STRUCTURE PREDICTION AND VARIANT INTERPRETATION OF MEMBRANE PROTEINS AIDED BY MACHINE LEARNING ALGORITHMS

By

Bian Li

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

May 11, 2018

Nashville, Tennessee

Approved:

Jens Meiler, Ph.D.

Terry Lybrand, Ph.D.

Clare McCabe, Ph.D.

Terunaga Nakagawa, Ph.D.

To my infinitely supportive parents Dingou Li and Yinglian Li

and

To my beloved wife Xuan Zhang

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest appreciation to my advisor and committee chair, Dr. Jens Meiler, who has the attitude and expertise of a great scientist and the charisma of an inspiring mentor. I feel extremely fortunate that my application to the Graduate Program in Chemistry at Vanderbilt five years ago was among the few selected by Dr. Meiler for further consideration. He interviewed me and made my application case a strong one to the program. His endorsement was undeniably one of the decisive reasons that convinced the program to give me an offer. I have also had the fortune of having him as my advisor. His ingenious insight, invaluable scientific guidance, and incessant encouragement are so critical that without them nothing in this dissertation would have been possible.

I'm also extremely grateful to Dr. Terry Lybrand, Dr. Clare McCabe, and Dr. Terunaga Nakagawa, who are my committee members. It's difficult to find a better committee for one who works on problems in computational structural biology. They have continuously made suggestions to improve my dissertation research and have never hesitated to give me career advice and to write strong letters of reference for me. Dr. Lybrand also gave me invaluable advices on things that need particular attention when looking for postdoctoral positions and making the most of a postdoctoral training.

I'm indebted to all of those whom I have had the privilege of working with in Dr. Meiler's research group. Particularly, I would like to thank Jeffrey Mendenhall, who is an extremely talented programmer and scientist in the group. He is my officemate and sits next to me. The skills, both computational and life, that I learned from him has proven to be indispensable for me to survive as a scientist and a sojourner in the U.S. I would also like to express my thanks to Dr. Axel Fischer, who was my mentor when I was a rotation student. He helped me tremendously to learn C++.

Graduate school is a huge take both for me personally and for my family. It is a tunnel that is so long and so stochastic that the chance of reaching the other end and seeing the light is nothing but an abyss without the infinite love and support from my family. In the past five years, even though I've only seen my parents once, I know they are always there whenever I need them. Finally, I would like to give my special thanks to my beloved wife for her love, support, patience, sacrifice, and everything else.

SUMMARY

Protein folding is a process of molecular self-assembly during which a disordered polypeptide chain collapses to form a compact and well-defined three-dimensional (3D) tertiary structure. A grand challenge in biochemistry has been to understand the process by which proteins fold into their functional tertiary structure (folding mechanism) and to predict this tertiary structure from amino acid sequence (structure prediction), two tasks that are collectively known as "the protein folding problem". Solving this problem is of far-reaching impact as it will not only reveal the missing link between sequence and structure but also provide molecular biologists with a theoretical framework and practical tools for applications such as drug design and protein engineering. Chapter I of this dissertation gives a comprehensive review of the computational techniques developed in the past half century or so for studying the protein folding problem.

Helical membrane proteins (HMPs) play essential roles in various biological processes, including signal transduction, ionic and molecular transportation across the membrane, and energy generation. It was estimated that HMPs constitute about 20% to 30% of the human genome. Frequently, these transmembrane proteins do not function as monomers but undergo concerted interactions to form either homo-oligomers or interacting with other transmembrane proteins to form hetero-oligomers. Despite their prevalence in the genome, a very small portion of structures in the Protein Data Bank are HMPs due to the experimental difficulties in determining structures of HMPs and their complexes. Therefore, accurate and efficient computational methods would be valuable tools to complement existing experimental techniques. Chapters II, III, and IV describe a novel computational approach developed in this work for improving the tertiary structure prediction of HMPs and the quaternary structure prediction of HMP complexes.

In chapter II, the concept of residue weighted contact number (WCN) is introduced and a method is developed, using state-of-the-art machine learning techniques, for predicting WCNs from amino acid sequence alone. The WCN of an amino acid residue is defined as the number of neighboring residues weighted by their proximity to the focal residue. It measures the local packing degree of residues within the protein tertiary structure. In helical membrane proteins, every transmembrane helix has a characteristic profile of WCNs and this profile is strongly coupled with native contacts between helices. This implies that WCNs can be incorporated as restraints in the prediction of helix-helix packing. In chapter III, it is demonstrated that residues' WCNs predicted

by the method developed in chapter II are effective restraints for improving the fraction of native contacts in predicted tertiary structure models of HMPs. Chapter IV concerns with the characterization of interfaces between HMPs and the prediction of quaternary structures of HMP complexes via protein-protein docking. First, the physicochemical characteristics and evolutionary conservation of interface residues are compared with residues on the rest of the surface, a machine learning-based method is then developed for predicting the WCNs of interface and surface residues. Finally, it is showed that predicted interface residues and their WCNs can be used to derive a powerful score for selecting native-like docking candidates of HMP complexes

Proteins mutate in response to change in environment or errors in gene replication. A lot of diseases are caused by dysfunctional variants of HMPs. Mapping the relationship between variants and their functional impact is an essential step toward precision medicine. Ideally, except for certain well-established disease-causing cases, variants should be evaluated by physiologically relevant experimental functional assays, but experimental characterization remains labor-intensive and costly to scale. Variant interpretation is bound to present an increasingly daunting challenge in the era of next-generation sequencing. Under such constraints, computational methods, which are usually machine learning-based, represent a common predictive approach.

Dysfunctional variants of the KCNQ1 potassium channel are associated with the congenital long QT syndrome. Chapter V describes a machine learning-based, protein-specific method developed in this work, that is capable of accurately classifying the functional impact of nonsynonymous variants of KCNQ1. This method was trained on a manually curated, functionally validated dataset to classify molecular functional impact. It showed superior performance when compared with eight previous methods tested in parallel.

Chapter VI concludes with a summary of the key contributions this work made to the relevant fields and some considerations on a few major limitations needed to be addressed in future work. It also points out some questions that are of significant interests for future work.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

I. INTRODUCTION

This chapter has been published under (Li *et al.*, 2018).

I-1 The protein folding problem

Protein folding is a process of molecular self-assembly during which a disordered polypeptide chain collapses to form a compact and well-defined three-dimensional (3D) tertiary structure. A grand challenge in biochemistry has been to understand the process by which proteins fold into their functional tertiary structure (folding mechanism) and to predict this tertiary structure from amino acid sequence (structure prediction), two tasks that are collectively known as "the protein folding problem" (Chan and Dill, 1993, Dill *et al.*, 2008, Dill and MacCallum, 2012). Solving this problem is of far-reaching impact as it will not only reveal the missing link between sequence and structure but also provide molecular biologists with a theoretical framework and practical tools for applications such as drug design and protein engineering. As a result, an enormous amount of effort has been contributed to study the protein folding problem by the scientific community. This is illustrated by Figure I-1, which shows the striking growth in the number of articles published each year on this problem since Anfinsen's "thermodynamic hypothesis" of protein folding, that protein native state resides in the global minimum of Gibbs free energy, was formally stated in 1973 (Anfinsen, 1973). A comprehensive review of the study of this problem is deemed impossible for an article of this kind. As many excellent review articles on the theories of protein folding and their experimental validation have been published over the years (Dill *et al.*, 1995, Onuchic *et al.*, 1997, Dobson *et al.*, 1998, Dobson and Karplus, 1999, Radford, 2000, Onuchic and Wolynes, 2004, Bartlett and Radford, 2009, Bowman *et al.*, 2011, Englander and Mayne, 2014, Wolynes, 2015), here we focus our discussion on computational methods for studying folding mechanisms and predicting tertiary structures. Specifically, we limit our discussion to protein folding simulations and *de novo* protein structure prediction at atomic detail, as methods based on coarse-grained representation of protein structures were recently comprehensively reviewed (Kmiecik *et al.*, 2016). In addition, due to space limitations, we are not able to cover the complete literature of this topic, and we apologize to those whose contributions have not received the deserved attention.

Nevertheless, the two key components of any folding simulation or structure prediction methods are efficient sampling of conformational space and accurate evaluation of the energy of

sampled conformations. Hence, the main body of this article is devoted to discussing different algorithms and their advances toward efficient sampling of conformational space followed by approaches and progress toward accurate energy functions. To put the discussion under the theoretical framework of protein folding, we first briefly summarize different views on mechanisms of protein folding. The interplay between sampling algorithms and energy functions is concretely illustrated by discussing some representative methods shown to be relatively successful in the Critical Assessment of protein Structure Prediction (CASP) experiment (Moult *et al.*, 1995, Tai *et al.*, 2014). Finally, we present a summary on the progress and outline specific challenges that future development in the field will likely overcome.

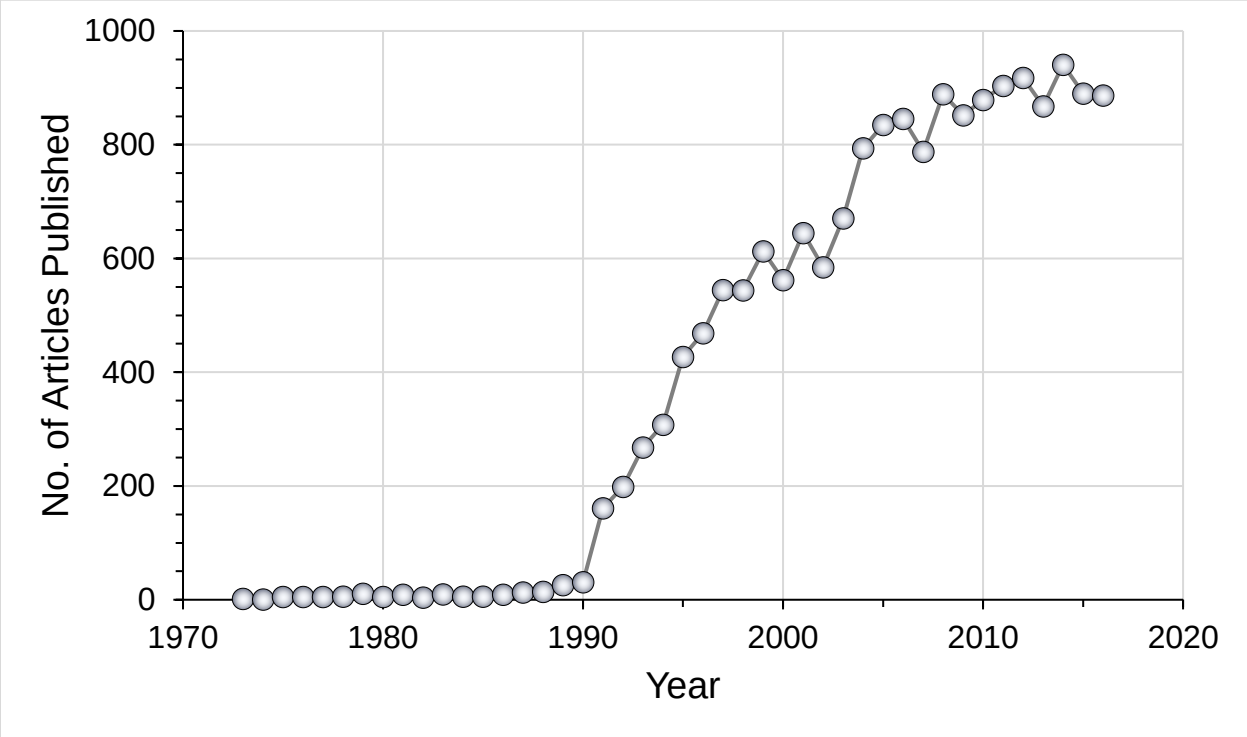


Figure I-1 Growth of the number of articles on the protein folding problem.

The number of articles published each year (1973-2016) with the phrase "protein structure prediction" or "protein folding" in either the title, or abstract, or author keywords. The data were taken from Web of Science.

I-1.1 Thermodynamics of protein folding

When a protein folds, it experiences constant counteractions between the effective energy, which favors the native state, and the configurational entropy, which favors unfolded states (Karplus, 2011). The term "effective energy" refers to the free energy of the system (protein plus solvent)

which consists of the intramolecular energy of the protein in vacuum plus the solvation free energy (the free energy of transfer of the protein from the gas phase to solution). The Gibbs free energy of the protein-solvent system is the sum of the effective energy and the configurational entropy (Lazaridis and Karplus, 1999, Lazaridis and Karplus, 2000, Lazaridis and Karplus, 2003) (Figure I-2). At equilibrium, both folded and unfolded states can be characterized by their Gibbs free energy. The difference in Gibbs free energies between the native state and unfolded states is termed the free energy of folding.

$$\Delta G_{folding} = \Delta H_{folding} - T\Delta S_{folding} + \Delta\Delta G_{solvation} \quad \textbf{I-1}$$

Both the enthalpic and entropic contributions to $\Delta G_{folding}$ mainly arise from intramolecular and protein-solvent non-bonded interactions and rearrangement of solvent molecules. As calculating the exact Gibbs free energy from first principles is prohibitive (Leach, 2001), a simplified energy function is used in practical computer simulations of protein dynamics, folding, and structure prediction. Broadly speaking, there are two different types of approaches to a simplified energy function. The first is a classical mechanical model that describes the potential energy, which is parameterized by analyzing the fundamental forces between particles; the second is a statistical model parameterized on data derived from statistical analysis of pair interactions and other properties in known protein structures (Lazaridis and Karplus, 2000). In this review, we will use the term "energy" frequently when we discuss various implementations of energy functions for evaluating the "energy" of sampled conformations, however, the reader is advised to keep in mind that such energy approximations are not physically realistic Gibbs free energies.

