the following page for more details on BCL/Rosetta modeling. The generated models using this approach were further refined in MODELLER.

[a] http://www.meilerlab.org/bclcommons

Model	GA341 score[A]	Ramachandran statistics favored[B]	Ramachandran statistics allowed[C]	Main-chain parameters[D]	C_α-RMSD[E] (Å)
pH 5 (H^+-bound)	0.84	73.1 %	94.0 %	−0.1	0.7
Refined X-ray (+TPP)	0.99	72.0 %	95.6 %	−0.2	4.7
X-ray	—	—	—	—	4.1

Table VIII.1.: Quality assessment of the generated models of EmrE. *(A) A model is predicted as reliable when the GA341 score is higher than a pre-specified cutoff* (0.7)*. A reliable model has a probability of the correct fold that is larger than* 95 %*. (B) Percentage of residues in favored regions. (C) Percentage of residues in allowed regions. (D) Overall G-factor. The stereochemical quality of the generated models was evaluated using PROCHECK.*[240] *G-factor-values below* −0.5 *are considered unusual. (E)* C_α*-RMSD between the two chains.*

VIII.2.5. Refinement of the X-ray structure based on DEER distance constraints

The X-ray structure of EmrE in TPP-bound state (PDB entry 3B5D) was refined in several iterations using MODELLER version 9.10.[239] A previously built complete atomistic model of dimeric EmrE was used.[233] *In silico* spin labeling of the protein structure using rotamer library approach was performed using MMM 2013.2 software package.[214] In each iteration, the rotamer ensemble was first calculated at 298 K on the spin labeled positions. A rotamer, which best fits the mean N-O midpoint position of the whole ensemble was attached to the template structure that was supplied to MODELLER. The X-ray structure was refined using the center of the Gaussian components corresponding to TPP-bound state. Secondary structure assignments deduced from our previous study were also considered as constraints.[233] Only models with a GA341 score of 0.75 or higher were included in the ensemble. GA341 specifies a threshold, below which models should be rejected (table VIII.1). After importing models into the MMM software package, rotamer ensembles were recalculated on them. The stereochemical quality of the generated models was evaluated using PROCHECK (table VIII.1).[240]

VIII.2.6. Modeling the EmrE structure at pH 5

The protein structure prediction protocol consisted of a two-step approach using BCL::Fold,[2] to assemble the SSEs in the three-dimensional space and subsequently Rosetta[94,149] to construct loop regions and predict side chain conformations. In one of the models, a two-fold symmetry was imposed such that both protomers have similar conformations. In another model, the symmetry was relaxed to obtain an asymmetric dimer. Over the course of the optimization, transformations are applied to the SSEs to sample different conformations and the free energy of the sampled conformations was approximated using knowledge-based potentials, evaluating properties like the burial of residues, residue-residue interactions, and steric interference between residues. The knowledge-based potentials were supplemented with a scoring function based on the CONE model to quantify the agreement of the conformation with the EPR distance data.[2] To achieve sufficient coverage of the conformational space, 100 000 models were sampled for the pH 5 state using this protocol. The resulting models were sorted by their pseudo-energy score encompassing the scores from the knowledge-based potentials and the EPR agreement. The model with the most favorable score was selected for loop construction and refinement with Rosetta. The resulting model was refined using Rosetta relax[94] to improve agreement with the EPR data and a CONE model-based scoring function was used to quantify the agreement. Upon construction of all loop regions, the model with the most favorable score was selected. The generated models using this approach were further refined in MODELLER as described in section VIII.2.5 on the previous page.

VIII.2.7. General methodology

To investigate the EmrE structure under ligand conditions that are expected to promote transition between transport intermediates, we used our library of cysteine mutants,[233] spanning almost the entire sequence of EmrE, to introduce single spin labels into the monomer leading to a doubly-labeled dimer. Spin labeled mutants, which didn't show significant structural and functional perturbation in previous assays,[233] were analyzed by Q-band (33.9 GHz) DEER; also called PELDOR spectroscopy.[75,241,242] Distance distributions were determined in n-dodecyl-β-D-maltopyranoside (β-DDM) detergent micelles, which maintain the structural and functional integrity of EmrE.[243] The simplest model of EmrE transport satisfying energetics and coupling considerations entails at least three distinct intermediates: substrate-bound, proton-bound, and ligand-free or apo. It is likely that the apo state is only transiently populated considering the high concentration of protons on the extracellular side and the obligatory exchange between protons and substrate. Therefore, DEER measurements were carried out at pH 5 to mimic the acidic environment of the periplasm and protonate acidic residues, at pH 8 to mimic the relatively higher pH of the cytoplasm and promote deprotonation, and at pH 8 in the presence of excess TPP to trap the substrate-bound conformation. For a limited number of mutants, we confirmed our interpretation in lipid bilayers through reconstitution of EmrE into nanodiscs.

VIII.3. Results

VIII.3.1. Substrate binding and protonation induce distinct conformations of EmrE

The DEER data set (figure VIII.1 on the following page and figure G.1 on page 198) was transformed into distance distributions characterizing the spatial relationships between pairs of TMHs in the dimer as described in the methods section. Overall, the distributions reveal a number of trends consistent with three distinct conformations corresponding to the proton-bound, substrate-bound and apo EmrE intermediates. First, we observed changes in the average distances as well as the width of the distributions between the three conditions unequivocally demonstrating extensive conformational rearrangements (figure VIII.1 on the following page). Second, the shape of these distributions suggests that the TPP-bound state is ordered in stark contrast with the highly dynamic apo state. Specifically, we observed broad distributions in TMHs 1 to 3 at pH 8. At a number of sites, the shape of the distributions and the changes induced by different ligand conditions suggested equilibrium between multiple states. Finally, there are extensive, ligand-dependent, rearrangements in the structure of the loops (L1 to L4; figure G.1 on page 198), particularly L1 and L3 connecting TMHs 1 and 2 and TMHs 3 and 4 respectively, suggesting an important role for these segments in the mechanism of transport.[244] Notably, binding of substrate increases order in these loops whereas protonation/deprotonation typically leads to broad distance distributions (figure G.1 on page 198).

To provide a global perspective on the structural rearrangements, we plotted the change in the distance as a function of residue number (figure VIII.2 on the next page and figure VIII.4 on page 120). This is necessarily an oversimplification as many distributions particularly for the apo conformation are broad and can't be rigorously characterized by a unique distance. Nevertheless, this representation allows the qualitative visualization of the regions of conformational changes thereby identifying a

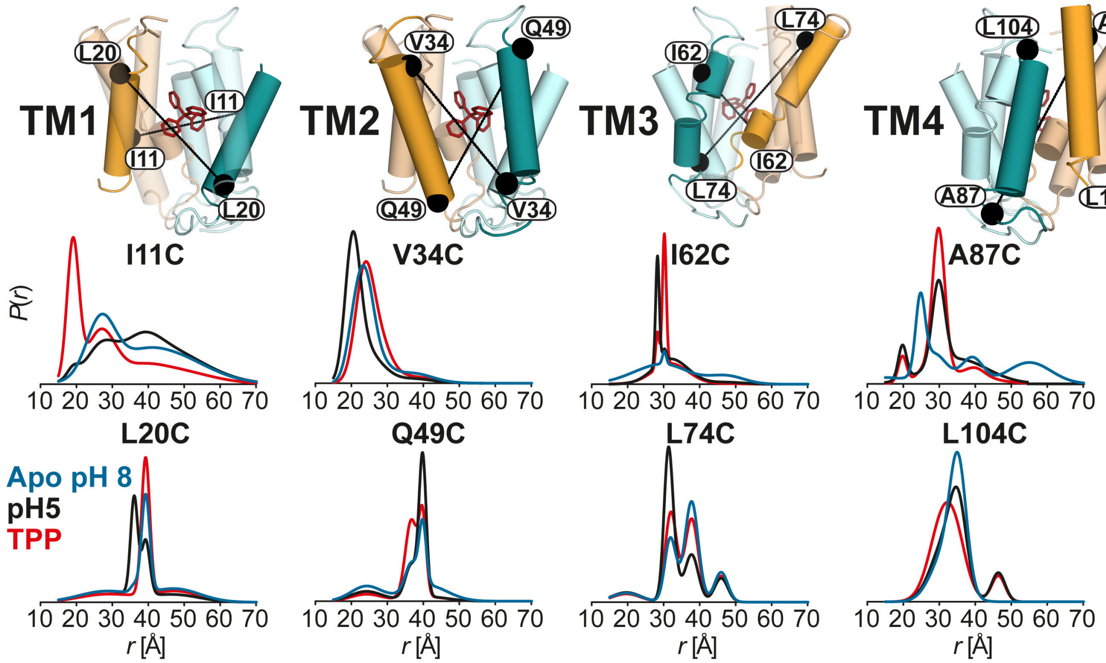

***Figure VIII.1.: Ligand-dependent conformational changes of EmrE in the transmembrane regions (TMHs 1 to 4.** Distance distributions depicting the probability of a distance P(r) versus distance (r) between identical positions in the doubly-labeled dimer. Distance distributions for each pair were obtained in the ligand-free (blue; apo pH 8), proton-bound (black; pH 5) and TPP-bound (red; TPP) intermediates.*

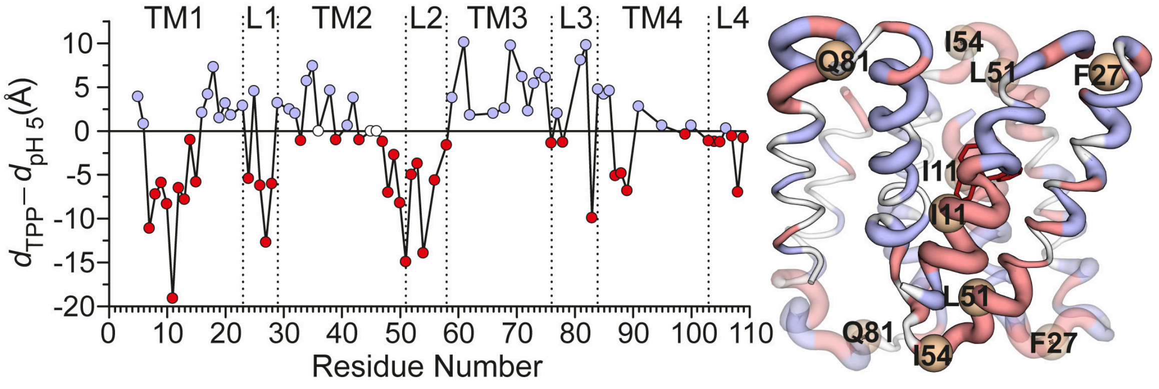

***Figure VIII.2.: Ligand-dependent changes in the distance as a function of residue number for TPP to proton-bound intermediates ($d_{TPP} - d_{pH5}$).** The absolute value of the change in the distance between the two states is displayed by the ribbon thickness on the TPP-bound crystal structure. Residues with positive and negative distance change are colored blue and red, respectively. Unchanged residues and those where no data was obtained are colored white. See figure VIII.4 on page 120 for more detail.*

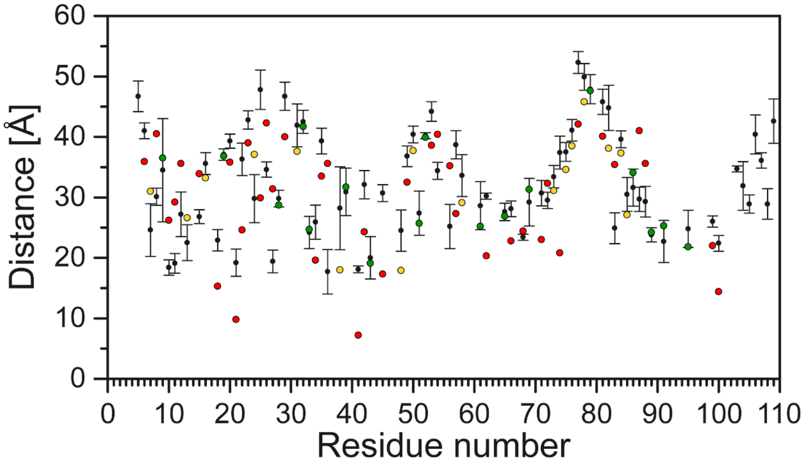

Figure VIII.3.: Comparison of the obtained experimental distances for the TPP-bound state of EmrE with distances predicted from the crystal structure. *The average distances (colored circles) predicted by the X-ray structure were calculated using the MMM 2013.2 software package.*[214] *The color code indicates whether distances are matched (green, within the standard deviation of the specified experimental distance), poorly matched (yellow, within twice the standard deviation), or not matched (red, deviating by more than twice the standard deviation). Only experimental distances between residues* 6 *and* 100 *are compared since the rest are not resolved for both monomers in the structure.*

complex web of structural rearrangements focused on TMHs 1 to 3 and the loops as highlighted by the width of the ribbon representation of EmrE in figure VIII.2 on the previous page.

VIII.3.2. The TPP-bound conformation

Because the current structural data is exclusively for the TPP-bound state, we will consider it as a reference for the purpose of interpreting the distance distributions in a structural context. Previous work from our laboratory provided a complete view of the accessibility and mobility profile of spin labels in this state.[233] The data pointed to extensive disagreement between the crystal structure and the conformation in liposomes. The nature of these disagreements and the low resolution of the structure lead to the conclusion that there are substantial issues with the orientation of helices in the crystal structure.[233]

The isomerization between inward- and outward-facing conformations detected by NMR is not expected to lead to changes in the distance distributions since the packing of the two protomers in the asymmetric dimer is identical.[234] Therefore, we compared the experimental distances to those predicted from the crystal structure and found discrepancies that extend across the entire structure (figure VIII.3) further confirming the conclusion deduced from the accessibility and mobility analysis regarding the accuracy of the structure. Nevertheless, the crystal structure (PDB entry 3B5D) will be presented in the figures to provide a general reference for the location of the labels. A model of the TPP-bound conformation, refined to agree with the experimental distances, will be presented in the discussion section.

VIII.3.3. Conformational changes induced by protonation

Comparison of the pH 5 and TPP-bound distance distributions reveal extensive structural changes upon protonation. These are primarily observed at residues in TMHs 1 and 3, which are directly involved in substrate binding, and TMH2, which borders the bilayer. In contrast, minor rearrangements along the interface of TMH4 are reported by the spin labels. Large changes in the average distances and the width of the distance distributions are observed along the N-terminal part of TMH1 (residues 8 to 15) with the largest changes occurring at sites with restricted mobility previously assigned to helix/helix interfaces (*i.e.*, residues 11 and 15). Therefore, we interpret these distance changes as reflecting a degree of rotation rather than a simple tilt of the helices. Rotation is expected to alter spin label mobility at helical interfaces and consequently lead to large distance changes due to the repacking of spin label side chains. In contrast, helix tilt would be expected to yield smaller amplitude distance changes and to be observed concurrently at lipid-facing sites.

Conformational changes at pH 5 include the C-terminal end of TMH 1 predominantly in the form of helix tilt. The different nature of rearrangements along the two segments of TMH 1 presumably necessitates a hinge point or an unwinding in the helix. Consistent with this conclusion, the distance distributions at 12 and 13 are broad suggesting flexibility of the backbone.

Large amplitude distance changes are observed in loop 1 (figure G.1 on page 198 and figure VIII.2 on page 116) indicating extensive ligand-dependent repacking. While the exact nature of the underlying structural rearrangements is difficult to infer from the data, they suggest that the loop plays a central role in the occlusion/exposure of the substrate binding site. Notable is residue 27, where a short distance component is observed in the apo and in the TPP-bound state whereas the pH 5 distribution suggests a highly disordered conformation. The penetration of this residue towards the binding site, implied by the short distance component, presumably stabilizes the bound substrate.

The N-terminal segment of TMH 2, consisting of residues 31 to 35 to, undergoes a closing motion upon protonation evidenced by the shift in the average distance. A residue by residue analysis of the 37 to 44 stretch is hindered by the close proximity of spin labels (less than 20 Å).[75] However, the distributions at sites 41 and 42, where DEER decays can be analyzed, show the persistence of the closing trend. This movement is attenuated near the end of the helix although changes in the distance distributions are detected at residues 48 and 49 (figure VIII.1 on page 116). As TMH 2 merges into loop 2, we observed evidence of ligand-induced changes in order manifested by large changes in the width of the distributions at residues 50 and 51.

Except for residues 52 and 53, most residues in loop 2 were characterized by broad distributions indicating a highly dynamic backbone (figure G.1 on page 198 and figure VIII.2 on page 116). The distributions consist of a well-defined component and a broad underlying component, which we interpret as reflecting the existence of one conformation wherein the loop backbone is rigid. This conformation is stabilized by ligand binding and is reduced in the apo intermediate.

The N-terminus of TMH 3 is in direct contact with the bound substrate in the crystal structure. Moreover, previous mutagenesis studies implicated residues in this helix in substrate binding.[245] Distributions at residues 62 and 68 and CW-EPR spectrum at residue 64 suggest that the N-terminal part of TMH 3 undergoes repacking between proton- and TPP-bound states but a quantitative interpretation is hindered by the broad distributions and the close distances (figure VIII.1 on page 116). Beginning at residue 71, protonation invariably leads to a distinct short distance component not observed in the TPP-bound state suggesting a closing movement in this region of the transporter (figure VIII.1 on page 116). This movement is likely facilitated by the GVG motif in TMH 3, which displays changes in its dynamics in the absence of substrate.[233,246]

The conformational changes at the C-terminal half of TMH 3 propagates to the loop linking it to TMH 4 (figure G.1 on page 198 and figure VIII.2 on page 116). Remarkably, we observed distinct evidence of short components in the pH 5 state (e.g. W76, Q81, R82, and D84) that imply a large scale closing movement of the loop indicative of an occlusion of the substrate binding cavity in the absence of substrate.

Although noticeable ligand-induced changes in the distributions are observed at several sites in TMH 4 (figure VIII.1 on page 116), they are generally smaller in magnitude and no discernable pattern was evident from comparison of changes at successive sites. Given that this TMH is involved in dimer formation, it is not unexpected that the rearrangements at TMHs 1 and 3 necessitate repacking at the dimer interface.

VIII.3.4. The apo state

The profile of the apo state that emerges from the distributions at pH 8 is that of a highly dynamic conformation (figure VIII.4 on the next page). Broad distributions indicative of conformational sampling are observed along the N-terminal parts of TMHs 1 and 3 as well as in the loops (figure G.1 on page 198). However, there are notable exceptions that occur at functionally important residues. Specifically, the apo state distributions at residues 14 and 18 are narrower than in the ligand-bound state suggesting that the substrate binding site may become occluded in the absence of ligands.

Unexpected short distance components at residues in loop 3 (figure G.1 on page 198) are indicative of a large amplitude excursion for this segment, which would be at either side of the membrane in an antiparallel structure. Such movement may be associated with an occlusion of the substrate binding region near TMH 3 similar to what is observed at pH 5. These results rationalize previous accessibility data that revealed simultaneous large exposure of spin labels in loop 3 to NiEDDA and molecular oxygen.[233]

VIII.3.5. Conformational dynamics in lipid bilayers

To investigate ligand-induced conformational changes in lipid bilayers, we carried out DEER measurements on representative spin labeled EmrE mutants, which have been reconstituted in lipid nanodiscs (figure VIII.5 on page 121). Comparison of the distance distributions demonstrates pH- and substrate-dependent rearrangements along the same structural elements as in detergents. More importantly the sign of the distance changes is identical in detergent micelles and lipid bilayers suggesting that similar conformations are stabilized by protonation and substrate binding in the two environments. However, we found that, for the majority of the residues investigated, the width of the distributions was narrower in lipid bilayers indicating a more ordered/less dynamic structure. These were particularly notable for residues in the center of TMHs 1 and 2 and the C-terminal part of TMH 3 (e.g. residue 12 in TMH 1; figure VIII.5 on page 121). In contrast, reconstitution in lipid bilayers either didn't affect the disorder of the loops or, more notably, promoted fluctuations by loop 3 as evident by the increase in the population of the short component at residue 76 (as seen in figure VIII.5 on page 121).

VIII.3.6. The proton sensor of EmrE

We determined the pK_a of the conformational transition of ligand-free EmrE using G26 as a reporter (figure G.2 on page 199 and figure G.3 on page 200). A titration curve was constructed by determination

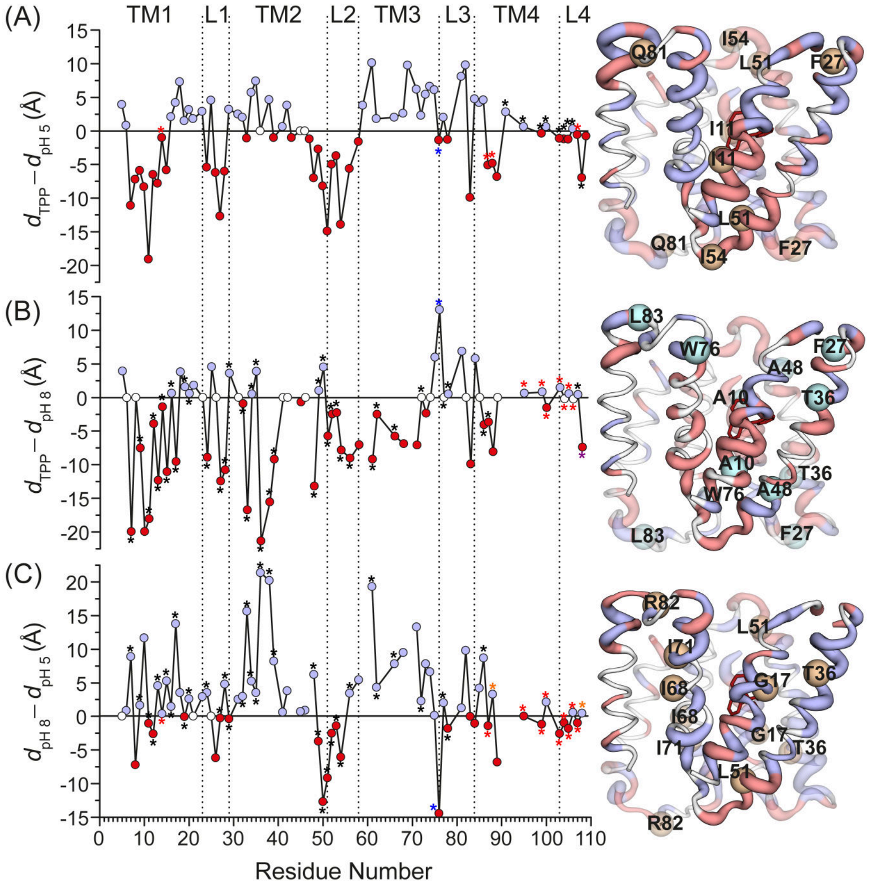

Figure VIII.4.: Ligand-dependent changes of EmrE in the distance as a function of residue number. *The absolute value of the change in the distance between the two states is displayed by the ribbon thickness on the TPP-bound crystal structure. Residues with positive and negative distance change are colored blue and red, respectively. Unchanged residues and those where no data was obtained are colored white. (A) TPP- to proton-bound* ($d_{TPP} - d_{pH5}$)*. Residues 14, 87, 88 and 107 (red asterisk): for protonated (pH 5) state, weighted average distance of the distribution is used and for TPP-bound state the center of the Gaussian component corresponding to this state is used. Residues 91, 95, 99, 100, 103 to 106, and 108 (black asterisk): for both states weighted average distance of the distribution is used.(B) TPP-bound to apo* ($d_{TPP} - d_{pH8}$)*. Residues 7, 9, 11 to 17, 19, 20, 24, 27 to 29, 32 to 36, 38, 39, 48 to 54, 56, 61, 62, 66, 72, 77, 78, 86, 87, and 107 (black asterisk): for apo (pH 8) state, weighted average distance of the distribution and for TPP-bound state, the center of the Gaussian component corresponding to this state are used. Residue 108 (purple asterisk): for TPP-bound state, weighted average distance of the distribution and for apo state, the center of the Gaussian component corresponding to this state is used. Residues 95, 99, 100, and 103 to 106 (red asterisk): for both states, weighted average distance of the distribution is used. (C) apo to proton-bound* ($d_{pH8} - d_{pH5}$)*. Residues 7, 9, 11 to 13, 15 to 17, 19, 20, 24, 27 to 29, 32 to 36, 38, 39, 48 to 54, 56, 61, 62, 66, 72, 77, 78, and 86 (black asterisk): for apo (pH 8) state, weighted average distance of the distribution and for protonated (pH 5) state the center of the Gaussian component corresponding to this state is used. Residues 88 and 108 (yellow asterisk): for protonated (pH 5) state, weighted average distance of the distribution and for apo state the center of the Gaussian component corresponding to this state is used. Residues 14, 87, 95, 99, 100, and 103 to 107 (red asterisk): for both states the weighted average distance of the distribution is used.*

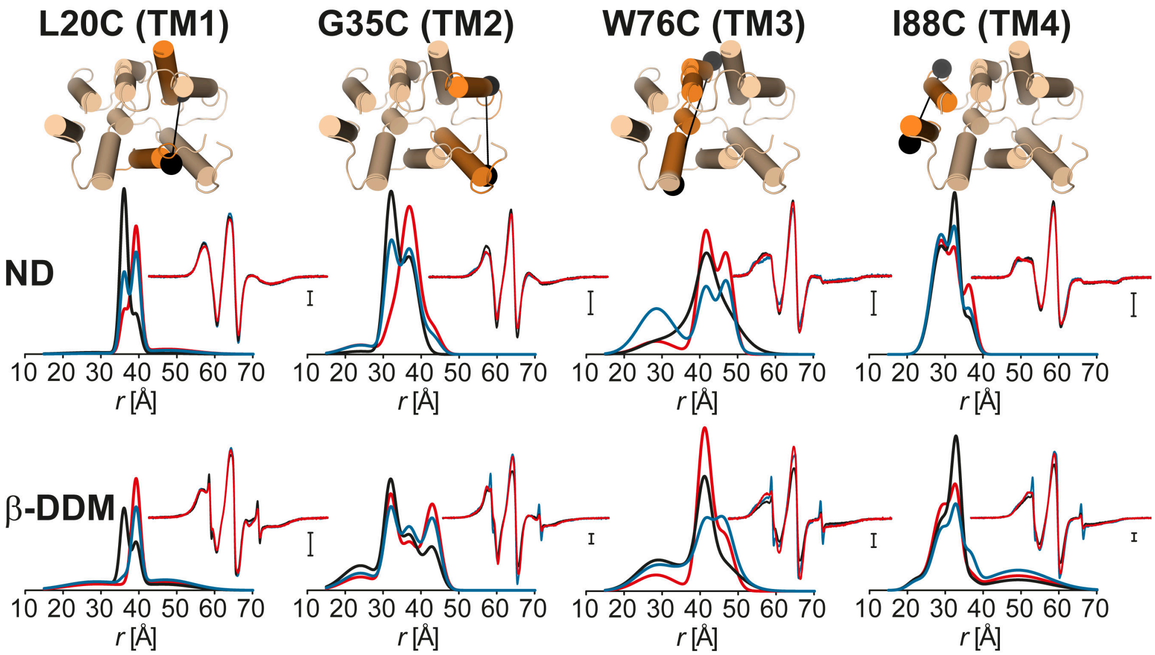

Figure VIII.5.: Ligand-dependent conformational changes of EmrE in nanodiscs composed of **Escherichia coli** ***polar lipids.*** *The distributions in DDM micelles are shown for reference. Apo (blue), proton-bound (black;* pH *5 for β-DDM,* pH *6 for nanodiscs), and TPP-bound (red). The CW-EPR spectra are shown in the insets.*

of the amplitudes of the distance components in the pH 5 to pH 10 range. pK_a-values of approximately 7.7 and 7.4 were obtained in detergent micelles and lipid bilayers, respectively, indicating that the environment of E14 was not substantially altered in lipids. Previous kinetic analysis yielded a pK_a-value for E14 of about 7.3 although a wider range was reported from steady state analysis.[25] The similar values of the pK_a suggest that the conformational changes detected by EPR are involved in the antiport mechanism of EmrE.

Consistent with this conclusion, we found that the pH-induced distances changes in TMH 1 are primarily associated with the protonation of E14 (figure G.3 on page 200). Substitution of E14 with glutamine designed to mimic protonation abrogates the distance changes due to the shift from pH 8 to pH 5 in TMH 1 (e.g. residues 11, 15 and 20; figure G.2 on page 199 and figure G.4 on page 201). More importantly, the distributions at pH 8 in the E14Q background are similar to those at pH 5 in the wild type (WT), suggesting that the E14Q mimics protonation of E14.

In TMH 2, the E14Q substitution also attenuates the pH-dependent distance changes but unlike TMH 1 the shape of the distribution at pH 5 is noticeably broader than in the WT background (figure G.4 on page 201). To identify the residues involved in modifying the effects of E14 protonation, we introduced the D84N/E25Q substitutions while monitoring selected sites (figure G.2 on page 199 and figure G.4 on page 201). We observed an increase in disorder primarily for the pH 8 conformation and to a lesser extent the pH 5 conformation in this background. Consistent with a role for the protonation of these residues in the transport mechanism, we found that while E14Q substitution attenuated the distance changes in TMH 3, it also introduced a large degree of disorder reflected in the width of the distance distributions (e.g. I62; figure G.2 on page 199 and figure G.4 on page 201). We interpret this result as

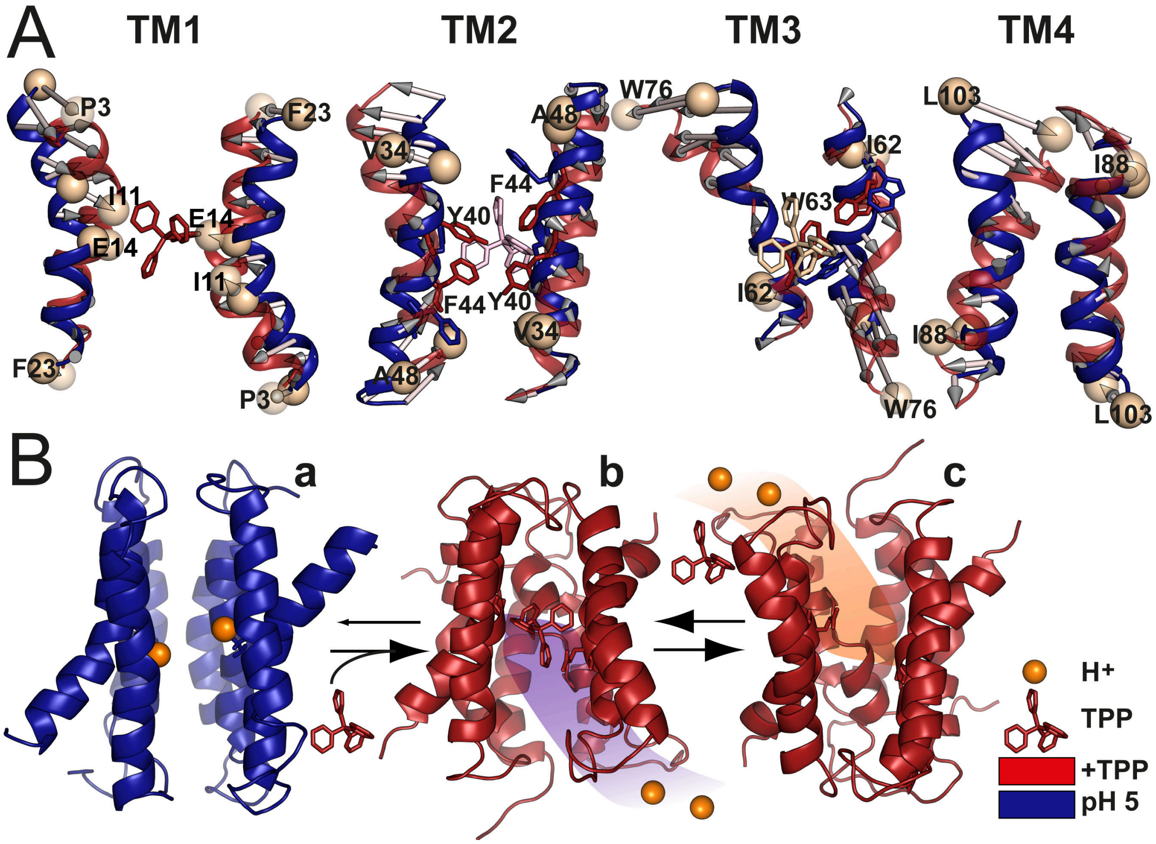

Figure VIII.6.: ***Model of EmrE transport derived from EPR data.*** *(A) Conformational changes between protonated (pH 5) and TPP-bound intermediates. Overall alignment between respective TMH pairs is shown; for TMH 3, alignment was based on residues 58 to 64. (B) The resting state is a protonated but water-occluded conformation of EmrE,represented here with the symmetric model generated by BCL::Fold/Rosetta and refined in MODELLER by pH 5 distance restraints (a). Subsequent binding of the substrate from inner membrane leaflet promotes release of the protons yielding the refined TPP-bound crystal structure (b). Conformational exchange of the monomers enables alternating access to the extracellular milieu and exchange of substrate with protons resumes the cycle (c).*

reflecting the change in protonation state of D84 and E25. Importantly, conformational changes in loop 3 are primarily determined by the protonation of D84 rather than E14 (figure G.4 on page 201) consistent with previous reports.[244] Together the data reveal that pH-induced structural changes, while primarily mediated by E14, are affected by other acidic residues such as D84.