Solvation can be accounted for by either immersing the protein into explicit solvent molecules or including in the energy function a term that implicitly models solvation free energy. The former approach is often adopted in molecular dynamics simulations and is desirable especially in cases where the purpose is to study structural details about protein-solvent interaction. Two major limitations of this approach are that the computational expense is high, and the effective energy of a protein conformation is not known. The latter approach, often referred to as implicit solvation, is typically orders of magnitude faster and compatible with more sampling techniques than corresponding simulations with explicit solvent (Lazaridis and Karplus, 2003).

I-1.2 Simulation of protein folding, and tertiary structure prediction are very different subproblems

While both prediction of protein tertiary structure and simulation of folding require efficient search of conformational space and accurate evaluation of the energy of sampled conformations, it needs to be emphasized that these two subproblems are rather different, with distinct solutions and limitations. Methods for tertiary structure prediction generally create 3D models by assembling small structural fragments or motifs, quite often, with physically unrealistic trajectory of conformational search and evaluate the energy of sampled conformations using statistical potentials. While this approach has worked quite successfully in creating models that are close to native structures (Bradley *et al.*, 2005a, Zhang, 2009, Moult *et al.*, 2016), it has very little chance of giving insight into the mechanisms of folding. It is doubtless that if one could simulate actual folding processes, both subproblems would be solved. However, as will be explained in later sections, this is only possible for relatively small proteins using molecular dynamics simulations. Thus, methods for simulating folding mechanisms, while often employ physically realistic energy functions and can reveal important thermodynamics and kinetics about folding, are generally not useful for predicting structures for all but only small proteins.

I-1.3 Chemical kinetics of protein folding: mechanisms and pathways

The conformational space accessible to a polypeptide chain is astronomically large; a systematic search for the functional structure of a polypeptide chain with 100 residues would take an amount of time even longer than the age of the universe. The fact that proteins fold on a biologically meaningful timescale, with some attaining their functional structures in just a few microseconds, led Levinthal to conclude that there must be well-defined folding mechanisms and pathways to the native state (Levinthal, 1968, Levinthal, 1969), so that protein folding is under "kinetic control". A full characterization of the folding process requires elucidation of the mechanisms by which transition states and intermediates, if any, are formed and the determination of whether there is a single defined pathway or multiple pathways to the native state.

The "classical view" of protein folding assumes a sequential model and postulates a well-defined sequence of intermediates which follow one to carry the protein from the unfolded random coil to a uniquely folded native state (Levinthal, 1968, Kim and Baldwin, 1982, Kim and Baldwin, 1990). In the search for such a single mechanism of protein folding, several models have been

proposed about how folding gets started and native contacts and structure are subsequently formed (Baldwin, 1989, Fersht, 1997, Daggett and Fersht, 2003a, Daggett and Fersht, 2003b). The "nucleation" model postulated that a folding-initiating local secondary structure, or nucleus, is formed slowly followed by the rapid propagation of native structure in a stepwise manner (Wetlaufer, 1973). However, this model was dropped from favor as it predicts the absence of folding intermediates. The "framework" model and the related "diffusion-collision" model proposed that secondary structures segments are preformed independently of tertiary structure before they diffuse and collide to give stable tertiary structure (Kim and Baldwin, 1990, Karplus and Weaver, 1994). The "hydrophobic collapse" model hypothesized that folding starts with a rapid collapse around hydrophobic residues to the molten globule state (compact denatured state), which narrows down the conformational exploration to the native state significantly (Baldwin, 1989, Ptitsyn, 1996). An essential feature of these latter models is that they predict the presence of folding intermediates. However, the fact that some proteins fold by simple two-state kinetics, without the accumulation of folding intermediates, and that secondary and tertiary structure form simultaneously led to the formulation of the "nucleation-condensation" model (Fersht, 1997, Daggett and Fersht, 2003a, Daggett and Fersht, 2003b). This model assumed the concerted formation of local and nonlocal structures and was considered a "unifying" mechanism of protein folding (Daggett and Fersht, 2003b). It should be notes, however, that the "nucleation-condensation" model does not preclude the presence of folding intermediates (Daggett and Fersht, 2003a, Daggett and Fersht, 2003b).

The observation that molten globules form asynchronously over a range of timescales fostered the concept of protein folding funnel (Frauenfelder *et al.*, 1991, Bryngelson *et al.*, 1995, Dill and Chan, 1997, Onuchic *et al.*, 1997, Dobson *et al.*, 1998, Brooks *et al.*, 2001, Wolynes, 2015) (Figure 2). In this "new view", it is inferred that proteins must fold into their unique native state through multiple unpredictable pathways that involve the progressive organization of an ensemble of partially folded intermediates on a rugged effective energy hypersurface that resembles a funnel. The funnel shape arises from the fact that the number of accessible configurations, which determine the configurational entropy, decreases as the energy decreases (Karplus, 2011) . A more recent formulation of the mechanism of protein folding is centered around of concept of foldons (Englander *et al.*, 2007, Englander and Mayne, 2014). In what's called the foldon-based hypothesis, a protein starts folding by forming an initial seed foldon through unguided search, and

it follows a foldon-determined folding pathway as the seed foldon guides subsequent foldons in a "folding upon binding" way. While this hypothesis states that proteins fold along a definite path after formation of the initial foldon, the foldon formation at the initial stage is assumed to be accomplished through a disordered multitrack search (Englander and Mayne, 2014).

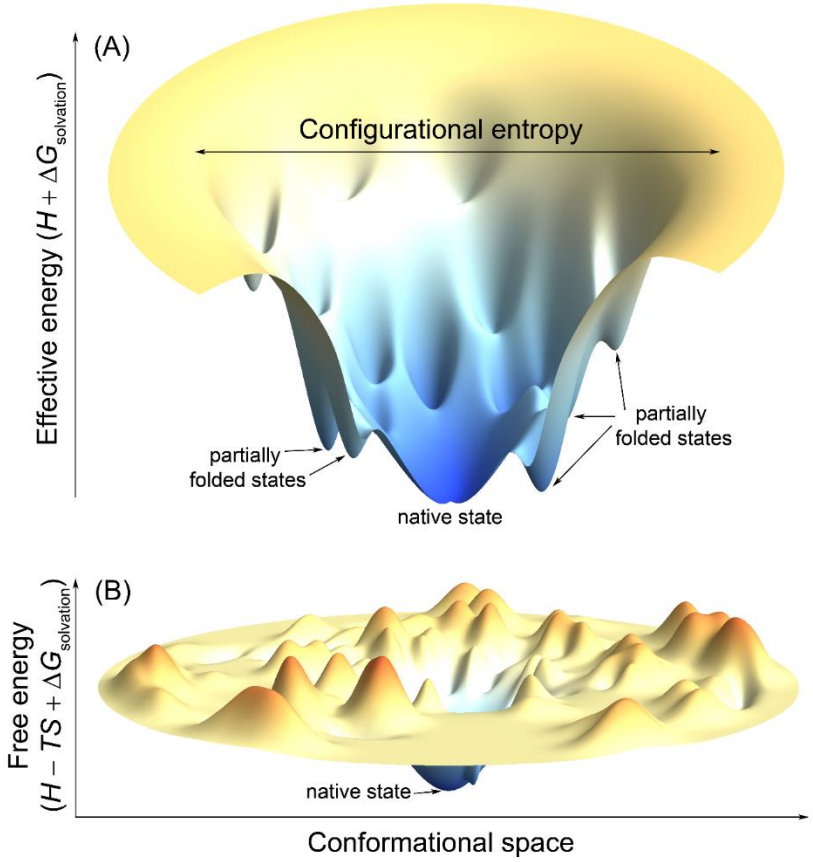


Figure I-2 Schematic three-dimensional surface rendering of a hypothetical folding funnel diagram and a (Gibbs) free energy landscape to reference state.

(A) A folding funnel diagram is a pictorial representation of the counteracting nature of the two thermodynamic variables, effective energy and configurational entropy, in protein folding and explains how the Levinthal paradox is resolved (Karplus, 2011). The effective energy is plotted vertically and the configurational entropy horizontally. The funneled shape stems from the fact that the number of accessible configurations, which determines the configurational entropy, decreases as the native state of a protein is approached (Karplus, 2011). (B) A free energy landscape maps between conformations and free energies. The global minimum on the landscape corresponds to the conformation of the native state and local minima correspond to partially unfolded states, which are separated by free energy barriers from the native state. Note that real free energy landscapes are high-dimensional and extremely rugged.

I-1.4 Prediction of protein folding rates

Folding rate is an essential parameter for characterizing protein folding kinetics. What factors determine whether a protein will be a slow or fast folder? Can we predict folding rates from amino acid sequences as well? Theoretical studies have suggested that the size, native topology, and stability of a protein influence the rate and mechanisms by which it folds. In searching for a causal relationship, a key advance made in 1998 was that in a set of 12 non-homologous single domain proteins folding rate shows a significant correlation with a simple measure of topological complexity of the native fold, the so-called contact order, which is defined as the average sequence separation between all pairs of native contacts normalized by sequence length (Plaxco *et al.*, 1998). In contrast, the correlations between the size or native state stability and folding rate are weak to non-existent (Plaxco *et al.*, 1998). Based on this observation, another parameter called long-range order, which counts long-range contacts (contacts that are close in space but distant in sequence), was proposed and found to be a strong predictor of the folding rates of two-state proteins (Gromiha and Selvaraj, 2001). Contact order and long-range order have also been combined to form a parameter called total contact distance that has better correlation with folding rates (Zhou and Zhou, 2002b). Folding rates were also found to be inversely correlated with a parameter called multiple contact index which measures the number of residues with multiple long-range contacts (Gromiha, 2009). The correlations between the various topological parameters just discussed and folding rates suggest that it is viable to predict folding rates from amino acid sequences because native topologies are determined by amino acid sequences (Baker, 2000). In fact, several bioinformatics tools have been developed for this purpose (Ivankov and Finkelstein, 2004, Gromiha *et al.*, 2006, Ouyang and Liang, 2008, Chou and Shen, 2009, Guo and Rao, 2011), and two notable web servers are FOLD-RATE (Gromiha *et al.*, 2006), FoldRate (Chou and Shen, 2009).

I-2 Conformational sampling is a bottleneck

A polypeptide chain with a typical size can adopt an astronomical number of conformations. It is agreed that conformational sampling remains to be a bottleneck of *de novo* structure prediction (Jones, 1997a, Baker and Sali, 2001, Bradley *et al.*, 2005b, Zhang, 2008, Kim *et al.*, 2009, Maximova *et al.*, 2016). Nevertheless, there has been exciting improvement in sampling algorithms based on statistical mechanical principles or guided by experimental or predicted

restraints, all of which are further accelerated by improvements in hardware speed and power (Maximova *et al.*, 2016). For the convenience of discussion, we divide conformational search methods into the following three broad categories: molecular dynamics simulations, Monte Carlo simulations, and genetic algorithms. For each category of algorithms, we give a general formulation of the algorithm and a summary of the latest studies in which the algorithm was applied to study protein folding mechanism or *de novo* protein structure prediction.

I-2.1 Unbiased molecular dynamics simulations

Molecular dynamics (MD) simulation is a widely used computational technique for exploring the macroscopic properties of molecular systems through explicit computation of microscopic particle motions. MD has had enormously influential applications in biomolecular systems and has been heavily used to study motion-related phenomena such as protein folding, conformational flexibility, protein structure determination from NMR, ligand-protein interaction, and protein-membrane interaction (Karplus and Petsko, 1990, van Gunsteren and Berendsen, 1990, Karplus and McCammon, 2002, Gumbart *et al.*, 2005, Karplus and Kuriyan, 2005, Lindahl and Sansom, 2008, Klepeis *et al.*, 2009, Durrant and McCammon, 2011, Periole, 2017). The two essential elements of a MD simulation are the interaction potential for the particles and the equations of motion governing the dynamics of the particles (Leach, 2001, Rapaport, 2004). Interaction potentials will be discussed in the section: Energy functions are evolving objects. Here, we describe how MD simulations explore the phase space of a molecular system.

A typical MD run involves generation of successive microstates of a molecular system by solving Newton's equations of motion for all atoms simultaneously with femtosecond timesteps (Eq. I-2).

$$m_i \frac{d^2 \boldsymbol{r}_i}{dt^2} = \boldsymbol{F}_i = -\frac{\partial U(\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_N)}{\partial \boldsymbol{r}_i} \qquad \textbf{I-2}$$

where $\boldsymbol{r}_i$ and $U(\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_N)$ denote position vector and potential energy of point mass i, respectively. $\boldsymbol{F}_i$ denotes the force acted upon point mass i. The result of the simulation is a trajectory of microstates that specify how the system evolves in phase space (Leach, 2001). In principle, equilibrium properties can be computed by averaging over the trajectory if it is of sufficient length to give a representative ensemble of the microstates of the system. Unfortunately,

the usefulness of MD in studying long timescale biological phenomena is often limited due to inadequate sampling of all relevant conformational states of a system. Even when the energy barriers separating two topologically different low energy regions of the conformational space are of order k_BT, traversing them by random thermal fluctuation cannot be achieved within a reasonable amount of time.