VIII.4. Discussion

The results presented above reveal extensive, functionally relevant, conformational changes as a consequence of protonation/deprotonation of glutamate 14. Rotation and tilting of TMHs 1 to 3, which together form the substrate/proton binding site, not only reconfigure access to the cavity but presumably modulate substrate affinity through reorientation of side chains. The inter-helical loops emerge as central elements in this protonation switch undergoing extensive repacking. The large, ligand-dependent distance changes cannot arise from the isoenergetic alternating access through simple conformational exchange of each protomer because the two resulting dimer structures are identical except for their orientation relative to the bilayer. Rather the distance changes reflect the population of novel conformations that are primarily stabilized by the binding/dissociation and protonation/deprotonation of E14. In addition, the data uncovered a contribution of E25 and/or D84 to the conformational switch primarily through extensive repacking of loop 3. Finally, deprotonation and substrate release appears to induce a high level of disorder suggesting large amplitude equilibrium fluctuations in multiple structural elements notably TMH 3. However, structural interpretation of the distance distributions in this apo state is hindered by the breadth of these distributions.

To highlight the conformational changes induced by protonation in a structural context, we carried out detailed *de novo* modeling of the pH 5 conformation using BCL::Fold and Rosetta.[2,94,149] The conformational search was restrained by the experimental distances at pH 5.[2] A two-fold symmetry was imposed such that both protomers have similar conformations in contrast to the crystal structure where the two protomers have distinct conformations and form an asymmetric dimer. However despite the extensive nature of the distance restraints and their coverage along the protein sequence, compatibility of the DEER data with an asymmetric dimer cannot be excluded; primarily due to the uncertainty in translation of measured distances between the spin labels into backbone structural restraints.[2,149] However, as discussed below, we consider the asymmetric dimer to be less mechanistically plausible for the protonated state. The TPP-bound crystal structure was refined using MODELLER[239] restrained by the distances obtained in the TPP-bound conformation following a protocol introduced by Jeschke *et al.* (figure VIII.6 on the preceding page).[214] To ensure that the spin label side chains were treated similarly in both models, the *de novo* pH 5 models were further refined using the MODELLER protocol (figure VIII.6 on the previous page, figure G.5 on page 202, and figure VIII.7 on the next page).

The conformational changes gleaned from comparing these models suggest plausible and previously unappreciated mechanistic elements of EmrE transport (figure VIII.6 on the preceding page, figure G.5 on page 202, and figure VIII.7 on the following page). The flexibility of TMH 1, presumably a consequence of the two consecutive glycines at positions 8 and 9, enables large scale reconfiguration upon concurrent substrate dissociation and protonation of E14. Not only does the distance between the two TMHs 1 increases (figure VIII.6 on the previous page), but extensive rotation of the N-terminal part alters the side chains orientations in the substrate binding site (figure VIII.6 on the preceding page). The rearrangements of the backbone and side chains of TMHs 1 and 3 may provide the mechanism to lower the affinity to substrate in conjunction with the competition by protons for binding to E14 (figure G.5 on page 202).

Although substrate access to the binding site is typically represented as occurring from the cytoplasm, a competing model posits that for hydrophobic substrates, such as those of EmrE, partitioning is likely to occur from the inner leaflet of the bilayer.[247] Consistent with this model, we observed that rotation/tilting of TMH 2 (figure VIII.6 on the preceding page) swings open a gate consisting of Tyr 40 and Phe 44 thereby enabling direct access to and from the bilayer to the binding site.

Substrate dissociation and protonation induces repacking of the N-terminus of TMH 3, which participates in the coordination of the substrate as shown in figure VIII.6 on the previous page, and figure VIII.7 on the following page). In concordance with the conclusion from the EPR accessibility data,[233] the C-terminal part of TMH 3 undergoes large amplitude movement coupled to extensive repacking of loop 3 (figure VIII.6 on the previous page). This movement is controlled by the protonation state of E25 and/or D84 suggesting a degree of decoupling of this loop from the protonation state of E14.

How may these conformational changes enable coupled transport? Because of the high concentration

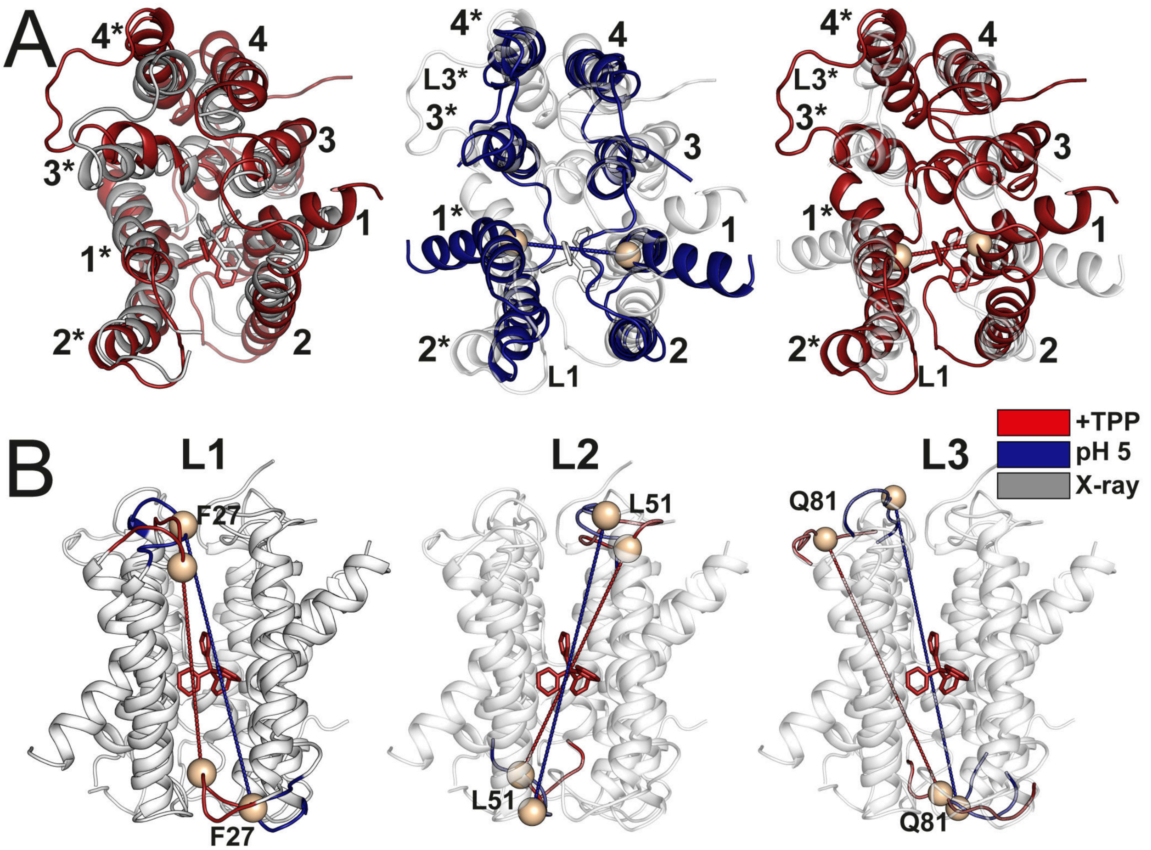

Figure VIII.7.: Structural comparison of the generated models of EmrE and the X-ray structure. *(A) Overall structural alignments between X-ray structure and the refined model (right), refined X-ray structure and the symmetric protonated model (middle and left) are shown. (B) Conformational changes in the loop regions (L1 to L3) between TPP-bound and protonated models.*

of protons on the extracellular side, we envision that in the absence of substrate, E14 is protonated and the transporter is in the protonated conformation (corresponding to pH 5 here; figure VIII.6 on page 122). Measured values of E14's pK_a would imply proton leakage unless the structure is proton-occluded, which for EmrE would require a symmetric conformation wherein E14 is not exposed to the pH of the cytoplasm. Therefore, we propose that the protonated conformation does not undergo the isoenergetic alternating access and thus is symmetric as shown in figure VIII.6 on page 122. Substrate binding to this conformation, which occurs from the inner leaflet of the bilayer through the TMH 2 fenestration, releases the two protons by stabilizing the asymmetric TPP-bound conformation. Through its isoenergetic alternating access,[234,236] this state exposes the substrate to the extracellular milieu at which point protons displace the substrate enabling a new cycle of transport (figure VIII.6 on page 122).

VIII.5. Acknowledgments

This work is dedicated to Dr. Hanane A. Koteiche who initiated this research direction in the laboratory. We especially thank Dr. Sepan T. Amadi and Sanjay Mishra for their preceding studies on EmrE which paved the way for this work. We thank Drs. Richard Stein and Derek Claxton for critical reading and editing of the manuscript. This work was supported by NIH Grants R01 GM077659 (to H.S.M.) and R01 GM080403 (to J.M.).

CHAPTER IX
DISCUSSION AND FUTURE DIRECTIONS

The body of the work presented in this dissertation details the development and application of methods to predict protein structures and ensemebles either purely *de novo* or in conjunction with limited experimental data from EPR and XL-MS experiments. Some limited evaluation was also performed in conjunction with spectroscopic data from NMR-NOE experiments.

This chapter is focused on the discussion of this work and the future directions suggested by the presented results. The specific findings include the architecture of the employed prediction pipelines, which is discussed in section IX.1, the influence of limited experimental data on the prediction accuracy, which is discussed in section IX.2 on page 128, and the general importance of computational modeling for structural biology, which is discussed in section IX.3 on page 130. The conclusions are followed by a discussion of the future directions for protein ensemble prediction in section IX.4 on page 131, which focuses on the potential of *in silico* simulation of spectroscopic data from EPR experiments to estimate a conformation's population size in the protein's equilibrium.

IX.1. Hierarchical structure prediction pipelines provide an efficient approach for structure and ensemble prediction

One of the main challenges of protein structure and protein ensemble prediction is the vast size of the conformational space. With the exception of proline, the backbone of each canonical amino acid exhibits two rotational degrees of freedom. If the side chains are also considered, two additional rotatable have to be modeled for each amino acid on average. In consequence, four rotatable bonds have to be modeled per amino acid on average, what would consume an unfeasible amount of computational resources.

In this work, different approaches to design a protein structure or ensemble prediction pipeline were evaluated. The study detailed in chapter II on page 12 was conducted in the context of the CASP experiment and demonstrated that exhaustive sampling of all possible conformations is not necessary for achieving sufficient coverage of the conformational space. Instead, the developed pipeline relies on the assumption that it is possible to predict inflexible dihedral angles with sufficient accuracy. In particular, it is possible to reliably predict the SSE type of a specific sequence span and its environment — if it is membrane-spanning or located in the cytosol. Using these assumptions, the size of the sampling space can be significantly reduced by allowing only limited deviations from idealized dihedral angles for sequence spans predicted to be α-helices or β-strands. Using such a rather coarse-grained representation of the protein's tertiary structure also allows further simplifications like an implicit representation of the residues' side chains through "superatoms", which eliminates exhaustive sampling of the side chain conformations. Although this approach is not able to account for substantial deviations from idealized dihedral angles and therefore sample a protein's tertiary structure at atomic detail, it enables the creation of a hierarchical prediction pipeline.

A general hierarchical protein structure and ensemble prediction pipeline employs different sampling and scoring module going from low-resolution to high-resolution. The pipeline detailed in this work consists of three modules: 1.) a low-resolution topology sampling module, 2.) a high-resolution sampling and refinement module, and 3.) an MD module for further refinement and stability evaluation of the sampled protein structures.

The low-resolution topology sampling module consists of the BCL::Fold algorithm[59] that assembles predicted SSEs in the three-dimensional space using an MCM algorithm. By only allowing limited deviations from idealized dihedral angles, this module is capable of sampling more topologies within the same number of CPU cycles. A low-resolution scoring function relying on knowledge-based potentials is employed to correspond with the coarse-grained sampling..[60] This module produces a low-resolution model only, without side chains or pronounced deviations from idealized dihedral angles.

The high-resolution sampling and refinement module consists of the Rosetta software suite[94] for molecular modeling. This module also does not employ exhaustive sampling of the dihedral angles, but uses structure fragments collected from the PDB to explore the conformational space. In this context, sequence spans in the protein model are replaced with the collected fragments to sample different conformations while favoring conformations that have been observed in experimentally determined structures before. The scoring functions employed by this module perform a high-resolution approach, evaluating atomic details like Van der Waals interactions that are evaluated using the Lennard-Jones potential. The conformations of side chains are sampled using a rotamer library.

The MD module used the Amber package[102] in conjunction with the ffSB98ildn force field.[104] For each sampled protein structure, a 50 ns NPT production run using Langevin dynamics was performed. Subsequently, hierarchical clustering with complete-linkage was used to identify all sub-states. From each cluster, one representative was selected and reevaluated using the Rosetta scoring function. Subsequent selection of the final models was based on their Rosetta pseudo-energy score.

These three prediction modules were connected through a clustering module that employed a *k*-means approach[100] using the RMSD as dissimilarity metric. From each cluster, the medoid was selected for further processing. Using this approach, the number of models subjected to high-resolution refinement and MD simulations could be substantially reduced without losing representations of the topological space. Topological duplicates were only present at the low-cost first prediction module. The second and third modules that require significantly more computational resources were only applied to topologically distinct protein models.

Employing this pipeline enabled the *de novo* prediction of more than 55 proteins — the majority of them in two to five setups with different sets of experimental data from predicted contacts, from XL-MS experiments, and from NMR-NOE experiments — in a period of less than three months. For the majority of the proteins, the pipeline was able to sample the correct topology of the protein and therefore providing a starting point for the computationally more expensive refinement methods. However, for several targets it was not possible to select the most accurate models. This was especially pronounced in cases when no additional structural information in form of experimental data was available for the prediction.

In conclusion, development of a protein structure and ensemble prediction pipeline that can deal with large proteins makes it necessary to account for the limited computational resources available. Full-atom prediction and simulation of protein structures can take days or even months, depending on the size of the protein, when done *de novo*. A hierarchical prediction pipeline reduces the required number of CPU cycles by applying high-cost algorithms only to a small and distinct subset of the conformational space. However, additional improvements are needed for model selection, which currently cannot be performed reliably. Improvements in the incorporation of experimental data might mitigate this problem. This is further discussed in section IX.2 on the following page and section IX.4 on page 131 suggests a new approach through *in silico* simulation of the experimental data.

IX.2. Incorporation of limited experimental data improves the accuracy of structure prediction

The free energy of a conformation is determined by a complex set of interactions that happen within the protein itself or between the protein and its environment — e.g. hydrogen bonding between the protein's residues and the cytosol or hydrophobic interactions between the protein's residues and a cell membrane. An accurate computation of a conformation's free energy would require a comprehensive simulation of the crowded cell and a representation of the cell's molecules at an atomic level. Additional simulations would have to be performed to obtain an estimate of the system's conformational entropy. For the resulting atomic system, a quantum mechanical evaluation would have to be performed. Since these calculations are not feasible within an acceptable time span, approximations to the system representation have to be applied. Typical approximations to the system are implicit representations of the protein's environment like the cytosol and membrane or even representation of a residue's side chains through a single "superatom". These simplifications require an adaptation of the approach to compute the conformation's free energy since not all information is present to perform a full-atom evaluation. Typically, these adaptations result in ambiguities making it difficult to distinguish thermodynamically stable from unstable conformations. Examples of this were observed in chapter II on page 12, chapter IV on page 48, chapter VI on page 83, and chapter VII on page 99. For those studies, pure *de novo* prediction frequently resulted in ambiguous results and uncertainty about the most stable conformations.

Besides techniques like X-ray crystallography or NMR spectroscopy there are other spectroscopic techniques that can provide insight into the structure of a protein. Two examples are EPR spectroscopy and XL experiments in conjunction with MS. EPR spectroscopy can yield two types of information: the distance between the two spin-labeled residues in form of a distance distribution and the accessibility of the spin-labeling site to a paramagnetic relaxation agent. XL experiments in conjunction with MS yield the maximum distance between the two cross-linked residues. Although providing important structural information about the protein, both techniques are typically not able to unambiguously determine its structure.

However, incorporating the limited experimental data from these techniques into computational structure and ensemble prediction methods solves two problems. The computational approach can fill the information gaps in the experimental data and the geometric interpretation of the experimental data resolves ambiguities in the free energy approximation of the computational approach. The resulting questions concern the way how to incorporate the experimental data since neither the *in vitro* nor the *in vivo* system can be simulated comprehensively

For experimental data from EPR experiments, I developed and evaluated pipelines incorporating distance and accessibility measurements. The distance data was incorporated through a knowledge-based scoring function that quantifies the agreement of a protein model with the EPR data. The scoring function was based on the CONE model[78] that translates the difference between the observed spin-spin distance (D_{SL}) and the distance of the C_β-atoms at the spin-labeling site (D_{BB}) into an agreement score. The agreement score is defined on the $D_{\mathrm{SL}} - D_{\mathrm{BB}}$ range −12.5 Å to 12.5 Å. Data from EPR accessibility measurements was incorporated using an approach that is similar to the neighbor vector approximation of the SASA. For the spin-labeled residue, the exposure moment $\boldsymbol{E}_{\mathrm{w}}$ is determined from its neighbor count e and its normalized $\mathrm{C}_\beta - \mathrm{C}_\alpha$ vector $\boldsymbol{s}$ by computing $\boldsymbol{E}_{\mathrm{w}} = \sum_{n=1}^{N} e_n \cdot \boldsymbol{s}_n$ over overlapping windows of length n. The overlapping windows became necessary because of the simplified representation of the protein's side chains. The agreement S_{orient} with the experimentally observed accessibility was subsequently computed from the torsion angle θ between the exposure moments as

$S_{\text{orient}} = -0.5 \cdot cos(\theta)$. Accordingly, a torsion angle of π corresponds to no agreement and a torsion angle of 0 corresponds to full agreement.

In this work, the influence of distance and accessibility restraints derived from EPR experiments on *de novo* protein structure prediction was evaluated in different studies (chapter III on page 29 and chapter VI on page 83). In each case, inclusion of experimental data from EPR measurements significantly improved the sampling accuracy as well as the discrimination of inaccurate models. In the most extensive study (chapter III on page 29), EPR-derived distance and accessibility restraints were used to predict the tertiary structure of twenty-nine monomeric and oligomeric membrane proteins. Inclusion of both types of restraints improved the sampling accuracy by 1.0 Å on average. Additionally, the discrimination of inaccurate models was improved, which was demonstrated by an improvement of the enrichment from 1.3 to 2.5. An additional finding was that incorporation of EPR accessibility restraints did not have a significant influence on the sampling accuracy. However, accessibility restraints proved useful for determining the rotation states of α-helices, which was demonstrated by an improvement of the contact recovery from 30 % to 39 % on average. In the other study that concerned the prediction of the soluble monomeric and membrane-associated homodimeric states of BAX (chapter VI on page 83), incorporation of EPR distance restraints enabled the prediction algorithm to overcome scoring problems resulting from α-helix 9 of BAX. The sampling accuracy and enrichment improved from 5.9 Å and 0.4 to 3.9 Å and 1.5 for soluble monomeric BAX and from 5.7 Å and 1.3 to 3.3 Å and 2.1 for membrane-associated homodimeric BAX.

XL-MS experiments yield the maximum distance between the two cross-linked residues. The maximum distance is defined by the used cross-linking reagent and the length of its spacer. A naïve approach for exploiting this information would be to just use it as a hard cutoff for the Euclidean distance between the two cross-linked residues in the protein model. However, this approach discards useful information. The cross-linking reagent is typically added to the protein in its folded state, making a fully extended conformation of the cross-link unlikely. Instead, the cross-link is more likely to follow a path along the surface of the protein. The shortest surface path between two residues can be computed but the necessary calculations are computationally too expensive for usage within a *de novo* protein structure prediction method. Instead, the geometrical center of the protein was used as the center of a sphere and the surface path was approximated using the arc length d_{arc} between the cross-linked residues (chapter IV on page 48). The agreement of a protein model with the experimental from XL-MS experiments was then quantified through summation of d_{arc} and distances of the cross-linked residues from the surface of the sphere, l_{SS1} and l_{SS2}, and subsequent comparison with the spacer length l_{XS}. An additional cosine transition region was introduced to account for the inaccuracies of this approximation.

Despite the rather coarse-grained approximation, using this method significantly improved the prediction accuracy and the ability of the scoring function to distinguish between accurate and inaccurate models. Cross-linkers with different spacer lengths were tested on a benchmark set consisting of fifteen soluble proteins. Inclusion of distance restraints derived from XL-MS improved the prediction accuracy by about 1.0 Å on average and up to 2.2 Å for some test cases. This corresponds to an improvement by two standard deviations, making the improvement statistically significant. The discrimination of inaccurate models could also be improved to an enrichment of 2.1.

The second focus of the cross-linking study was evaluating the influence of the spacer length on the prediction accuracy (chapter IV on page 48). Choosing a cross-linker with a longer spacer results in a larger number of cross-linked residues and therefore in a larger number of distance restraints. However, these distance restraints can be satisfied by a larger number of possible conformations; therefore reducing the discrimination power of the restraint. To determine the optimal spacer length, the number of potential cross-links was simulated on a set of 2055 non-redundant protein folds and fitted against a regression model. Cross-linkers with a spacer length estimated to be optimal were then compared to cross-linkers with shorter and longer spacers by carrying out protein structure prediction runs and comparison of the prediction accuracies. The results demonstrated the existence on an optimal spacer length l_{opt} in dependence on the lengths of the cross-linked side chains ($SS1$ and $SS2$) and the MW of the protein that can be predicted as $l_{\text{opt}} = k \cdot \sqrt[3]{MW} + \sqrt[3]{SS1 + SS2}$, with $k \approx \frac{1}{3}$.

In conclusion, limited experimental data from EPR or XL-MS experiments increases the likelihood of sampling models that are structurally similar to the major population in the protein's equilibrium. It was demonstrated that distance restraints derived from EPR or XL-MS experiments focus the sampling on native-like conformations, with EPR accessibility restraints being useful for determining the rotation states of helices. Additionally, an optimal cross-linker spacer length for maximizing the contained structural information was found. However, the results also demonstrate that incorporation of experimental restraints does not completely solve the problems in *de novo* protein structure prediction. For six of the twenty-nine membrane proteins predicted from EPR distance and accessibility data, it was not possible to sample models with an RMSD100 less than 6 Å relative to the experimentally determined reference structure. Although these results were achieved with low-resolution prediction only, the sampled models are probably too inaccurate for high-resolution refinement using Rosetta and MD simulations.

IX.3. Computational modeling provides an orthogonal approach for structure determination and validation

The vast majority of protein structures deposited in the PDB was determined using X-ray crystallography or NMR spectroscopy — although recently there has been an increasing number of protein structures determined using EM. However, those techniques are not applicable to all proteins. In particular, membrane proteins pose challenges to both X-ray crystallography and NMR spectroscopy. This is reflected by the constitution of the PDB — only 2.5 % of the deposited structures are membrane proteins. Different reasons contribute to this. Besides the difficulty of obtaining a sufficient amount of protein for X-ray crystallography or NMR spectroscopy, the two-dimensional nature of the membrane also hinders crystallization in the three-dimensional crystal lattice. For X-ray crystallography, stabilizing mutations might be necessary if the protein exhibits a substantial amount of flexibility. NMR spectroscopy on the other hand is typically hindered by the size of membrane proteins to which the membrane mimic also needs to be added.

Per se computational protein structure prediction methods are not hindered by any of these limitations. In reality, however, the biological system is too complex to be simulated comprehensively on currently available computer hardware (see section IX.1 on page 126 for a detailed discussion). Through complementation with limited experimental data, intrinsic limitations of the computational method can be compensated for and the computational method can fill information gaps in the experimental data. In this work, that was demonstrated for data from EPR and XL-MS experiments. In the context of membrane proteins, especially EPR data is very valuable since EPR measurements neither require an inflexible protein nor have a size limit. This makes EPR spectroscopy a suitable tool for studying the structure and dynamics of membrane proteins. Especially in conjunction with computational modeling, comprehensive mechanistic models describing the function and dynamics of membrane proteins can be developed.

In chapter VIII on page 111, a combined approach of EPR distance and accessibility measurements and computational modeling was employed to determine the protonation-dependent conformational dynamics of the SMR transporter protein EmrE. EPR measurements were conducted separately at pH 5 and with the ligand TPP bound to the protein. For the computational modeling, a two-pronged approach was performed. The tertiary structure of EmrE at pH 5 was predicted *de novo* from the EPR distance and accessibility data. The tertiary structure of TPP-bound EmrE was predicted using an X-ray-derived model of EmrE as starting point for the simulation. The *de novo* prediction started from predicted SSEs using a prediction pipeline consisting of BCL::Fold[2] and Rosetta.[94] The EPR distance and accessibility data was incorporated into the pipeline using the CONE model[78] and the approach derived from the neighbor vector method. The agreement of the final model with the EPR data was further evaluated using MMM software.[214] The prediction of the TPP-bound structure refined the X-ray-derived model of EmrE for improved agreement with the EPR data using the MODELLER software. By using this approach, two models could be generated that explained the EPR data and provided a mechanical model describing the conformational changes required for substrate binding and release.

Besides the structure prediction of membrane proteins, computational modeling can also be used for soluble proteins. Although soluble proteins are more readily accessible for techniques like X-ray crystallography or NMR spectroscopy, there are still cases where the protein is either too flexible for X-ray crystallography or too large for NMR spectroscopy. Other use cases include the validation of already determined structures through an orthogonal technique. In chapter VII on page 99, the tertiary structure of the C-terminal domain of ExoU was determined using a combined approach of computational modeling and EPR spectroscopy. The prediction of the tertiary structure of the C-terminal domain was performed using an iterative pipeline consisting of BCL::Fold[59] and *k*-means clustering.[100] The agreement of the final structure with the EPR data was performed through explicit simulation of the MTSL spin-labels. The endogenous residues at the spin-labeling sites were replaced with a rotamer library of MTSL and their dynamics were simulated using Rosetta.[94] Subsequently, the simulated distance distributions were compared to the experimentally determined distance distributions. Using this approach, the tertiary structure of the C-terminal domain of ExoU could be predicted in the absence of its chaperone SpcU. This enabled the validation of the X-ray-derived model of ExoU and the comparison of the structures in the presence and in the absence of their chaperone.

In conclusion, computational modeling provides an orthogonal approach for structure determination and validation, especially in conjunction with limited experimental data. Proteins that are too dynamic or too large for X-ray crystallography or NMR spectroscopy can be investigated using a combined approach of EPR spectroscopy and computational modeling to determine conformational dynamics or to validate existing structures. The additional advantage of computational modeling is the ability to readily explore alternative conformations that are not accessible to X-ray crystallography or NMR spectroscopy. Especially in conjunction with EPR spectroscopy, this could provide a way to determine a protein's equilibrium constitution — the conformations with relevant populations along with an estimate of their relative population size. This potential future direction is discussed in detail in section IX.4.

IX.4. Using *in silico* simulation of spectroscopic data should be the next step for ensemble prediction

In order to obtain a comprehensive understanding of a protein, it is not sufficient to know the conformation at the free energy minimum. In the equilibrium the protein exists is multiple different conformations where the probability of each conformation is determined by its free energy relative to alternative conformations. Ideally, a protein modeling method would be able to determine each conformation that accounts for a significant population size and give and estimate of the latter. However, exhaustive sampling of all possible conformations and accurate computation of their free energies is unfeasible (see section IX.1 on page 126 for a detailed discussion of this). Consequently, an alternative approach with reduced computational demand is needed. As discussed in section IX.2 on page 128, EPR spectroscopy is able to determine the distance distribution between two spin-labeled residues that is observed on the protein's equilibrium. If it would be possible to extend the size of the sampling space and recompute the EPR data *in silico*, the predicted ensemble could be fitted against the experimentally observed spectra and the population sizes could be extracted from the fitting parameters.

To achieve this goal, several problems need to be overcome: i) The sampling of conformations needs to be accelerated. Especially flexible regions of the protein have to be sampled comprehensively to ensure that all conformation that account for significant population sizes are contained in the predicted ensemble. ii) The simulation of the spin-label dynamics has to be accelerated. Although there are existing approaches like MMM[214] that can reliably simulate spin-label dynamics and therefore recompute the EPR-derived distance distributions, these approaches are too slow for usage within a *de novo* prediction. iii) An approach is needed to fit the simulated distance distributions of the ensemble representatives against the experimentally determined spectra and derive a population size for each representative from the fitting parameters.

Acceleration of the conformation sampling could be achieved through orthogonal approaches. Whereas the application of structural perturbations to the intermediate protein model could be further parallelized and accelerated by performing the calculations on a graphics processing unit (GPU), as orthogonal approach the protein model could also be assembled from structural fragments in a sequence-independent manner. Chapter V on page 67 details a study of the latter that focused on the rapid sampling of loop conformations using structural fragments. This study was focused on loop regions, because in the absence of periodic hydrogen bonds, these regions are more likely to exhibit substantial conformational flexibility and comprehensive sampling of them consequently is crucial for protein ensemble prediction. For this study, a template library of loop conformations was collected from about 87 000 protein structures deposited in the PDB. The loop templates were parameterized according to their geometric properties that consisted of the sequence length d, the translation vector $\boldsymbol{t}$ between the two anchor points of the loop, and the Euler angles $\boldsymbol{E}$ between the tow anchor points of the loop. A hash function $f: d \times \boldsymbol{t} \times \boldsymbol{E} \to k$ was subsequently used to compute a one-dimensional hash key k for each loop template. The loop sampling algorithm then computes for each loop in the intermediate protein model the hash key k and selects a matching template from the library within constant time complexity. The loop's sequence was then fitted against the conformation of the selected template and inserted into the protein model. The algorithm was complemented with a CCD[101] implementation to ensure that loops can also be sampled correctly if no complete matching template was available. The benchmarking of the algorithm was performed on a set consisting of eighteen soluble proteins and eleven membrane proteins that contained 296 non-terminal loop regions. The algorithm achieved a closure rate of 100 %, while only requiring a CPU time of (161 ± 13) ms per loop on average. On contrast, the Rosetta "loophash" algorithm,[192] which was used as reference point for the benchmark, required about 160 s per loop on average. For 94 % of all benchmark loops, the experimentally determined conformation could be sampled within an accuracy limit of 2 Å. These results demonstrate that the rapid loop sampling algorithm is suitable for being used within an ensemble prediction pipeline.

The simulation of the spin-label dynamics needs to be able to estimate the likelihood of different spin-label conformations, given a specific conformation of the protein. Hence, the interactions of the spin-label with the side chains of the protein need to be evaluated and an approximation of the free energy needs to be derived for each spin-label conformation. In a previous study, Alexander *et al.* created a rotamer library of the spin-label MTSL from high-resolution structures deposited in the PDB. This rotamer library could be used in conjunction with rotamer libraries for the canonical amino acids to sample different rotamer combinations. Exhaustive sampling of all combinations will likely be unfeasible but integration into an MCM algorithm could identify the most stable combinations. Additional porting to a GPU could further reduce the time required for the computations. Using this approach enables two approaches: i) computation of the distance distribution derived from the distances between the free electrons or ii) recomputation of the primary EPR data using the formula $\sum_{n=1}^{k} \cos(t \cdot (1 - 3 \cdot \sqrt{0.005 \cdot n} \cdot \kappa / r^3))$, where r is the distance between the free electrons, κ is the EPR constant, and k is the number of integration points. Either approach provides a simulated spectrum that can be compared to the experimentally determined spectrum.

Estimation of the population sizes for the sampled conformations could be achieved by comparison of the simulated spectra of the sampled conformations to the experimentally determined spectra. For this purpose, the simulated spectra s_i could be fitted against the experimentally determined spectrum s_e using multiple linear regression to determine the weighting factors w_i for the simulated spectra so that the distance d, between the weighted simulated spectra and the experimentally determined spectrum, $d = \sum_i w_i \cdot s_i - s_\mathrm{e}$, becomes minimal. Consequently, the weighting factors w_i could be interpreted as relative population size of the corresponding conformation. However, additional thought needs to be put in approaches with very low weighting factors. It is conceivable that not all sampled conformations of the protein are thermodynamically stable and therefore contribute with any significance to the fitting. Those conformations should be filtered out in this step.

In conclusion, experimental data from techniques like EPR spectroscopy provides information about the structural constitution of a protein's equilibrium. Current approaches like the CONE model[78] do not use this information to its full potential. Besides throwing away information about alternative conformations, this also results in the problem that one protein might not be able to explain the experimental data. Hence, I suggest an ensemble approach, where the experimental data is simulated for the whole ensemble and then fitted against the experimentally determined data. This approach will provide information about the contribution of each ensemble element and therefore provide insight into alternative conformations.

REFERENCES

[1] Axel Walter Fischer, Sten Heinze, Daniel K. Putnam, Bian Li, James C Pino, Yan Xia, Carlos F Lopez, and Jens Meiler. "CASP11 – An Evaluation of a Modular BCL::Fold-Based Protein Structure Prediction Pipeline". In: *PLOS ONE* 11.4 (Apr. 2016). Ed. by Yang Zhang, e0152517. DOI: 10.1371/journal.pone.0152517.