A wide range of biologically interesting phenomena occurs over timescales on the order of milliseconds, several orders of magnitude beyond the reach of conventional MD simulations. As a result, studying processes that involve major conformational changes, such as protein folding, activation, and deactivation, by MD simulations has been traditionally challenging (Gruebele, 2002). The very first protein folding simulation via MD at the microsecond timescale was notably made by Duan and Kollman (1998), who simulated the folding process of the villin headpiece (a 36-mer) in explicit solvent for two months on parallel supercomputers. The simulation showed a mechanism for the peptide to reach a marginally stable state with a main chain RMSD of 5.7 Å from the native state (Duan and Kollman, 1998). This peptide was later *de novo* folded by Zagrovic and coworkers (2002) to an ensemble of states whose average Cα RMSD is 1.7 Å from the native state. The total simulation time was 300 μs or approximately 1000 CPU years with the help of worldwide-distributed computers (Zagrovic *et al.*, 2002).

Substantial progress has been made during the past decade or so to extend the folding times accessible by conventional MD simulations through efficient parallelization of MD codes or MD-specialized hardware (Lane *et al.*, 2013) (Figure I-3). The MD-specialized software package Desmond and the massively parallelized machine Anton, both developed recently at D.E. Shaw Research, have allowed for conducting millisecond timescale MD simulations of systems with tens of thousands of atoms in just a few weeks (Bowers *et al.*, 2006, Shaw *et al.*, 2007, Shaw *et al.*, 2014). Desmond is a collection of codes that implement novel parallel algorithms and numerical techniques to perform high-throughput and accurate MD simulations on conventional computational clusters, general-purpose supercomputers, and GPUs (Bowers *et al.*, 2006). Anton is built on MD-specific ASICs (application-specific integrated circuits) that interact in a tightly coupled manner using a high-speed communication network. Its ability to efficiently perform simulations on the timescales over which many physiologically relevant processes take place expands the set of problems for which the use of MD is tractable (Shaw *et al.*, 2007, Shaw *et al.*,

2014). Armed with this specialized set of software and hardware, researchers at D.E. Shaw Research have been able to simulate protein folding from extended random coils (Shaw *et al.*, 2010, Lindorff-Larsen *et al.*, 2011) and study structural origin of slow diffusion in protein folding (Chung *et al.*, 2015), protein-ligand recognition (Dror *et al.*, 2011, Shan *et al.*, 2011), mechanism of nucleotide exchange in G proteins (Dror *et al.*, 2015), and mechanisms of kinase activation and inhibition (Shan *et al.*, 2014, Ingram *et al.*, 2015) at realistic timescales. The *de novo* folding simulations conducted at D.E. Shaw Research generated computational insights in favor of the single-pathway view of protein folding (Figure 2). For example, equilibrium simulations of WW domain captured multiple folding and unfolding events that consistently follow a well-defined folding pathway (Shaw *et al.*, 2010). However, subsequent folding simulations of 12 fast-folding proteins showed that although a majority of them fold along a single dominant route, differing "transition state classes" were observed for two proteins (Lindorff-Larsen *et al.*, 2011).

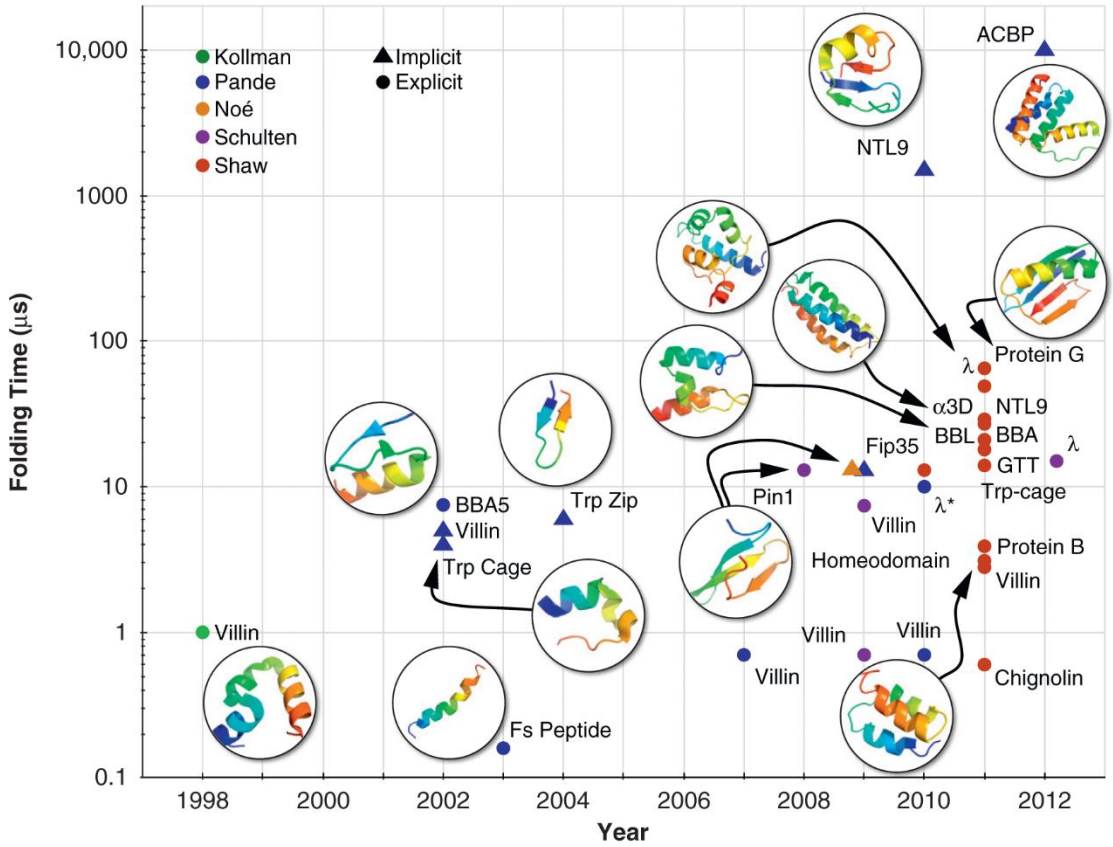


Figure I-3 Folding time scales accessible to MD simulations

Folding time scales accessible to MD simulations have increased exponentially since Duan and Kollman used MD simulations in explicit solvent to study the process through which the villin headpiece reaches a marginally state (Duan and Kollman, 1998). Shown are proteins simulated using unbiased, all-atom MD simulations in empirical force fields reported in the literature. Here, an accessible folding time scale is defined as one within which folding events are observed in MD simulations of folding from unfolded states. According to this definition, whether the ~10 ms folding time of ACBP is already accessible needs to be confirmed by further simulations as no folding events were observed in any of the trajectories used to construct a Markov state model of the ACBP folding reaction (Voelz *et al.*, 2012). Adapted, with permission, from reference (Lane *et al.*, 2013). See reference (Lane *et al.*, 2013) for reference to each folding simulation highlighted in the figure.

A different approach to overcome the sampling challenge of MD is through statistical analysis of multiple independent trajectories or aggregating independent short simulations using Markov state models (MSM) to make a complete model of system dynamics (Pande *et al.*, 2010, Prinz *et al.*, 2011, Lane *et al.*, 2013). The MSM effectively pieces together this complete model from independent trajectories, allowing for prediction of kinetic phenomena on timescales much longer than the individual trajectories used to construct the model (Lane *et al.*, 2013). While the MSM-based "multi-trajectory" approach has some advantages over the reaction coordinate-based single trajectory analysis, such as identifying areas of phase space for adaptive sampling (Bowman *et al.*, 2010, Weber and Pande, 2011), insights gained from MSM analysis does not always agree with with the single pathway view of folding. For example, while it was shown via single-trajectory analysis that folding of the WW domain follows a definite pathway where the first hairpin folds first (Shaw *et al.*, 2010), a parallel statistically significant pathway where the second hairpin of the WW domain folds first was detected using MSM to analyze the same simulation trajectories (Lane *et al.*, 2011). Similar analysis conducted on the MD trajectories of 12 small fast-folding proteins (Beauchamp *et al.*, 2012), while showed that two-state model is inadequate for the same set of systems as described by a previous study (Lindorff-Larsen *et al.*, 2011), revealed a richer picture of populated states for some more complicated systems.

I-2.2 Enhanced sampling techniques in MD

The ruggedness of energy landscapes with many local minima separated by high-energy barriers makes adequate conformational sampling a challenging task. MD trajectories often do not reach all biologically relevant conformations, a problem that can be addressed by employing enhanced sampling algorithms (Okamoto, 2004, Bernardi *et al.*, 2015). Two popular enhanced sampling techniques in simulations of biological systems are replica exchange molecular dynamics (REMD)

and metadynamics (Bernardi *et al.*, 2015). While we focus our discussion on the REMD along temperature, which is also known as parallel tempering, several variants of replica exchange protocols have also been reported (Fukunishi *et al.*, 2002, Itoh *et al.*, 2011, Wu *et al.*, 2012).

The replica-exchange method was developed to overcome the multitude of local minima separated by high energy barriers (Sugita and Okamoto, 1999). Many molecular simulation scenarios require ergodic sampling of energy landscapes that feature many minima, and barriers between minima can be difficult to overcome at ambient temperatures over accessible simulation timescales. Replica-exchange simulations seek to enhance the sampling in such scenarios by running n non-interacting copies of the system C_i $(i = 1, \dots, n)$ in parallel each at a different temperature T_i in the canonical ensemble (Figure I-4). The non-interacting nature of this artificial compound system $(C_1, C_2, \dots, C_n)$ ensures that each state's weight factor is given by the product of Boltzmann factors of each copy.

$$w = \exp\left\{-\sum_{i=1}^{n} \beta_i U_i\right\} \qquad \textbf{I-3}$$

Compared to a standard Monte Carlo simulation, which affects the conformation of only one copy, REMD explores the energy landscape by periodically exchanging the conformations of replicas. The probability of transition of a compound system such that the conformations between a pair of copies (C_i, C_j) are exchanged is

$$p = \min(1, e^{\Delta}) \qquad \textbf{I-4}$$

where

$$\Delta = \left(\beta_j - \beta_i\right)\left(U_j - U_i\right) \qquad \textbf{I-5}$$

In most cases, exchange of the conformations of replicas decreases auto-correlation, thus enabling replicas to reach thermal equilibrium faster than without exchange. However, for protein folding simulation, a recent study showed that the efficiency of REMD is not much higher than that of conventional MD if the folding rate is not very temperature-dependent (Rosta and Hummer, 2009). While it is not necessary to restrict the exchange to copies with neighboring temperature (e.g. $j = i + 1$), doing so will be optimal, since the transition probability decreases exponentially with the

difference in temperature between copies (Hansmann, 1997). It is also worth noting that while exchange of conformations between copies must be conducted in a Monte Carlo way, there is no restriction on which algorithms are used for updating the conformation of an individual copy locally. In fact, several variants of REMD have been developed (Mori *et al.*, 2016). For example, a replica-exchange Monte Carlo (REMC) technique was implemented in the threading-based structure prediction pipeline QUARK and tested in CASP11 (Zhang *et al.*, 2016).

Metadynamics is a class of methods that eases sampling by introducing a time-dependent biasing potential that acts on a selected number of coarse-grained order parameters, often referred to as collective variables (CVs) (Laio and Parrinello, 2002, Piana and Laio, 2007, Barducci *et al.*, 2011, Valsson *et al.*, 2016). CVs are generally nonlinear functions of the atomic positions of the simulated system that should ideally distinguish between all relevant metastable states. Some simple but informative CVs used in protein folding simulations are number of Cα contacts, number of backbone H-bonds, and helicity of the backbone, and the free energy surface is usually plotted as a function of these CVs (Piana and Laio, 2007). The added biasing potential is introduced through successive addition of small repulsive Gaussian kernels deposited along the system trajectory in CV space (Figure I-4) (Barducci *et al.*, 2011, Valsson *et al.*, 2016). The added Gaussian kernel is a function of the current position and the previous position of the system in the CV space, and its intended purpose is to discourage the system from revisiting configurations that have already been sampled, thus accelerating sampling. The final summation of the deposited Gaussian kernels also gives an unbiased estimate of the free energy landscape of the system. In contrast to these advantages, it is, however, far from trivial to decide when to stop a simulation and find a set of CVs proper for describing the process of interest (Barducci *et al.*, 2011, Valsson *et al.*, 2016).

Both REMD and metadynamics have been used to *de novo* fold several small peptides and proteins. The first example of using REMD to sample a folded structure starting from a completely unfolded state is probably the study of Rhee *et al.* (Rhee and Pande, 2003) where a 23-residue BBA5 protein was folded by what's called multiplexed REMD. Using REMD simulations in implicit solvent, Pitera *et al.* (Pitera and Swope, 2003) folded a 20-residue designed Trp-cage peptide starting from an extended coil to a state < 1.0 Å Cα RMSD from conformations in the NMR ensemble. Recently, Jiang *et al.* (Jiang and Wu, 2014) folded a diverse set of 14 fast folding

proteins from their unfolded states using REMD with a residue-specific force field. A similar study by Nguyen *et al.* (Nguyen *et al.*, 2014) included a larger set of 17 proteins; while they successfully folded most proteins, misfolded structures are thermodynamically preferred for 3 proteins.