[2] Axel Walter Fischer, Nathan Scott Alexander, Nils Woetzel, Mert Karakas, Brian E. Weiner, and Jens Meiler. "BCL::MP-Fold: Membrane protein structure prediction guided by EPR restraints". In: *Proteins: Structure, Function, and Bioinformatics* 83.11 (Nov. 2015), pp. 1947–1962. DOI: 10.1002/prot.24801.

[3] Tommy Hofmann, Axel W Fischer, Jens Meiler, and Stefan Kalkhof. "Protein structure prediction guided by crosslinking restraints – A systematic evaluation of the impact of the crosslinking spacer length". In: *Methods* 89 (Nov. 2015), pp. 79–90. DOI: 10.1016/j.ymeth.2015.05.014.

[4] Axel Walter Fischer, Rocco Moretti, Nathan Scott Alexander, Jeffrey Mendenhall, Nicholas J Hyman, and Jens Meiler. "Efficient sampling of loop conformations using conformation hashing in conjunction with cyclic coordinate descent". In: *PloS one (submitted)* (2018).

[5] Axel W Fischer, Enrica Bordignon, Stephanie Bleicken, Ana J. García-Sáez, Gunnar Jeschke, and Jens Meiler. "Pushing the size limit of de novo structure ensemble prediction guided by sparse SDSL-EPR restraints to 200 residues: The monomeric and homodimeric forms of BAX". In: *Journal of Structural Biology* 195.1 (July 2016), pp. 62–71. DOI: 10.1016/j.jsb.2016.04.014.

[6] Axel Walter Fischer, David M Anderson, Maxx H Tessmer, Dara W. Frank, Jimmy B Feix, and Jens Meiler. "Structure and Dynamics of Type III Secretion Effector Protein ExoU As determined by SDSL-EPR Spectroscopy in Conjunction with De Novo Protein Folding". In: *ACS Omega* 2.6 (June 2017), pp. 2977–2984. DOI: 10.1021/acsomega.7b00349.

[7] Reza Dastvan, Axel W Fischer, Smriti Mishra, Jens Meiler, and Hassane S Mchaourab. "Protonation-dependent conformational dynamics of the multidrug transporter EmrE". In: *Proceedings of the National Academy of Sciences* 113.5 (Feb. 2016), pp. 1220–1225. DOI: 10.1073/pnas.1520431113.

[8] Marci Surpin and Natasha Raikhel. "Plant cell biology: Traffic jams affect plant development and signal transduction". In: *Nature Reviews Molecular Cell Biology* 5.2 (Feb. 2004), pp. 100–109. DOI: 10.1038/nrm1311.

[9] Enrique Rojo, C.Stewart Gillmor, Valentina Kovaleva, Chris R. Somerville, and Natasha V. Raikhel. "VACUOLELESS1 Is an Essential Gene Required for Vacuole Formation and Morphogenesis in Arabidopsis". In: *Developmental Cell* 1.2 (Aug. 2001), pp. 303–310. DOI: 10.1016/S1534-5807(01)00024-7.

[10] Kaspar Hollenstein, Roger JP Dawson, and Kaspar P. Locher. "Structure and mechanism of ABC transporter proteins". In: *Current Opinion in Structural Biology* 17.4 (Aug. 2007), pp. 412–418. DOI: 10.1016/j.sbi.2007.07.003.

[11] W. A. Catterall. "Structure and Function of Voltage-Gated Ion Channels". In: *Annual Review of Biochemistry* 64.1 (Jan. 1995), pp. 493–531. DOI: 10.1146/annurev.biochem.64.1.493.

[12] B. -M. Sjöberg. "Ribonucleotide reductases — a group of enzymes with different metallosites and a similar reaction mechanism". In: *Metal Sites in Proteins and Models*. Vol. 88. 1997, pp. 139–173. DOI: 10.1007/3-540-62870-3_5.

[13] Carolyn R Bertozzi. “Chemical Glycobiology”. In: *Science* 291.5512 (Mar. 2001), pp. 2357–2364. DOI: 10.1126/science.1059820.

[14] Peter M.J. Burgers and Thomas A Kunkel. “Eukaryotic DNA Replication Fork”. In: *Annual Review of Biochemistry* 86.1 (June 2017), pp. 417–438. DOI: 10.1146/annurev-biochem-061516-044709.

[15] Xiao Jian Sun, Paul Rothenberg, C Ronald Kahn, Jonathan M Backer, Eiichi Araki, Peter a Wilden, Deborah a Cahill, Barry J Goldstein, and Morris F White. “Structure of the insulin receptor substrate IRS-1 defines a unique signal transduction protein”. In: *Nature* 352.6330 (July 1991), pp. 73–77. DOI: 10.1038/352073a0.

[16] R J Davis. “The mitogen-activated protein kinase signal transduction pathway.” In: *The Journal of biological chemistry* 268.20 (July 1993), pp. 14553–6.

[17] Kai Simons and Derek Toomre. “Lipid rafts and signal transduction”. In: *Nature Reviews Molecular Cell Biology* 1.1 (Oct. 2000), pp. 31–39. DOI: 10.1038/35036052.

[18] M. Simon, M. Strathmann, and N. Gautam. “Diversity of G proteins in signal transduction”. In: *Science* 252.5007 (May 1991), pp. 802–808. DOI: 10.1126/science.1902986.

[19] J. H C Wang and B. P. Thampatty. “An Introductory Review of Cell Mechanobiology”. In: *Biomechanics and Modeling in Mechanobiology* 5.1 (Mar. 2006), pp. 1–16. DOI: 10.1007/s10237-005-0012-z.

[20] P Liang and T H MacRae. “Molecular chaperones and the cytoskeleton.” In: *Journal of cell science* 110 (Pt 1.13 (July 1997), pp. 1431–40.

[21] T. D. Pollard and J. A. Cooper. “Actin, a Central Player in Cell Shape and Movement”. In: *Science* 326.5957 (Nov. 2009), pp. 1208–1212. DOI: 10.1126/science.1175862.

[22] T.J Mitchison and L.P Cramer. “Actin-Based Cell Motility and Cell Locomotion”. In: *Cell* 84.3 (Feb. 1996), pp. 371–379. DOI: 10.1016/S0092-8674(00)81281-7.

[23] Hagit Yerushalmi and Shimon Schuldiner. “A common binding site for substrates and protons in EmrE, an ion-coupled multidrug transporter”. In: *FEBS Letters* 476.1-2 (June 2000), pp. 93–97. DOI: 10.1016/S0014-5793(00)01677-X.

[24] Hagit Yerushalmi, Mario Lebendiker, and Shimon Schuldiner. “EmrE, an Escherichia coli 12-kDa Multidrug TrRansporter, Exchanges Toxic Cations and H+ and Is Soluble In Organic Solvents”. In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1555.1-3 (2002), pp. 1–7.

[25] Shimon Schuldiner. “EmrE, a model for studying evolution and mechanism of ion-coupled transporters”. In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1794.5 (May 2009), pp. 748–762. DOI: 10.1016/j.bbapap.2008.12.018.

[26] Shimon Schuldiner. “Competition as a Way of Life for H+-Coupled Antiporters”. In: *Journal of Molecular Biology* 426.14 (July 2014), pp. 2539–2546. DOI: 10.1016/j.jmb.2014.05.020.

[27] David Baker and David A. Agard. “Kinetics versus Thermodynamics in Protein Folding”. In: *Biochemistry* 33.24 (June 1994), pp. 7505–7509. DOI: 10.1021/bi00190a002.

[28] Yi Fang. “Thermodynamic Principle Revisited: Theory of Protein Folding”. In: *Advances in Bioscience and Biotechnology* 06.01 (2015), pp. 37–48. DOI: 10.4236/abb.2015.61005.

[29] C B Anfinsen. “Principles that govern the folding of protein chains.” In: *Science (New York, N.Y.)* 181.96 (1973), pp. 223–230. DOI: 10.1126/science.181.4096.223.

[30] Daniel E. Koshland. “The Key–Lock Theory and the Induced Fit Theory”. In: *Angewandte Chemie International Edition in English* 33.2324 (Jan. 1995), pp. 2375–2378. DOI: 10.1002/anie.199423751.

[31] Rommie E. Amaro, Riccardo Baron, and J. Andrew McCammon. “An improved relaxed complex scheme for receptor flexibility in computer-aided drug design”. In: *Journal of Computer-Aided Molecular Design* 22.9 (Sept. 2008), pp. 693–705. DOI: 10.1007/s10822-007-9159-2.

[32] William Sinko, Steffen Lindert, and J. Andrew McCammon. “Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design”. In: *Chemical Biology & Drug Design* 81.1 (Jan. 2013), pp. 41–49. DOI: 10.1111/cbdd.12051.

[33] Julie R. Schames, Richard H. Henchman, Jay S. Siegel, Christoph A. Sotriffer, Haihong Ni, and J. Andrew McCammon. “Discovery of a Novel Binding Trench in HIV Integrase”. In: *Journal of Medicinal Chemistry* 47.8 (Apr. 2004), pp. 1879–1881. DOI: 10.1021/jm0341913.

[34] Roberto Di Santo, Roberta Costi, Alessandra Roux, Marino Artico, Antonio Lavecchia, Luciana Marinelli, Ettore Novellino, Lucia Palmisano, Mauro Andreotti, Roberta Amici, Clementina Maria Galluzzo, Lucia Nencioni, Anna Teresa Palamara, Yves Pommier, and Christophe Marchand. “Novel Bifunctional Quinolonyl Diketo Acid Derivatives as HIV-1 Integrase Inhibitors: Design, Synthesis, Biological Activities, and Mechanism of Action”. In: *Journal of Medicinal Chemistry* 49.6 (Mar. 2006), pp. 1939–1945. DOI: 10.1021/jm0511583.

[35] Jacob D. Durrant, Michael D. Urbaniak, Michael A J Ferguson, and J. Andrew McCammon. “Computer-Aided Identification of Trypanosoma brucei Uridine Diphosphate Galactose 4-Epimerase Inhibitors: Toward the Development of Novel Therapies for African Sleeping Sickness”. In: *Journal of Medicinal Chemistry* 53.13 (July 2010), pp. 5025–5032. DOI: 10.1021/jm100456a.

[36] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. “The Protein Data Bank.” In: *Nucleic acids research* 28.1 (Jan. 2000), pp. 235–242. DOI: 10.1093/nar/28.1.235.

[37] Helen M Berman. “The Protein Data Bank: a historical perspective.” In: *Acta crystallographica. Section A, Foundations of crystallography* 64.Pt 1 (Jan. 2008), pp. 88–95. DOI: 10.1107/S0108767307035623.

[38] Kiyofumi Takaba, Kazuki Takeda, Masayuki Kosugi, Taro Tamada, and Kunio Miki. “Distribution of valence electrons of the flavin cofactor in NADH-cytochrome b5 reductase”. In: *Scientific Reports* 7 (Feb. 2017), p. 43162. DOI: 10.1038/srep43162.

[39] Marta C. Marques, Cristina Tapia, Oscar Gutiérrez-Sanz, Ana Raquel Ramos, Kimberly L. Keller, Judy D. Wall, Antonio L. De Lacey, Pedro M. Matias, and Inês A C Pereira. “The direct role of selenocysteine in [NiFeSe] hydrogenase maturation and catalysis”. In: *Nature Chemical Biology* 13.5 (Mar. 2017), pp. 544–550. DOI: 10.1038/nchembio.2335.

[40] Roslyn M Bill, Peter J F Henderson, So Iwata, Edmund R S Kunji, Hartmut Michel, Richard Neutze, Simon Newstead, Bert Poolman, Christopher G Tate, and Horst Vogel. “Overcoming barriers to membrane protein structure determination.” In: *Nature biotechnology* 29.4 (Apr. 2011), pp. 335–340. DOI: 10.1038/nbt.1833.

[41] J. Lepault, F. P. Booy, and J. Dubochet. “Electron microscopy of frozen biological suspensions”. In: *Journal of Microscopy* 129.1 (Jan. 1983), pp. 89–102. DOI: 10.1111/j.1365-2818.1983.tb04163.x.

[42] Susan D. Saban, Ronald R. Nepomuceno, Lance D. Gritton, Glen R. Nemerow, and Phoebe L. Stewart. “CryoEM Structure at 9Å Resolution of an Adenovirus Vector Targeted to Hematopoietic Cells”. In: *Journal of Molecular Biology* 349.3 (June 2005), pp. 526–537. DOI: 10.1016/j.jmb.2005.04.034.

[43] Alan Merk, Alberto Bartesaghi, Soojay Banerjee, Veronica Falconieri, Prashant Rao, Mindy I. Davis, Rajan Pragani, Matthew B. Boxer, Lesley A. Earl, Jacqueline L.S. Milne, and Sriram Subramaniam. “Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery”. In: *Cell* 165.7 (June 2016), pp. 1698–1707. DOI: 10.1016/j.cell.2016.05.040.

[44] A. Bartesaghi, A. Merk, S. Banerjee, D. Matthies, X. Wu, Jacqueline L S Milne, and Sriram Subramaniam. “2.2 A resolution cryo-EM structure of -galactosidase in complex with a cell-permeant inhibitor”. In: *Science* 348.6239 (June 2015), pp. 1147–1151. DOI: 10.1126/science.aab1576.

[45] K Wüthrich. “Protein structure determination in solution by NMR spectroscopy”. In: *Journal of Biological Chemistry* 265.36 (1990), pp. 22059–22062.

[46] Gunnar Jeschke. “DEER Distance Measurements on Proteins”. In: *Annual Review of Physical Chemistry* 63.1 (Jan. 2012), pp. 419–446. DOI: 10.1146/annurev-physchem-032511-143716.

[47] Johann P. Klare. “Site-Directed Spin Labeling and Electron Paramagnetic Resonance (EPR) Spectroscopy: A Versatile Tool to Study Protein-Protein Interactions”. In: *Protein Interactions*. InTech, Mar. 2012, pp. 427–447. DOI: 10.5772/37209.

[48] James L. Baber, John M. Louis, and G. Marius Clore. “Dependence of Distance Distributions Derived from Double Electron-Electron Resonance Pulsed EPR Spectroscopy on Pulse-Sequence Time”. In: *Angewandte Chemie International Edition* 54.18 (Apr. 2015), pp. 5336–5339. DOI: 10.1002/anie.201500640.

[49] C. Nick Pace, J. Martin Scholtz, and Gerald R. Grimsley. “Forces stabilizing proteins”. In: *FEBS Letters* 588.14 (June 2014), pp. 2177–2184. DOI: 10.1016/j.febslet.2014.05.006.

[50] Brian E Weiner, Nils Woetzel, Mert Karakaş, Nathan Alexander, and Jens Meiler. “BCL::MP-fold: Folding membrane proteins through assembly of transmembrane helices”. In: *Structure* 21.7 (July 2013), pp. 1107–1117. DOI: 10.1016/j.str.2013.04.022.

[51] Angel E. Garcia and Kevin Y Sanbonmatsu. “Exploring the energy landscape of a ? hairpin in explicit solvent”. In: *Proteins: Structure, Function, and Genetics* 42.3 (Feb. 2001), pp. 345–354. DOI: 10.1002/1097-0134(20010215)42:3<345::AID-PROT50>3.0.CO;2-H.

[52] Hugh Nymeyer and A. E. Garcia. “Simulation of the folding equilibrium of -helical peptides: A comparison of the generalized Born approximation with explicit solvent”. In: *Proceedings of the National Academy of Sciences* 100.24 (Nov. 2003), pp. 13934–13939. DOI: 10.1073/pnas.2232868100.

[53] L Forrest. “Membrane simulations: bigger and better?” In: *Current Opinion in Structural Biology* 10.2 (Apr. 2000), pp. 174–181. DOI: 10.1016/S0959-440X(00)00066-X.

[54] E LINDAHL and M SANSOM. “Membrane proteins: molecular dynamics simulations”. In: *Current Opinion in Structural Biology* 18.4 (Aug. 2008), pp. 425–431. DOI: 10.1016/j.sbi.2008.02.003.

[55] Philippe Ferrara, Joannis Apostolakis, and Amedeo Caflisch. “Evaluation of a fast implicit solvent model for molecular dynamics simulations”. In: *Proteins: Structure, Function, and Genetics* 46.1 (Jan. 2002), pp. 24–33. DOI: 10.1002/prot.10001.

[56] Jianhan Chen, Charles L. Brooks, and Jana Khandogin. “Recent advances in implicit solvent-based methods for biomolecular simulations”. In: *Current Opinion in Structural Biology* 18.2 (Apr. 2008), pp. 140–148. DOI: 10.1016/j.sbi.2008.01.003.

[57] Min-yi Shen and Karl F. Freed. “Long Time Dynamics of Met-Enkephalin: Comparison of Explicit and Implicit Solvent Models”. In: *Biophysical Journal* 82.4 (Apr. 2002), pp. 1791–1808. DOI: 10.1016/S0006-3495(02)75530-6.

[58] Ruhong Zhou. “Free energy landscape of protein folding in water: Explicit vs. implicit solvent”. In: *Proteins: Structure, Function, and Genetics* 53.2 (Nov. 2003), pp. 148–161. DOI: 10.1002/prot.10483.

[59] Mert Karakaş, Nils Woetzel, Rene Staritzbichler, Nathan Alexander, Brian E Weiner, and Jens Meiler. “BCL::Fold - De Novo Prediction of Complex and Large Protein Topologies by Assembly of Secondary Structure Elements”. In: *PLoS ONE* 7.11 (Jan. 2012), e49240. DOI: 10.1371/journal.pone.0049240.

[60] Nils Woetzel, Mert Karakaş, Rene Staritzbichler, Ralf Müller, Brian E Weiner, and Jens Meiler. “BCL::Score-Knowledge Based Energy Potentials for Ranking Protein Models Represented by Idealized Secondary Structure Elements”. In: *PLoS ONE* 7.11 (Jan. 2012), e49242. DOI: 10.1371/journal.pone.0049242.

[61] Carol A. Rohl, Charlie E.M. Strauss, Kira M.S. Misura, and David Baker. “Protein Structure Prediction Using Rosetta”. In: *Methods in Enzymology*. Vol. 383. 2003. 2004, pp. 66–93. DOI: 10.1016/S0076-6879(04)83004-0.

[62] Yang Zhang. “Template-based modeling and free modeling by I-TASSER in CASP7”. In: *Proteins: Structure, Function, and Bioinformatics* 69.S8 (2007), pp. 108–117. DOI: 10.1002/prot.21702.

[63] Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. “The I-TASSER Suite: protein structure and function prediction”. In: *Nature Methods* 12.1 (Dec. 2014), pp. 7–8. DOI: 10.1038/nmeth.3213.

[64] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. “Critical assessment of methods of protein structure prediction: Progress and new directions in round XI”. In: *Proteins: Structure, Function, and Bioinformatics* 84.April (Sept. 2016), pp. 4–14. DOI: 10.1002/prot.25064.

[65] Lisa N. Kinch, Wenlin Li, R. Dustin Schaeffer, Roland L. Dunbrack, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Nick V. Grishin. “CASP 11 target classification”. In: *Proteins: Structure, Function, and Bioinformatics* 84.January (Sept. 2016), pp. 20–33. DOI: 10.1002/prot.24982.

[66] Andriy Kryshtafovych, Bohdan Monastyrskyy, Krzysztof Fidelis, John Moult, Torsten Schwede, and Anna Tramontano. “Evaluation of the template-based modeling in CASP12”. In: *Proteins: Structure, Function, and Bioinformatics* January (Nov. 2017), pp. 1–33. DOI: 10.1002/prot.25425.

[67] Lisa N. Kinch, Wenlin Li, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Nick V. Grishin. “Evaluation of free modeling targets in CASP11 and ROLL”. In: *Proteins: Structure, Function, and Bioinformatics* 84.December (Sept. 2016), pp. 51–66. DOI: 10.1002/prot.24973.

[68] Steffen Lindert, Phoebe L Stewart, and Jens Meiler. "Hybrid approaches: applying computational methods in cryo-electron microscopy." In: *Current opinion in structural biology* 19.2 (Apr. 2009), pp. 218–25. DOI: 10.1016/j.sbi.2009.02.010.

[69] Steffen Lindert, René Staritzbichler, Nils Wötzel, Mert Karakaş, Phoebe L Stewart, and Jens Meiler. "EM-Fold: De Novo Folding of α-Helical Proteins Guided by Intermediate-Resolution Electron Microscopy Density Maps". In: *Structure* 17.7 (July 2009), pp. 990–1003. DOI: 10.1016/j.str.2009.06.001.

[70] Steffen Lindert, Tommy Hofmann, Nils Wötzel, Mert Karakaş, Phoebe L Stewart, and Jens Meiler. "Ab initio protein modeling into CryoEM density maps using EM-Fold." In: *Biopolymers* 97.9 (Sept. 2012), pp. 669–77. DOI: 10.1002/bip.22027.

[71] Nils Woetzel, Steffen Lindert, Phoebe L Stewart, and Jens Meiler. "BCL::EM-Fit: Rigid body fitting of atomic structures into density maps using geometric hashing and real space refinement". In: *Journal of Structural Biology* 175.3 (Sept. 2011), pp. 264–276. DOI: 10.1016/j.jsb.2011.04.016.

[72] Steffen Lindert and J. Andrew McCammon. "Improved cryoEM-Guided Iterative Molecular Dynamics–Rosetta Protein Structure Refinement Protocol for High Precision Protein Structure Prediction". In: *Journal of Chemical Theory and Computation* 11.3 (Mar. 2015), pp. 1337–1346. DOI: 10.1021/ct500995d.

[73] Keren Lasker, Andrej Sali, and Haim J. Wolfson. "Determining macromolecular assembly structures by molecular docking and fitting into an electron density map". In: *Proteins: Structure, Function, and Bioinformatics* 78.15 (Nov. 2010), pp. 3205–3211. DOI: 10.1002/prot.22845.

[74] Elina Tjioe, Keren Lasker, Ben Webb, Haim J. Wolfson, and Andrej Sali. "MultiFit: a web server for fitting multiple protein structures into their electron microscopy density map". In: *Nucleic Acids Research* 39.suppl (July 2011), W167–W170. DOI: 10.1093/nar/gkr490.

[75] Gunnar Jeschke. "Modeling of protein structural transitions from sparse distance constraints and homology information". In: *Journal of Structural Biology* (2012), pp. 1–44.

[76] Andrej Sali. "Comparative protein modeling by satisfaction of spatial restraints". In: *Molecular Medicine Today* 1.6 (Sept. 1995), pp. 270–277. DOI: 10.1016/S1357-4310(95)91170-7.

[77] Benjamin Webb and Andrej Sali. "Comparative Protein Structure Modeling Using MODELLER". In: *Current Protocols in Bioinformatics*. Vol. 47. Hoboken, NJ, USA: John Wiley & Sons, Inc., Sept. 2014, pp. 5.6.1–5.6.32. DOI: 10.1002/0471250953.bi0506s47.

[78] Nathan Alexander, Ahmad Al-Mestarihi, Marco Bortolus, Hassane Mchaourab, and Jens Meiler. "De Novo High-Resolution Protein Structure Determination from Sparse Spin-Labeling EPR Data". In: *Structure* 16.2 (Feb. 2008), pp. 181–195. DOI: 10.1016/j.str.2007.11.015.

[79] Y. Goldgur, R. Craigie, G. H. Cohen, T. Fujiwara, T. Yoshinaga, T. Fujishita, H. Sugimoto, T. Endo, H. Murai, and D. R. Davies. "Structure of the HIV-1 integrase catalytic domain complexed with an inhibitor: A platform for antiviral drug design". In: *Proceedings of the National Academy of Sciences* 96.23 (Nov. 1999), pp. 13040–13043. DOI: 10.1073/pnas.96.23.13040.

[80] Maria L. Barreca, Keun Woo Lee, Alba Chimirri, and James M. Briggs. "Molecular Dynamics Studies of the Wild-Type and Double Mutant HIV-1 Integrase Complexed with the 5CITEP Inhibitor: Mechanism for Inhibition and Drug Resistance". In: *Biophysical Journal* 84.3 (Mar. 2003), pp. 1450–1463. DOI: 10.1016/S0006-3495(03)74958-3.

[81] José Nelson Onuchic and Peter G. Wolynes. “Theory of protein folding”. In: *Current Opinion in Structural Biology* 14.1 (2004), pp. 70–75. DOI: 10.1016/j.sbi.2004.01.009.

[82] Corey Hardin, Michael P Eastwood, Michael C Prentiss, Zadia Luthey-Schulten, and Peter G Wolynes. “Associative memory Hamiltonians for structure prediction without homology: alpha/beta proteins.” In: *Proceedings of the National Academy of Sciences of the United States of America* 100.4 (2003), pp. 1679–1684. DOI: 10.1073/pnas.252753899.

[83] Corey Hardin, Michael P. Eastwood, Michael Prentiss, Z. Luthey-Schulten, and Peter G. Wolynes. “Folding funnels: The key to robust protein structure prediction”. In: *Journal of Computational Chemistry* 23.1 (2002), pp. 138–146. DOI: 10.1002/jcc.1162.

[84] Joseph D Bryngelson and Peter G Wolynes. “Spin glasses and the statistical mechanics of protein folding.” In: *Proceedings of the National Academy of Sciences of the United States of America* 84.21 (1987), pp. 7524–7528. DOI: 10.1073/pnas.84.21.7524.

[85] Joseph Bryngelson and Peter G Wolynes. “Intermediates and barrier crossing in a random energy model (with applications to protein folding)”. In: *The Journal of Physical Chemistry* 93.19 (1989), pp. 6902–6915. DOI: doi:10.1021/j100356a007.

[86] Sten Heinze, Daniel K. Putnam, Axel W. Fischer, Tim Kohlmann, Brian E. Weiner, and Jens Meiler. “CASP10 – BCL::Fold efficiently samples topologies of large proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 83.3 (Mar. 2015), pp. 547–563. DOI: 10.1002/prot.24733.

[87] Steffen Lindert, Nathan Alexander, Nils Wötzel, Mert Karakaş, Phoebe L Stewart, and Jens Meiler. “EM-Fold: De novo atomic-detail protein structure determination from medium-resolution density maps”. In: *Structure* 20.3 (Mar. 2012), pp. 464–478. DOI: 10.1016/j.str.2012.01.023.

[88] Daniel K. Putnam, Brian E. Weiner, Nils Woetzel, Edward W. Lowe, and Jens Meiler. “BCL::SAXS: GPU accelerated Debye method for computation of small angle X-ray scattering profiles”. In: *Proteins: Structure, Function, and Bioinformatics* 83.8 (Aug. 2015), pp. 1500–1512. DOI: 10.1002/prot.24838.

[89] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. “Critical assessment of methods of protein structure prediction (CASP) - round x”. In: *Proteins: Structure, Function and Bioinformatics* 82.SUPPL.2 (2014), pp. 1–6. DOI: 10.1002/prot.24452.

[90] Sergey Ovchinnikov, David E Kim, Ray Yu-Ruei Wang, Yuan Liu, Frank DiMaio, and David Baker. “Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta”. In: *Proteins: Structure, Function, and Bioinformatics* 84.January (Sept. 2016), pp. 67–75. DOI: 10.1002/prot.24974.

[91] Wenxuan Zhang, Jianyi Yang, Baoji He, Sara Elizabeth Walker, Hongjiu Zhang, Brandon Govindarajoo, Jouko Virtanen, Zhidong Xue, Hong-Bin Shen, and Yang Zhang. “Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11”. In: *Proteins: Structure, Function, and Bioinformatics* 84.August (Sept. 2016), pp. 76–86. DOI: 10.1002/prot.24930.

[92] Jianyi Yang, Wenxuan Zhang, Baoji He, Sara Elizabeth Walker, Hongjiu Zhang, Brandon Govindarajoo, Jouko Virtanen, Zhidong Xue, Hong-Bin Shen, and Yang Zhang. “Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade”. In: *Proteins: Structure, Function, and Bioinformatics* 84.August (Sept. 2016), pp. 233–246. DOI: 10.1002/prot.24918.

[93] Hyungrae Kim and Daisuke Kihara. “Protein structure prediction using residue- and fragment-environment potentials in CASP11”. In: *Proteins: Structure, Function, and Bioinformatics* 84.August (Sept. 2016), pp. 105–117. DOI: 10.1002/prot.24920.

[94] Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, Oliver F. Lange, James Thompson, Ron Jacak, Kristian Kaufman, P. Douglas Renfrew, Colin a. Smith, Will Sheffler, Ian W. Davis, Seth Cooper, Adrien Treuille, Daniel J. Mandell, Florian Richter, Yih En Andrew Ban, Sarel J. Fleishman, Jacob E. Corn, David E. Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popović, James J. Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker, and Philip Bradley. “Rosetta3: An object-oriented software suite for the simulation and design of macromolecules”. In: *Methods in Enzymology* 487.C (2011), pp. 545–574. DOI: 10.1016/B978-0-12-381270-4.00019-6.

[95] Richard Bonneau, Ingo Ruczinski, Jerry Tsai, and David Baker. “Contact order and ab initio protein structure prediction.” In: *Protein science : a publication of the Protein Society* 11.8 (2002), pp. 1937–1944. DOI: 10.1110/ps.3790102.

[96] Elizabeth Durham, Brent Dorr, Nils Woetzel, René Staritzbichler, and Jens Meiler. “Solvent accessible surface area approximations for rapid and accurate protein structure prediction”. In: *Journal of Molecular Modeling* 15.9 (Sept. 2009), pp. 1093–1108. DOI: 10.1007/s00894-009-0454-9.

[97] Julia Koehler Leman, Ralf Mueller, Mert Karakas, Nils Woetzel, and Jens Meiler. “Simultaneous prediction of protein secondary structure and transmembrane spans”. In: *Proteins: Structure, Function and Bioinformatics* 81.7 (July 2013), pp. 1127–1140. DOI: 10.1002/prot.24258.

[98] D T Jones. “Protein secondary structure prediction based on position-specific scoring matrices.” In: *Journal of molecular biology* 292.2 (Sept. 1999), pp. 195–202. DOI: 10.1006/jmbi.1999.3091.

[99] Jeffrey Mendenhall and Jens Meiler. “Prediction of Transmembrane Proteins and Regions using Fourier Spectral Analysis and Advancements in Machine Learning”. In: *SERMACS 2014*. 2014. DOI: 10.13140/RG.2.1.2545.8724.

[100] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*. 2015.

[101] Adrian AA Canutescu and Roland L Dunbrack. “Cyclic coordinate descent: A robotics algorithm for protein loop closure.” In: *Protein science : a publication of the Protein Society* 12.5 (May 2003), pp. 963–972. DOI: 10.1110/ps.0242703.

[102] D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, III T.E. Cheatham, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, and P.A. Kollman. *AMBER 14*. 2014.

[103] Oliviero Carugo and S Pongor. “A normalized root-mean-square distance for comparing protein three-dimensional structures.” In: *Protein science : a publication of the Protein Society* 10.7 (2001), pp. 1470–1473. DOI: 10.1110/ps.690101.of.

[104] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. “Improved side-chain torsion potentials for the Amber ff99SB protein force field”. In: *Proteins: Structure, Function and Bioinformatics* 78.8 (2010), pp. 1950–1958. DOI: 10.1002/prot.22711.

[105] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. “Comparison of simple potential functions for simulating liquid water”. In: *The Journal of Chemical Physics* 79.2 (1983), p. 926. DOI: 10.1063/1.445869.

[106] Arvind Ramanathan and Pratul K. Agarwal. “Computational identification of slow conformational fluctuations in proteins”. In: *Journal of Physical Chemistry B* 113.52 (2009), pp. 16669–16680. DOI: 10.1021/jp9077213.

[107] Richard J Loncharich, Bernard R Brooks, and Richard W Pastor. “Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N-methylamide”. In: *Biopolymers* 32.5 (May 1992), pp. 523–535. DOI: 10.1002/bip.360320508.

[108] Romelia Salomon-Ferrer, Andreas W. Götz, Duncan Poole, Scott Le Grand, and Ross C. Walker. “Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald”. In: *Journal of Chemical Theory and Computation* 9.9 (Sept. 2013), pp. 3878–3888. DOI: 10.1021/ct400314y.

[109] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. “Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes”. In: *Journal of Computational Physics* 23.3 (Mar. 1977), pp. 327–341. DOI: 10.1016/0021-9991(77)90098-5.

[110] Shuichi Miyamoto and Peter a. Kollman. “Molecular dynamics studies of calixspherand complexes with alkali metal cations: calculation of the absolute and relative free energies of binding of cations to a calixspherand”. In: *Journal of the American Chemical Society* 114.10 (May 1992), pp. 3668–3674. DOI: 10.1021/ja00036a015.

[111] Daniel R. Roe and Thomas E. Cheatham. “PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data”. In: *Journal of Chemical Theory and Computation* 9.7 (July 2013), pp. 3084–3095. DOI: 10.1021/ct400341p.

[112] Adam Zemla. “LGA: A method for finding 3D similarities in protein structures”. In: *Nucleic Acids Research* 31.13 (July 2003), pp. 3370–3374. DOI: 10.1093/nar/gkg571.

[113] W Kabsch and C Sander. “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.” In: *Biopolymers* 22.12 (1983), pp. 2577–2637. DOI: 10.1002/bip.360221211.

[114] K W Plaxco, K T Simons, and D Baker. “Contact order, transition state placement and the refolding rates of single domain proteins.” In: *Journal of molecular biology* 277.4 (1998), pp. 985–994. DOI: 10.1006/jmbi.1998.1645.

[115] S.T. Rao and Michael G. Rossmann. “Comparison of super-secondary structures in proteins”. In: *Journal of Molecular Biology* 76.2 (May 1973), pp. 241–256. DOI: 10.1016/0022-2836(73)90388-4.

[116] Israel Hanukoglu. “Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites”. In: *Biochemistry and Molecular Biology Education* 43.3 (May 2015), pp. 206–209. DOI: 10.1002/bmb.20849.

[117] F. M G Pearl. “The CATH database: an extended protein family resource for structural and functional genomics”. In: *Nucleic Acids Research* 31.1 (Jan. 2003), pp. 452–455. DOI: 10.1093/nar/gkg062.

[118] J. Ko, H. Park, L. Heo, and C. Seok. “GalaxyWEB server for protein structure prediction and refinement”. In: *Nucleic Acids Research* 40.W1 (May 2012), W294–W297. DOI: 10.1093/nar/gks493.

[119] L. Heo, H. Park, and C. Seok. “GalaxyRefine: protein structure refinement driven by side-chain repacking”. In: *Nucleic Acids Research* 41.W1 (July 2013), W384–W388. DOI: 10.1093/nar/gkt458.

[120] Alpan Raval, Stefano Piana, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. “Refinement of protein structure homology models via long, all-atom molecular dynamics simulations”. In: *Proteins: Structure, Function, and Bioinformatics* 80.8 (2012), pp. 2071–2079. DOI: 10.1002/prot.24098.

[121] Matthew P. Jacobson, David L. Pincus, Chaya S. Rapp, Tyler J.F. Day, Barry Honig, David E. Shaw, and Richard A. Friesner. “A hierarchical approach to all-atom protein loop prediction”. In: *Proteins: Structure, Function, and Bioinformatics* 55.2 (Mar. 2004), pp. 351–367. DOI: 10.1002/prot.10613.

[122] Michael Feig and Vahid Mirjalili. “Protein structure refinement via molecular-dynamics simulations: What works and what does not?” In: *Proteins: Structure, Function, and Bioinformatics* 84.August (Sept. 2016), pp. 282–292. DOI: 10.1002/prot.24871.

[123] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[124] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. “UCSF Chimera - A visualization system for exploratory research and analysis”. In: *Journal of Computational Chemistry* 25.13 (Oct. 2004), pp. 1605–1612. DOI: 10.1002/jcc.20084.

[125] John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. “How many drug targets are there?” In: *Nature reviews. Drug discovery* 5.12 (Dec. 2006), pp. 993–996. DOI: 10.1038/nrd2199.

[126] Gábor E Tusnády, Zsuzsanna Dosztányi, and István Simon. “Transmembrane proteins in the Protein Data Bank: Identification and classification”. In: *Bioinformatics* 20.17 (Nov. 2004), pp. 2964–2972. DOI: 10.1093/bioinformatics/bth340.

[127] Christopher G Tate and Gebhard FX Schertler. *Engineering G protein-coupled receptors to facilitate their structure determination*. Aug. 2009. DOI: 10.1016/j.sbi.2009.07.004.

[128] Isabelle Mus-Veteau. “Heterologous Expression of Membrane Proteins”. In: *Heterologous Expression of Membrane Proteins: Methods and Protocols*. Methods in Molecular Biology 601 (2009). Ed. by Isabelle Mus-Veteau, p. 272. DOI: 10.1007/978-1-60761-344-2.

[129] Brian K Kobilka. “G protein coupled receptor structure and activation.” In: *Biochimica et biophysica acta* 1768.4 (Apr. 2007), pp. 794–807. DOI: 10.1016/j.bbamem.2006.10.021.

[130] CongBao Kang and Qingxin Li. *Solution NMR study of integral membrane proteins*. Aug. 2011. DOI: 10.1016/j.cbpa.2011.05.025.

[131] Hak Jun Kim, Stanley C Howell, Wade D Van Horn, Young Ho Jeon, and Charles R Sanders. "Recent advances in the application of solution NMR spectroscopy to multi-span integral membrane proteins". In: *Progress in Nuclear Magnetic Resonance Spectroscopy* 55.4 (Nov. 2009), pp. 335–360. DOI: 10.1016/j.pnmrs.2009.07.002.

[132] Nathan S Alexander, Anita M Preininger, Ali I Kaya, Richard a Stein, Heidi E Hamm, and Jens Meiler. "Energetic analysis of the rhodopsin-G-protein complex links the α5 helix to GDP release." In: *Nature structural & molecular biology* 21.1 (2014), pp. 56–63. DOI: 10.1038/nsmb.2705.

[133] Wayne L Hubbell and Christian Altenbach. *Investigation of structure and dynamics in membrane proteins using site-directed spin labeling*. 1994. DOI: 10.1016/S0959-440X(94)90219-4.

[134] Jinhui Dong, Guangyong Yang, and Hassane S McHaourab. "Structural basis of energy transduction in the transport cycle of MsbA." In: *Science (New York, N.Y.)* 308.5724 (May 2005), pp. 1023–1028. DOI: 10.1126/science.1106592.

[135] Aleksander Czogalla, Aldona Pieciul, Adam Jezierski, and Aleksander F Sikorski. "Attaching a spin to a protein – site-directed spin labeling in structural biology." In: *Acta biochimica Polonica* 54.2 (Jan. 2007), pp. 235–44.

[136] Hanane A Koteiche, Anderee R Berengian, and Hassane S Mchaourab. "Identification of protein folding patterns using site-directed spin labeling. Structural characterization of a beta-sheet and putative substrate binding regions in the conserved domain of alpha A-crystallin." In: *Biochemistry* 37.37 (1998), pp. 12681–12688. DOI: 10.1021/bi9814078.

[137] Hanane A Koteiche and Hassane S Mchaourab. "Folding pattern of the alpha-crystallin domain in alphaA-crystallin determined by site-directed spin labeling." In: *Journal of molecular biology* 294.2 (Nov. 1999), pp. 561–577. DOI: 10.1006/jmbi.1999.3242.

[138] Christian Altenbach, T Marti, H G Khorana, and Wayne L Hubbell. "Transmembrane protein structure: spin labeling of bacteriorhodopsin mutants." In: *Science (New York, N.Y.)* 248.4959 (June 1990), pp. 1088–1092. DOI: 10.1126/science.2160734.

[139] Michael a Lietzow and Wayne L Hubbell. "Motion of Spin Label Side Chains in Cellular Retinol-Binding Protein: Correlation with Structure and Nearest-Neighbor Interactions in An Antiparallel ??-Sheet". In: *Biochemistry* 43.11 (Mar. 2004), pp. 3137–3151. DOI: 10.1021/bi0360962.

[140] Christian Altenbach, Ke Yang, David L Farrens, Zohreh T Farahbakhsh, H. Gobind Khorana, and Wayne L Hubbell. "Structural features and light-dependent changes in the cytoplasmic interhelical E-F loop region of rhodopsin: A site-directed spin-labeling study". In: *Biochemistry* 35.38 (Sept. 1996), pp. 12470–12478. DOI: 10.1021/bi960849l.

[141] L Salwiński and Wayne L Hubbell. "Structure in the channel forming domain of colicin E1 bound to membranes: the 402-424 sequence." In: *Protein science : a publication of the Protein Society* 8.3 (Mar. 1999), pp. 562–572. DOI: 10.1110/ps.8.3.562.

[142] Petr P Borbat, Hassane S Mchaourab, and Jack H Freed. "Protein structure determination using long-distance constraints from double-quantum coherence ESR: Study of T4 lysozyme". In: *Journal of the American Chemical Society* 124.19 (May 2002), pp. 5304–5314. DOI: 10.1021/ja020040y.

[143] Gunnar Jeschke and Yevhen Polyhach. "Distance measurements on spin-labelled biomacromolecules by pulsed electron paramagnetic resonance." In: *Physical chemistry chemical physics : PCCP* 9.16 (Apr. 2007), pp. 1895–1910. DOI: 10.1039/b614920k.

[144] Ping Zou, Marco Bortolus, and Hassane S Mchaourab. “Conformational Cycle of the ABC Transporter MsbA in Liposomes: Detailed Analysis Using Double Electron-Electron Resonance Spectroscopy”. In: *Journal of Molecular Biology* 393.3 (Oct. 2009), pp. 586–597. DOI: 10.1016/j.jmb.2009.08.050.

[145] Christian Altenbach, Ana Karin Kusnetzow, Oliver P Ernst, Klaus Peter Hofmann, and Wayne L Hubbell. “High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation.” In: *Proceedings of the National Academy of Sciences of the United States of America* 105.21 (May 2008), pp. 7439–7444. DOI: 10.1073/pnas.0802515105.

[146] Derek P Claxton, Matthias Quick, Lei Shi, Fernanda Delmondes de Carvalho, Harel Weinstein, Jonathan A Javitch, and Hassane S McHaourab. “Ion/substrate-dependent conformational dynamics of a bacterial homolog of neurotransmitter:sodium symporters.” In: *Nature structural & molecular biology* 17.7 (July 2010), pp. 822–829. DOI: 10.1038/nsmb.1854.

[147] Sudha Chakrapani, Pornthep Sompornpisut, Pathumwadee Intharathep, Benoît Roux, and Eduardo Perozo. “The activated state of a sodium channel voltage sensor in a membrane environment.” In: *Proceedings of the National Academy of Sciences of the United States of America* 107.12 (Mar. 2010), pp. 5435–5440. DOI: 10.1073/pnas.0914109107.

[148] Valeria Vásquez, Marcos Sotomayor, D. Marien Cortes, Benoît Roux, Klaus Schulten, and Eduardo Perozo. “Three-Dimensional Architecture of Membrane-Embedded MscS in the Closed Conformation”. In: *Journal of Molecular Biology* 378.1 (Apr. 2008), pp. 55–70. DOI: 10.1016/j.jmb.2007.10.086.

[149] Stephanie Hirst, Nathan Alexander, Hassane S. Mchaourab, and Jens Meiler. “ROSETTAEPR: An Integrated Tool for Protein Structure Determination From Sparse EPR Data”. In: *Biophysical Journal* 100.3 (2011), 216a. DOI: 10.1016/j.bpj.2010.12.1390.

[150] Ian Mitchelle S de Vera, Mandy E. Blackburn, Luis Galiano, and Gail E. Fanucci. “Pulsed EPR distance measurements in soluble proteins by Site-Directed Spin Labeling (SDSL)”. In: *Current Protocols in Protein Science* SUPPL.74 (2013), pp. 1–29. DOI: 10.1002/0471140864.ps1717s74.

[151] Yunhuang Yang, Theresa a Ramelot, Robert M McCarrick, Shuisong Ni, Erik a Feldmann, John R Cort, Huang Wang, Colleen Ciccosanti, Mei Jiang, Haleema Janjua, Thomas B Acton, Rong Xiao, John K Everett, Gaetano T Montelione, and Michael a Kennedy. “Combining NMR and EPR methods for homodimer protein structure determination”. In: *Journal of the American Chemical Society* 132.34 (Sept. 2010), pp. 11910–11913. DOI: 10.1021/ja105080h.

[152] Vladimir Yarov-Yarovoy, Jack Schonbrun, and David Baker. “Multipass membrane protein structure prediction using Rosetta.” In: *Proteins* 62.4 (Mar. 2006), pp. 1010–1025. DOI: 10.1002/prot.20817.

[153] P Barth, B Wallner, and David Baker. “Prediction of membrane protein structures with complex topologies using limited constraints.” In: *Proceedings of the National Academy of Sciences of the United States of America* 106.5 (Feb. 2009), pp. 1409–1414. DOI: 10.1073/pnas.0808323106.

[154] Brian E Weiner, Nathan S Alexander, Louesa R Akin, Nils Wötzel, Mert Karakaş, and Jens Meiler. “BCL::Fold–protein topology determination from limited NMR restraints.” In: *Proteins* 82.4 (Apr. 2014), pp. 587–95. DOI: 10.1002/prot.24427.

[155] Kelli Kazmier, Nathan S Alexander, Jens Meiler, and Hassane S Mchaourab. “Algorithm for selection of optimized EPR distance restraints for de novo protein structure determination”. In: *Journal of Structural Biology* 173.3 (Mar. 2011), pp. 549–557. DOI: 10.1016/j.jsb.2010.11.003.

[156] D Eisenberg, R M Weiss, and T C Terwilliger. “The hydrophobic moment detects periodicity in protein hydrophobicity.” In: *Proceedings of the National Academy of Sciences of the United States of America* 81.1 (Jan. 1984), pp. 140–144. DOI: 10.1073/pnas.81.1.140.

[157] Sudha Chakrapani, Luis G Cuello, D Marien Cortes, and Eduardo Perozo. “Structural Dynamics of an Isolated Voltage-Sensor Domain in a Lipid Bilayer”. In: *Structure* 16.3 (Mar. 2008), pp. 398–409. DOI: 10.1016/j.str.2007.12.015.

[158] Olivier Dalmas, Luis G Cuello, Vishwanath Jogini, D Marien Cortes, Benoit Roux, and Eduardo Perozo. “Structural Dynamics of the Magnesium-Bound Conformation of CorA in a Lipid Bilayer”. In: *Structure* 18.7 (July 2010), pp. 868–878. DOI: 10.1016/j.str.2010.04.009.

[159] Ping Zou and Hassane S Mchaourab. “Alternating Access of the Putative Substrate-Binding Chamber in the ABC Transporter MsbA”. In: *Journal of Molecular Biology* 393.3 (Oct. 2009), pp. 574–585. DOI: 10.1016/j.jmb.2009.08.051.

[160] P Barth, Jack Schonbrun, and David Baker. “Toward high-resolution prediction and design of transmembrane helical protein structures.” In: *Proceedings of the National Academy of Sciences of the United States of America* 104.40 (Oct. 2007), pp. 15682–15687. DOI: 10.1073/pnas.0702515104.

[161] Christian Altenbach, D a Greenhalgh, H G Khorana, and Wayne L Hubbell. “A collision gradient method to determine the immersion depth of nitroxides in lipid bilayers: application to spin-labeled mutants of bacteriorhodopsin.” In: *Proceedings of the National Academy of Sciences of the United States of America* 91.5 (Mar. 1994), pp. 1667–1671. DOI: 10.1073/pnas.91.5.1667.

[162] April A Frazier, Mark A Wisner, Nathan J Malmberg, Kenneth G Victor, Gail E Fanucci, Eric A Nalefski, Joseph J Falke, and David S Cafiso. “Membrane orientation and position of the C2 domain from cPLA2 by site-directed spin labeling”. In: *Biochemistry* 41.20 (May 2002), pp. 6282–6292. DOI: 10.1021/bi0160821.

[163] Robert D Nielsen, Kepeng Che, Michael H Gelb, and Bruce H Robinson. “A ruler for determining the position of proteins in membranes”. In: *Journal of the American Chemical Society* 127.17 (May 2005), pp. 6430–6442. DOI: 10.1021/ja042782s.

[164] Håkan Viklund and Arne Elofsson. “OCTOPUS: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar”. In: *Bioinformatics* 24.15 (Aug. 2008), pp. 1662–1668. DOI: 10.1093/bioinformatics/btn221.

[165] David Baker and Andrej Sali. “Protein structure prediction and structural genomics.” In: *Science (New York, N.Y.)* 294.5540 (Oct. 2001), pp. 93–96. DOI: 10.1126/science.1065659.

[166] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. “The Pfam protein families database”. In: *Nucleic Acids Research* 40.D1 (2011), pp. D290–D301. DOI: 10.1093/nar/gkr1065.

[167] Kamil Khafizov, Carlos Madrid-Aliste, Steven C Almo, and Andras Fiser. “Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative.” In: *Proceedings of the National Academy of Sciences of the United States of America* 111.10 (2014), pp. 3733–8. DOI: 10.1073/pnas.1321614111.

[168] John Moult. *A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction*. June 2005. DOI: 10.1016/j.sbi.2005.05.011.

[169] Richard Bonneau, Jerry Tsai, Ingo Ruczinski, Dylan Chivian, Carol Rohl, Charlie E M Strauss, and David Baker. "Rosetta in CASP4: Progress in ab initio protein structure prediction". In: *Proteins: Structure, Function and Genetics* 45.SUPPL. 5 (Jan. 2001), pp. 119–126. DOI: 10.1002/prot.1170.

[170] Richard Bonneau, Charlie E M Strauss, Carol a Rohl, Dylan Chivian, Phillip Bradley, Lars Malmström, Tim Robertson, and David Baker. "De novo prediction of three-dimensional structures for major protein families". In: *Journal of Molecular Biology* 322.1 (Sept. 2002), pp. 65–78. DOI: 10.1016/S0022-2836(02)00698-8.

[171] Peter M Bowers, C E Strauss, and David Baker. "De novo protein structure determination using sparse NMR data." In: *Journal of biomolecular NMR* 18.4 (2000), pp. 311–318. DOI: 10.1023/A:1026744431105.

[172] Evgeniy V Petrotchenko and Christoph H Borchers. "ICC-CLASS: isotopically-coded cleavable crosslinking analysis software suite." In: *BMC bioinformatics* 11 (Jan. 2010), p. 64. DOI: 10.1186/1471-2105-11-64.

[173] Andrea Sinz. "Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions." In: *Mass spectrometry reviews* 25.4 (2006), pp. 663–82. DOI: 10.1002/mas.20082.

[174] Juri Rappsilber. "The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes." In: *Journal of structural biology* 173.3 (Mar. 2011), pp. 530–40. DOI: 10.1016/j.jsb.2010.10.014.

[175] M M Young, N Tang, J C Hempel, C M Oshiro, E W Taylor, I D Kuntz, B W Gibson, and G Dollinger. "High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry." In: *Proceedings of the National Academy of Sciences of the United States of America* 97.11 (2000), pp. 5802–5806. DOI: 10.1073/pnas.090099097.

[176] K. Lasker, F. Forster, S. Bohn, T. Walzthoeni, E. Villa, P. Unverdorben, F. Beck, R. Aebersold, A. Sali, and W. Baumeister. "Inaugural Article: Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach". In: *Proceedings of the National Academy of Sciences* 109.5 (2012), pp. 1380–1387. DOI: 10.1073/pnas.1120559109.

[177] Richard B Jacobsen, Kenneth L Sale, Marites J Ayson, Petr Novak, Joohee Hong, Pamela Lane, Nichole L Wood, Gary H Kruppa, Malin M Young, and Joseph S Schoeniger. "Structure and dynamics of dark-state bovine rhodopsin revealed by chemical cross-linking and high-resolution mass spectrometry." In: *Protein science : a publication of the Protein Society* 15.6 (2006), pp. 1303–1317. DOI: 10.1110/ps.052040406.

[178] Stefan Kalkhof, Christian Ihling, Karl Mechtler, and Andrea Sinz. "Chemical cross-linking and high-performance Fourier transform ion cyclotron resonance mass spectrometry for protein interaction analysis: Application to a calmodulin/target peptide complex". In: *Analytical Chemistry* 77.2 (2005), pp. 495–503. DOI: 10.1021/ac0487294.

[179] Andrea Sinz. “Investigation of protein-protein interactions in living cells by chemical crosslinking and mass spectrometry”. In: *Analytical and Bioanalytical Chemistry* 397.8 (2010), pp. 3433–3440. DOI: 10.1007/s00216-009-3405-5.

[180] Verena Tinnefeld, Albert Sickmann, and Robert Ahrends. “Catch me if you can: Challenges and applications of cross-linking approaches”. In: *European Journal of Mass Spectrometry* 20.1 (2014), pp. 99–116. DOI: 10.1255/ejms.1259.

[181] Abdullah Kahraman, Franz Herzog, Alexander Leitner, George Rosenberger, Ruedi Aebersold, and Lars Malmström. “Cross-Link Guided Molecular Modeling with ROSETTA”. In: *PLoS ONE* 8.9 (2013). DOI: 10.1371/journal.pone.0073411.

[182] Stefan Kalkhof, Sebastian Haehn, Mats Paulsson, Neil Smyth, Jens Meiler, and Andrea Sinz. “Computational modeling of laminin N-terminal domains using sparse distance constraints from disulfide bonds and chemical cross-linking”. In: *Proteins: Structure, Function and Bioinformatics* 78.16 (2010), pp. 3409–3427. DOI: 10.1002/prot.22848.

[183] Alexander Leitner, Thomas Walzthoeni, Abdullah Kahraman, Franz Herzog, Oliver Rinner, Martin Beck, and Ruedi Aebersold. “Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics.” In: *Molecular & cellular proteomics : MCP* 9.8 (2010), pp. 1634–1649. DOI: 10.1074/mcp.R000001-MCP201.

[184] Boris L Zybailov, Galina V Glazko, Mihir Jaiswal, and Kevin D Raney. “Large Scale Chemical Cross-linking Mass Spectrometry Perspectives.” In: *Journal of proteomics & bioinformatics* 6.Suppl 2 (2013), p. 001. DOI: 10.4172/jpb.S2-001.

[185] Guoli Wang and Roland L. Dunbrack. “PISCES: A protein sequence culling server”. In: *Bioinformatics* 19.12 (2003), pp. 1589–1591. DOI: 10.1093/bioinformatics/btg224.

[186] Abdullah Kahraman, Lars Malmström, and Ruedi Aebersold. “Xwalk: Computing and visualizing distances in cross-linking experiments”. In: *Bioinformatics* 27.15 (Aug. 2011), pp. 2163–2164. DOI: 10.1093/bioinformatics/btr348.

[187] Daniela M. Schulz, Stefan Kalkhof, Andreas Schmidt, Christian Ihling, Christoph Stingl, Karl Mechtler, Olaf Zschörnig, and Andrea Sinz. “Annexin A2/P11 interaction: New insights into annexin A2 tetramer structure by chemical crosslinking, high-resolution mass spectrometry, and computational modeling”. In: *Proteins: Structure, Function and Genetics* 69.2 (2007), pp. 254–269. DOI: 10.1002/prot.21445.

[188] Timothy F Havel, Gordon M Crippen, and D Irwin. “Effects of Distance Constraints on Macromolecular Conformation . 11 . Simulation of Experimental Results and Theoretical Predictions”. In: *Biopolymers* 18 (1979), pp. 73–81. DOI: 10.1002/bip.1979.360180108.

[189] Eric D Merkley, Steven Rysavy, Abdullah Kahraman, Ryan P Hafen, Valerie Daggett, and Joshua N Adkins. “Distance restraints from crosslinking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine-lysine distances.” In: *Protein science : a publication of the Protein Society* 23.6 (June 2014), pp. 747–59. DOI: 10.1002/pro.2458.

[190] N S Green, E Reisler, and K N Houk. “Quantitative evaluation of the lengths of homobifunctional protein cross-linking reagents used as molecular rulers.” In: *Protein science : a publication of the Protein Society* 10.7 (2001), pp. 1293–1304. DOI: 10.1101/ps.51201.The.

[191] Carol A. Rohl, Charlie E.M. Strauss, Dylan Chivian, and David Baker. “Modeling structurally variable regions in homologous proteins with rosetta”. In: *Proteins: Structure, Function, and Bioinformatics* 55.3 (Apr. 2004), pp. 656–677. DOI: 10.1002/prot.10629.

[192] MD Tyka, Kenneth Jung, and David Baker. “Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers”. In: *Journal of computational chemistry* 33.31 (2012), pp. 2483–2491. DOI: 10.1002/jcc.23069.Efficient.

[193] Daniel J Mandell, Evangelos a Coutsias, and Tanja Kortemme. “Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling”. In: *Nature Methods* 6.8 (Aug. 2009), pp. 551–552. DOI: 10.1038/nmeth0809-551.

[194] Frances C Bernstein, Thomas F Koetzle, Graheme J.B. Williams, Edgar F Meyer, Michael D Brice, John R Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. “The protein data bank: A computer-based archival file for macromolecular structures”. In: *Journal of Molecular Biology* 112.3 (May 1977), pp. 535–542. DOI: 10.1016/S0022-2836(77)80200-3.

[195] Guoli Wang and Roland L. Dunbrack. “PISCES: Recent improvements to a PDB sequence culling server”. In: *Nucleic Acids Research* 33.SUPPL. 2 (2005), pp. 94–98. DOI: 10.1093/nar/gki402.

[196] Richard Pio. “Euler angle transformations”. In: *IEEE Transactions on Automatic Control* 11.4 (Oct. 1966), pp. 707–715. DOI: 10.1109/TAC.1966.1098430.

[197] Gregory Slabaugh. *Computing Euler angles from a rotation matrix*. Tech. rep. 1999.

[198] J. J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones. “Secondary structure prediction with support vector machines”. In: *Bioinformatics* 19.13 (2003), pp. 1650–1655. DOI: 10.1093/bioinformatics/btg223.

[199] R. M. Fine, H. Wang, P. S. Shenkin, D. L. Yarmush, and C. Levinthal. “Predicting antibody hypervariable loop conformations II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations”. In: *Proteins: Structure, Function, and Genetics* 1.4 (Apr. 1986), pp. 342–362. DOI: 10.1002/prot.340010408.

[200] Peter S. Shenkin, David L. Yarmush, Richard M. Fine, Huajun Wang, and Cyrus Levinthal. “Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures”. In: *Biopolymers* 26.12 (Dec. 1987), pp. 2053–2085. DOI: 10.1002/bip.360261207.

[201] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega”. In: *Molecular Systems Biology* 7.1 (Apr. 2014), pp. 539–539. DOI: 10.1038/msb.2011.75.

[202] Stephanie Bleicken, Gunnar Jeschke, Carolin Stegmueller, Raquel Salvador-Gallego, Ana J. García-Sáez, and Enrica Bordignon. “Structural Model of Active Bax at the Membrane”. In: *Molecular Cell* 56.4 (Oct. 2014), pp. 496–505. DOI: 10.1016/j.molcel.2014.09.022.

[203] M. E. Bamberger and G. E. Landreth. “Inflammation, Apoptosis, and Alzheimer’s Disease”. In: *The Neuroscientist* 8.3 (June 2002), pp. 276–283. DOI: 10.1177/1073858402008003013.

[204] Katsumi Eguchi. “Apoptosis in autoimmune diseases.” In: *Internal medicine (Tokyo, Japan)* 40.4 (2001), pp. 275–284. DOI: 10.2169/internalmedicine.40.275.

[205] B. Favaloro, N. Allocati, V. Graziano, C. Di Ilio, and V. De Laurenzi. “Role of apoptosis in disease”. In: *Aging* 4.5 (2012), pp. 330–349.

[206] Brian Leber, Jialing Lin, and David W Andrews. “Still embedded together binding to membranes regulates Bcl-2 protein interactions.” In: *Oncogene* 29.38 (Sept. 2010), pp. 5221–5230. DOI: 10.1038/onc.2010.283.

[207] Andreas Strasser, L O’Connor, and Vishva M Dixit. “Apoptosis signaling.” In: *Annual review of biochemistry* 69.0066-4154 LA - eng PT - Journal Article PT - Review PT - Review, Academic (2000), pp. 217–245. DOI: 10.1146/annurev.biochem.69.1.217.

[208] Peter E Czabotar, Guillaume Lessene, Andreas Strasser, and Jerry M Adams. “Control of apoptosis by the BCL-2 protein family: implications for physiology and therapy.” In: *Nature reviews. Molecular cell biology* 15.1 (2014), pp. 49–63. DOI: 10.1038/nrm3722.

[209] Richard J Youle and Andreas Strasser. “The BCL-2 protein family: opposing activities that mediate cell death.” In: *Nature reviews. Molecular cell biology* 9.1 (Jan. 2008), pp. 47–59. DOI: 10.1038/nrm2308.

[210] Motoshi Suzuki, Richard J Youle, and Nico Tjandra. “Structure of Bax: coregulation of dimer formation and intracellular localization.” In: *Cell* 103.4 (Nov. 2000), pp. 645–654. DOI: http://dx.doi.org/10.1016/S0092-8674(00)00167-7.

[211] Peter E Czabotar, Dana Westphal, Grant Dewson, Stephen Ma, Colin Hockings, W Douglas Fairlie, Erinna F Lee, Shenggen Yao, Adeline Y Robin, Brian J Smith, David C S Huang, Ruth M Kluck, Jerry M Adams, and Peter M Colman. “Bax crystal structures reveal how BH3 domains activate Bax and nucleate its oligomerization to induce apoptosis”. In: *Cell* 152.3 (Jan. 2013), pp. 519–531. DOI: 10.1016/j.cell.2012.12.031.

[212] Dana Westphal, Grant Dewson, Marie Menard, Paul Frederick, Sweta Iyer, Ray Bartolo, Leonie Gibson, Peter E Czabotar, Brian J. Smith, Jerry M Adams, and Ruth M Kluck. “Apoptotic pore formation is associated with in-plane insertion of Bak or Bax central helices into the mitochondrial outer membrane.” In: *Proceedings of the National Academy of Sciences of the United States of America* 111.39 (2014), E4076–E4085. DOI: 10.1073/pnas.1415142111.

[213] M Pannier, S Veit, A Godt, Gunnar Jeschke, and H W Spiess. “Dead-time free measurement of dipole-dipole interactions between electron spins.” In: *Journal of magnetic resonance (San Diego, Calif. : 1997)* 142.2 (Feb. 2000), pp. 331–340. DOI: 10.1016/j.jmr.2011.08.035.

[214] Yevhen Polyhach, Enrica Bordignon, and Gunnar Jeschke. “Rotamer libraries of spin labelled cysteines for protein studies”. In: *Phys. Chem. Chem. Phys.* 13.6 (2011), pp. 2356–2366. DOI: 10.1039/C0CP01865A.

[215] Claire Gendrin, Carlos Contreras-Martel, Stéphanie Bouillot, Sylvie Elsen, David Lemaire, Dimitrios A. Skoufias, Philippe Huber, Ina Attree, and Andréa Dessen. “Structural Basis of Cytotoxicity Mediated by the Type III Secretion Toxin ExoU from Pseudomonas aeruginosa”. In: *PLoS Pathogens* 8.4 (Apr. 2012). Ed. by Mark A. Saper, e1002637. DOI: 10.1371/journal.ppat.1002637.

[216] Andrei S. Halavaty, Dominika Borek, Gregory H. Tyson, Jeff L. Veesenmeyer, Ludmilla Shuvalova, George Minasov, Zbyszek Otwinowski, Alan R. Hauser, and Wayne F. Anderson. “Structure of the Type III Secretion Effector Protein ExoU in Complex with Its Chaperone SpcU”. In: *PLoS ONE* 7.11 (Nov. 2012). Ed. by Gunnar F. Kaufmann, e49388. DOI: 10.1371/journal.pone.0049388.

[217] Rebecca M. Phillips. “In Vivo Phospholipase Activity of the Pseudomonas aeruginosa Cytotoxin ExoU and Protection of Mammalian Cells with Phospholipase A2 Inhibitors”. In: *Journal of Biological Chemistry* 278.42 (July 2003), pp. 41326–41332. DOI: 10.1074/jbc.M302472200.

[218] H Sato. “The mechanism of action of the Pseudomonas aeruginosa-encoded type III cytotoxin, ExoU”. In: *The EMBO Journal* 22.12 (June 2003), pp. 2959–2969. DOI: 10.1093/emboj/cdg290.

[219] Ciara M Shaver and Alan R Hauser. “Relative Contributions of Pseudomonas aeruginosa ExoU, ExoS, and ExoT to Virulence in the Lung”. In: *Infection and Immunity* 72.12 (Dec. 2004), pp. 6969–6977. DOI: 10.1128/IAI.72.12.6969-6977.2004.

[220] Arup Roy-Burman, Richard H. Savel, Sara Racine, Britta L. Swanson, Neelambika S. Revadigar, Junichi Fujimoto, Teiji Sawa, Dara W. Frank, and Jeanine P. Wiener-Kronish. “Type III Protein Secretion Is Associated with Death in Lower Respiratory and Systemic Pseudomonas aeruginosa Infections”. In: *The Journal of Infectious Diseases* 183.12 (June 2001), pp. 1767–1774. DOI: 10.1086/320737.

[221] Markus Allewelt, Fadie T. Coleman, Martha Grout, Gregory P. Priebe, and Gerald B. Pier. “Acquisition of Expression of the Pseudomonas aeruginosa ExoU Cytotoxin Leads to Increased Bacterial Virulence in a Murine Model of Acute Pneumonia and Systemic Spread”. In: *Infection and Immunity* 68.7 (July 2000), pp. 3998–4004. DOI: 10.1128/IAI.68.7.3998-4004.2000.

[222] Gregory H. Tyson, Andrei S. Halavaty, Hyunjin Kim, Brett Geissler, Mallory Agard, Karla J. Satchell, Wonhwa Cho, Wayne F. Anderson, and Alan R. Hauser. “A Novel Phosphatidylinositol 4,5-Bisphosphate Binding Domain Mediates Plasma Membrane Localization of ExoU and Other Patatin-like Phospholipases”. In: *Journal of Biological Chemistry* 290.5 (Jan. 2015), pp. 2919–2937. DOI: 10.1074/jbc.M114.611251.