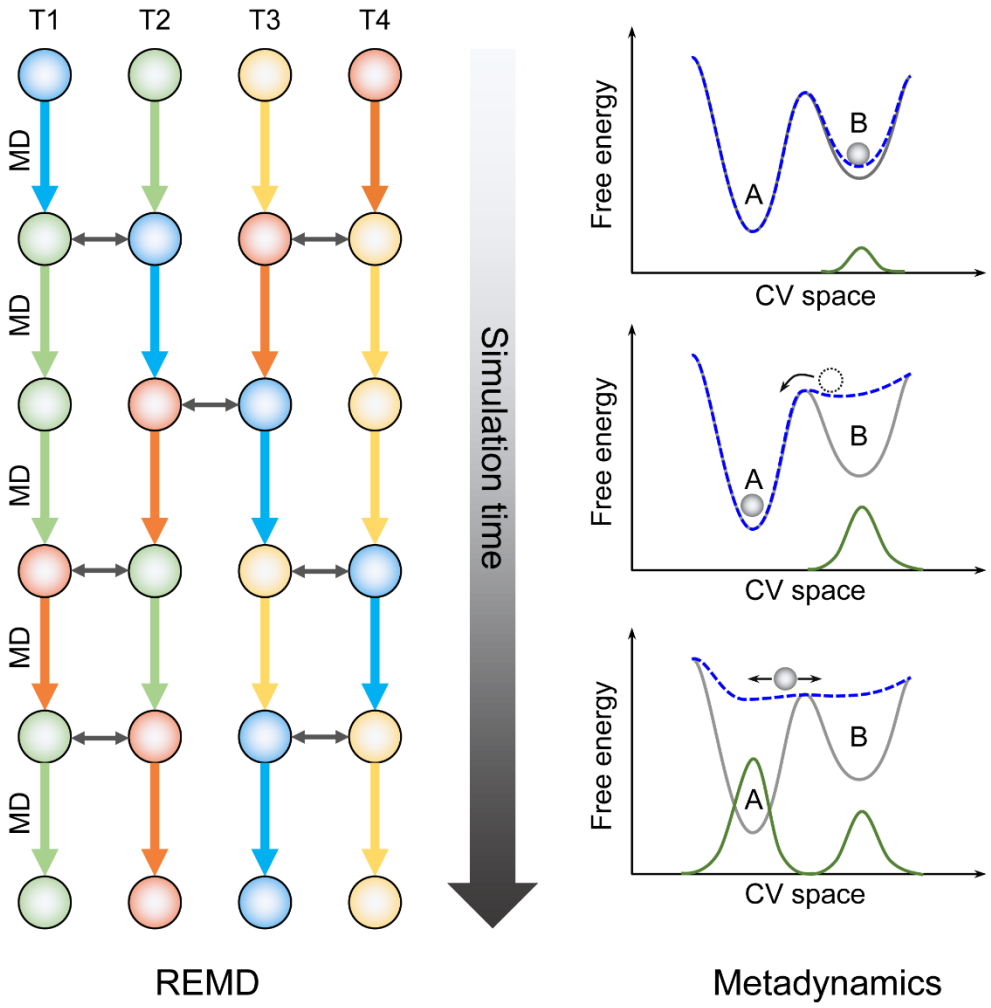


Figure I-4 A sketch of the process of REMD and that of metadynamics

REMD: a set of non-interacting replicas (T1 though T4 in this illustration), each runs at a different temperature. Each color represents a single replica. As the simulation proceeds, each replica walks up and down in temperature. In an efficient REMD, replicas at neighboring temperatures are swapped (shown as double-headed arrows) based on Metropolis criterion and all replicas will experience swapping. Metadynamics: this illustrative system has two minima A and B (gray curve). The system trapped in B is lifted by progressive deposition of repulsive Gaussian kernels (green curve) and the free energy landscape changes accordingly (blue dashed curve). After B is filled up, the system moves into A which is filled up similarly. When the simulation completes, the green curve gives a first rough negative estimate of the free energy landscape of the system.

I-2.3 Monte Carlo simulation

MD simulation is without a doubt a required technique if one wishes to study folding pathway or kinetics computationally. However, for tertiary structure prediction of large proteins whose energy landscapes are populated with many local minima separated by high barriers, Monte Carlo (MC) simulation can be much more efficient (Figure 5 (A)). It is, in fact, the underlying search engine of some of the most successful *de novo* tertiary structure prediction methods (Simons *et al.*, 1997, Bradley *et al.*, 2005a, Xu and Zhang, 2012, Zhang *et al.*, 2016) and our method BCL::Fold (Karakas *et al.*, 2012). Unlike MD simulations where successive conformations of the system are connected through time, in a MC simulation, each new conformation of the system depends only upon its immediate predecessor. The technique of MC simulation was introduced as the first computer simulation of a molecular system in 1952 (Metropolis *et al.*, 1953). Nowadays, the term "Monte Carlo" is often used to describe a simulation whenever random sampling is performed.

A MC simulation explores the phase space of a system by randomly perturbing the current conformation by actions such as moving a single atom or molecule or adjusting dihedral angles. The energy of the new conformation is then evaluated using an energy function. If the new conformation is lower in energy than its predecessor, it is accepted as a starting conformation for the next iteration. If the energy is higher, the new conformation is accepted with a probability based on the famous Metropolis criterion (Metropolis *et al.*, 1953) (Eq. I-4). This is often done by comparing the Boltzmann factor of the new conformation to a random number between 0 and 1, and the new conformation is accepted if its Boltzmann factor is greater than the random number and rejected otherwise. While the essential search algorithm of MC-based structure prediction methods is the same, they differ in the starting components for assembling 3D models and in the repertoire of MC moves implemented for perturbing the model (Vitalis and Pappu, 2009).

Primitive MC sampling can be computationally expensive and thus inefficient at finding global energy minimum. Typically, these methods are coupled with some optimization technique that vastly decreases computational expense by directing the progression of the MC simulation toward global energy minimum. One optimization technique is gradient-based sampling, where MC iterations are directed down local property gradients, i.e. the potential next state with the lowest energy is selected. For instance, gradients can be calculated based on side chain rotameric states (Xiangian Hu, 2010) or, in the HP-lattice model (Dill *et al.*, 1995), the movement of a residue

in various directions (Hu *et al.*, 2009). However, when the conformational space is continuous rather than discrete, gradient descent becomes unfeasible because the energy cannot be calculated for every step forward. The most popular optimization approach shown to effectively accelerate the convergence of a MC simulation is probably simulated annealing (Kirkpatrick *et al.*, 1983). The essential feature of this technique is that it combines MC sampling of conformational space at an initially elevated temperature with a proper cooling scheme over the course of the simulation. The cooling scheme, if gentle enough, theoretically ensures the system will reach the global minimum. In turn, the probability of a higher energy step being accepted decreases over time, and models are directed toward the global energy minimum (Tsallis and Stariolo, 1996). Many powerful *de novo* tertiary structure prediction methods integrate this MC simulated annealing approach (Kmiecik *et al.*, 2016); we include a detailed discussion on some selected examples of such methods (see Examples of methods for *de novo* tertiary structure prediction).

I-2.4 Genetic algorithms

Genetic algorithms (GAs) are an optimization procedure based on the process of evolution that occurs in nature. GAs have been used in a variety of applications. Some prominent ones include automatic programming, machine learning, and population genetics (Goldberg, 1989). Generally, a GA initializes the optimization process by randomly generating an initial population of trial solutions each encoded as a string of bits, also called a chromosome (Figure 5(B)). Offspring are produced by applying nature-inspired operations, namely mutations and crossovers on bit strings. Mutations are introduced into strings by flipping one or more bits, whereas crossovers between two individuals consist of randomly selecting a crossover site and exchanging the left segment of one string with the right segment of the other (Figure 5(B)). The fittest offspring are selected for continual refinement via the iteration of multiple generations (Schulze-Kremer, 2000).

A large number of studies on the use of GAs for *de novo* protein structure prediction and protein folding simulation have been made (Pedersen and Moult, 1996, Cui *et al.*, 1998, Schulze-Kremer, 2000, Custodio *et al.*, 2004, Unger, 2004, Hoque *et al.*, 2009, Huang *et al.*, 2010, Zhang *et al.*, 2010, Custodio *et al.*, 2014, Bošković and Brest, 2016, Rashid *et al.*, 2016) since the pioneering work of Dendekar and Argos (Dandekar and Argos, 1992) on *de novo* folding simulation of a model protein of a four β-strand bundle and that of Unger and Moult (Unger and Moult, 1993) on searching for global energy minimum on the 2D HP lattice model. The simplest

protein representations used in GAs is the 2D HP model developed by Lau and Dill (Lau and Dill, 1989). In this model, amino acids are of only two types: hydrophobic (H) or polar (P). The sequence is folded on a 2D square lattice on which bonds are orthogonal to each other. Folded structures are evaluated by a so-called "hydrophobic potential" where each pair of non-bonded direct hydrophobic contact (occupying neighboring non-diagonal lattice vertices) receives -1. Using HP lattice models avoids the computational cost needed for all-atom models while still capturing the general principles that govern protein folding, and they can be extended to account for physicochemical characteristics of individual residues such as size, hydrophobicity, and charge. In more detailed models, proteins can be represented as a sequence of pairs of dihedral angles that describe the backbone degrees of freedom of each residue. Mutations can be introduced simply by changing the dihedral angle of a residue and crossovers by swapping randomly assigned sections of two sequences (Schulze-Kremer, 2000, Unger, 2004).

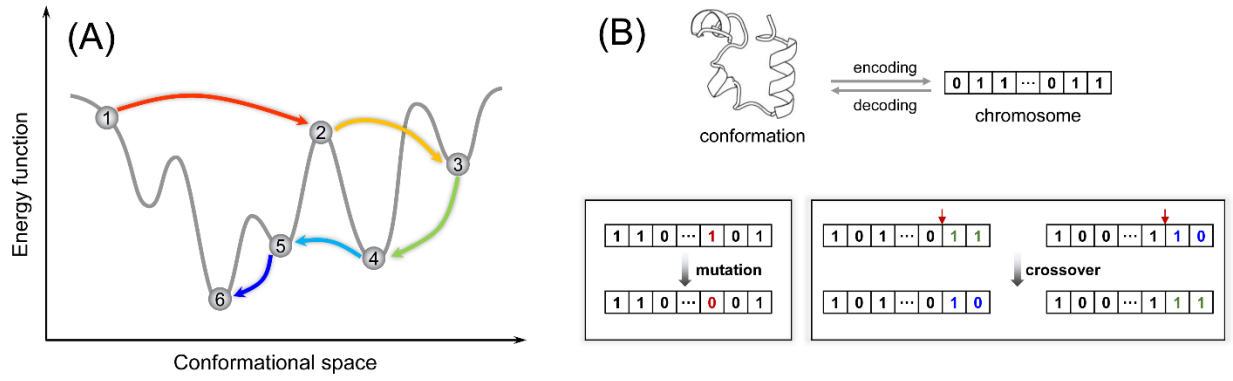


Figure I-5 Monte Carlo simulated annealing and genetic operations in genetic algorithms

(A) A Monte Carlo simulated annealing procedure allows the system to "freely" navigate on the free energy surface. For example, transition from state 4 to 5 would be prohibitive to MD simulations due to the high-energy barrier separating them. (B) In genetic algorithms, conformations are encoded as bit strings (or real-valued arrays) called chromosomes. A mutation operation flips the bit value at a randomly selected site, whereas a crossover operation takes a pair of chromosomes and exchanges parts of chromosomes split at a randomly selected crossover site.

I-3 Energy functions are evolving objects

An essential part of almost all successful protein folding simulations or protein tertiary structure predictions is an energy function that is a good approximation to the energy landscape of real proteins. Energy functions can be roughly divided into two classes: physics-based force fields and

knowledge-based potentials (Lazaridis and Karplus, 2000). Historically, physics-based forces fields are coupled with MD or MC simulations to study protein dynamics or calculate free energies (Wang *et al.*, 2001, Ponder and Case, 2003, Mackerell, 2004, Lopes *et al.*, 2015), whereas knowledge-based potentials are mostly used for fold recognition or tertiary structure prediction (Sippl, 1995, Godzik, 1996, Skolnick, 2006). Before we give a detailed account on them, we remind the reader that both of these two types of energy functions are evolving objects. To improve accuracy, further parameter optimization for physics-based force fields is required and statistics need to be rederived for knowledge-based potentials when energy function deficiencies are identified or data sets of better qualities become available.

I-3.1 Physics-based force fields

Physics-based force fields are classical mechanical models that approximate the potential energy of chemical systems. Force field models ignore the electronic motions in a system and only consider interactions among nuclei. Compared to *ab initio* quantum mechanical methods, force fields are much more computationally efficient while giving an acceptable level of accuracy. A force field has a functional form and a (usually very large) set of associated parameters that, taken together, model bonded and non-bonded interactions in a system. The functional form of a force field is often a compromise between accuracy and computational efficiency and depends on the level of resolution (all-atom or coarse-grained), chemical nature (inorganic, small organic, or biomolecular), and target properties of the systems to be modeled. Nevertheless, most force fields have five components (Eq. I-6). The first three of them, so-called bond stretching, angle bending, and torsion, model bonded interactions. The last two components describe electrostatic and van der Waals non-bonded interactions (Leach, 2001).

$$U(\boldsymbol{r}^N) = \sum_{\text{bonds}} \frac{k_b}{2}(l - l_0)^2 + \sum_{\text{angles}} \frac{k_\theta}{2}(\theta - \theta_0)^2 + \sum_{\text{torsions}} \frac{k_\varphi}{2}[1 + \cos(n\varphi - \gamma)] + \sum_{\text{electrostatics}} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \sum_{\text{VDW}} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \quad \textbf{I-6}$$

The form of this energy function may look simple, but we must keep in mind that the set of parameters associated with it is very large. For example, the term that models bond stretching (a harmonic potential) has a different force constant k_b and an equilibrium bond length l_0 for each bond type. These parameters must be determined by fitting the force field to a given set of data obtained from experiments or quantum mechanical calculations. Depending on the size of the data set, parameter optimization may be conducted in a number of ways: trial and error, least-squares fitting (Lifson and Warshel, 1968), or, recently, machine-learning algorithms (Behler, 2016).