[223] Maxx H Tessmer, David M Anderson, Adam Buchaklian, Dara W. Frank, and Jimmy B Feix. “Cooperative substrate-cofactor interactions and membrane localization of the bacterial PLA2 enzyme, ExoU”. In: *Journal of Biological Chemistry* 414 (Jan. 2017), jbc.M116.760074. DOI: 10.1074/jbc.M116.760074.

[224] Marc A. Benson, Steven M. Komas, Katherine M. Schmalzer, Monika S. Casey, Dara W. Frank, and Jimmy B. Feix. “Induced Conformational Changes in the Activation of the Pseudomonas aeruginosa type III Toxin, ExoU”. In: *Biophysical Journal* 100.5 (Mar. 2011), pp. 1335–1343. DOI: 10.1016/j.bpj.2011.01.056.

[225] L. J. McGuffin, Kevin Bryson, and David T. Jones. “The PSIPRED protein structure prediction server”. In: *Bioinformatics* 16.4 (Apr. 2000), pp. 404–405. DOI: 10.1093/bioinformatics/16.4.404.

[226] Peter J Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[227] Renato Cordeiro de Amorim and Christian Hennig. “Recovering the number of clusters in data sets with noise features using feature rescaling factors”. In: *Information Sciences* 324 (Dec. 2015), pp. 126–145. DOI: 10.1016/j.ins.2015.06.039.

[228] Kristian W. Kaufmann, Gordon H. Lemmon, Samuel L. DeLuca, Jonathan H. Sheehan, and Jens Meiler. “Practically Useful: What the R <scp>osetta</scp> Protein Modeling Suite Can Do for You”. In: *Biochemistry* 49.14 (Apr. 2010), pp. 2987–2998. DOI: 10.1021/bi902153g.

[229] Gunnar Jeschke, V Chechik, P Ionita, A. Godt, H. Zimmermann, J. Banham, C. R. Timmel, D. Hilger, and H. Jung. "DeerAnalysis2006 - a comprehensive software package for analyzing pulsed ELDOR data". In: *Applied Magnetic …* 498 (2006), pp. 473–498. DOI: 10.1007/BF03166213.

[230] C G Tate. "The projection structure of EmrE, a proton-linked multidrug transporter from Escherichia coli, at 7 A resolution". In: *The EMBO Journal* 20.1 (Jan. 2001), pp. 77–81. DOI: 10.1093/emboj/20.1.77.

[231] Yen-Ju Chen, Owen Pornillos, Samantha Lieu, Che Ma, Andy P Chen, and Geoffrey Chang. "X-ray structure of EmrE supports dual topology model." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.48 (2007), pp. 18999–19004. DOI: 10.1073/pnas.0709387104.

[232] Ines Lehner, Daniel Basting, Bjoern Meyer, Winfried Haase, Theofanis Manolikas, Christoph Kaiser, Michael Karas, and Clemens Glaubitz. "The Key Residue for Substrate Transport (Glu 14) in the EmrE Dimer Is Asymmetric". In: *Journal of Biological Chemistry* 283.6 (Feb. 2008), pp. 3281–3288. DOI: 10.1074/jbc.M707899200.

[233] Sepan T. Amadi, Hanane a. Koteiche, Sanjay Mishra, and Hassane S. Mchaourab. "Structure, dynamics, and substrate-induced conformational changes of the multidrug transporter EmrE in liposomes". In: *Journal of Biological Chemistry* 285.34 (2010), pp. 26710–26718. DOI: 10.1074/jbc.M110.132621.

[234] Emma A Morrison, Gregory T DeKoster, Supratik Dutta, Reza Vafabakhsh, Michael W Clarkson, Arjun Bahl, Dorothee Kern, Taekjip Ha, and Katherine A Henzler-Wildman. "Antiparallel EmrE exports drugs by exchanging between asymmetric structures". In: *Nature* 481.7379 (Dec. 2011), pp. 45–50. DOI: 10.1038/nature10703.

[235] Min-Kyu Cho, Anindita Gayen, James R Banigan, Maureen Leninger, and Nathaniel J Traaseth. "Intrinsic Conformational Plasticity of Native EmrE Provides a Pathway for Multidrug Resistance". In: *Journal of the American Chemical Society* 136.22 (June 2014), pp. 8072–8080. DOI: 10.1021/ja503145x.

[236] Sarel J Fleishman, Susan E Harrington, Angela Enosh, Dan Halperin, Christopher G Tate, and Nir Ben-Tal. "Quasi-symmetry in the Cryo-EM Structure of EmrE Provides the Key to Modeling its Transmembrane Domain". In: *Journal of Molecular Biology* 364.1 (Nov. 2006), pp. 54–67. DOI: 10.1016/j.jmb.2006.08.072.

[237] Katherine Henzler-Wildman. "Analyzing conformational changes in the transport cycle of EmrE". In: *Current Opinion in Structural Biology* 22.1 (Feb. 2012), pp. 38–43. DOI: 10.1016/j.sbi.2011.10.004.

[238] Smriti Mishra, Brandy Verhalen, Richard A Stein, Po-Chao Wen, Emad Tajkhorshid, and Hassane S Mchaourab. "Conformational dynamics of the nucleotide binding domains and the power stroke of a heterodimeric ABC transporter". In: *eLife* 3 (May 2014), e02740. DOI: 10.7554/eLife.02740.

[239] Andrej Šali and Tom L. Blundell. "Comparative Protein Modelling by Satisfaction of Spatial Restraints". In: *Journal of Molecular Biology* 234.3 (Dec. 1993), pp. 779–815. DOI: 10.1006/jmbi.1993.1626.

[240] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. "PROCHECK: a program to check the stereochemical quality of protein structures". In: *Journal of Applied Crystallography* 26.2 (Apr. 1993), pp. 283–291. DOI: 10.1107/S0021889892009944.

[241] Olav Schiemann and Thomas F. Prisner. "Long-range distance determinations in biomacromolecules by EPR spectroscopy". In: *Quarterly Reviews of Biophysics* 40.01 (Feb. 2007), p. 1. DOI: 10.1017/S003358350700460X.

[242] Hassane S. Mchaourab, P. Ryan Steed, and Kelli Kazmier. "Toward the Fourth Dimension of Membrane Protein Structure: Insight into Dynamics from Spin-Labeling EPR Spectroscopy". In: *Structure* 19.11 (Nov. 2011), pp. 1549–1561. DOI: 10.1016/j.str.2011.10.009.

[243] Misha Soskine, Shirley Mark, Naama Tayer, Roy Mizrachi, and Shimon Schuldiner. "On Parallel and Antiparallel Topology of a Homodimeric Multidrug Transporter". In: *Journal of Biological Chemistry* 281.47 (Nov. 2006), pp. 36205–36212. DOI: 10.1074/jbc.M607186200.

[244] James R Banigan, Anindita Gayen, Min-Kyu Cho, and Nathaniel J Traaseth. "A Structured Loop Modulates Coupling between the Substrate-binding and Dimerization Domains in the Multidrug Resistance Transporter EmrE". In: *Journal of Biological Chemistry* 290.2 (Jan. 2015), pp. 805–814. DOI: 10.1074/jbc.M114.601963.

[245] Hagit Yerushalmi, Mario Lebendiker, and Shimon Schuldiner. "Negative Dominance Studies Demonstrate the Oligomeric Structure of EmrE, a Multidrug Antiporter from Escherichia coli". In: *Journal of Biological Chemistry* 271.49 (Dec. 1996), pp. 31044–31048. DOI: 10.1074/jbc.271.49.31044.

[246] Anindita Gayen, James R. Banigan, and Nathaniel J. Traaseth. "Ligand-Induced Conformational Changes of the Multidrug Resistance Transporter EmrE Probed by Oriented Solid-State NMR Spectroscopy". In: *Angewandte Chemie International Edition* 52.39 (Sept. 2013), pp. 10321–10324. DOI: 10.1002/anie.201303091.

[247] Gerrit J Poelarends, Piotr Mazurkiewicz, and Wil N Konings. "Multidrug transporters and antibiotic resistance in Lactococcus lactis". In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1555.1-3 (Sept. 2002), pp. 1–7. DOI: 10.1016/S0005-2728(02)00246-3.

APPENDICES

APPENDIX A
EVALUATION OF BCL::FOLD IN THE CASP11 EXPERIMENT

This appendix is based on the publication "CASP11 – An Evaluation of a Modular BCL::Fold-Based Protein Structure Prediction Pipeline".[1] It provides supplementary data and procedures for chapter II on page 12. The supplementary data in section A.1 contains additional analysis of potential correlations between structural features the benchmark proteins and their respective prediction accuracy using the protein structure prediction pipeline detailed in chapter II on page 12. The procedures described in section A.2 on page 159 detail the protocol employed during the CASP experiment.

A.1. Supplementary data

This section provides plots regarding possible correlations between model discrimination, as quantified through the enrichment metric, the proteins CO and secondary structure content. Additionally, plots regarding the score-accuracy correlation in BCL::Fold are shown for all proteins in the CASP benchmark set that was evaluated over the course of the experiment. The provided tables list the sampling accuracy- and enrichment-values for all benchmark proteins and pipeline modules in the presence and in the absence of limited experimental data.

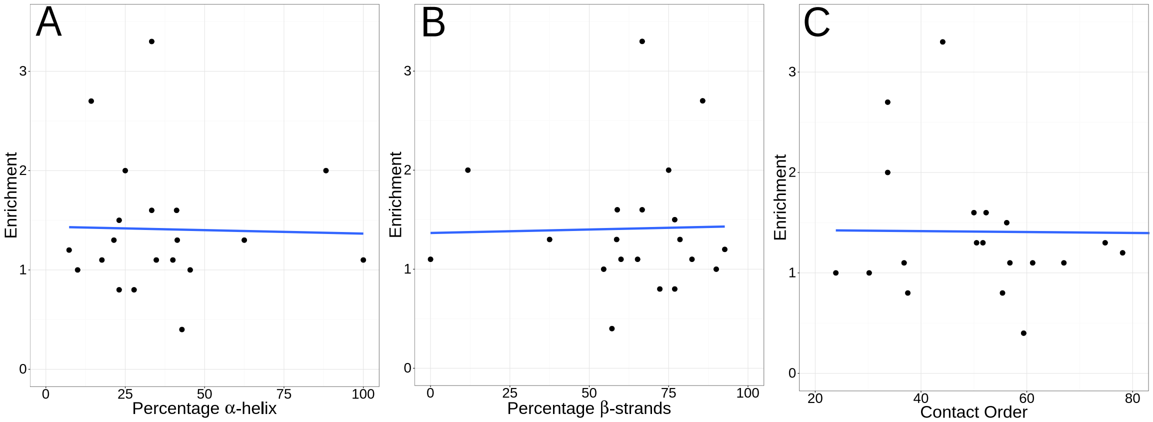

Figure A.1.: No dependence of the enrichment on secondary structure content or CO. *No correlation between the enrichment and the percentage of α-helices (A), β-strands (B), or fold complexity as quantified through the CO metric (C) could be observed. In each case, the absolute value of the R-value was less than* 0.1.

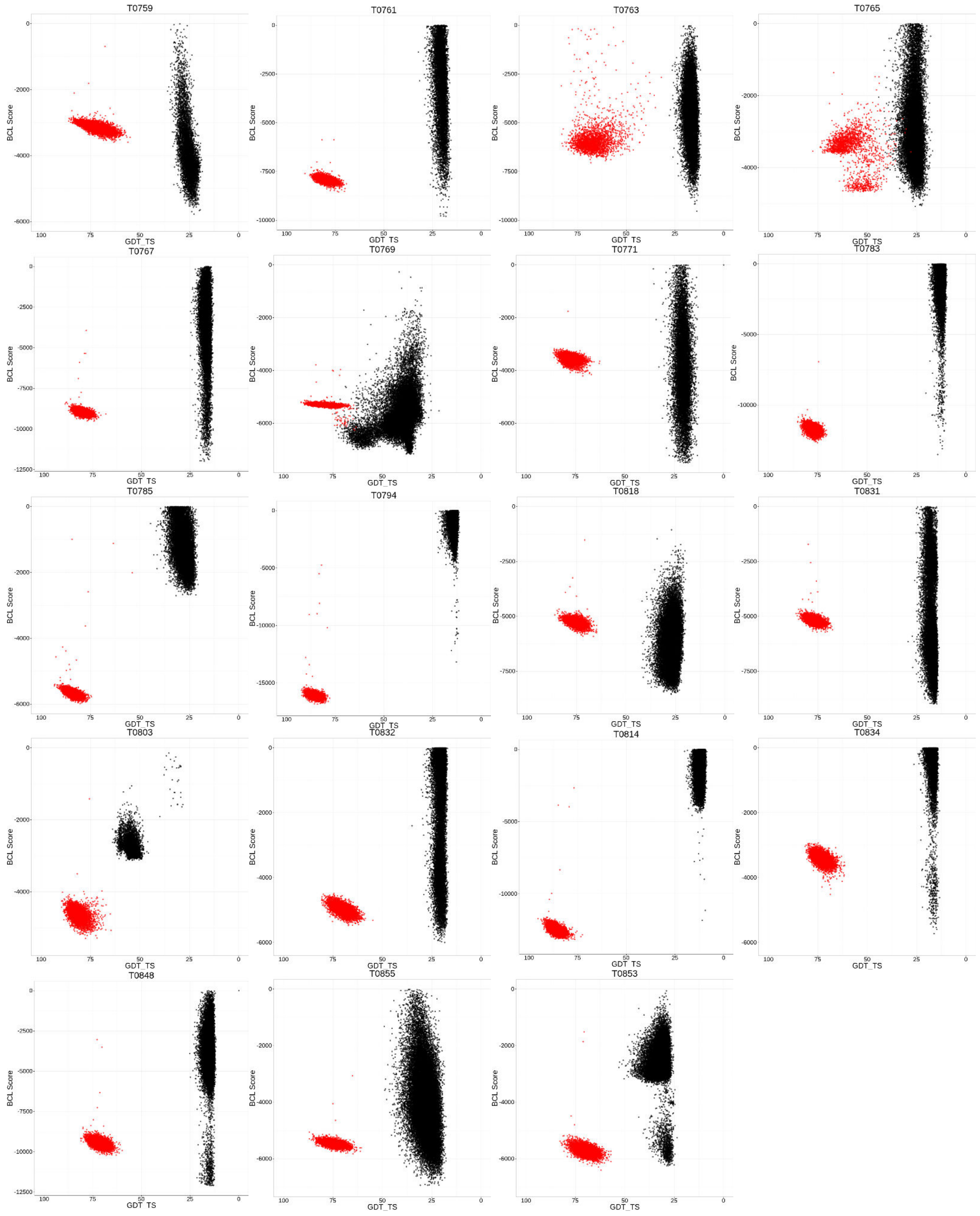

Figure A.2.: Score-accuracy correlations of* de novo *folded models and relaxed experimentally determined structures. *Shown are the BCL-score of the models (y-axis) and the GDT_TS of the models relative to the experimentally determined structure (x-axis).* De novo *folded models are depicted as black dots and models sampled through relaxation of the experimentally determined structure are shows as red dots.*

	Fold		Cluster	Loop	Cluster	MD
Target	GDT_TS (%)	e	GDT_TS (%)	GDT_TS (%)	GDT_TS (%)	GDT_TS (%)
T0759	36	1.0	29	26	21	22
T0761	32	0.8	29	19	15	17
T0763	28	1.3	20	24	19	19
T0765	46	0.4	32	50	26	26
T0767	25	1.1	20	15	13	13
T0769	74	3.3	66	69	66	77
T0771	36	1.5	27	22	18	18
T0781	20	1.3	15	12	9	9
T0783	16	1.6	10	13	10	9
T0785	46	1.0	40	26	22	22
T0794	24	1.1	14	11	10	8
T0803	64	2.7	13	26	21	16
T0814	15	1.2	10	9	7	8
T0818	42	2.0	37	34	24	18
T0831	30	2.0	21	17	14	12
T0832	35	1.1	25	22	19	19
T0834	26	1.3	19	15	12	10
T0848	26	1.6	20	12	11	11
T0853	53	0.8	42	27	23	17
T0855	49	1.1	40	33	28	18
mean	36	1.4	26	24	20	18

Table A.1.: Model accuracy decay over the course of the pipeline. *The quality of the most accurate models decayed over the course of the protein structure prediction pipeline. For each pipeline module, the GDT_TS-value of the most accurate model is shown. For the low-resolution topology search, also the enrichment (e) is shown.*

	T0		TP		TC		TS		TX	
Target	μ_{10} (%)	e	μ_{10} (%)	e	μ_{10} (%)	e	μ_{10} (%)	e	μ_{10} (%)	e
T0761	30	0.8	32	0.8	37	0.8	30	1.0	—	—
T0763	26	1.3	29	1.1	42	1.0	36	0.9	—	—
T0767	24	1.1	27	0.9	29	1.0	25	1.3	26	1.2
T0785	44	1.0	43	1.4	46	0.6	36	1.2	—	—
T0794	22	1.1	24	1.5	21	1.5	22	1.1	—	—
T0814	10	1.2	19	1.1	23	1.3	17	1.0	—	—
T0818	41	2.0	42	1.9	52	2.0	43	1.9	—	—
T0831	29	2.0	37	1.8	35	3.5	—	—	—	—
T0832	31	1.1	29	0.4	46	3.0	29	2.7	—	—
T0834	24	1.3	24	1.5	25	2.1	—	—	—	—
T0848	24	1.6	35	1.2	36	1.3	—	—	—	—
T0853	50	0.8	49	0.7	60	1.9	—	—	—	—
mean	30	1.3	33	1.2	38	1.7	30	1.4	26	1.2

Table A.2.: Protein structure prediction results from limited experimental data. *The average GDT_TS-values of the ten most accurate models (μ_{10}) and the enrichment (e) are shown for prediction from the primary structure alone (T0), from predicted residue-residue contacts (TP), only correct residue-residue contacts (TC), NMR-NOE restraints (TS), and XL-MS restraints (TX).*

A.2. Protein structure prediction protocol for the CASP11 experiment

The following protocol requires an installation of the BCL, Rosetta,[94] and R[a] with the cluster package[100] installed. The BCL license can be obtained for academic and business purposes on the website of the Meiler Laboratory.[b] The Rosetta license can be obtained at `http://www.rosettacommons.org`.

The protocol for protein structure prediction in this study encompassed three modules that were connected through clustering. The following sections detail the protocols for the low-resolution topology search (section A.2.1), clustering (section A.2.2 on the following page), and high-resolution refinement and loop construction (section A.2.3 on page 161). The protocol for clustering was performed twice — first, between the low-resolution topology search and the high-resolution refinement and second, between the high-resolution refinement and the MD refinement.

A.2.1. Low-resolution topology search

The low-resolution topology search was performed using BCL::Fold, which is part of the BCL. The BCL::Fold algorithm assembles SSEs in the three-dimensional space using an MCM algorithm. Consequently, an SSE pool needs to be defined of predicted from which the MCM algorithm can draw the SSE. The SSE prediction were performed using the algorithms PSIPRED,[98] Jufo9D,[97] and MASP.[99] In a first step, the BCL was used to create an SSE pool from the SSE predictions. The directory `<seq_dir>` has to contain the SSE prediction files generated by PSIPRED, Jufo9D, and MASP.

```
bcl.exe CreateSSEPool -prefix <seq_dir> -pool_min_sse_length 5 3 -ssmethods JUFO9D PSIPRED MASP
 ↪  -sse_threshold 0.4 0.4 0.4 -factory SSPredThreshold
```

The SSEs in the SSE pool are subsequently arranged in the three-dimensional space using the BCL::Fold algorithm.[59] The command line below uses the previously generated SSEs pool in conjunction with an MCM algorithm to sample twenty models. Required input files are the SSE pool that was generated with the previous command line, the secondary structure predictions, and the stage file, which configures the MCM algorithm.

```
bcl.exe Fold -fasta <protein>.fasta -sequence_data <seq_dir> <protein> -sspred JUFO9D PSIPRED
 ↪  -pool <protein>.pool -pool_separate -stages_read stages.txt -protein_storage <output_dir>
 ↪  -nmodels 20 -opencl Disable
```

The flag `-opencl Disable` disables the initialization of OpenCL, which is not used by this algorithm and can lead to problems on some systems if left enabled. The stage file configures the MCM algorithm — it sets what number of MC steps to perform, which transformations to apply, and which scoring terms to use. Its format is shown in the following example:

```
STAGE Stage_assembly_1
TYPE MCM
```

[a] `https://www.r-project.org`

[b] `http://www.meilerlab.org/bclcommons`

```
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_01.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_2
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_02.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_3
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_03.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_4
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_04.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_5
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_05.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_refinement_1
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE refinement_01.scoreweights
MUTATE_PROTOCOLS Default Refinement
NUMBER_ITERATIONS 2000 400
STAGE_END
```

A.2.2. Clustering for model selection

The clustering was performed using the R package with cluster.[100] As a preparatory step, the pairwise dissimilarities between the models have to be quantified by computing the pairwise RMSD100 distances between the sampled models have to be computed. For the following command, the file `pdbs.ls` contains the file paths to the protein structure files in the PDB format. The following command will create an upper triangle-matrix that contains the pairwise RMSD100-values. This application is capable of multi-threading, which can be enabled by setting the flag `-scheduler PThread <num_threads>` to the desired value, which should correspond to the number of available CPU cores in the system.

```
bcl PDBCompare -pdb_list pdbs.ls -quality RMSD -norm100
```

The command line above results in a distance matrix with each row and column corresponding to one model. Accordingly, each element of the matrix corresponds to one model-model distance. The flag `-quality` defines, which dissimilarity metric to use. In the command live above, the RMSD100 is used, as indicated by the `-norm100` flag. The models from the low-resolution topology search are subsequently clustered based on their pairwise distances using R in conjunction with the cluster package:

```
# Load the cluster library.
library(cluster)

# Load the dissimilarity matrix that was created created with the BCL.
data_mat <- as.matrix(read.table(”distance_matrix.tbl”, header = T))

# Create a full matrix from the triangle matrix.
data_mat <- data_mat + t(data_mat)

# Convert into a dissimilarity matrix.
data_mat <- as.dist(data_mat)

# Perform clustering with k cluster centers.
clusters <- pam(data_mat, k)

# Display information about the clustering.
clusters$clusinfo

# Display cluster medoids.
clusters$medoids
```

The medoids were subsequently selected for further refinement. The command `cluster$clusinfo` outputs information about the clustering like the average cohesion of the clusters and their minimum separations. Depending on those values, number of cluster centers should be adjusted and the clustering repeated.

A.2.3. Addition of loop and side chain coordinates

The construction of loop regions and side chains as well as the high-resolution refinement were performed using Rosetta. A prerequisite is the creation of the fragment files, which can be obtained from the Robetta server.[c] Subsequently the loop regions are constructed. The command line below will sample ten models:

```
loopmodel -database database/ -in:file:fasta <protein>.fasta -in:file:s <protein>.pdb
 ↪  -loops:loop_file <protein>.loops -out:prefix <output_path> -nstruct 10 -loops:fa_input
 ↪  -loops:frag_sizes 9 3 1 -loops:frag_files 9.bin 3.bin none -loops:build_initial true
 ↪  -loops:remodel quick_ccd -loops:refine refine_ccd -loops:extended true -loops:relax relax
 ↪  -ex1 -ex2 -out:output true -out:pdb true
```

The options file used is given below. The files `9.bin` and `3.bin` are the fragment files for the respective fragment lengths of 9 and 3 residues. This configuration file uses the CCD algorithm for loop construction and loop refinement. The resulting protein model is subsequently subjected to "relaxation" and optimization using the Rosetta application.

```
 -loops:fa_input
 -loops:frag_sizes 9 3 1
 -loops:frag_files 9.bin 3.bin none
 -loops:build_initial true
 -loops:remodel quick_ccd
 -loops:refine refine_ccd
 -loops:extended true
 -loops:relax relax
 -ex1
 -ex2
 -out:output true
 -out:pdb true
```

[c] http://robetta.bakerlab.org

APPENDIX B
MEMBRANE PROTEIN STRUCTURE PREDICTION FROM EPR DATA

This appendix is based on the publication “BCL::MP-Fold: Membrane protein structure prediction guided by EPR restraints”.[2] It provides supplementary data and procedures for chapter III on page 29. The supplementary data in section B.1 contains tables listing the prediction results in terms of model accuracy, as quantified through the RMSD100 metric, and model discrimination, as quantified through the enrichment metric, for combinations of SDSL-EPR distance and accessibility data. Additional data evaluates dependencies of model accuracy and model discrimination on the number of SDSL-EPR distance restraints.The procedures described in section B.2 on page 169 detail the protocol employed for the study.

B.1. Supplementary data

This section contains additional data evaluating the influence of limited experimental data from EPR distance and accessibility measurements on sampling accuracy and model discrimination. For four membrane proteins, the influence of the number of distance restraints on the sampling accuracy was evaluated and is shown in a plot. The tables contain the sampling accuracy- and enrichment-values for all benchmark proteins and benchmark setups.

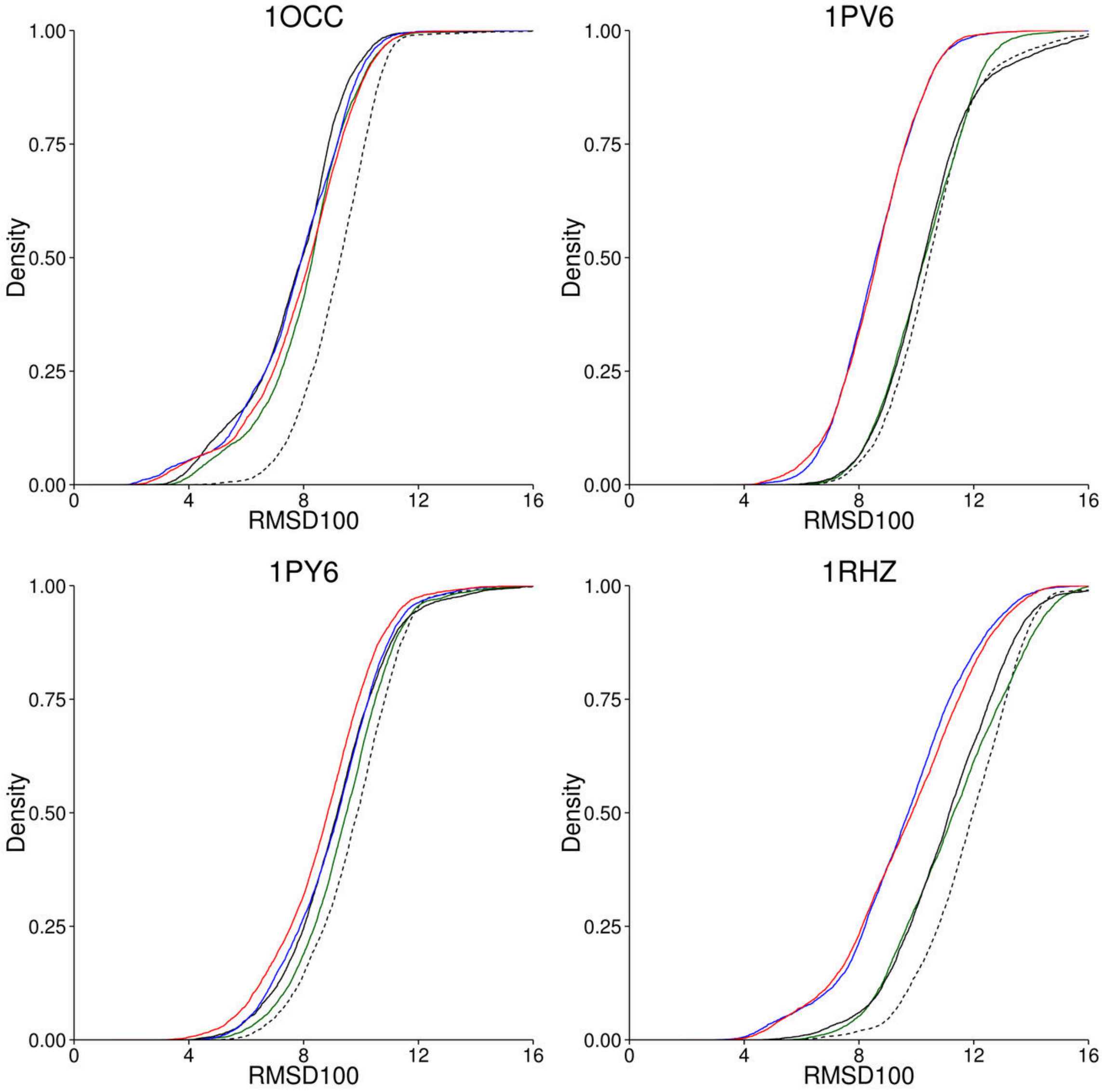

Figure B.1.: Influence of the number of EPR restraints on the prediction accuracy.*The tertiary structure of four proteins was predicted with varying numbers of EPR distance restraints. Without restraints (dashed black), one restraint per ten residues with SSEs (green), one restraint per five residues (solid black), one restraint per three residues (blue), and one restraint per two residues (red). Shown is the cumulative density (y-axis) of models with respect to their RMSD100-values (x-axis).*

	None				Accessibility				Distance				Accessibility & Distance			
Protein	best (Å)	μ_4 (Å)	τ_4 (%)	τ_8 (%)	best (Å)	μ_4 (Å)	τ_4 (%)	τ_8 (%)	best (Å)	μ_4 (Å)	τ_4 (%)	τ_8 (%)	best (Å)	μ_4 (Å)	τ_4 (%)	τ_8 (%)
1IWG	4.2	4.9	0.0	19.1	3.9	4.3	0.1	26.3	3.8	4.4	0.0	22.1	3.6	4.2	0.1	25.8
1J4N	5.2	5.3	0.0	18.0	4.2	4.7	0.0	22.5	4.4	4.8	0.0	21.0	3.8	4.3	0.0	25.4
1KPL	10.1	10.3	0.0	0.0	8.8	9.6	0.0	0.0	8.5	9.4	0.0	0.0	8.5	9.3	0.0	0.0
1OCC	3.7	4.5	0.0	20.1	2.2	2.9	0.8	30.2	2.4	2.8	3.1	44.9	2.0	2.3	7.0	50.8
1OKC	6.4	6.8	0.0	1.2	8.2	8.8	0.0	0.0	7.1	7.8	0.0	0.1	6.9	7.7	0.0	0.2
1PV6	5.2	6.1	0.0	5.3	5.3	5.8	0.0	6.0	5.2	5.7	0.0	8.5	4.9	5.5	0.0	10.5
1PY6	4.4	5.1	0.0	15.2	3.9	4.3	0.0	17.5	3.6	4.2	0.1	24.2	3.3	3.8	0.2	26.7
1RHZ	5.6	5.9	0.0	2.1	5.5	5.7	0.0	1.9	4.2	4.7	0.0	5.6	3.9	4.4	0.0	7.0
1U19	5.3	5.7	0.0	4.7	5.4	6.1	0.0	3.0	5.5	6.2	0.0	3.6	5.2	5.9	0.0	4.1
1XME	8.0	8.7	0.0	0.0	8.1	8.7	0.0	0.0	7.3	7.8	0.0	0.2	7.2	7.8	0.0	0.2
2BG9	2.3	2.6	10.8	52.7	2.2	2.3	28.1	51.3	2.2	2.3	28.1	57.7	2.1	2.2	32.6	61.3
2BL2	7.5	8.4	0.0	0.0	7.8	8.4	0.0	0.0	5.0	6.1	0.0	0.9	5.1	6.2	0.0	0.8
2BS2	6.1	6.3	0.0	2.4	6.3	6.6	0.0	1.8	5.3	5.9	0.0	3.5	5.2	5.8	0.0	3.5
2IC8	5.5	6.1	0.0	4.7	6.0	6.2	0.0	4.2	5.0	5.5	0.0	9.8	5.1	5.5	0.0	9.4
2K73	3.3	3.6	0.5	24.8	3.3	3.5	0.7	21.4	2.7	3.0	3.4	30.1	2.8	3.1	3.3	30.0
2KSF	4.1	4.4	0.0	22.3	2.6	3.0	2.6	21.9	2.9	3.1	5.2	25.9	2.7	2.9	8.3	25.1
2KSY	4.9	5.3	0.0	11.2	3.7	4.5	0.0	13.4	3.5	4.1	0.1	21.4	3.6	4.1	0.1	20.4
2NR9	5.6	6.0	0.0	5.8	6.0	6.6	0.0	3.4	6.4	6.9	0.0	2.5	6.4	6.9	0.0	2.5
2XUT	7.7	8.5	0.0	0.0	7.5	8.4	0.0	0.0	7.7	8.2	0.0	0.1	7.7	8.2	0.0	0.1
3GIA	10.0	10.2	0.0	0.0	9.5	9.8	0.0	0.0	9.1	9.6	0.0	0.0	9.1	9.6	0.0	0.0
3KCU	7.0	7.5	0.0	0.3	7.7	8.1	0.0	0.0	6.3	7.2	0.0	0.6	6.5	7.3	0.0	0.4
3KJ6	6.6	7.0	0.0	0.7	4.9	6.4	0.0	1.0	5.2	5.9	0.0	3.0	4.9	5.8	0.0	3.0
3P5N	4.6	5.8	0.0	5.5	5.3	5.8	0.0	4.5	4.8	5.6	0.0	10.2	4.8	5.6	0.0	10.0
2BHW	7.4	7.8	0.0	0.2	6.9	7.4	0.0	0.3	3.0	3.4	0.9	31.7	2.9	3.4	0.9	31.2
2H8A	3.3	3.9	0.1	32.5	3.1	3.8	0.2	34.5	1.9	2.1	6.4	41.7	1.8	2.0	7.6	44.3
2HAC	1.3	1.4	75.3	89.6	1.3	1.4	83.7	94.8	1.3	1.5	71.8	92.1	1.3	1.5	87.0	96.8
2L35	2.6	2.9	4.5	19.8	1.7	1.8	17.1	22.4	1.9	2.1	31.0	48.3	1.9	2.1	31.0	48.3
2ZY9	4.7	5.7	0.0	3.4	4.5	5.5	0.0	4.4	2.7	3.3	0.4	21.4	2.8	3.4	0.3	20.9
3CAP	7.4	8.0	0.0	0.1	7.3	7.9	0.0	0.1	4.9	5.4	0.0	5.8	4.7	5.3	0.0	4.1
mean	5.5	6.0	3.0	12.5	5.3	5.8	4.6	13.3	4.6	5.1	5.2	18.0	4.5	5.0	6.2	19.4