Well-known examples of force fields intended for modeling proteins include CHARMM (Gelin and Karplus, 1979, Brooks *et al.*, 1983, MacKerell *et al.*, 1998, Mackerell *et al.*, 2004, Brooks *et al.*, 2009, Best *et al.*, 2012), AMBER (Weiner and Kollman, 1981, Weiner *et al.*, 1984, Li and Bruschweiler, 2010, Lindorff-Larsen *et al.*, 2010), OPLS (Jorgensen and Tirado-Rives, 1988, Robertson *et al.*, 2015), GROMOS (Van Gunsteren and Berendsen, 1987, van Gunsteren *et al.*, 1998), MARTINI (Marrink *et al.*, 2007, Monticelli *et al.*, 2008). These force fields were previously compared in-depth (Ponder and Case, 2003), we note here that while the functional forms of these force fields invariably contain the five terms of Eq. I-6, some of them or their different versions may differ in specifics in the treatment of non-bonded interactions and the levels of resolution covered. For example, although more recent versions of the CHARMM and AMBER force fields do not model hydrogen-bonding energetics explicitly, originally CHARMM and AMBER force fields both incorporated a 12-10 Lennard-Jones potential to model hydrogen-bonding (Gelin and Karplus, 1979, Weiner *et al.*, 1984). The need for more efficient evaluation of non-bonded interactions arises when the number of interaction sites is large. One straightforward way to improve efficiency is to absorb aliphatic hydrogens into the carbon atom to which they are bonded to form 'united atoms' as was done in the united-atom version of the CHARMM and OPLS force fields, or to use a coarse-graining approach where a group of heavy atoms are combined to form a representative virtual interaction site. The MARTINI force field aims at providing a simple model that is computationally fast and easy to use, and it adopted a 'four-to-one' coarse-graining scheme, meaning that on average four heavy atoms are represented by one interaction site (Marrink *et al.*, 2007, Monticelli *et al.*, 2008, Marrink and Tieleman, 2013). Although all-atom simulations are often more desirable, if special care is taken during calibration of the building blocks and parameterization, a level of accuracy comparable to all-atom simulations may be possible in reproducing some thermodynamic properties with reduced representations while achieving

considerable computational savings (Baron *et al.*, 2006a, Baron *et al.*, 2006b, Baron *et al.*, 2007, Marrink *et al.*, 2007, Monticelli *et al.*, 2008, Marrink and Tieleman, 2013). Coarse-grained protein models and their applications was recently reviewed in detail (Kmiecik *et al.*, 2016).

Physics-based force fields are traditionally coupled with MD in simulating protein dynamics and folding (McCammon *et al.*, 1977). There have been a plethora of such studies where the utility of force fields for protein tertiary structure prediction or the accuracy of reproducing experimental data were reported (Duan and Kollman, 1998, Zagrovic *et al.*, 2002, Pande *et al.*, 2003, Summa and Levitt, 2007, Lindorff-Larsen *et al.*, 2011, Patapati and Glykos, 2011, Lindorff-Larsen *et al.*, 2012, Huang and MacKerell, 2013, Piana *et al.*, 2013). However, no agreement has been reached regarding whether force fields are sufficiently robust for these applications (Lee *et al.*, 2009, Piana *et al.*, 2014). Early analysis concluded that MD simulations under physics-based force fields are not particularly successful in structure prediction (Lee *et al.*, 2009). However, for small, fast folding proteins that are also very stable, evidence has been accumulating that demonstrates that physics-based force fields are sufficiently accurate for predicting native-state structures and folding rates (Shaw *et al.*, 2010, Lindorff-Larsen *et al.*, 2011, Piana *et al.*, 2012, Piana *et al.*, 2013, Piana *et al.*, 2014, Chung *et al.*, 2015). In particular, it was pointed out the prediction of tertiary structures, folding rates, and melting temperatures appears to be more robust than the prediction of the enthalpy and heat capacity of folding or that of the radii of gyration of unfolded states (Piana *et al.*, 2014). It needs to be pointed out, however, that whether these force fields hold accurate for simulating larger proteins remains to be studied.

I-3.2 Knowledge-based potentials

Unlike physics-based force fields, which model interactions found in the most basic molecular systems using fundamental laws of physics explicitly and separately, knowledge-based potentials (KBPs) are energy functions derived from statistical analyses of known protein structures and the application of the inverse Boltzmann relation to the probability distribution of geometries (Wodak, Sippl, 1993, Sippl, 1995). The physical meaning of KBPs has been under vigorous debate since their introduction (Finkelstein *et al.*, 1995, Thomas and Dill, 1996, Ben-Naim, 1997, Moult, 1997, Shortle, 2003, Hamelryck *et al.*, 2010), although justifications of KBPs as "potentials of mean force" have been provided by analogy to the reversible work theorem in statistical thermodynamics (Sippl *et al.*, 1996) or on the basis of probabilistic arguments (Simons *et al.*, 1997, Hamelryck *et*

al., 2010). Nevertheless, KBPs are widely used and surprisingly effective in scenarios including but not limited to protein structure prediction (Simons *et al.*, 1997, Lu and Skolnick, 2001, Shen and Sali, 2006, Xu and Zhang, 2012), refinement of NMR structures (Kuszewski *et al.*, 1996, Yang *et al.*, 2012), fold recognition (Kocher *et al.*, 1994, Majek and Elber, 2009), protein-ligand or protein-protein interactions (Gohlke *et al.*, 2000, Zhang *et al.*, 2005, Huang and Zou, 2006a, Huang and Zou, 2006b), and protein design (Poole and Ranganathan, 2006). Thus, in this article, we summarize the formalism of KBPs, specific implementations of different types of potentials, and their applications instead of concerning about the physical interpretation of KBPs.

A KBP energy function is a linear combination of individual potentials with each capturing a specific type of interaction. The most common formulation of such energy functions is:

$$E(C|S) = \sum_{ij} w_{ij} \left(-kT \ln \frac{p(c_j|s_i)}{p(c_j)} \right) \qquad \textbf{I-7}$$

where $E(C|S)$ is the energy of conformation C given that the underlying amino acid sequence is S. $p(c_j|s_i)$ is the probability that a given sequence s_i adopts conformation c_j, whereas $p(c_j)$ is an unconditional probability that any sequence fragment adopts conformation c_j. $\frac{p(c_j|s_i)}{p(c_j)}$ can be thought of as an "equilibrium constant" of a hypothetical chemical reaction: $random\ sequence, unique\ conformation \rightarrow unique\ sequence, unique\ conformation$ (Shortle, 2003). In addition to the above inverse Boltzmann formulation, other formulations of individual KBP terms have also been widely used. For example, the KBP under the modeling package Rosetta was formulated based on the Bayes' theorem (Simons *et al.*, 1997). This approach was also adopted by Woetzel *et al.* recently to derive the KBP for a SSE-based protein structure prediction algorithm (Karakas *et al.*, 2012, Woetzel *et al.*, 2012, Weiner *et al.*, 2013, Fischer *et al.*, 2016). In their Discrete Optimized Protein Energy, or DOPE, Shen and Sali computed the negative logarithm of the joint probability density function of a given protein (Shen and Sali, 2006).

The types of individual potentials incorporated into a KBP energy function are essentially only limited by the type of statistical relations that can be practically extracted from known protein structures. Depending on its intended purpose, a KBP may include individual potentials that fall into one or several categories. We elaborate three such potentials in the following and refer the

reader to references (Simons *et al.*, 1997, Woetzel *et al.*, 2012, Xu and Zhang, 2012) for examples of other potentials.

1) *pairwise distance-dependent potential* that approximates residue contact energies (Wodak, Sippl, 1990, Sippl, 1993, Sippl, 1995). Such contact potentials are based on native inter-residue contacts which play a key role in determining folding kinetics and native state stability (Gromiha and Selvaraj, 2004). The concept of pairwise distance-dependent potentials was first introduced in the pioneering work of Tanaka and Scheraga (Tanaka and Scheraga, 1976), who related residue contact frequencies to the free energies of formation of corresponding interactions using the simple relationship between free energy and equilibrium constant. Their work was followed by that of Miyazawa and Jernigan (Miyazawa and Jernigan, 1985, Miyazawa and Jernigan, 1996), who formalized the theory of residue contact potentials using quasi-chemical approximation. However, these early implementations of contact potentials are not, in fact, distance-dependent, except that a single cutoff distance was used to define residue contact. A real pairwise distance-dependent potential was first introduced by Sipp (Sippl, 1990), and this was followed by an explosion of different statistical potentials (Hendlich *et al.*, 1990, Kocher *et al.*, 1994, Park and Levitt, 1996, Bahar and Jernigan, 1997, Melo and Feytmans, 1997, Park *et al.*, 1997, Reva *et al.*, 1997, Rooman and Gilis, 1998, Samudrala and Moult, 1998, Betancourt and Thirumalai, 1999, Lu and Skolnick, 2001, Zhou and Zhou, 2002a, Fang and Shortle, 2005, Qiu and Elber, 2005, Summa *et al.*, 2005, Dehouck *et al.*, 2006, Shen and Sali, 2006, Woetzel *et al.*, 2012). Such pair potentials are usually formulated at residue level, where inter-residue distances are measured between C_β atoms or sidechain centroids in reduced representation of amino acid residues to promote computational efficiency. However, atomic-level formulation usually gives better discriminatory power albeit at the cost of more computational resource (Sippl, 1996, Sippl *et al.*, 1996, Melo and Feytmans, 1997, Samudrala and Moult, 1998, Lu and Skolnick, 2001, Shen and Sali, 2006).

2) *solvent accessibility-based environment potentials* that represent the interactions of individual residues with their local environment (Bowie *et al.*, 1991, Kocher *et al.*, 1994, DeLuca *et al.*, 2011, Xu and Zhang, 2012). Residue environment potentials are often included to account for solvation effects. Precise calculation of solvent accessibility requires full atomic structure and is time-consuming. In tertiary structure prediction scenarios where reduced representations of residues are used, good approximations to solvent accessibility, such as residue contact numbers, provide

significant computational savings (Durham *et al.*, 2009, Woetzel *et al.*, 2012, Fischer *et al.*, 2015, Li *et al.*, 2016). It should be noted that in addition to transforming solvent accessibility statistics to energy-like potentials using the inverse Boltzmann relation, they have also been incorporated into KBP energy functions as a penalty term to disfavor models where residue-specific solvent accessibilities disagree with expected solvent accessibilities (Xu and Zhang, 2012, Li *et al.*, 2017a).

3) *potentials of torsion angles* that evaluate backbone ϕ, ψ torsion angles and/or the preference of side-chain rotamers (Kocher *et al.*, 1994, Kuszewski *et al.*, 1996, Betancourt and Skolnick, 2004, Fang and Shortle, 2005, Amir *et al.*, 2008, Yang *et al.*, 2012, Kim *et al.*, 2013). It is well known that only certain combinations of ϕ, ψ torsion angles are populated in proteins (Ramakrishnan and Ramachandran, 1965) and significant correlations exist between side-chain torsion angle probabilities and backbone ϕ, ψ angles (Dunbrack and Karplus, 1993). Including such potentials has been shown to enable the energy function to exclude conformations that have unlikely combinations of torsion angles. In a study by Kocher *et al.* (Kocher *et al.*, 1994) where several types of potentials were tested to recognize protein native folds, potentials representing backbone torsion angle preferences recognized as many as 68 protein chains out of a total of 74. This result was striking given the fact that backbone torsion potentials consider solely local interactions along the chain and are well known to be incapable of determining the full 3D fold (Kocher *et al.*, 1994). Potentials of torsion angles have also been used to refine structures generated from NMR data (Kuszewski *et al.*, 1996, Yang *et al.*, 2012). Kuszewski *et al.* (Kuszewski *et al.*, 1996) incorporated a database-derived torsion angle potential into the target function for NMR structure refinement, resulting in a significant improvement in various quantitative measures of quality (Ramachandran plot, side-chain torsion angles, and overall packing. In a similar way, Yang *et al.* (Yang *et al.*, 2012) constructed a database of 2405 refined NMR structures.

I-4 Improving sampling and scoring with restraints

Due to their intrinsic inaccuracies, a common issue with energy functions is that incorrect conformations may be scored comparably to (or even better than) the native state (Skolnick, 2006), lending the energy function inability to recognize the native state (Figure 6(A)). This issue be remedied by incorporating sparse experimental data as restraints, which offers some structural

information that by itself is insufficient to completely determine the protein's structure (Figure I-6 (B), (C)).