Table B.1.: Sampling accuracy comparison for folding with and without EPR restraints. *Results for folding the proteins based on predicted SSEs without restraints, with accessibility restraints only, with distance restraints only, and with accessibility and distance restraints. Shown are the RMSD100 of the most accurate model sampled (best), the average RMSD100 of the ten most accurate models (μ_{10}) as well as the percentage of the sampled models with RMSD100-values of less than 4 Å and 8 Å (τ_4 and τ_8). Using protein specific data derived from EPR experiments significantly improves the sampling accuracy. The improvement is mainly caused by using EPR distance restraints; accessibility restraints have a minor effect on the sampling accuracy. The proteins above the separating line are monomeric proteins; below the separating line are multimeric proteins.*

	None				Accessibility				Distance				Accessibility & Distance			
Protein	best (%)	ϕ_{10} (%)	γ_{20} (%)	γ_{40} (%)	best (%)	ϕ_{10} (%)	γ_{20} (%)	γ_{40} (%)	best (%)	ϕ_{10} (%)	γ_{20} (%)	γ_{40} (%)	best (%)	ϕ_{10} (%)	γ_{20} (%)	γ_{40} (%)
1IWG	23.7	21.3	0.2	0.0	33.5	29.4	2.0	0.0	31.3	26.9	1.2	0.0	33.8	28.3	2.9	0.0
1J4N	31.6	28.7	2.7	0.0	45.9	39.7	13.5	0.1	40.5	36.8	11.2	0.0	44.8	39.8	22.5	0.3
1KPL	22.8	14.4	0.0	0.0	18.7	15.5	0.0	0.0	19.0	15.6	0.0	0.0	18.4	16.1	0.0	0.0
1OCC	44.0	34.3	1.7	0.0	77.1	57.8	20.7	1.1	63.0	55.9	15.4	1.5	66.1	59.4	25.6	4.6
1OKC	15.4	11.0	0.0	0.0	22.0	16.4	0.0	0.0	19.5	16.4	0.0	0.0	19.8	17.6	0.0	0.0
1PV6	19.6	13.9	0.0	0.0	23.6	20.6	0.1	0.0	22.1	19.0	0.1	0.0	22.7	19.5	0.1	0.0
1PY6	21.0	19.1	0.0	0.0	35.1	30.4	2.2	0.0	32.8	27.4	1.4	0.0	36.0	31.1	3.5	0.0
1RHZ	33.0	29.2	1.6	0.0	38.1	35.1	6.2	0.0	41.9	36.0	7.0	0.1	41.8	37.2	11.3	0.2
1U19	25.4	17.3	0.0	0.0	25.8	22.7	0.2	0.0	21.9	18.4	0.0	0.0	23.8	19.0	0.1	0.0
1XME	10.0	8.5	0.0	0.0	10.6	9.2	0.0	0.0	9.6	8.3	0.0	0.0	10.1	8.8	0.0	0.0
2BG9	79.5	73.9	37.5	5.4	95.5	90.9	93.0	52.2	84.1	78.2	51.2	16.5	92.5	87.8	78.3	40.6
2BL2	18.3	14.8	0.0	0.0	25.7	20.9	0.1	0.0	31.7	23.9	0.2	0.0	28.7	23.0	0.3	0.0
2BS2	23.9	21.7	0.3	0.0	32.5	28.9	1.0	0.0	27.6	24.8	0.4	0.0	34.2	28.8	1.2	0.0
2IC8	27.4	21.8	0.2	0.0	29.9	26.2	1.0	0.0	26.2	23.4	0.3	0.0	34.5	29.0	2.1	0.0
2K73	44.2	39.8	6.5	0.1	65.1	60.3	37.2	3.7	52.3	47.8	14.0	0.5	65.8	60.7	38.5	6.7
2KSF	47.2	37.0	4.1	0.0	73.6	66.6	28.6	3.4	58.5	53.0	16.6	0.8	71.5	65.2	35.9	6.6
2KSY	29.6	22.2	0.2	0.0	37.6	31.5	3.3	0.0	38.8	28.9	1.4	0.0	37.7	31.5	2.9	0.0
2NR9	24.1	20.0	0.1	0.0	27.4	22.7	0.4	0.0	25.1	20.9	0.1	0.0	24.4	23.4	0.3	0.0
2XUT	9.8	9.0	0.0	0.0	12.3	10.0	0.0	0.0	10.0	9.7	0.0	0.0	10.1	8.9	0.0	0.0
3GIA	7.6	6.5	0.0	0.0	8.3	7.0	0.0	0.0	7.6	6.8	0.0	0.0	7.8	6.5	0.0	0.0
3KCU	19.1	15.5	0.0	0.0	21.5	18.4	0.0	0.0	21.2	17.6	0.0	0.0	20.3	19.0	0.0	0.0
3KJ6	18.4	15.9	0.0	0.0	29.5	22.1	0.1	0.0	21.7	19.6	0.1	0.0	25.4	21.8	0.2	0.0
3P5N	29.8	21.2	0.1	0.0	33.0	29.2	1.2	0.0	30.9	25.3	0.2	0.0	30.9	30.0	0.9	0.0
2BHW	42.2	36.8	2.3	0.0	46.8	40.0	3.8	0.2	49.5	43.1	4.2	0.2	48.9	42.3	4.4	0.2
2H8A	28.8	19.5	0.0	0.0	30.9	28.2	1.1	0.0	44.8	40.9	7.2	0.2	49.8	45.7	10.7	0.7
2HAC	100.0	98.9	87.0	80.4	100.0	98.0	90.9	83.1	99.4	98.0	81.7	73.1	99.4	97.7	82.6	75.2
2L35	53.2	49.2	11.1	0.5	76.6	73.8	53.2	15.2	49.4	47.3	26.6	0.9	76.0	70.6	79.0	18.9
2ZY9	16.9	14.2	0.0	0.0	27.1	23.7	0.3	0.0	29.6	24.7	0.3	0.0	36.8	30.9	1.4	0.0
3CAP	12.7	11.9	0.0	0.0	17.9	14.0	0.0	0.0	14.4	12.3	0.0	0.0	17.5	14.2	2.8	0.0
mean	30.3	25.8	5.4	3.0	38.7	34.1	12.4	5.5	35.3	31.3	8.3	3.2	38.9	35	14.1	5.3

Table B.2.: Contact recovery comparison for folding with and without EPR restraints. *The usage of EPR accessibility restraints significantly increases the percentage of recovered contacts (amino acids that are separated by at least six residues and have a maximum Euclidean distance of* 8 Å *in the experimentally determined structure). Shown are the highest contact recovery achieved (best), the average contact recovery of the ten models with the highest contact recovery (ϕ_{10}) and the percentage of models for which more than* 20 % *and* 40 % *of the contacts were recovered (γ_{20} and γ_{40}). The proteins above the separating line are monomeric proteins; below the separating line are multimeric proteins.*

	Predicted Pool						Native Pool					
Protein	E_{None}	E_{Acc}	E_{Dist}	σ_{Dist}	E_{EPR}	σ_{EPR}	E_{None}	E_{Acc}	E_{Dist}	σ_{Dist}	E_{EPR}	σ_{EPR}
1IWG	1.1	1.5	2.2	0.2	2.3	0.2	0.4	0.9	1.2	0.4	2.2	0.3
1J4N	0.4	0.3	1.7	0.2	1.8	0.2	0.5	0.6	0.9	0.2	1.7	0.3
1KPL	2.2	2.4	2.4	0.1	2.4	0.3	0.3	0.3	2.4	0.4	2.4	0.2
1OCC	1.5	1.5	4.0	1.0	5.1	1.0	1.8	1.8	3.9	1.0	5.0	1.0
1OKC	1.3	1.4	1.7	0.2	1.7	0.2	1.3	1.4	1.5	0.1	1.7	0.2
1PV6	1.2	1.3	1.9	0.2	1.9	0.3	0.5	1.0	1.7	0.4	2.0	0.3
1PY6	2.1	2.0	3.2	0.3	3.3	0.3	2.3	2.3	3.2	0.3	3.3	0.3
1RHZ	1.1	1.3	1.8	0.4	2.1	0.6	1.3	1.4	1.6	0.4	1.4	0.7
1U19	1.6	2.2	2.6	0.2	2.6	0.2	2.0	1.7	2.0	0.3	2.9	0.7
1XME	1.4	1.3	1.7	0.3	1.7	0.3	1.1	1.0	1.7	0.3	3.6	0.5
2BG9	0.9	1.8	2.5	0.5	2.5	0.4	1.6	2.7	1.6	0.6	1.9	0.8
2BL2	0.5	0.5	1.4	0.3	1.4	0.3	1.1	0.7	3.1	0.5	4.9	1.0
2BS2	2.0	2.0	2.6	0.2	2.6	0.1	1.9	2.3	2.4	0.3	2.8	0.3
2IC8	1.0	1.2	1.6	0.3	1.7	0.2	1.2	1.2	1.7	0.3	2.8	0.7
2K73	2.1	1.5	1.6	0.6	2.1	0.2	2.2	2.4	2.8	0.6	7.8	0.7
2KSF	2.6	2.5	2.4	0.8	2.6	0.7	3.8	3.2	2.1	0.7	2.3	0.7
2KSY	1.7	2.2	2.6	0.7	3.0	0.6	2.3	2.0	3.3	0.6	3.8	1.1
2NR9	0.8	0.9	1.0	0.2	1.1	0.2	1.6	1.2	1.5	0.3	1.9	0.2
2XUT	1.2	1.3	1.6	0.2	1.6	0.2	0.5	0.5	1.3	0.2	2.1	0.3
3GIA	0.7	0.7	1.2	0.2	1.2	0.2	0.5	0.6	0.8	0.2	0.5	0.1
3KCU	0.8	1.5	1.6	0.2	1.7	0.2	1.3	1.1	1.6	0.2	1.8	0.5
3KJ6	1.6	2.0	2.3	0.2	2.2	0.2	1.4	1.7	2.0	0.3	4.2	0.8
3P5N	0.8	1.3	1.5	0.3	1.6	0.3	1.3	1.1	1.7	0.3	2.3	0.8
2BHW	0.8	0.9	2.1	0.6	2.3	0.5	1.2	1.7	3.9	1.2	4.4	1.7
2H8A	2.1	2.4	6.3	0.4	6.2	0.7	1.6	2.1	5.8	0.9	7.2	0.9
2HAC	0.8	2.8	3.4	0.6	4.0	0.9	0.5	2.4	1.2	0.2	2.5	0.7
2L35	0.8	1.7	1.9	0.7	1.9	0.7	0.8	1.0	2.3	0.9	2.6	1.2
2ZY9	0.7	2.3	2.4	0.2	2.9	0.3	1.0	1.4	2.0	0.4	3.1	0.4
3CAP	1.3	1.8	3.3	0.4	3.7	0.4	1.7	2.2	3.1	0.8	3.8	0.2
mean	1.3	1.6	2.3	0.3	2.5	0.4	1.4	1.5	2.2	0.5	3.1	0.6

Table B.3.: Enrichments achieved for folding with and without EPR restraints. *EPR restraints significantly improve our ability to select the most accurate models among the sampled ones. When using EPR distance and accessibility restraints, the enrichment (E_{EPR}) could be improved in each case compared to structure prediction without EPR data (E_{None}). To be independent from specific spin labeling patterns ten different EPR distance restraint sets were used and the standard deviation regarding enrichment computed (σ_{EPR}). The experiment was also conducted using accessibility (E_{Acc}) and distance restraints (E_{Dist} and σ_{Dist}) only. In addition to using predicted SSEs (predicted pool), the experiment was repeated using SSEs obtained from the experimentally determined structure (native pool). The proteins above the separating line are monomeric proteins; below the separating line are multimeric proteins.*

	None				Accessibility				Distance				Accessibility & Distance			
Protein	best (%)	ϕ_{10} (%)	γ_{20} (%)	γ_{40} (%)	best (%)	ϕ_{10} (%)	γ_{20} (%)	γ_{40} (%)	best (%)	ϕ_{10} (%)	γ_{20} (%)	γ_{40} (%)	best (%)	ϕ_{10} (%)	γ_{20} (%)	γ_{40} (%)
1IWG	34.5	30.5	3.3	0.0	44.8	42.2	19.2	0.3	43.5	37.0	7.4	0.0	48.9	44.5	22.0	1.0
1J4N	38.8	35.3	7.8	0.0	51.0	42.9	26.5	0.2	47.8	42.8	17.5	0.2	51.6	46.0	29.0	0.5
1KPL	18.4	15.1	0.0	0.0	21.7	17.8	0.0	0.0	23.0	18.9	0.0	0.0	20.7	17.7	0.0	0.0
1OCC	44.0	38.5	4.7	0.1	65.7	61.9	32.2	2.9	65.5	59.6	19.9	2.4	74.3	68.5	32.5	7.5
1OKC	15.1	13.2	0.0	0.0	18.9	15.8	0.0	0.0	20.9	17.4	0.0	0.0	20.2	17.9	0.0	0.0
1PV6	18.7	17.5	0.0	0.0	27.1	23.8	0.3	0.0	28.2	22.6	0.2	0.0	28.6	23.1	0.2	0.0
1PY6	23.2	22.0	0.2	0.0	49.1	36.1	4.0	0.0	36.0	30.4	2.8	0.0	38.2	33.6	4.2	0.0
1RHZ	32.0	26.2	0.6	0.0	40.2	34.6	4.1	0.0	42.1	35.2	5.2	0.0	47.2	41.2	13.0	0.1
1U19	25.0	20.5	0.1	0.0	29.9	27.7	2.3	0.0	30.1	25.6	0.9	0.0	40.2	34.0	3.9	0.0
1XME	10.3	9.2	0.0	0.0	13.1	11.3	0.0	0.0	11.1	9.4	0.0	0.0	11.1	9.2	0.0	0.0
2BG9	75.0	69.8	48.5	5.2	95.5	92.0	96.7	77.6	88.6	84.0	70.3	21.5	95.5	93.0	96.3	72.2
2BL2	45.9	43.0	20.0	0.3	46.3	45.6	45.8	2.2	46.5	44.5	31.7	3.3	50.7	49.1	59.9	14.2
2BS2	32.5	23.1	0.3	0.0	42.9	31.4	1.3	0.0	34.1	27.2	0.9	0.0	41.6	35.2	3.1	0.0
2IC8	28.0	26.3	1.3	0.0	37.2	30.7	3.0	0.0	36.8	31.2	4.5	0.0	43.4	38.8	11.8	0.1
2K73	53.5	45.6	7.7	0.2	69.8	65.5	45.6	5.8	63.0	55.7	25.3	2.4	74.2	70.2	54.8	13.3
2KSF	45.3	40.8	6.4	0.1	84.9	76.6	43.5	7.5	67.0	60.5	29.7	2.8	85.7	81.3	61.8	16.1
2KSY	29.2	24.5	0.5	0.0	38.8	35.3	6.6	0.0	37.8	33.2	3.6	0.0	49.2	43.7	10.5	0.3
2NR9	62.3	55.2	10.7	0.9	77.9	73.6	56.0	15.9	61.7	56.8	43.1	4.3	68.6	65.4	90.1	33.3
2XUT	75.0	69.8	48.5	5.2	21.7	19.8	0.1	0.0	21.9	20.0	0.1	0.0	24.4	21.6	0.3	0.0
3GIA	14.4	10.5	0.0	0.0	12.1	10.9	0.0	0.0	12.0	10.3	0.0	0.0	11.2	10.2	0.0	0.0
3KCU	8.9	7.6	0.0	0.0	10.0	8.5	0.0	0.0	9.2	7.6	0.0	0.0	8.8	7.3	0.0	0.0
3KJ6	18.1	16.0	0.0	0.0	21.5	18.9	0.0	0.0	22.0	18.6	0.0	0.0	29.9	23.2	0.2	0.0
3P5N	25.3	21.4	0.1	0.0	30.9	27.2	1.1	0.0	34.0	27.9	1.2	0.0	37.1	33.6	3.3	0.0
2BHW	25.7	23.0	0.4	0.0	33.5	27.5	1.5	0.0	30.7	27.1	1.6	0.0	36.0	31.9	3.7	0.0
2H8A	100.0	99.7	85.2	77.1	100.0	99.8	89.0	81.5	100.0	99.6	86.1	79.0	100.0	99.8	91.7	83.5
2HAC	49.2	45.6	8.7	0.3	56.5	45.9	7.5	0.3	51.6	45.1	7.2	0.2	66.3	57.3	8.1	1.4
2L35	30.1	24.7	0.6	0.0	42.3	37.7	6.3	0.1	35.9	27.8	2.4	0.0	56.2	52.7	20.9	4.3
2ZY9	19.8	17.1	0.0	0.0	27.7	23.1	0.3	0.0	23.5	19.0	0.0	0.0	29.6	25.8	0.8	0.0
3CAP	20.9	17.7	0.0	0.0	26.7	21.7	0.2	0.0	22.2	18.8	0.0	0.0	26.9	22.0	0.1	0.0
mean	35.1	31.3	8.8	3.1	42.7	38.1	17.0	6.7	39.5	35.0	12.5	4.0	45.4	41.3	21.5	8.5

Table B.4.: Structure prediction results when using SSEs derived from the native structure. *Results for folding the proteins based on SSEs obtained from their native structure without restraints, with accessibility restraints, with distance restraints and with accessibility and distance restraints. Shown are the contact recovery of the most accurate model sampled (best), the average of the ten most accurate models (ϕ_{10}) as well as the percentage of the sampled models with contact recovery-values greater than 20 % and 40 % (γ_{20} and γ_{40}). The proteins above the separating line are monomeric proteins; below the separating line are multimeric proteins.*

B.2. Protein structure prediction protocol for membrane proteins from EPR data

This protocol requires an installation of the BCL, which is available on the website of the Meiler Laboratory.[a] The tertiary structure of the membrane proteins discussed in this work was predicted from their primary structure and experimentally determined and simulated SDSL-EPR distance restraints. This protocol describes the steps necessary to simulate SDSL-EPR distance restraints based on the CONE model and usage of SDSL-EPR distance restraints to predict the tertiary structure of membrane proteins using the *de novo* protein structure prediction algorithm BCL::Fold, which is part of the BCL.

B.2.1. Simulation of SDSL-EPR distance restraints

The protocol for the simulation of SDSL-EPR distance restraints consists of two steps: 1. selection of the spin labeling sites and 2. simulation of the experimental uncertainty. Both steps are described in this section.

The following command line selects spin labeling sites for SDSL-EPR distance measurements. The algorithm optimizes the distributions of the measurements using an MCM algorithm.[155]

```
bcl.exe OptimizeDataSetPairwise -fasta 1IWGA.fasta -pool_min_sse_lengths 0 0 -pool 1IWG.pool
 ↪  -distance_min_max 15 50 -nc_limit 10 -ensembles 1IWG_ensembles.ls -mc_number_iterations
 ↪   100000 100000 -prefix 1IWG_ -nmodels 500 -read_scores_optimization opt_score_weights.wts
 ↪  -read_mutates_optimization mutate_weights.wts -data_set_size_fraction_of_sse_resis 0.2
 ↪  -random_seed
```

The following command line adds a CONE model-based uncertainty related to the translation from backbone distance into spin-spin distance[78] to simulate the uncertainty accompanying SDSL-EPR distance measurements.

```
bcl.exe SimulateDistanceRestraints -pdb 1IWGA.pdb -simulate_distance_restraints -output_file
 ↪   1IWG.epr_cst_bcl -min_sse_size 0 0 0 -add_distance_uncertainty sl_cb.histograms
 ↪  -restraint_list 1IWG.epr 0 1 5 6 -random_seed
```

B.2.2. Prediction of the tertiary structure of membrane proteins from SDSL-EPR data

Tertiary structure prediction with BCL::Fold is a two-stage process. The SSEs of the protein are predicted using machine learning methods and the predicted SSEs are subsequently arranged in the three-dimensional space using an MCM algorithm.

The following command line will predict transmembrane SSEs of the protein 1IWG using the prediction method OCTOPUS.[164] The "input" folder must contain the fasta and OCTOPUS prediction files for 1IWG.

```
bcl.exe CreateSSEPool -ssmethods OCTOPUS -pool_min_sse_lengths 5 3 -sse_threshold 0.5 0.5 0.5
 ↪  -prefix 1IWG -join_separate -factory SSPredThreshold
```

[a] http://www.meilerlab.org/bclcommons

The predicted SSEs are subsequently arranged in the three-dimensional space using an MCM algorithm. The following command line will sample 40 models for the protein 1IWG from predicted SSEs using BCL::Fold. The SSE pool created with the previous command is used as input data.

```
bcl.exe protein:Fold -native 1IWGA.pdb -function_cache -pool_separate -min_sse_size 5 3 -quality
 ↪   RMSD GDT_TS -superimpose RMSD -sspred OCTOPUS JUFO9D -pool
 ↪   1IWGA.SSPredHighest_JUFO9D_OCTOPUS.pool -stages_read stages.txt -pool_prefix 1IWGA -nmodels
 ↪   40 -prefix 1IWG_dist_acc_pred_ -membrane -protein_storage pdbs/ -tm_helices
 ↪   1IWGA.SSPredHighest_OCTOPUS.pool -sequence_data sspred/ 1IWG -opencl Disable
 ↪  -restraint_types DistanceEPR AccessibilityEPR -restraint_prefix restraints/1 -random_seed
```

APPENDIX C
PROTEIN STRUCTURE PREDICTION FROM CROSS-LINKING DATA

This appendix is based on the publication "Protein structure prediction guided by crosslinking restraints – A systematic evaluation of the impact of the crosslinking spacer length".[3] It provides supplementary data and procedures for chapter IV on page 48. The supplementary data in section C.1 contains tables listing the prediction results in terms of model accuracy and model discrimination for different XL spacer lengths and reactivities. The procedures described in section C.2 on page 176 detail the computational procedures employed in this study.

C.1. Supplementary data

This section contains additional data detailing the determination of the optimal cross-linker spacer length, translation of experimental data from XL-MS experiments into structural restraints, and evaluation of the cross-linker spacer length on the sampling accuracy and model discrimination. The cross-link yield per spacer length is listed for each of the benchmark proteins and the sampling accuracy, as quantified through the RMSD100 metric, and the model discrimination, as quantified through the enrichment metric, are listed in dependence on the cross-linker spacer length. Additional figures depict the translation from the experimental data into structural restraints and show the distance distributions for Lys-Lys pairs.

	optimal		short1		short2		long1		long2	
Protein	l (Å)	#rest	l (Å)	#rest	l (Å)	#rest	l (Å)	#rest	l (Å)	#rest
1HRC	10.2	13	2.5	0	7.5	7	17.5	27	30	107
3IV4	10.4	5	2.5	2	7.5	2	17.5	7	30	13
1BGF	10.7	6	2.5	3	7.5	4	17.5	10	30	13
1T3Y	10.9	35	2.5	9	7.5	20	17.5	42	30	63
3M1X	10.9	1	2.5	0	7.5	0	17.5	5	30	19
1X91	11.0	2	2.5	0	7.5	1	17.5	8	30	27
1JL1	11.2	7	2.5	0	7.5	3	17.5	11	30	24
1MBO	11.3	9	2.5	0	7.5	3	17.5	23	30	77
2QNL	11.5	6	2.5	4	7.5	4	17.5	8	30	15
2AP3	12.1	53	2.5	0	7.5	19	17.5	136	30	427
1J77	12.2	29	2.5	7	7.5	16	17.5	36	30	70
1ES9	12.5	8	7.5	0	17.5	1	37.5	17	45	20
3B5O	12.7	15	7.5	2	17.5	8	37.5	21	45	25
1XQ0	13.3	9	7.5	0	17.5	4	37.5	14	45	44
2IXM	13.5	41	7.5	20	17.5	41	37.5	49	45	57

Table C.1.: Lys-Lys cross-links yielded by different spacer lengths. *Cross-links obtained for the benchmark proteins. Simulated and experimentally determined cross-links were obtained for the fifteen benchmark proteins. For each protein, an optimal spacer length l was determined (optimal). Additional cross-links were simulated for two shorter (short1 and short2) and two longer (long1 and long2) spacer lengths. The number of yielded cross-links (#rest) is shown for each spacer length. For the two proteins 1HRC and 1MBO, experimentally determined cross-links were published.*

	Without restraints				Optimal Lys/Lys		All Lys/Lys		All reactivities	
Protein	best (Å)	σ_{best} (Å)	e	σ_{e}	best (Å)	e	best (Å)	e	best (Å)	e
1HRC	4.5	0.3	0.8	0.1	3.8	2.0	3.8	2.0	3.7	5.9
3IV4	6.7	0.2	1.2	0.3	5.7	2.5	5.3	2.5	5.2	1.9
1BGF	6.6	0.4	1.0	0.2	5.7	2.1	4.9	2.4	6.2	1.6
1T3Y	7.0	0.7	1.7	0.4	6.4	2.9	5.7	3.0	6.2	2.3
3M1X	3.8	0.1	0.7	0.2	3.8	0.7	3.6	1.5	3.6	1.7
1X91	4.8	0.2	2.0	0.5	4.8	2.0	2.0	3.2	2.1	3.5
1JL1	6.4	0.4	1.2	0.1	5.6	2.1	5.3	2.8	5.1	2.7
1MBO	7.1	0.6	0.8	0.3	6.4	2.0	6.5	1.6	4.2	2.5
2QNL	7.0	0.6	1.0	0.3	4.8	1.9	4.1	2.1	6.1	2.1
2AP3	2.5	0.1	1.6	0.5	2.0	3.0	1.6	3.1	2.2	2.0
1J77	6.8	0.3	0.5	0.2	5.0	2.0	4.0	2.4	3.8	3.2
1ES9	7.3	0.8	1.1	0.6	5.7	2.1	5.6	2.8	6.3	2.9
3B5O	9.2	0.9	1.4	0.2	8.6	1.9	9.0	2.6	7.1	1.9
1XQ0	9.9	1.0	1.1	0.3	8.3	1.9	8.5	2.4	7.4	2.1
2IXM	9.4	0.9	1.1	0.4	7.9	1.7	8.5	1.7	7.0	1.9
mean	6.6	0.5	1.1	0.3	5.6	2.1	5.2	2.4	5.1	2.6

Table C.2.: Protein structure prediction results for different cross-linker reactivities. *Comparison between structure prediction results with and without cross-linking restraints. By using geometrical restraints obtained from cross-linking experiments, the size of the sampling space can be reduced resulting in an improved sampling accuracy. This is shown by significant improvements in the RMSD100-value of the most accurate model (best). Furthermore, cross-linking restraints provide geometrical information, which improves the discrimination power of the scoring function, leading to an improvement in the enrichment (e). Without restraints, ten independent prediction trajectories were conducted and the standard deviation of the accuracy of the best model (σ_{best}) and the enrichment (σ_{e}) are reported.*

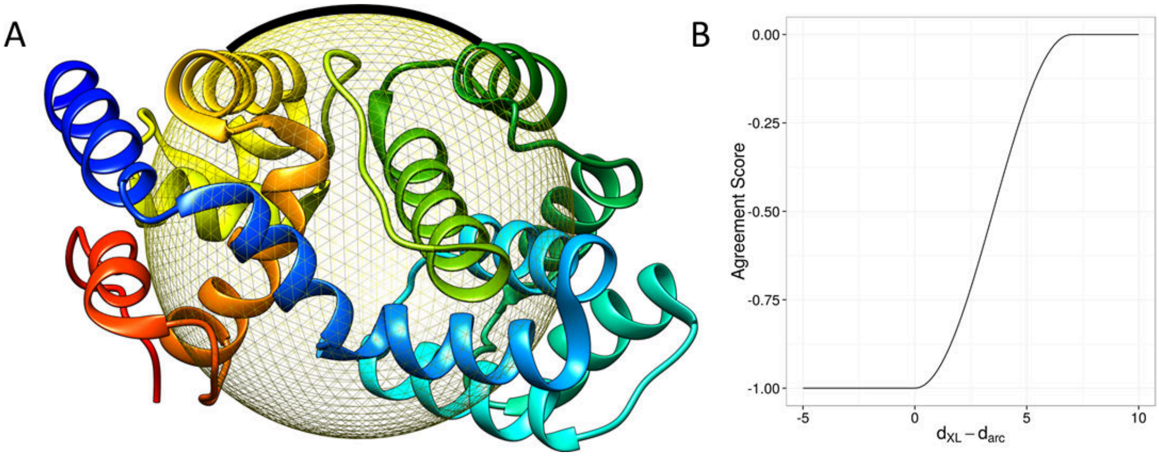

Figure C.1.: Implicit translation from cross-linking data into structural restraints. *Explicit simulation of the cross-linker conformation is computationally expensive and prohibitive for use in a rapid scoring function required for protein structure prediction. Instead, the cross-linker conformation and the path crossed by the cross-linker were approximated through computing the arc length connecting the two cross-linked residues (A). The agreement of a model with cross-linking data was evaluated by computing the difference between the arc length (*$\mathrm{d}_{\mathrm{arc}}$*) and the cross-linker length (*d_{xl}*). The agreement of the model with the cross-linking data is quantified with a score between* -1 *and* 0*, with* -1 *being the best agreement and* 0 *being the worst agreement (B).*

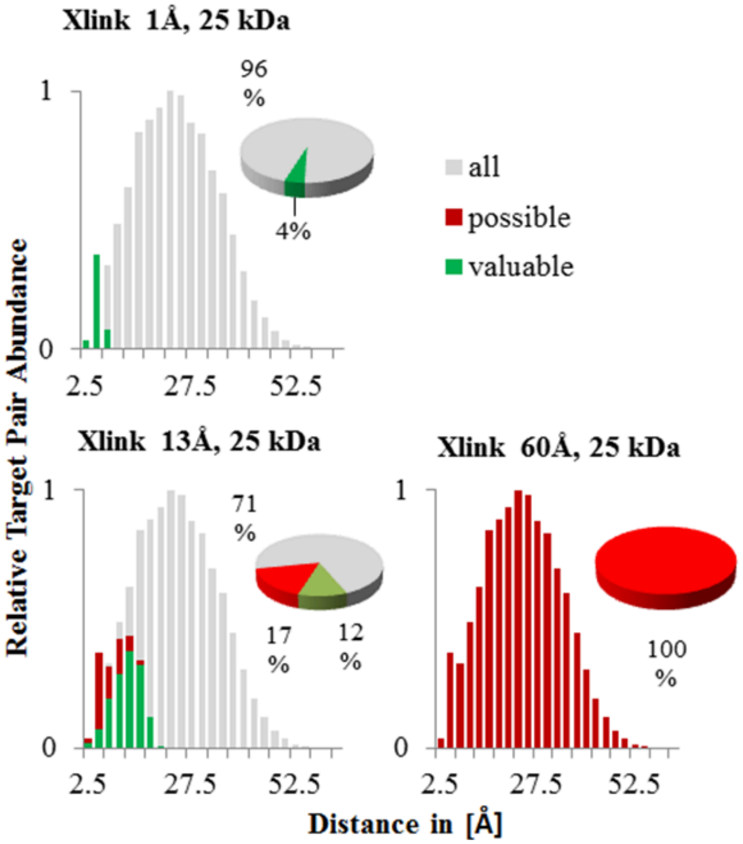

Figure C.2.: Lys-Lys pair distributions. *Distribution of all possible and valuable Lys-Lys pairs for a* 25 kDa *to* 27.5 kDa *weight bin. Gray bars show all theoretical pairs in their specific distance cluster of* ± 2.5 Å. *Red bars show pairs that could be connected with respect to their surface distance by a specific cross-link (here* 1 Å, 13 Å *and* 60 Å*) always including the side chain contribution to the overall length. Green bars show pairs that are considered valuable by our proposed scoring function. The pie charts show the accumulated number of cross-links for every spacer length.*

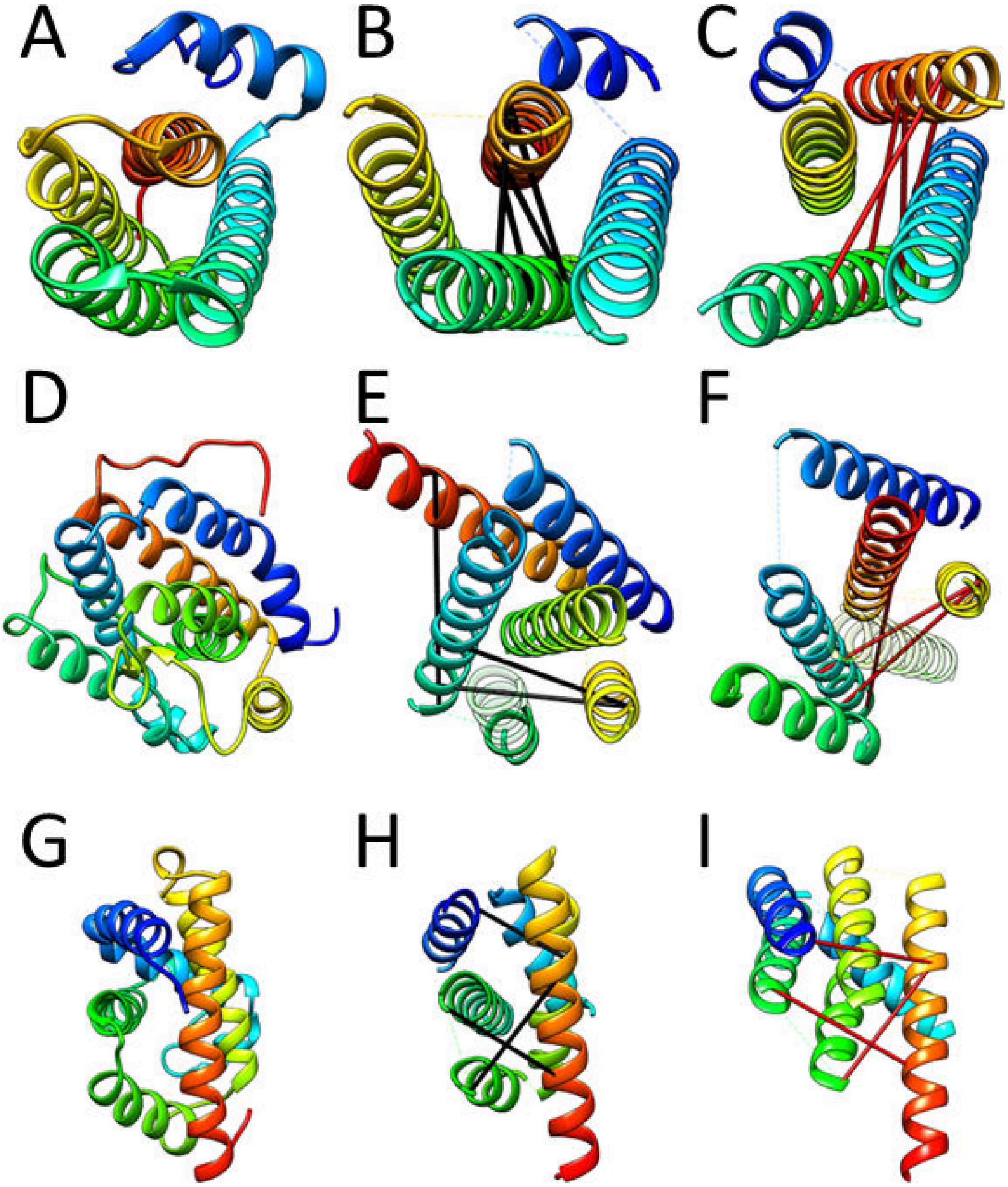

Figure C.3.: Selected prediction results from cross-linking data. *Most accurate models sampled with and without using cross-linking restraints. The RMSD100-values of the most accurate models sampled for 1X91, 1J77, and 1MBO were* 4.8 Å, 6.8 Å *and* 7.1 Å. *By using restraints yielded by Lys-Lys/Asp/Glu reactive cross-linkers, the accuracy could be improved to* 2.7 Å, 5.0 Å *and* 4.2 Å. *Shown are the native structures of 1X91, 1J77, and 1MBO (A, D, G), the most accurate models sampled without cross-linking restraints (B, E, H), and the most accurate models sampled with cross-linking restraints (C, F, I). Selected restraints are shown that are not fulfilled in the model predicted without cross-linking data (red bars), but that are fulfilled in the model predicted with cross-linking data (black bars).*

C.2. Protein structure prediction protocol from cross-links of various lengths

The following protocol capture requires an installation of the BCL, which can be obtained from the website of the Meiler Laboratory.[a] The protein structure prediction protocol from cross-linking data consisted of a three-stage approach:

1. Prediction of the secondary structure from the primary structure (section C.2.1).
2. Prediction of the tertiary structure from the secondary structure and cross-linking data (section C.2.2).
3. Analysis of the results (section C.2.3).