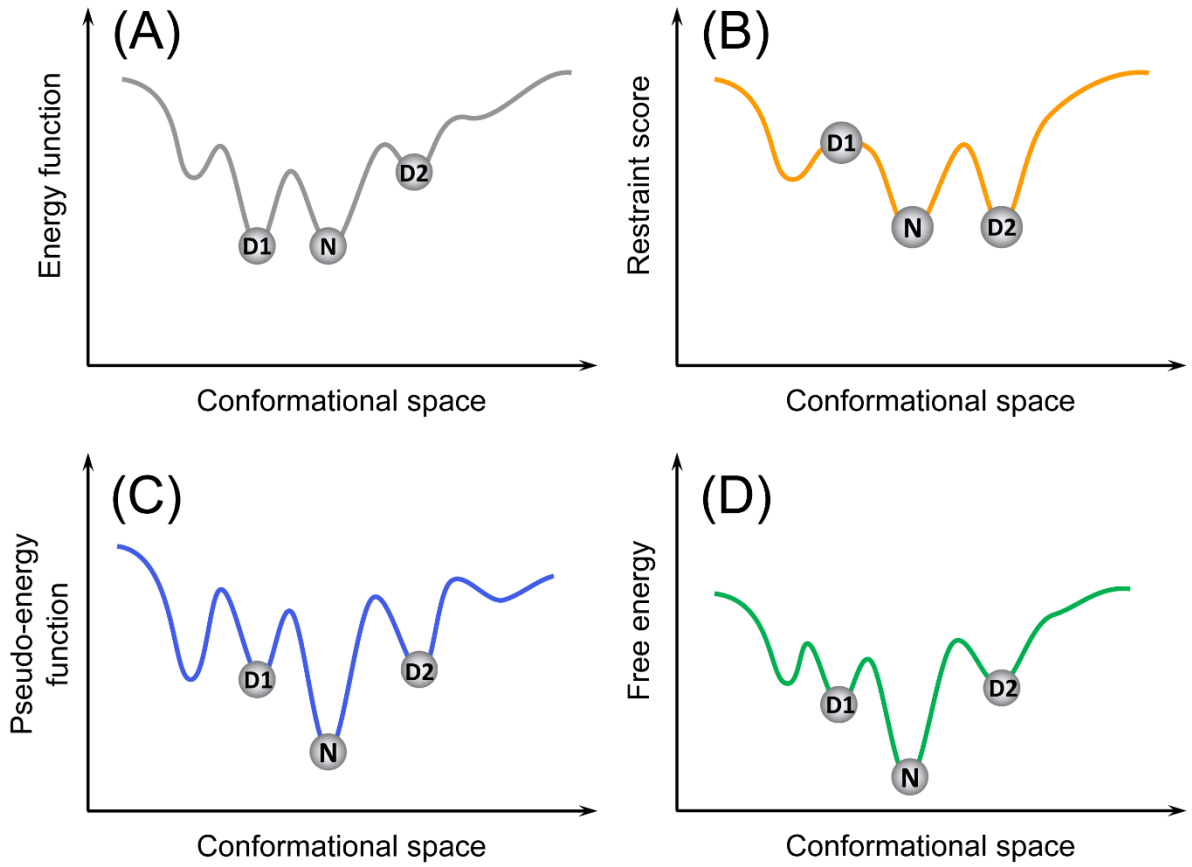


Figure I-6 Cooperative effects of energy functions and sparse restraints on a hypothetical protein

(A) the energy function has two comparable minima, lending itself the inability to tell decoy D1 from the native state N; (B) a scenario where decoy D1 violates some restraints and is thus penalized by the restraint score. However, as sparse restraints by themselves are insufficient to completely determine the protein's structure, there exists decoys, such as D2, that satisfy the restraints as well as the native state N does; (C) Adding a restraint score to the energy function results in what's called a pseudo-energy function which, in an ideal scenario, would be able to tell decoys apart from the native state; (D) the real free energy surface of the protein.

I-4.1 Sparse experimental data as restraints

Restraints from sparse experimental data drastically decrease the conformational space that needs to be sampled to only those structures consistent with the data. Many software suites implement algorithms to couple their *de novo* prediction methods with limited experimental data, including

those from nuclear magnetic resonance (NMR), electron paramagnetic resonance (EPR), cross linking mass spectrometry (XL-MS), and electron microscopy (EM).

NMR rivals X-ray crystallography as a technique by which an entire protein structure can be unambiguously determined. Solution-state NMR can determine the structure of relatively small proteins (< ~ 20 kDa), but intensive experimental techniques and analysis of NMR spectra are required to determine a high-quality structure of a protein. Each residue typically requires upwards of 15 constraints. Oftentimes, NMR spectroscopy can provide some degree of low-resolution information about the global conformation of a protein, even for larger proteins (Venters *et al.*, 1995, Battiste and Wagner, 2000). These sparse restraints, including chemical shifts (CSs), Nuclear Overhauser Enhancements (NOEs), and residual dipolar couplings (RDCs), do not provide enough information to fully determine the structure of a protein, but they can be used in conjunction with computational protein structure prediction software. CSs provide information about the protein backbone conformation, while NOEs and RDCs give information about the global fold of the protein. *De novo* protein structure prediction software can take advantage of just CSs (Latek *et al.*, 2007), CSs and NOEs (Bowers, 2000), or all three types of restraints (Weiner *et al.*, 2014).

Site-directed spin labeling (SDSL) and EPR can be used to glean information about proteins of nearly any size in their native environments. In addition, only a small amount of sample is required for structural interrogation by EPR. The accessibility and mobility of the spin labels can be used to determine the exposure and topology of SSEs (Farahbakhsh *et al.*, 1992, Altenbach *et al.*, 2005). Distances between spin labels can be detected up to 60 Å, and can give insight into the overall fold of the protein as well as different conformational states (Rabenstein and Shin, 1995, Borbat *et al.*, 2002). However, it is not feasible to use EPR to determine the full structure of a protein. EPR is experimentally intensive, as it requires the introduction of unpaired electrons at selected sites within proteins. This is usually done by cysteine substitution mutagenesis followed by modification of the sulfhydryl group with a nitroxide reagent. However, nonsense suppressor methodology, solid-phase peptide synthesis, or "click-chemistry" have also been used (Klare and Steinhoff, 2009). This technique will only give a small part of structural information about the protein, so these sparse EPR data can be used in conjunction with computational protein structure prediction methods (Alexander *et al.*, 2008, Hirst *et al.*, 2011, Fischer *et al.*, 2015). The selection

of sites to spin label is integral to the efficacy of structure determination by EPR (Alexander *et al.*, 2008).

Similarly, XL-MS experiments can be used to determine inter-atomic distances that serve as experimental restraints. XL-MS can be used with proteins in their native states, and it has proven to be compatible with relatively large proteins, flexible proteins, and membrane proteins (Kalkhof *et al.*, 2005, Jacobsen *et al.*, 2006, Lasker *et al.*, 2012). In addition, the samples used can be heterogeneous and dynamic, as the output of XL-MS experiments is an average. The basis of XL-MS is the ability of two functional groups of a protein to form covalent bonds if they are within a certain distance of one another. These cross links can occur both inter- and intramolecularly. The proteins are then enzymatically digested, and MS is used to identify these cross links and surface labels (Young *et al.*, 2000, Back *et al.*, 2003, Sinz, 2003).

EM provides data similar in format to that of X-ray crystallography, that is, a density map of a protein or complex. The data are thus less sparse than many of the aforementioned experimental techniques, but EM has historically provided lower-resolution density maps, from which an atomic structure cannot be gleaned. However, even low-resolution EM density maps are integral for identifying the overall organization of large molecular complexes. In recent years, EM technologies have progressed such that density maps with resolutions in the range of 4 – 8 Å can regularly be attained, at which level SSEs can be visualized and even some side chain character can be visualized (Bihnstein, 2015). Many computational modeling methods have been developed that work with EM density maps (Lindert *et al.*, 2009b), either in fitting previously solved structures into density maps, determining the topology and location of SSEs (Jiang *et al.*, 2001, Abeysinghe *et al.*, 2008), performing comparative modeling, and *de novo* protein structure prediction (Lindert *et al.*, 2009a, Lindert *et al.*, 2009b, Woetzel *et al.*, 2011, Lindert *et al.*, 2012a, Lindert *et al.*, 2012b).

Most *de novo* protein structure prediction algorithms require the use of a segmented density map, which can be accomplished with the use of various segmentation algorithms (Baker *et al.*, 2006, Pintilie *et al.*, 2010, Burger V, 2011). Then, SSEs can be extracted from the density map either manually or with the use of algorithms that automate the selection of helices and/or sheets from a segmented density map (Jiang *et al.*, 2001, Kong and Ma, 2003, Kong *et al.*, 2004, Baker *et al.*, 2007). Next, *de novo* modeling algorithms can use these data with the density map and

primary sequence of the protein in order to create a full structural model either via optimization (Chen *et al.*, 2016) or using Monte Carlo methods (Lindert *et al.*, 2009a, Lindert *et al.*, 2012a, Wang *et al.*, 2015).

I-4.2 Predicted contacts as restraints

If no experimental restraints are available for the protein, secondary and tertiary structural restraints can be predicted from an amino acid sequence based on existing structures. Secondary structures can be predicted using machine learning methods. Artificial neural networks (ANNs) can be used to predict secondary structures from position-specific scoring matrices (Jones, 1999, Yan *et al.*, 2013), reduced amino acid representation (Leman *et al.*, 2013), or multiple sequence alignments (MSAs) (Rost and Sander, 1993, Rost *et al.*, 1993). Methods have also been developed specifically to predict membrane protein topology from amino acid sequence using ANNs (Viklund *et al.*, 2008, Viklund and Elofsson, 2008, Leman *et al.*, 2013), support vector machines (SVMs) (Nugent and Jones, 2009), or Hidden Markov Models (HMMs) (Krogh *et al.*, 2001, Kahsay *et al.*, 2005).

It is a long-standing observation that 3D protein folds can be predicted from sufficient information regarding the protein's inter-residue contacts (Göbel *et al.*, 1994, Olmea and Valencia, 1997, de Juan *et al.*, 2013); the addition of even relatively sparse information about tertiary contacts into an algorithm's scoring function can help improve protein models (Kim *et al.*, 2014). Recently, the incorporation of long range contact predictions has resulted in some of the the most effective *de novo* protein structure prediction algorithms (Monastyrskyy *et al.*, 2015, Moult *et al.*, 2016). Several algorithms have been devised to predict these contacts using the principle of correlated mutations (de Juan *et al.*, 2013). In general, amino acid contacts that stabilize the protein fold are assumed to evolve complementarily – if one residue of a contact is mutated, the other will likely also mutate to a reasonable interaction partner.

In order to identify pairs of correlated mutations, amino acid pairs can be scored based on their physicochemical similarity using the McLachlan matrix (McLachlan, 1971), which is based on the frequencies of observed mutations in homologous proteins. Correlated mutations can also be scored by mutual information between MSAs based on the equation

$$I = \sum_{ab} f(a_i b_j) \log \frac{f(a_i b_j)}{f(a_i) f(b_j)} \qquad \textbf{I-8}$$

The above equation indicates that the mutual information between two protein sites *i* and *j* is computed by summing over amino acid pairs *ab* for every amino acid type *a* and *b*, where $f(a_i b_j)$ is the observed relative frequency of *ab* at columns *ij* and $f(a_i)$ is the observed relative frequency of amino acid type *a* at position *i*. The identification of these correlated mutations is used in many methods of multiple sequence alignment (Göbel *et al.*, 1994, Neher, 1994, Pollock and Taylor, 1997, Ashkenazy and Kliger, 2010, Hopf *et al.*, 2014), from which tertiary contact predictions can be extrapolated.

In recent years, numerous algorithms have come out that account for covariance caused by indirect inter-residue coupling effects, which has led to improvement in prediction of correlated mutations (Burger and van Nimwegen, 2010, Marks *et al.*, 2011, Morcos *et al.*, 2011, Jones *et al.*, 2012, Marks *et al.*, 2012, Kamisetty *et al.*, 2013, Skwark *et al.*, 2013, Ekeberg *et al.*, 2014, Hopf *et al.*, 2014, Kaján *et al.*, 2014, Michel *et al.*, 2014, Ovchinnikov *et al.*, 2014, Skwark *et al.*, 2014, Jones *et al.*, 2015). These methods were developed to resolve the issue that two residues aligned in multiple sequence alignments may exhibit statistical dependencies even though they are distant in physical space, which usually arises from chains of interacting pairs of residues. Also, information regarding the conservation of certain residues regardless of their tertiary contacts must be considered for correlated mutations to properly represent actual 3D contacts. Many methods have been devised that decouple direct from indirect residue coevolution, primarily based on statistical methods. Covariation-based contact prediction has also proven successful as a scoring metric for *de novo* folding (Morcos *et al.*, 2011, Kamisetty *et al.*, 2013).

Machine learning methods, including ANNs (Fariselli and Casadio, 1999, Fariselli *et al.*, 2001, Shackelford and Karplus, 2007, Tegge *et al.*, 2009, Xue *et al.*, 2009), genetic algorithms (MacCallum, 2004, Chen and Li, 2010), random forests (Li *et al.*, 2011), HMMs (Bjorkholm *et al.*, 2009, Lippi and Frasconi, 2009), and SVMs (Cheng and Baldi, 2007, Wu and Zhang, 2008), have also arisen as successful methods to predict 3D contacts. These methods use various features to predict contact maps. Some of the most successful of these machine learning methods for contact prediction are hybrid methods that predict contacts based on both physicochemical features and

evolutionary features, using MSAs as part of their training data sets (Wallner and Elofsson, 2006, Stout *et al.*, 2008, Ma *et al.*, 2013, Kosciolek and Jones, 2015).

I-5 Examples of methods for *de novo* tertiary structure prediction

Protein structure prediction methods can be broadly grouped into template-based modeling, where construction of target models involves threading the target sequence through the structure of homologous proteins (templates), and *de novo* structure prediction, where target models are constructed from sequence alone, without relying on similarity at fold level between the target sequence and any of the known structures (Baker and Sali, 2001, Bonneau and Baker, 2001, Hardin *et al.*, 2002, Lee *et al.*, 2009). Template-based modeling is based on the premise that tertiary structures of proteins in the same family are more conserved than their primary sequences (Chothia and Lesk, 1986, Fiser *et al.*, 2002, Illergard *et al.*, 2009). While it can produce accurate models for target sequences if templates with sequence identity > 25% are used (Cavasotto and Phatak, 2009) and can be practically useful (Xiong *et al.*, 2011, Zhan *et al.*, 2011, Li *et al.*, 2012), it is nevertheless purely mechanical in that it does not provide a general understanding of the role of particular interactions in maintaining the stability of protein structure (Baker and Sali, 2001, Cavasotto and Phatak, 2009). Thus, one could not gain insights into the physicochemical principles underlying protein folding (Pillardy *et al.*, 2001, Lee *et al.*, 2009). On the contrary, *de novo* methods sample and energy-evaluate the folded conformations as thoroughly as computational resource permits, and they assume the native conformation is the one with the lowest energy. Logically, two of the most crucial factors that dictate whether a *de novo* tertiary structure prediction method will be successful are its coverage of the conformational space and how accurate its energy function is. In this section, we discuss in detail some selected examples of *de novo* tertiary structure prediction methods and highlight some successful cases from the history of CASP (Figure I-7). Note that this selected set of methods is by no means exhaustive. The interested reader is referred to proceedings of CASP experiments (http://predictioncenter.org/index.cgi?page=proceedings), which cover a wider spectrum of methods and in more detail.

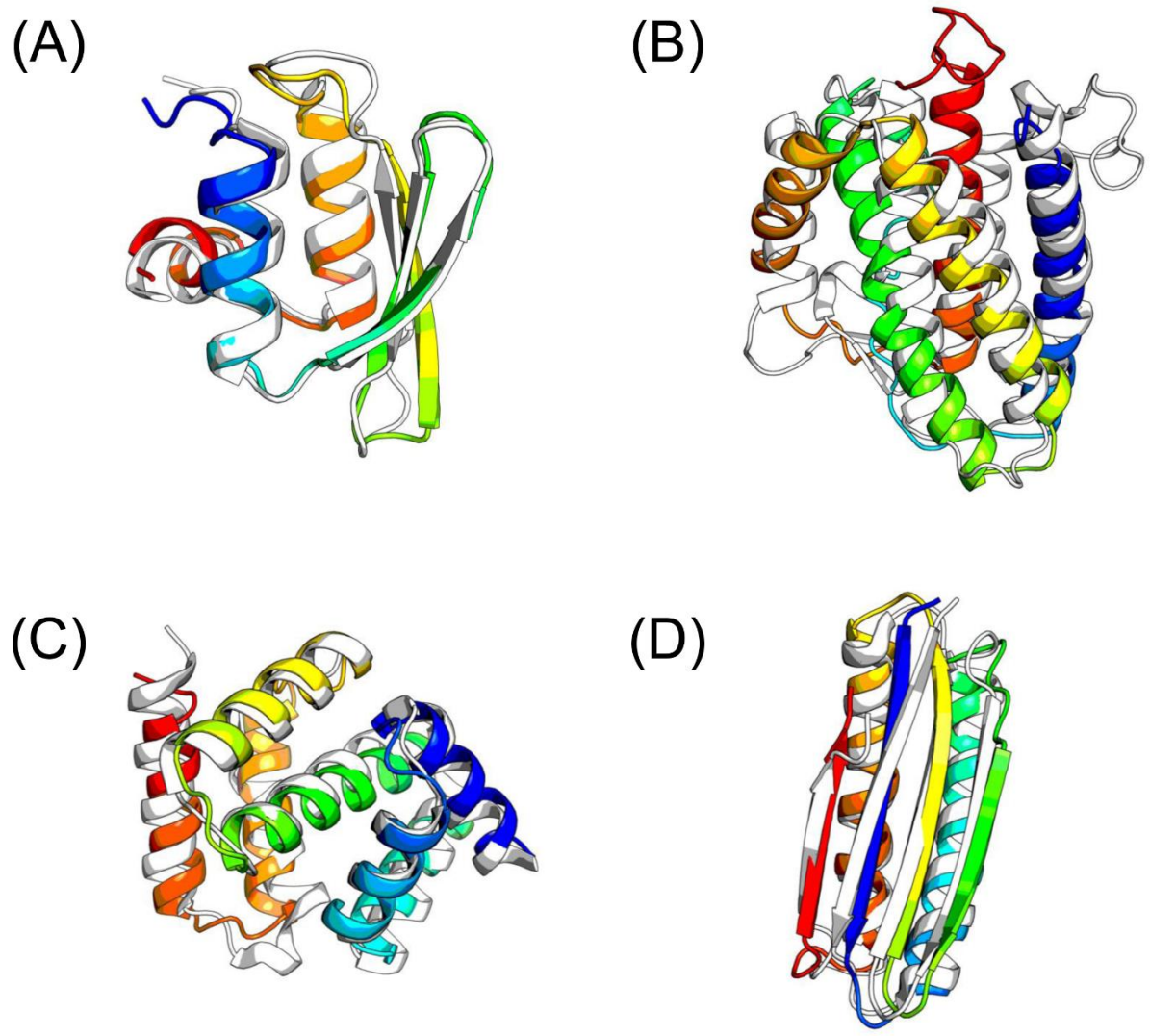


Figure I-7 Highlights of de novo structure prediction in CASP experiments

Predicted structure models (rainbow) are superimposed with the crystal structures (gray). (A) Rosetta-predicted structure model superimposed with a crystal structure (PDB code: 1whz) of CASP6 target T0281, hypothetical protein from *Thermus thermophilus* Hb8. This model is astonishingly close to the crystal structure, with a Cα-RMSD of 1.6 Å. (B) I-TASSER-predicted structure model superimposed with a crystal structure (PDB code: 4dkc) for the CASP10 ROLL target R0007, interleukin-34 protein from *Homo sapiens*. (C) Superposition of a QUARK-predicted structure model with a crystal structure (PDB code: 5tf3) of the CASP11 target T0837, hypothetical protein YPO2654 from *Yersinia pestis*. This model has a Cα-RMSD of 2.9 Å from the crystal structure. (D) Superposition of a BCL::Fold-predicted structure model with a solution NMR structure (PDB code: 2mq8) of CASP11 target T0769, a de novo designed protein LFR11 with ferredoxin fold. While this target is in the category template-based modeling, BCL::Fold assembled models for it without relying on any homologous templates.

I-5.1 FRAGFOLD

FRAGFOLD was developed based on the rationale that proteins tend to have common structural motifs at the super-secondary structural level (Jones, 1997b, Jones, 2001, Jones and McGuffin, 2003). In FRAGFOLD, 3D models are built by assembling super-secondary structural fragments from a library of highly resolved protein structures with MC simulated annealing and evaluated with a knowledge-based energy function. FRAGFOLD was initially tested in CASP2 (Jones, 1997b), and later in CASP4 (Jones, 2001) and CASP5 (Jones and McGuffin, 2003). Its success in predicting the fold of NK-Lysin marked the first correct *de novo* blind prediction of a protein's fold (Jones, 1997b).

The super-secondary structural fragments considered by FRAGFOLD include α-hairpin, α-corner, β-hairpin, β-corner, β-α-β unit, and split β-α-β unit. Favorable super-secondary structural fragments are selected based on the quality of threading. Threads that contradict the reliable regions of predicted secondary structure by PSIPRED (Jones, 1999) are skipped. In addition to this sequence-specific fragment list, a general fragment list that consists of all tripeptide, tetrapeptide, and pentapeptide fragments is also constructed from a library of highly resolved protein structures. The knowledge-based energy function in FRAGFOLD initially consists of a set of pairwise potentials, a solvation potential, a term for penalizing non-compact folds, a term for penalizing steric clashes, and a term that accounts for hydrogen-bonding (Jones, 1997b). This energy function was recently complemented with predicted contacts as restraints (Kosciolek and Jones, 2014). Kosciolek and coworkers (Kosciolek and Jones, 2014) found that combining statistical potentials with contacts predicted by PSICOV(Jones *et al.,* 2012) is significantly better than either statistical potentials or predicted contacts alone.

I-5.2 Rosetta

The Rosetta algorithm for *de novo* protein structure prediction employs MC simulated annealing to assemble protein-like 3D models from fragments of unrelated protein structures with similar local sequences using an energy function based on Bayes' theorem (Simons *et al.,* 1997, Rohl *et al.,* 2004). The algorithm is based on the experimental observation that local sequence preferences bias, but do not uniquely determine, the local structure of a protein (Rohl *et al.,* 2004). Rosetta has turned out to be one of the most successful methods indicated by results from CASP experiments

(Bradley *et al.*, 2003, Bradley *et al.*, 2005a, Jauch *et al.*, 2007) and several other studies (Bradley *et al.*, 2005b, Ovchinnikov *et al.*, 2017) (see Figure I-7(A) for an example).

Model construction in Rosetta is performed via a sequence of fundamental conformation modification operations termed "fragment insertion". For each fragment insertion, a sequence segment of three or nine residues is selected, and the torsion angles of these residues are replaced with the torsion angles of a homologous fragment selected from a ranked list of fragments of known structure (Simons *et al.*, 1997). Fragment insertions that decrease the energy of the resulting conformation are accepted and those that increase the energy are accepted according to the Metropolis criterion (Metropolis *et al.*, 1953). Derivation of the Rosetta energy function was based on a Bayesian separation of the total energy into components that describe the likelihood of a particular structure, independent of sequence, and those that describe the fitness of the sequence given a particular structure (Simons *et al.*, 1997).

$$P(structure|sequence) = \frac{P(sequence|structure)P(structure)}{P(sequence)} \quad \textbf{I-9}$$

The original Rosetta energy function is coarse-grained: terms corresponding to solvation and electrostatic effects are based on observed residue distributions derived from known protein structure databases, and hydrogen bonding is not explicitly described. However, preferences of β-strand pairing geometries and β-sheet patterns are included. Steric clashes are penalized, while van der Waals interactions are not explicitly modeled. A more physically realistic, atomic-level energy function was developed later for applications requiring more detailed structurally information. In this "fine-grained" version of the energy function, van der Waals interactions are modeled with a 6-12 Lennard-Jones potential. Solvation effects are included, using the Lazaridis-Karplus model (Lazaridis and Karplus, 1999), and hydrogen-bonding is explicitly accounted for using a secondary structure- and orientation-dependent potential derived from high-resolution protein structures (Kortemme *et al.*, 2003). Energetics of local interactions are described using an amino acid- and secondary structure-dependent potential for backbone torsion angles. The reader is referred to reference (Rohl *et al.*, 2004) for a more mathematically detailed description of the Rosetta energy function.

I-5.3 I-TASSER

Recent CASP experiments have shown significant advantages of integrating various techniques such as threading, *de novo* modeling and atomic-level structure refinement approaches into a single pipeline of tertiary structure prediction (Battey *et al.*, 2007, Jauch *et al.*, 2007, Zhang, 2009, Kinch *et al.*, 2011, Tai *et al.*, 2014, Kinch *et al.*, 2016). The I-TASSER method,(Wu *et al.*, 2007, Roy *et al.*, 2010, Yang *et al.*, 2015) which implements TASSER (Zhang and Skolnick, 2004) in an iterative mode, is one example of the composite approaches. I-TASSER has been particularly successful as shown by recent CASP experiments (Zhang, 2009, Roy *et al.*, 2010, Yang *et al.*, 2015, Zhang *et al.*, 2016) (see Figure I-7(B) for an example).

I-TASSER uses a sophisticated threading scheme, which compares the target sequence with template structures using profile-profile alignment, for selection of the most probable structure fragments. Aligned regions of the target sequence are modeled by connecting template fragments through a random walk of Cα–Cα bond vectors of variable lengths. Unaligned regions are simulated on a cubic lattice system for computational efficiency. Initial full-length coarse-grained models are refined via REMC simulation where two kinds of moves are implemented: off-lattice rigid fragment translations and rotations of the aligned regions and on-lattice 2–6 bond movements and multi-bond sequence shifts of unaligned regions (Zhang and Skolnick, 2004). The models of the first-round TASSER simulation are clustered and the cluster centroids are submitted to a second-round TASSER simulation to remove physically unrealistic interactions. Finally, backbone atoms and sidechain rotamers are added to the model with the lowest energy from the second round (Wu *et al.*, 2007). The energy function of I-TASSER includes the original TASSER knowledge-based potential and a new burial potential based on neural network-predicted accessible surface area (ASA) (Wu *et al.*, 2007). The original TASSER potential consists of long-range pair interactions of sidechain centers of mass, local Cα correlations, hydrogen-bond, hydrophobic burial interactions, propensities for predicted secondary structures, protein specific pair potentials of sidechain centers of mass, and tertiary contact restraints extracted from the threading templates (Zhang *et al.*, 2003).