C.2.1. Prediction of the secondary structure

The secondary structure of the benchmark proteins was predicted using the methods PSIPRED[98] and Jufo9D.[97] The prediction files generated by those two methods were subsequently used to create an SSE pool using the following command line:

```
bcl.exe CreateSSEPool -ssmethods PSIPRED JUFO9D -sse_threshold 0.4 0.4 0.4 -pool_min_sse_lengths 5
 ↪  3 -prefix <data_folder> -factory SSPredHighest -output_prefix <output_folder>
```

C.2.2. Prediction of the tertiary structure

The SSEs from the SSE pool are subsequently arranged in the three-dimensional space using the BCL::Fold algorithm, which is part of the BCL. The required input files are a file containing the protein's primary structure in the fasta format, the secondary structure prediction files from PSIPRED and Jufo9D, and the SSE pool generated in section C.2.1. The number of protein models to be sampled can be adjusted by setting the argument of the flag `-nmodels <num_models>` to the desired value.

```
bcl.exe protein:Fold -fasta <target.fasta> -sequence_data <sequence_data_folder> <target> -sspred
 ↪  JUFO9D PSIPRED -pool <sse_pool> -pool_separate -stages_read <stages_file> -protein_storage
 ↪  <output_folder> -prefix <seed> -nmodels <num_models> -random_seed <seed> -histogram_path
 ↪  <path_to_histogram_folder>
```

For predicting the tertiary structure from cross-linking restraints, the following flags have to be added: `-restraint_types Xlink` and `-restraint_prefix <path_to_restraint_file>`. The command line above will result in the desired number of protein models to be sampled and written out in the PDB format.

C.2.3. Analysis of the results

The resulting protein models were analyzed and ranked using the application BCL::Score,[60] which is part of the BCL. The following command line requires the model paths to be listed in an input file and creates a table with one row per protein model, detailing the value of each scoring term, the agreement with the XL-MS restraints, and the RMSD100 relative to the specified reference structure.

[a] http://www.meilerlab.org/bclcommons

```
bcl.exe protein:Score -pdblist <list_of_pdbs> -native <native_structure> -quality RMSD GDT_TS
 ↪  -weight_set <score_weigths> -sspred PSIPRED JUFO9D -sequence_data <sequence_data> <target>
 ↪  -pool <sse_pool> -score_table_write <output_file>
```

APPENDIX D
EFFICIENT SAMPLING OF LOOP CONFORMATIONS

This appendix is based on the publication "Efficient sampling of loop conformations using conformation hashing in conjunction with cyclic coordinate descent".[4] It provides supplementary data and procedures for chapter V on page 67. The computational procedures described in section D.1 detail the necessary steps to sample loop conformations using a combination of conformation hashing and CCD.

D.1. Computational procedures for sampling loop conformations using conformation hashing in conjunction with CCD

This section lists the computational procedures to sample an ensemble of loop conformations using conformation hashing in conjunction with CCD. The employed algorithms were implemented as part of the BCL. The following protocol capture detailing the procedures employed in the manuscript consists of multiple steps:

1. Selection of a set of protein structures in the PDB format to generate the library from.
2. Generation of the loop template library using the BCL.
3. Definition of the configuration files for the loop sampling algorithm to achieve a workflow consisting of conformation hashing and CCD.

The following sections describe each of the steps in detail. They require an installation of the BCL,[a] and, if needed, DSSP.[b] The selection and preparation of protein structures to generate the template library is discussed in section D.1.1. The necessary steps to generate the template library from the selected protein structures is detailed in section D.1.2 on the next page. The computational procedures to sample ensembles of loop conformations from the generated templated library is described in section D.1.3 on the following page.

D.1.1. Selection of set of protein structure to generate the template library

The initial set of protein models serves as seed for the loop template library; therefore determines the initial population of the library. The set of protein structures has to be chosen in a way to maximize the number of structurally dissimilar conformations while avoiding near-duplicates. A near-duplicate in this context is a structurally very similar loop conformation. Not excluding them will increase the size of the template library and therefore increase the time required for reading it in the algorithm. As of October 2017, the PDB contains about 124 000 protein structures. Of those, we only selected the ones within a resolution limit of 3 Å. Additionally, protein models consisting completely or partially of C_α-traces were excluded, since they do not contain a sufficient number of atom coordinates per residue to define a coordinate system for the anchor points. The exclusion of protein models below the chosen resolution limit and protein models containing C_α-traces can easily be achieved using the Dunbrack lab's PISCES server (available at `http://dunbrack.fccc.edu/PISCES.php`).[185,195]

[a] `http://www.meilerlab.org/bclcommons`

[b] `http://swift.cmbi.ru.nl/gv/dssp`

The remaining protein structures (about 85 000) were downloaded from the PDB and for structure groups with a common sequence identity above 25 %, only one representative remained in the set. The pairwise sequence identities were computed with Clustal Omega[201] using the following command:

```
clustalo -profile1 <seq_a> -profile2 <seq_b>
```

This command will result in an alignment of the two given sequences `<seq_a>` and `<seq_b>` in the FASTA-format. Subsequently, the sequence identity or similarity can readily be determined using a simple script.

D.1.2. Generation of the template library

From the remaining set of protein structures that was selected in section D.1.1 on the previous page, the loop template library was generated using the BCL. The BCL does not perform an analysis of the dihedral angles to distiguish between loop regions and helices or strands, but exclusively relies on the SSE definition provided in the PDB file. Although most protein structure deposited in the PDB contain SSE definitions, we chose to ignore those definitions and use the DSSP program[113] instead to define the specific secondary structure regions. This approach ensures that the same definition criteria are applied to all protein structures. The definitions can be obtained from DSSP using the following command, where `<input.pdb>` is the path to the PDB file:

```
dssp -i <input.pdb> -o <output.dssp>
```

The resulting definitions from the DSSP file at location `<output.dssp>` can then be inserted into the PDB file using the script `dssp2pdb` from James Stroud (available at `http://www.jamesstroud.com/software/dssp2pdb`) using the following command:

```
dssp2pdb -5 <input.dssp> <input.pdb> > <output.pdb>
```

The arguments `<input.dssp>`, `<input.pdb>`, and `<output.pdb>` specify the locations of the input DSSP and PDB files as well as to which location the resulting PDB file shall be written. The flag `-5` specifies that only -helices shall result in PDB-entries. To also consider 3(10)-helices, the flag `-3` needs to be set.

The resulting set of protein structures in the PDB format was then used to generate the loop template library. This step requires an installation of the BCL, which provides the `bcl` executable. The template library can be generated using the following command:

D.1.3. Sampling of loop conformations using conformation hashing complemented with CCD

The BCL protein modeling application is based on prediction pipelines, which can be defined and configured using text files. Each pipeline consists of a sequence of optimization modules that can be chained together to achieve the desired output. Each of the optimization modules in turn can be configured using a module-specific set of options. For the loop sampling using conformation hashing in conjunction with CCD, the pipeline consists of two modules — one for the conformation hashing and one for the CCD algorithm. An example pipeline is provided and discussed below.

```
EnsembleNode(
number trajectories=100,
optimizer=MCMOptimizer(
  score function=ProteinModelScoreSum(
    offset=0,
    terms(
    (
      weight=500,
      ProteinModelSSEPairs(
        score function=AASequencePair(
          scoring function=AAPairClash(
            sigmoid width=1,
            histogram file name=aa_distances_0.05.histograms
          ),
          normalize=0
        ),
        normalize=0
      )
    ),
    (
      weight=500,
      ProteinModelSSEPairs(
        score function=SSEPairsFragments(
          packer=SSEClash_NoCache,
          score function=SSEPairClash(min interface length=0,sigmoid width=1),
          normalize=0
        ),
        normalize=0
      )
    ),
    (
      weight=0,
      ProteinModelSSEPairs(
        score function=Loop(histogram filename=loop.histograms,max loop length=25),
        normalize=0
      )
    ),
    (
      weight=50000,
      ProteinModelSSEPairs(
        score function=LoopClosure(
          number excluded residues=1,
          sigmoid width=20,
          fraction allowed distance=1,
          exclude coil=1
        ),
        normalize=0
      )
    ),
    (weight=1000,ProteinModelCompleteness(ignore term loops=0)),
    (weight=1,ProteinModelGap)
    )
  ),
  mutates=MutateDecisionNode(
    (probability=0.5,MutateLoopAdd(loop library=<lib_path>)),
    (probability=0.1,MutateLoopRemove),
    (probability=0.2,MutateLoopReplace(loop library=<lib_path>)),
    (probability=0.3,MutateLoopAddResize(loop library=<lib_path>)),
    (probability=0.5,MutateLoopFragmentAdd(loop library=<lib_path>)),
    (probability=0.3,MutateLoopFragmentReplace(loop library=<lib_path>))
  ),
  termination criterion=Any(Iterations(10000),ConsecutiveUnimprovedSteps(1000)),
  metropolis=Metropolis(
    keep history=0,
    minimum change=0.0001,
    temperature control=DefaultTemperatureControl(temperature=300)
  )
)
)
```

The definition of the pipeline follows an `attribute=value` scheme. The configuration file provided above defines an MCM prediction pipeline performing conformational hashing for missing loop regions. Additional scoring terms and their corresponding weights can be added with the `score_function` block — the example includes scoring terms for evaluating steric interference between residues (`AAPairClash`), SSE-SSE interactions (`SSEPairsFragments`), and loop evaluation (`Loop` and `LoopClosure`). Additional mutates, functors that perturbate the protein model, can be added within the `mutates` block. The example includes the mutates for adding, removing, and replacing loops as well as mutates for fragment-based loop construction. Termination criteria for the MCM prediction can be added in the `termination criterion` block. The example defines termination after either a total of 10 000 MC steps or 1000 MC step without score improvement in a row.

After defining the pipeline in the configuration file `pipeline.conf`, it can be applied to a set of protein structures using the BCL "Optimize" application. This application accepts the pipeline configuration file and a text file `models.ls` as arguments. The file `models.ls` lists the paths to all input protein models in PDB format. The pipeline can then be applied to the input protein models using the following command:

```
  bcl Optimize -pdb_list models.ls -output_prefix <out_dir> -optimizer pipeline.conf
```

This command result in sampling the specified number of protein models using the provided prediction pipeline and write the sampled models to the location specified by the flag `-output_prefix`. By using the flag `-help`, a complete list of all command line options can be obtained.

APPENDIX E
PREDICTING THE MONOMERIC AND HOMODIMERIC FORMS OF BAX

This appendix is based on the publication "Pushing the size limit of de novo structure ensemble prediction guided by sparse SDSL-EPR restraints to 200 residues: The monomeric and homodimeric forms of BAX".[5] It provides supplementary data and procedures for chapter VI on page 83. The supplementary data in section E.1 lists the agreement of the X-ray-derived model of soluble monomeric BAX and membrane-associated homooligomeric BAX with the SDSL-EPR data. The procedures described in section E.2 on page 186 detail the computational procedures employed in this study.

E.1. Supplementary data

This section contains a plot depicting the dependence of the μ-value in dependence of the number of models taken into consideration. Additionally, tables are provided that quantify the agreement of the soluble monomeric and membrane-associated homodimeric BAX structures with the SDSL-EPR data. The agreement is shown according to geometrical aspects as well as in terms of the CONE model.

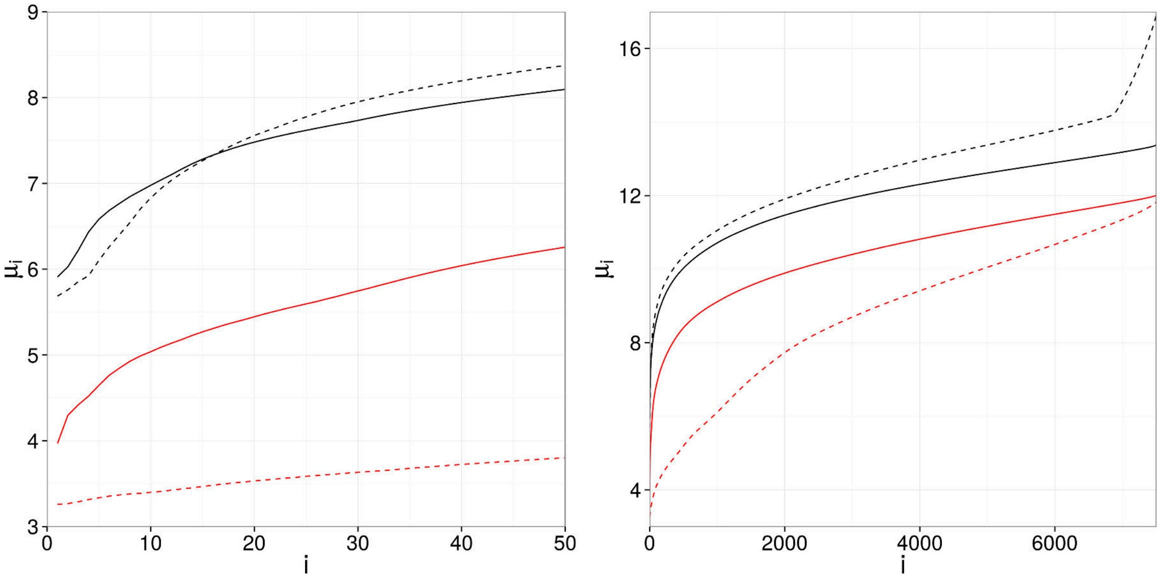

Figure E.1.: Dependence of the* μ*-value on the model numbers.*(A,B) The average RMSD100-values, μ_i, of the i most accurate models (y-axis) are shown in dependence of the number of considered models (x-axis) for soluble monomeric (solid lines) and homodimeric (dashed lines) BAX without (black) and with (red) using SDSL-EPR restraints.*

		NMR-derived model			Idealized NMR-derived model		
Restraint	D_{SL} (Å)	D_{BB} (Å)	Δ_{SL} (Å)	Score	D_{BB} (Å)	Δ_{SL} (Å)	Score
16:62	32.1	22.4	9.7	−0.68	21.9	10.2	−0.60
62:101	33.7	26.3	7.4	−0.88	26.7	7.0	−0.90
62:87	42.1	31.8	10.3	−0.59	32.1	10.0	−0.64
72:126	23.0	21.5	1.5	−1.00	18.4	4.6	−0.97
55:126	38.9	31.5	7.4	−0.87	32.7	6.2	−0.93
87:126	26.0	20.0	6.0	−0.94	17.6	8.4	−0.81
62:149	31.7	20.6	11.1	−0.40	21.3	10.4	−0.57
101:149	25.9	19.3	6.6	−0.91	19.4	6.5	−0.92
87:149	31.9	25.9	6.0	−0.94	28.0	3.9	−0.98
101:169	25.6	26.2	−0.6	−0.95	26.0	−0.4	−0.97
62:186	32.4	26.2	6.2	−0.93	28.0	4.4	−0.97
62:126	32.0	25.6	6.4	−0.92	24.9	7.1	−0.89
55:87	43.5	37.4	6.1	−0.93	40.0	3.5	−0.99
55:101	36.8	32.9	3.9	−0.98	34.2	2.6	−1.00
72:87	36.2	30.3	5.9	−0.94	26.1	10.1	−0.63
101:126	33.3	32.2	1.1	−1.00	30.5	2.8	−0.99
72:101	32.9	27.9	5.0	−0.96	24.6	8.3	−0.82
55:149	26.2	19.4	6.8	−0.91	22.7	3.5	−0.99
62:169	17.6	14.9	2.7	−1.00	13.8	3.8	−0.98
126:169	39.1	34.3	4.8	−0.97	33.7	5.4	−0.95
72:169	25.5	15.3	10.2	−0.60	17.3	8.2	−0.83
126:149	30.6	31.8	−1.2	−0.90	32.0	−1.4	−0.89
87:169	42.0	37.7	4.3	−0.98	37.9	4.1	−0.98
72:186	25.8	19.3	6.5	−0.92	17.8	8.0	−0.84
87:186	25.1	19.3	5.8	−0.94	19.6	5.5	−0.95
mean	—	—	5.6	−0.88	—	5.7	−0.88

Table E.1.: Agreement of the soluble monomeric BAX NMR-derived model with the SDSL-EPR data. *To study the suitability of the SDSL-EPR distance restraints in conjunction with the translation model for structure prediction, the agreement of the experimentally determined structure with the SDSL-EPR distance restraints was calculated with the best agreement having a score of* −1 *and the worst agreement having a value of* 0*. Some agreements might only be achievable through distinct bendings of the SSEs, which complicate structure prediction. Therefore, the agreement analysis is also shown for the NMR-derived model as well as the idealized structure. Also shown are the backbone* (D_{BB})*and observed spin-pin* (D_{SL})*.*

		X-ray-derived model			Idealized X-ray-derived model		
Restraint	D_{SL} (Å)	D_{BB} (Å)	Δ_{SL} (Å)	Score	D_{BB} (Å)	Δ_{SL} (Å)	Score
55A:55B	48.0	44.6	3.4	−0.99	42.6	5.4	−0.95
62A:62B	23.2	21.6	1.6	−1.00	22.8	0.4	−1.00
72A:72B	29.2	20.7	8.5	−0.81	17.9	11.3	−0.36
87A:87B	53.1	47.7	5.4	−0.95	46.6	6.5	−0.92
101A:101B	42.6	41.3	1.3	−1.00	41.0	1.6	−1.00
62A:87A	41.6	38.2	3.4	−0.99	37.8	3.8	−0.98
55A:87A	49.3	48.3	1.0	−1.00	46.6	2.7	−1.00
55A:101A	35.4	40.3	−4.9	−0.60	39.0	−3.6	−0.70
72A:87A	29.8	29.4	0.4	−1.00	28.1	1.7	−1.00
62A:101A	32.7	30.7	2.0	−1.00	31.3	1.4	−1.00
72A:101A	31.1	28.7	2.4	−1.00	27.2	3.9	−0.98
mean	—	—	3.1	−0.94	—	3.8	−0.90

Table E.2.: Agreement of the homodimeric BAX X-ray-derived model with the SDSL-EPR data. *To study the suitability of the SDSL-EPR distance restraints in conjunction with the translation model for structure prediction, the agreement of the experimentally determined structure with the SDSL-EPR distance restraints was calculated with the best agreement having a score of* -1 *and the worst agreement having a value of* 0*. Some agreements might only be achievable through distinct bendings of the SSEs, which complicate structure prediction. Therefore, the agreement analysis is also shown for the X-ray-derived model as well as the idealized structure. Also shown are the backbone* (D_{BB}*) and observed spin-pin* (D_{SL}*) distances.*

E.2. Procedures for the structure prediction of BAX

This protocol capture requires an installation of the BCL, which is available on the website of the Meiler Laboratory.[a] The following procedures describe all necessary steps to reproduce the data shown in the manuscript or to apply the protocol to predict the tertiary structure of other proteins from SDSL-EPR distance restraints. The described protocol requires the following files and applications, which are referenced in the following protocol capture:

- `target.fasta` — A file describing the primary structure of the protein in the fasta-format. For homooligomeric proteins, only one chain should be included.
- `target.ss` — A file containing secondary structure predictions from PSIPRED, which can be computed at http://bioinf.cs.ucl.ac.uk/psipred.
- `target.jufo` — A file containing secondary structure predictions from Jufo9D, which can be computed at http://www.meilerlab.org/index.php/servers/show?s_id=5.
- `stages.txt` — A file defining the setup of BCL::Fold. The file used for the EPR-guided prediction of BAX is provided in section E.2.1.
- `target.epr_cst_bcl` — A file containing the SDSL-EPR distance restraints. An example is provided in section E.2.2 on page 188.
- `bcl` — The BCL executable, which can be obtained free of charge for academic purposes at http://www.meilerlab.org/bclcommons.

The prediction of the tertiary structure from the input files listed above consists of multiple steps described in detail in the following sections:

1. Preparation of the input files `target.fasta`, `stages.txt`, and `target.epr_cst_bcl` (see the following section for details) and obtaining the BCL executable.
2. Obtaining the PSIPRED and Jufo9D secondary structure predictions (`target.ss` and `target.jufo`) from the respective online servers.
3. Prediction of the secondary structure of the protein, which is described in section E.2.3 on page 188.
4. Prediction of the tertiary structure of the protein, which is described in section E.2.4 on page 189.

E.2.1. Preparation of input files: the stage file

The stage file defines the setup of the BCL::Fold algorithm. Specifically, it defines how many MC steps to perform, which transformations to apply, and which scoring terms to use. The BCL::Fold algorithm is modular allows composition from multiple stages that form the resulting prediction pipeline. The stage file used in this study for soluble monomeric BAX was:

[a] http://www.meilerlab.org/bclcommons

```
STAGE Stage_assembly_1
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_01.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_2
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_02.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_3
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_03.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_4
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_04.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_5
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_05.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_refinement_1
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE refinement_01.scoreweights
MUTATE_PROTOCOLS Default Refinement
NUMBER_ITERATIONS 2000 400
STAGE_END
```

The listed stage file defines the configuration of the protein structure prediction pipeline. The pipeline consists of six modules, which are all MCM optimization algorithms. Using the options SCORE_PROTOCOLS and MUTATE_PROTOCOLS, the user can define which scoring terms to use and which transformations to apply during each module. The option NUMBER_ITERATIONS allows setting the maximum number of MC steps per module and the maximum number of consecutive MC steps without score improvement.

For homooligomeric proteins, the stage file has to be adjusted and the protocol Multimer has to be added to all SCORE_PROTOCOLS and MUTATE_PROTOCOLS to enable assembly and refinement in multimer mode. As of now, the BCL only supports assembly of monomeric proteins or symmetric multimeric proteins exhibiting cyclic or dihedral symmetry. The weights for the different scoring terms were kept constant over all stages besides the weights for the SSE clash (*sseclash*) and amino acid clash (*aaclash*) scores. They were 0 during assembly_1, 125 during assembly_2, 250 during assembly_3, 375 during assembly_4, and 500 during assembly_5 and refinement_1.

E.2.2. Preparation of input files: the restraint file

The BCL can use intra-protomer and inter-protomer distance restraints by specifying restraint endpoints using their respective chain and sequence identities. An example file defining SDSL-EPR distance restraints would have the following format:

```
Atom Distance Assigned
A 55 CB B 55  CB 48.0 100 1
A 62 CB A 102 CB 23.2 100 1
```

The list of residue-residue distances is preceded by the identifier Atom Distance Assigned and is followed by lines defining one restraint each. Each restraint line has to have the format <chain_id_1> <seq_id_1> CB <chain_id_2> <seq_id_2> CB distance 100 1. The chain and sequence identifiers refer to the spin labeling sites and distance refers to the experimentally observed spin-spin distance. The first line in the example above therefore defines an SDSL-EPR distance restraint with a spin-spin distance of 48.0 Å between residue 55 of chain A and residue 55 of chain B and an SDSL-EPR distance restraint with a spin-spin distance of 23.2 Å between residues 62 and 102 of chain A. The value CB refers to the C_β-atom that is used for calculating the residue-residue distance in the protein models. For glycine that does not have a C_β-atom, the value should still be set to CB, which will result in the $H_{\alpha 2}$-atom being used instead.

E.2.3. Prediction of the secondary structure

The BCL::Fold algorithm, which is part of the BCL, was used to generate SSE pools from the PSIPRED and Jufo9D secondary structure predictions. To generate the SSE pool for BAX, the following command line was used:

```
bcl.exe CreateSSEPool -ssmethods JUFO9D PSIPRED -pool_min_sse_lengths 5 3 -sse_threshold 0.5 0.5
 ↪  0.5 -prefix target -join_separate -factory SSPredThreshold
```

The command line listed above creates an SSE pool for the target protein. The working directory must contain the fasta and PSIPRED/Jufo9D prediction files for the target protein.

E.2.4. Prediction of the tertiary structure

Protein structure prediction was distributed onto the *Advanced Computing Center for Research and Education* at Vanderbilt University and the Titan cluster at Oak Ridge National Laboratory. The command line for predicting a number of structures for a monomeric protein using an SSE pool containing all predictions using EPR distance restraints is:

```
bcl.exe protein:Fold -fasta target.fasta -function_cache -pool_separate -min_sse_size 5 3 -sspred
 ↪   PSIPRED JUFO9D -pool target.pool -stages_read stages.txt -pool_prefix target -nmodels 40
 ↪  -prefix target -protein_storage pdbs/ -sequence_data . target -opencl Disable
 ↪  -restraint_types DistanceEPR -restraint_prefix target -random_seed
```

This command line assumes that the input files described in the previous sections are in the working directory. The command line for predicting the tertiary structure of homodimeric BAX has an additional flag for C2-symmetry: `-symmetry C2`. The BCL supports cyclic and dihedral symmetries with user-defined multiplicities that can be set using this flag with the corresponding argument (`Dn` for dihedral symmetry and `Cn` for cyclic symmetry, with `n` being the number of protomers in the protein).

E.2.5. Obtaining simulated EPR distance restraints

The BCL can also be used to simulate additional SDSL-EPR distance restraints as used in this study. The spin labeling sites were chosen using an algorithm attempting to distribute the spin labels over all SSEs.[155] The algorithm employs MCM sampling to find the optimal distribution of spin labeling sites and can be executed using the following command line:

```
bcl.exe OptimizeDataSetPairwise -fasta target.fasta -pool_min_sse_lengths 0 0 -pool target.pool
 ↪  -distance_min_max 15 50 -nc_limit 10 -ensembles target_ensembles.ls -mc_number_iterations
 ↪   100000 100000 -prefix target_ -nmodels 500 -read_scores_optimization opt_score_weights.wts
 ↪  -read_mutates_optimization mutate_weights.wts -data_set_size_fraction_of_sse_resis 0.2
 ↪  -random_seed
```

This command line creates 500 restraint sets for the target protein. Alternatively, simulated restraints can also be obtained from the Meiler Lab server.[b] The output of the previous algorithm will be pairs of spin labeling sites deemed optimal by the algorithm. In a second step, an uncertainty has to be added to simulate the uncertainty of the SDSL-EPR experiment:

```
bcl.exe SimulateDistanceRestraints -pdb target.pdb -simulate_distance_restraints -output_file
 ↪   target.epr_cst_bcl -min_sse_size 0 0 0 -add_distance_uncertainty sl_cb.histograms
 ↪  -restraint_list target.epr 0 1 5 6 -random_seed
```

[b] http://www.meilerlab.org/index.php/servers/show?s_id=16

E.2.6. Idealization of protein structures

The BCL can be used to idealize protein structures by setting the dihedral angles (ϕ, ψ) of SSEs to the idealized values of (−60°, −40°) for α-helices and to (−135°, 135°) for β-strands. The following command line will idealize the input PDB file and write out the idealized structure as PDB file:

```
bcl PDBConvert <input.pdb> -idealize -bcl_pdb
```

APPENDIX F
STRUCTURE AND DYNAMICS OF TYPE III SECRETION EFFECTOR PROTEIN EXOU

This appendix is based on the publication "Structure and Dynamics of Type III Secretion Effector Protein ExoU As determined by SDSL-EPR Spectroscopy in Conjunction with De Novo Protein Folding".[6] It provides supplementary data and procedures for chapter VII on page 99. The supplementary data in section F.1 lists the agreement of the X-ray-derived model of ExoU with the SDSL-EPR data and compares explicitly simulated EPR distance distributions to the experimentally determined distance distributions. The procedures described in section F.2 on page 194 detail the computational procedures employed in this study.

F.1. Supplementary data for ExoU

This section contains an analysis of the agreement of the X-ray-derived model of ExoU with the SDSL-EPR data under geometrical aspects and in the terms of the CONE model. The distance distributions derived from the DEER experiments are shown and compared to the distance distributions obtained from explicit spin label simulation on the X-ray-derived model.

Restraint	D_{SL} (Å)	D_{BB} (Å)	Score
629-645	19.4	23.3	−0.67
636-592	26.4	26.4	−0.99
636-645	20.6	16.1	−0.97
636-649	19.8	11.7	−0.83
636-657	20.0	14.2	−0.94
636-672	28.2	—	—
649-672	28.3	—	—

Table F.1.: Available SDSL-EPR data for the C-terminal domain of ExoU. *Seven SDSL-EPR measurements were available. Shown are the observed mean spin-spin distance (D_{SL}), the $C_\beta - C_\beta$ distance of the spin labeling sites in the X-ray-derived model (D_{BB}, PDB entry 3TU3), and the agreement score of the X-ray-derived model with the SDSL-EPR data according to the CONE model-based scoring function. Restraints without D_{BB} and score referred to spin labeling sites not resolved in the X-ray-derived model.*

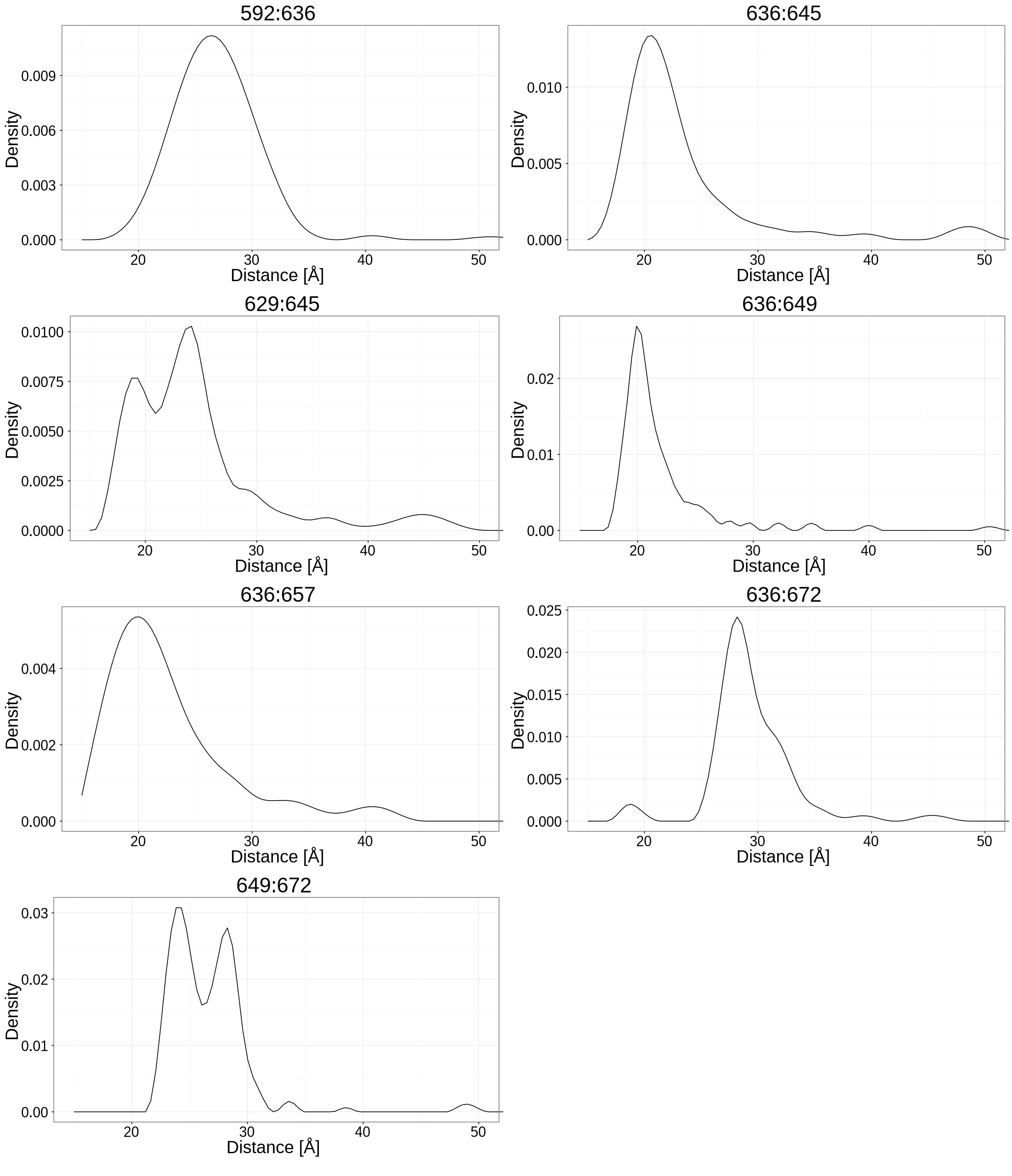

Figure F.1.: Residue-residue distance distributions derived from the DEER experiment for the C-terminal domain of ExoU.

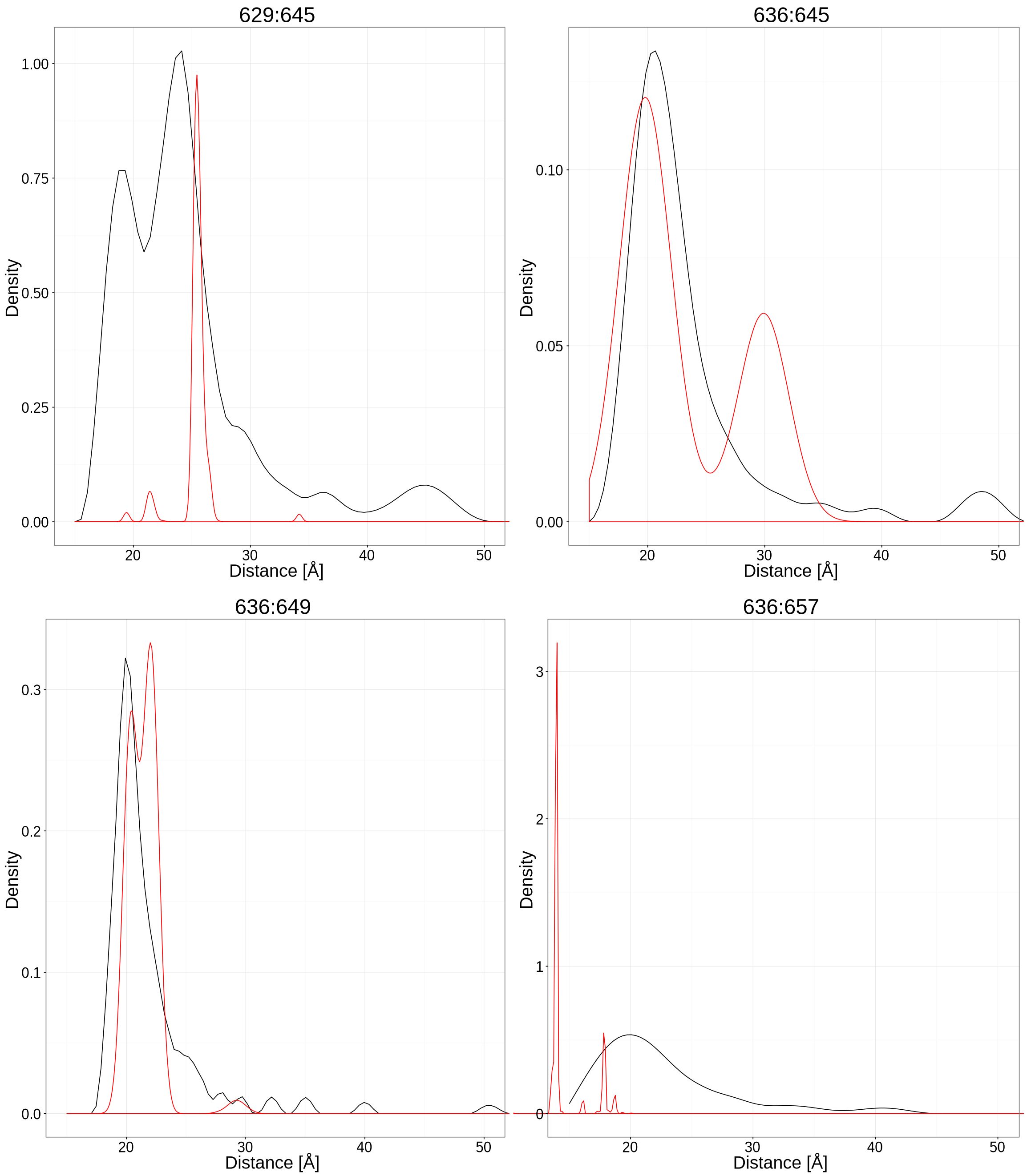

Figure F.2.: Explicit simulation of the spin labeling pairs used on ExoU. *The distance distributions arising from four spin labeling pairs were simulated explicitly (red) and compared to the experimentally determined distance distribution (black).*

F.2. Procedures for the EPR-guided structure and dynamics predictions of ExoU

This protocol capture requires an installation of the BCL, which is available on the website of the Meiler Laboratory.[a] The *de novo* protein structure prediction protocol for ExoU consisted of four modules: 1. secondary structure prediction (section F.2.1), 2. topology sampling (section F.2.2), 3. clustering (section F.2.3 on the following page), and 4. loop construction in conjunction with high-resolution refinement (section F.2.4 on the next page). The following sections provide a detailed protocol capture to reproduce the data reported in this manuscript.

F.2.1. Prediction of the secondary structure

The secondary structure of ExoU was predicted using PSIPRED and Jufo9D. Both approaches are accessible through their respective webservers at http://bioinf.cs.ucl.ac.uk/psipred (PSIPRED) and http://www.meilerlab.org/index.php/servers/show?s_id=5 (Jufo9D). The webservers require the primary structure of the protein in question and the predictions' output files for the topology search module. From the output files, an SSE pool can be created using the BCL with the following command line:

```
bcl CreateSSEPool -ssmethods JUFO9D PSIPRED -pool_min_sse_lengths 5 3 -sse_threshold 0.5 0.5 0.5
 ↪  -prefix <prefix> -join_separate -factory
```

The created SSE pool is a plain text file containing the start and end points (defined through their sequence and chain identifiers) of predicted SSEs. Because the three secondary structure prediction methods can produce conflicting results, the SSEs in the pool can be overlapping. The BCL::Fold algorithm ensures, that overlapping predictions are not inserted at the same time.

F.2.2. Topology sampling

In a preparation step, The EPR-derived distance restraints have to be formalized in a BCL-readable format like the following example, which defines two EPR-derived distance restraints:

```
  Atom Distance Assigned
  A 20 CB A 40 CB 26.4 100 1
  A 56 CB B 61 CB 20.6 100 1
```

The list of restraints has to be preceded by the line `Atom Distance Assigned`. Following this line, each line defines one residue-residue restraint. The residues affected by this restraint are identified by their respective chain and sequence identifiers. The example above defines two restraints — one intra-protomer restraint and one inter-protomer restraint.

After preparation of the restraint file and the SSE pool, the BCL can be used to predict the topology of the protein. The algorithm BCL::Fold is part of the BCL and employs an MCM algorithm to sample possible topologies. Knowledge-based potentials are used to approximate a topology's free energy. For the prediction of ExoU's topology, the following command line was used:

[a] http://www.meilerlab.org/bclcommons

```
bcl -stages_read stages.txt -restraint_types DistanceEPR -restraint_prefix <cst_prefix>
 ↪  -protein_storage <output_folder> -prefix <output_prefix> -sequence_data <input_prefix> 3tu3
 ↪  -sspred PSIPRED -opencl Disable -nmodels <num_models> -start_model <start_model>
```

The flag for providing a start model, `-start_model`, was only set for the second iteration of the topology sampling. The start models were the models selected through clustering (see section F.2.3) after the first round of topology sampling. The stage file provided through the flag `-stage_file` configures the MCM algorithm — it defines the number of MC steps to performs, which transformations to apply, and which scoring terms to use. The format is shown in section F.2.5 on the next page.

F.2.3. Clustering of the sampled models

Clustering was performed using a k-means implementation in R[100] in conjunction with using the RMSD as dissimilarity metric. Once a matrix containing the pairwise dissimilarities between models has been obtained, the clustering can be performed in R using the following sequence of commands:

```
# load libraries
library(cluster)

# load the dissimilarity matrix created with the BCL
data_mat <- as.matrix(read.table(”distance_matrix.tbl”, header = T))

# create a full matrix
data_mat <- data_mat + t(data_mat)

# convert into a dissimilarity matrix
data_mat <- as.dist(data_mat)

# cluster for k cluster centers
clusters <- pam(data_mat, k)

# display information about the clustering
clusters$clusinfo

# display cluster medoids
clusters$medoids
```

The silhouette score can be directly computed from the clusters object created above using the command `silhouette(clusters)`.

F.2.4. Loop construction and high-resolution refinement

The high-resolution refinement and loop construction using the CCD algorithm[101] was performed using the Rosetta software suite and can be repeated using the following command line:

```
loopmodel -loops:frag_sizes 9 3 1 -loops:frag_files <fragments_9> <fragments_3> none
 ↪  -loops:remodel quick_ccd -loops:refine refine_ccd -loops:extended -loops:relax fastrelax -ex1
 ↪  -ex2 -database <database> -nstruct <num_models> -in:file:s <start_model> -loops:loop_file
 ↪   <loops_file> -out:prefix <output_prefix> -constraints:cst_file <restraint_file>
 ↪  -constraints:epr_distance -score:weights
```

F.2.5. Configuration file for BCL::Fold

```
STAGE Stage_assembly_1
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_01.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_2
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_02.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_3
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_03.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_4
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_04.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_5
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE assembly_05.scoreweights
MUTATE_PROTOCOLS Default Assembly
NUMBER_ITERATIONS 2000 400
STAGE_END
```

```
STAGE Stage_refinement_1
TYPE MCM
SCORE_PROTOCOLS Default Restraint
SCORE_WEIGHTSET_FILE refinement_01.scoreweights
MUTATE_PROTOCOLS Default Refinement
NUMBER_ITERATIONS 2000 400
STAGE_END
```

APPENDIX G
PROTONATION-DEPENDENT CONFORMATIONAL DYNAMICS OF EMRE

This appendix is based on the publication “Protonation-dependent conformational dynamics of the multidrug transporter EmrE”.[7] It provides supplementary data and procedures for chapter VIII on page 111, describing the conformational dynamics of EmrE as determined through SDSL-EPR spectroscopy.

G.1. Supplementary data

This section contains additional data describing the conformational dynamics of EmrE. Distance distributions obtained through the DEER experiment detail the conformational dynamics in dependence on the pH-value and substrate-binding.

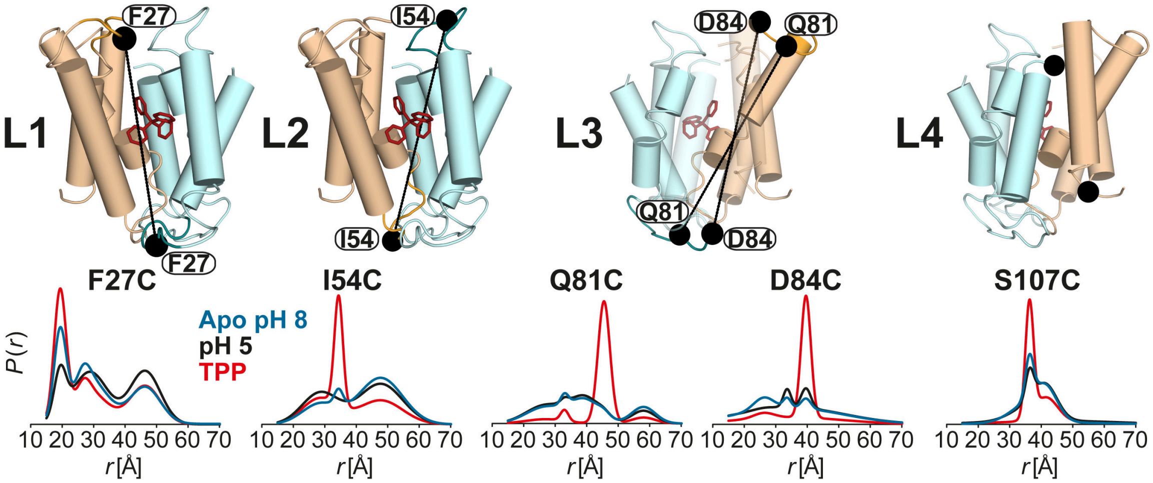

Figure G.1.: Ligand-dependent conformational changes of EmrE in the loop regions (L1 to L4). *Residue* 107 *is not resolved in the X-ray structure.*

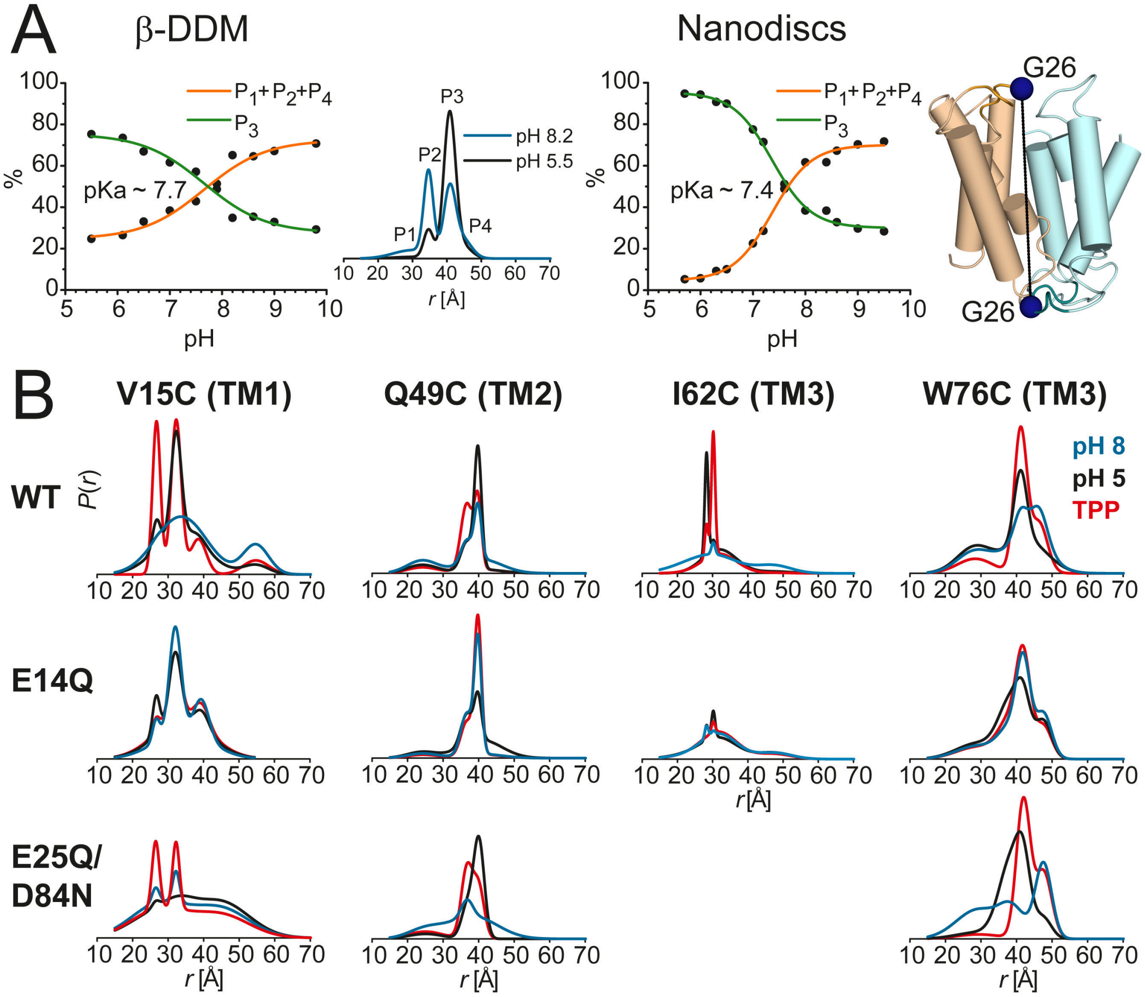

Figure G.2.: Equilibrium of EmrE. *(A)* p*H-dependent conformational equilibrium of EmrE in nanodiscs and* β*-DDM micelles. Distance distributions of the G26C pair were obtained at different* p*H-values ranging from* 5.5 *to* 10.0 *in* β*-DDM and* 5.7 *to* 9.5 *in nanodiscs (figure G.3 on the next page). The variation in population of rising* ($P_1 + P_2 + P_4$) *or equally decreasing* (P_3) *distance peaks (middle panel) as a function of* pH *was used to estimate the* pK_a*-value for conformational changes in EmrE. (B) Effect of protonation-mimetic mutation of acidic residues on conformational states in equilibrium (*β*-DDM micelles). The single (E14Q) or double (E25Q/D84N) mutations were combined with the single-cysteine mutations (see figure G.4 on page 201).*

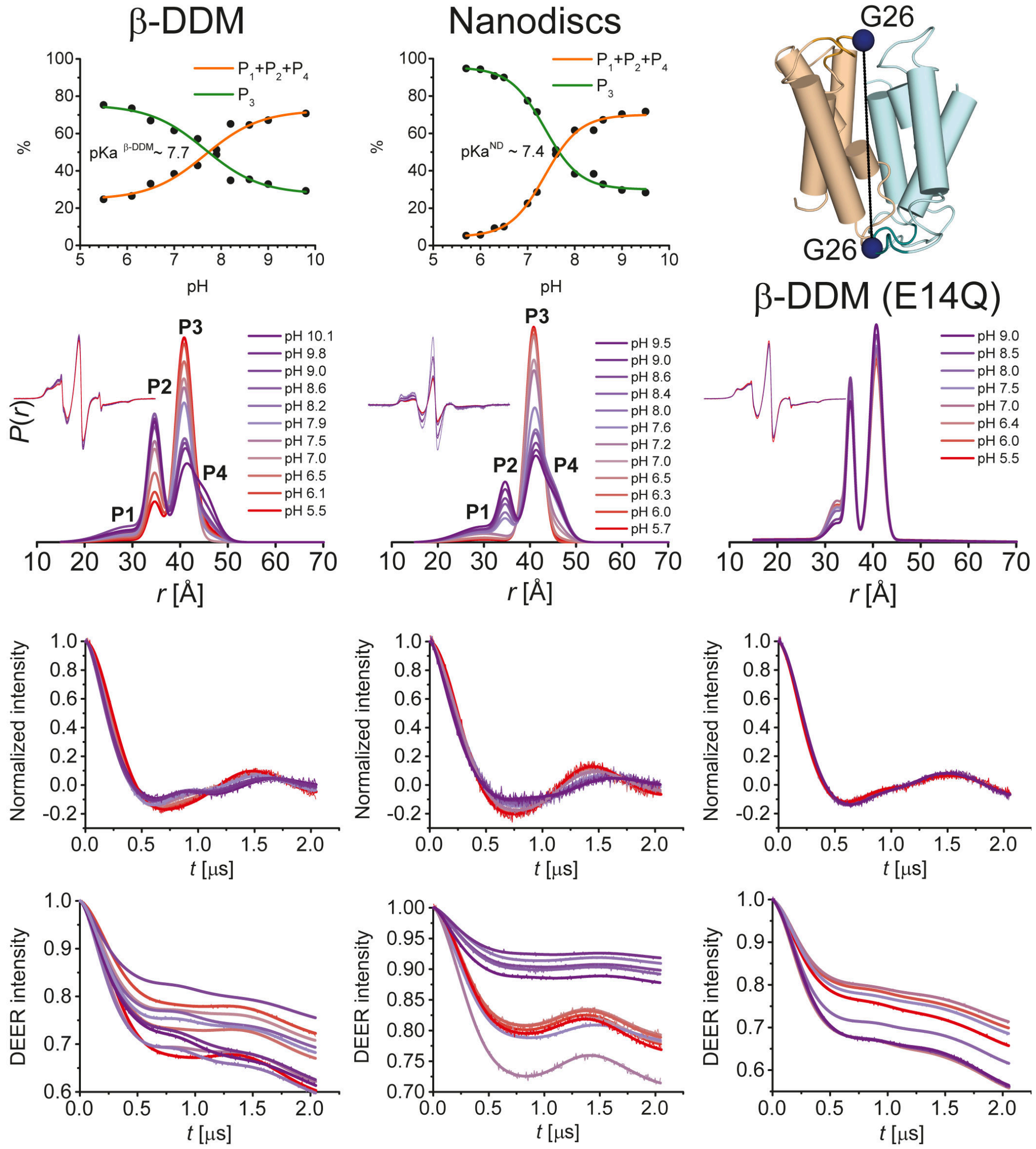

***Figure G.3.:* pH-dependent conformational equilibrium of EmrE in nanodiscs and β-DDM micelles.** *Distance distributions of the G26C pair were obtained at different* pH*-values ranging from* 5.5 *to* 10.0 *and* 5.7 *to* 9.5 *in* β*-DDM and nanodiscs respectively. In addition, distance distributions were obtained for the protonation-mimetic mutant G26C E14Q in* β*-DDM at* pH*-values ranging from* 5.5 *to* 9.0*. The variation in population of rising ($P_1 + P_2 + P_4$) or equally decreasing (P_3) distance peaks as a function of* pH *was used to estimate the* $\mathrm{p}K_\mathrm{a}$*-value for conformational changes in EmrE. From bottom to top, primary DEER traces with the corresponding fits, baseline-corrected and normalized DEER traces along with the fits, distance distributions and the CW-EPR spectra as insets, and* pH*-titration curves are shown. The E14Q mutation abrogates the* pH*-dependent changes in the amplitude of the distance components.*

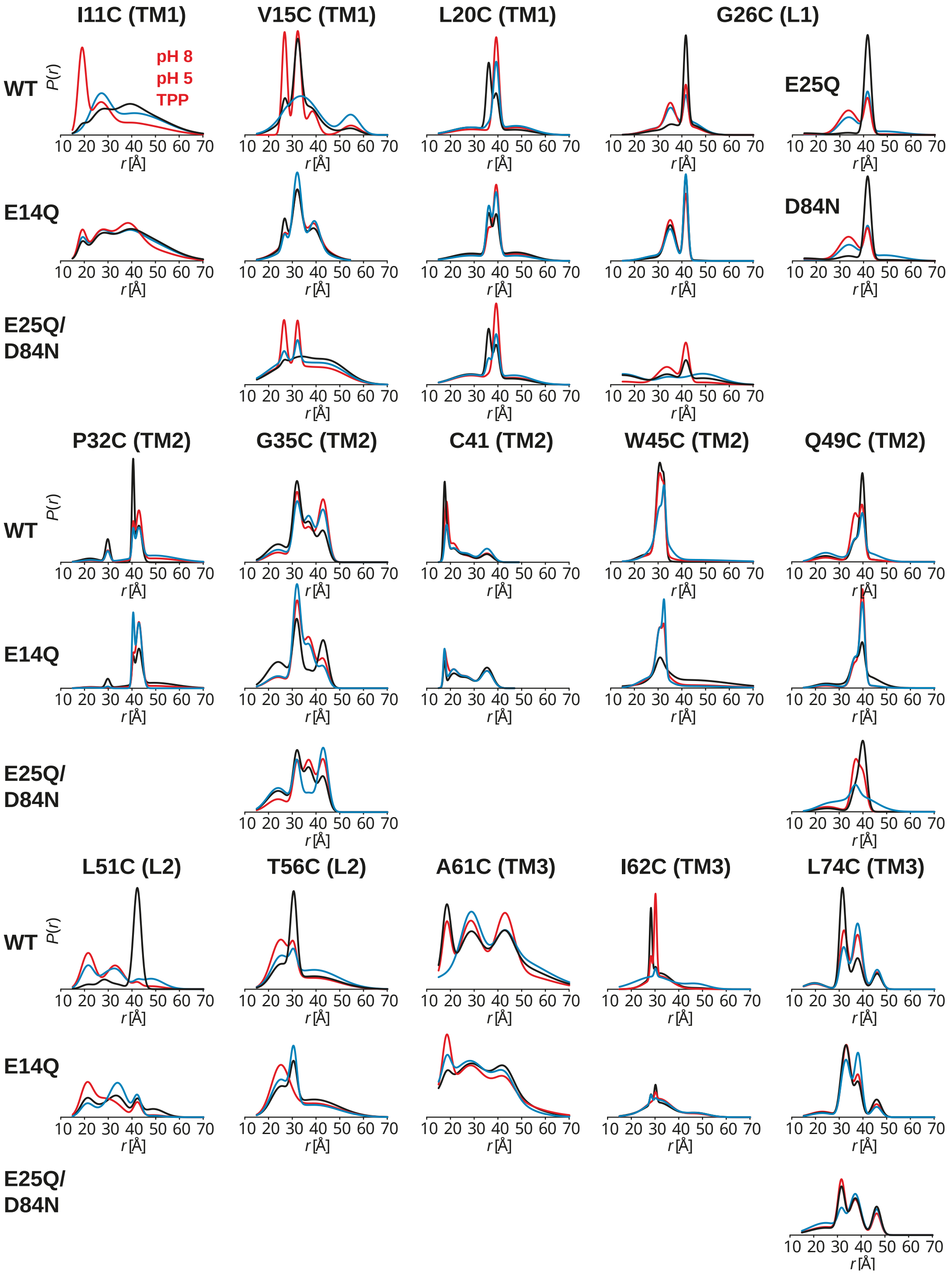

*Figure G.4.: **Effect of protonation-mimetic mutation of acidic residues on distance distributions in** β**-DDM micelles.** The single (E14Q, E25Q, D84N) or double (E25Q/D84N) mutations were combined with the single-cysteine mutations.*

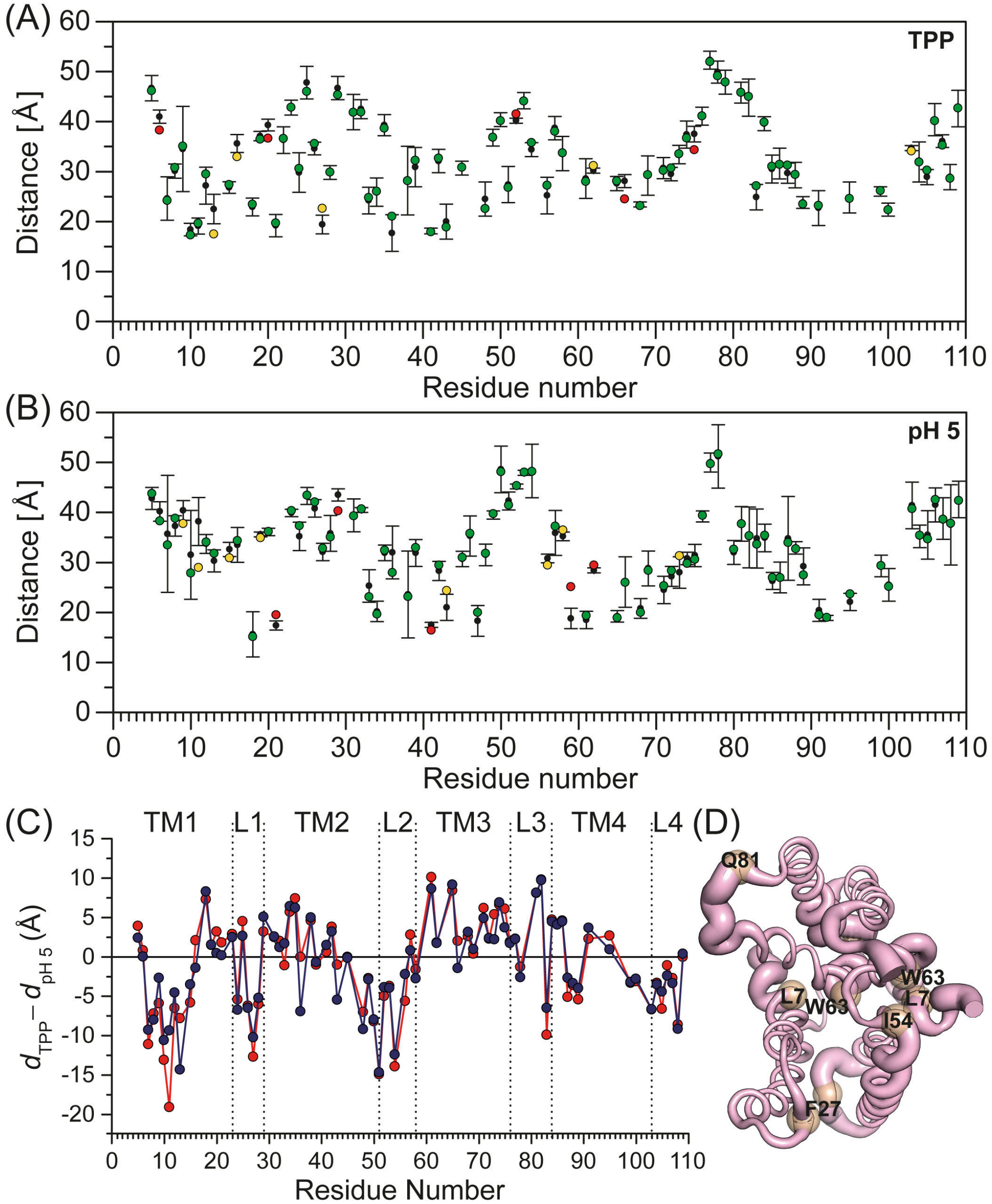

Figure G.5.: Comparison of the obtained experimental distances for the TPP-bound and protonated states with distances predicted on the generated models. *(A) Refined X-ray structure. (B) The symmetric protonated* (pH 5) *model. The average distances (colored circles) were predicted at* 298 K *using MMM 2013.2 software package.*[214] *The color code is described in figure VIII.3 on page 117. (C) Ligand-dependent changes (TPP-bound to protonated) in the experimental distances (red) vs. predicted ones on the TPP-bound and protonated models (royal blue). (D)* C_{α}*-RMSD between the refined TPP-bound X-ray structure and the symmetric protonated model is displayed by the ribbon thickness on the refined X-ray structure.*