I-5.4 QUARK

QUARK is an algorithm for *de novo* protein structure prediction using REMC simulations guided by a consensus knowledge-based energy function. In contrast with Rosetta and I-TASSER that

assemble fragments of fixed sizes, QUARK assembles 3D models from small structure fragments of multiple sizes from 1 to 20 residues. To increase the structural flexibility and the efficiency of conformational search, QUARK also implements a set of MC moves consisting of free-chain constructions and fragment substitutions between decoy and fragment structures (Xu and Zhang, 2012). The QUARK algorithm has been shown to be highly successful in recent CASP experiments (Xu and Zhang, 2012, Zhang *et al.*, 2016) (see Figure I-7(C) for an example).

QUARK generates structure fragments for target sequences by threading sequence segments through a library of non-homologous experimental structures. Multiple features such as solvent accessibility, real-value ϕ and ψ angles, and secondary structure types as predicted from back-propagation neural networks are used to improve generation of structure fragments. Optimization of 3D models is performed via REMC simulations that start with initial models assembled by chaining randomly selected fragments with varied sizes. Conformational sampling of each replica is done through residue-level, segment-level, and topology-level movements. After each running cycle, the conformations between every two adjacent replicas are exchanged according to the Metropolis criterion (Metropolis *et al.*, 1953). Protein structure models built by QUARK are evaluated by a composite knowledge-based energy function consisting of atomic-level pair potentials, hydrogen-bonding potential, SSE packing potentials, heuristic terms that account for excluded volumes, solvent accessibility, and radius of gyration (Xu and Zhang, 2012).

I-5.5 BCL::Fold

The BCL::Fold algorithm developed in our group seeks to overcome the limitations of protein size and fold complexity by assembling idealized SSEs (secondary structure elements) into 3D models. This algorithm was developed under the framework model of protein folding. As discussed previously, while the framework model is not always true, it is straightforward to implement. In addition, as shown by our benchmark study (Karakas *et al.*, 2012, Weiner *et al.*, 2013), BCL::Fold facilitates the sampling of non-local contacts. Thus, BCL::Fold may be a promising tool for structure prediction of proteins with high contact order (Plaxco *et al.*, 1998, Baker, 2000, Bonneau *et al.*, 2002). It's also worth mentioning, that in contrast to the other four methods, which heavily reply on the availability of homologous template structural fragments (short or long), BCL::Fold is "truly" *de novo* in the sense that no template structure is needed at any stage of the algorithm. While BCL::Fold was not ranked among the most successful methods, we would still like to

highlight the CASP11 target T0769. While this protein is in the category of template-based modeling, meaning that a suitable template can be identified that covers all or nearly all of the target, BCL::Fold predicted a model with a Cα-RMSD of 1.8 Å to the released solution NMR structure without relying on any homologous templates (Fischer *et al.*, 2016) (Figure I-7(D)).

In BCL::Fold, the necessary complexity reduction of the conformational space is achieved by assembling SSEs from a predetermined pool of SSEs using MC simulated annealing and omitting more flexible loop regions. A high-quality pool of SSEs can be readily created using machine learning-based secondary structure prediction methods such as PSIPRED (Jones, 1999). BCL::Fold implements a comprehensive list of SSE-based MC moves, which are categorized into six main categories: adding SSEs, removing SSEs, swapping SSEs, single SSE moves, SSE-pair moves, and moving domains consisting of multiple SSEs (Karakas *et al.*, 2012). Models generated by BCL::Fold are evaluated by a knowledge-based consensus energy function called BCL::Score (Woetzel *et al.*, 2012), which consists of potentials of residue pair interaction, residue environment, SSE packing, β-strand pairing, loop length, radius of gyration, contact order, secondary structure prediction agreement. Separate penalizing energy terms were also included to exclude conformations with clashes between amino acids or SSEs and loops that cannot be closed (Woetzel *et al.*, 2012). BCL::Score can also be complemented with experimental or predicted restraints to improve selection of native-like models (Weiner *et al.*, 2014, Fischer *et al.*, 2015, Li *et al.*, 2017a).

I-6 Outlook

In the past decade, we've seen hardware and algorithmic advances that enabled researchers to perform millisecond timescale simulations of protein folding, and we've also seen development of methodologies that predicted tertiary structure with better accuracy for proteins with larger size. Despite these achievements, there is still a long list of challenges on the way toward a solution to the protein folding problem.

On the folding mechanism side, even though long simulations have been available, unambiguous scientific results learned from such simulations have thus far been modest (Lane *et al.*, 2013). First, it is still being debated whether proteins fold via a single definite pathway or multiple parallel pathways. Although both views have received support from simulations and experiments (Englander and Mayne, 2014, Wolynes, 2015), additional simulations with more

robust trajectory analysis and experimental validation are required to disambiguate conflicting results. Second, realistic folding simulations have thus far been limited to small proteins (< 100 residues), it is questionable whether folding mechanisms revealed by these simulations are generalizable to larger proteins. Thus, simulating the folding of larger proteins will likely be a major trend for the next decade. Finally, as far as we are aware, a theory that is quantitative and makes specific prediction about how a protein would fold isn't yet available. The somewhat loosely defined models of hierarchical (framework) folding, nucleation-condensation, and foldons are difficult to validate or invalidate either by experiments or simulations. Nevertheless, closer interaction between simulations and experiments such that simulations be tested by experiments and in turn aid in the interpretation of experimental results and guide the design of future experiments will have greater impact on the field.

On the structure prediction side, larger proteins, especially those with multi-domains, stay a significant challenge to *de novo* structure prediction methodologies. These proteins are often characterized by their high contact order and long folding time (Plaxco *et al.*, 1998, Paci *et al.*, 2005). Conformational sampling of these proteins is usually inefficient and is complicated not only by protein size, but also by the considerable number of non-local contacts, which are formed by residues far apart in sequence but usually critical for structural stability (Moult, 2005, Kim *et al.*, 2009). Consequently, tools for *de novo* structure prediction are not likely to become practically useful for structure prediction for any but very small, sometimes medium-sized proteins (Jones, 1997a, Baker and Sali, 2001). Other challenging targets, especially for methods whose energy functions heavily rely on statistics extracted from known structures, may also include proteins with rare and unusual folds (Kinch *et al.*, 2016). Accurate prediction of tertiary structure for these challenging targets certainly requires the joined forces of high-performance hardware, efficient algorithms for conformational sampling, accurate energy functions, and, last but not least, valuable experimental restraints.

II. ACCURATE PREDICTION OF CONTACT NUMBERS FOR MULTI-SPANNING HELICAL MEMBRANE PROTEINS

This chapter has been published under (Li *et al.*, 2016).

II-1 Introduction

Helical membrane proteins (HMPs) play essential roles in various biological processes, including signal transduction, ionic and molecular transportation across the membrane, and energy generation. Due to their pharmacological relevance, about 50% of drugs on the market target HMPs (Overington *et al.*, 2006). It was estimated that HMPs constitute about 20% to 30% of the human genome (Krogh *et al.*, 2001). In spite of their prevalence in the genome, a very small portion of structures in the Protein Databank is HMPs due to the experimental difficulties in determining structures of HMPs. Therefore, accurate and efficient computational methods would be valuable tools to complement existing experimental techniques. One of the challenges in computational prediction of three-dimensional (3D) structure of HMPs is to predict helix-helix packing in which a transmembrane helix (TMH) either faces the lipids or is buried in the protein core. Knowing *a priori* whether an amino acid residue is exposed to the membrane lipid or buried inside the protein core provides valuable restraint information that can be incorporated to reduce the sampling space of helix-helix packing. As an intermediate step to the prediction of 3D structure of HMPs, it is worthwhile to develop reliable methods for predicting residue exposure.

Solvent accessibility is the most commonly used structural feature for characterizing the exposure environment of a residue (Lee and Richards, 1971). However, the applicability of solvent accessibility in helix-helix packing, or *de novo* 3D structure prediction, where an astronomical conformational space needs to be sampled is limited. Accurate computation of solvent accessibility needs full-atom representation of amino acid side chains. Therefore, it is computationally demanding. Residue weighted contact number (WCN), defined as the number of contacting residues of the residue of interest is another structural feature that reflects the exposure of a residue (Dill, 1999, Echave *et al.*, 2016). Computation of WCN does not require a full-atom representation of amino acid side chains and is numerically fast. Thus, WCN is more suitable for being incorporated into 3D structure prediction either in the form of restraints or knowledge-based potential. In addition, as WCN is negatively correlated with solvent accessibility (Durham *et al.*, 2009), it may as well be useful for addressing a spectrum of biological problems in which solvent

accessibility has been applied, such as epitope mapping (Haste Andersen *et al.*, 2006), hot spots detection (Martins *et al.*, 2014, Munteanu *et al.*, 2015), understanding of protein-protein interactions (Jones and Thornton, 1997b, Jones and Thornton, 1997a, Marsh and Teichmann, 2011), model quality assessment (Phatak *et al.*, 2011), and modeling of amino acid residue side-chain conformation (Eyal *et al.*, 2004).

Traditionally, prediction of WCN is treated as a two-state (higher or lower than the average WCN) or three-state (much higher, much lower, or close to average WCN) classification problem (Fariselli and Casadio, 2000, Pollastri *et al.*, 2001, Pollastri *et al.*, 2002). However, the applicability of classification approach is limited as it is difficult to use discrete exposure status for scoring in 3D structure prediction. Furthermore, subdividing residues into different states requires an arbitrary selection of a specific WCN as a cutoff. Therefore, real-value predictions should be preferred (Ahmad *et al.*, 2003). The problem of predicting WCNs for soluble proteins has been studied for more than a decade and promising results have been achieved (Kinjo *et al.*, 2005, Yuan, 2005). Even though a few attempts have been made to predict the burial status or real-value solvent accessibility of TMH residues (Beuming and Weinstein, 2004, Yuan *et al.*, 2006, Park *et al.*, 2007, Illergard *et al.*, 2010), given the fact that 3D structures of HMPs have long been desirably pursued, it is remarkable to notice that no work has been reported on predicting WCNs for HMPs.

Here, we present a dropout neural network-based method, termed TMH-Expo, for predicting WCNs for HMPs. We first curated a large non-redundant data set of HMPs with known structure based on which experimental WCNs were computed. Thereafter, we examined a set of feature vectors containing local sequence or evolutionary information for WCN prediction. Subsequently, a detailed analysis of the performance of TMH-Expo was conducted. Finally, we showed that predicted WCN reveals exposure patterns of TMHs and discussed the application of predicted WCN to 3D structure prediction and protein-protein docking.

II-2 Methods

II-2.1 Generation of data set

The data set of HMPs with known structures used in the current study was retrieved from the OPM (Orientation of Proteins in the Membrane) database (Lomize *et al.*, 2006). Peripheral HMPs and peptides were removed to obtain a set of "true" HMPs. A further refinement was carried out by

removing thylakoid HMPs as they have extreme topological complexity (Dekker and Boekema, 2005). The protein culling server PISCES (Wang and Dunbrack, 2003) was used to obtain a list of HMP chains that have a sequence length between 40 and 10000 residues, and pair-wise sequence identity of 25% or less. Non-X-ray structures, C_α-only structures, as well as X-ray structures with a resolution of > 3.0 Å or an R-factor > 0.3 were excluded. This culminated the final data set that consisted of 90 chains from 71 proteins from 33 OPM superfamilies for training the TMH-Expo. The complete list of protein chains used in this study can be found in Table A-1 in the APPENDICES. The transmembrane region for each protein chain was provided by OPM. The membrane normal aligns with the z-axis and the membrane center is positioned at z = 0. Secondary structure was assigned to each chain from the consensus identification of DSSP (Kabsch and Sander, 1983), Stride (Heinig and Frishman, 2004), and PALSSE (Majumdar *et al.*, 2005). A residue is considered as a TMH residue if it sits inside the membrane and the residue is part of a helical conformation.

II-2.2 Computation of WCN

The WCN of a residue i was originally defined as the number of C_α atoms of other residues within the sphere of radius d centered at the C_α atom of residue i (Nishikawa and Ooi, 1986). While this definition is straightforward, it has the disadvantage that each residue within the sphere is assigned an equal contribution to the total WCN. This is physically unrealistic because both van der Waals and electrostatic interactions are distance-dependent. To achieve a more physical approximation, we used a refined algorithm developed for WCN computation (Durham *et al.*, 2009). This algorithm is similar to that of Kinjo *et al.* (Kinjo *et al.*, 2005) where C_β atoms are used instead of the C_α atom and the boundary of the sphere is smoothened. Contribution to the total WCN is assigned to each residue inside the sphere in a distance-dependent way such that short-range contacting residues have higher contribution than long-range contacting ones (Kinjo *et al.*, 2005). Residues whose C_β atom is within 4.0 Å to the C_β atom of the residue of interest are assigned a weight of 1.0; those with a distance longer than 11.4 Å are assigned a weight of 0. Any residue in between is assigned a weight between 0.0 and 1.0 according to a smooth transition function. This scheme can be summarized into the following function:

$$w_{ij} = \begin{cases} 1, d_{ij} \le l \\ \frac{1}{2} cos\left(\frac{d_{ij} - l}{u - l} \times \pi\right) + \frac{1}{2}, l < d_{ij} < u \\ 0, d_{ij} \ge u \end{cases} \quad \textbf{II-1}$$

where w_{ij} is the contribution made by residue j to the total WCN of residue i, d_{ij} is the distance between the C_β atoms of residue i and residue j, l is the lower bound of d_{ij} within which $w_{ij} = 1.0$, and u is the upper bound of d_{ij} beyond which $w_{ij} = 0$. For glycine, $H_{\alpha 2}$ is used in place of C_β atom. The lower and upper bound are optimized values such that the correlation between WCN and the solvent accessible surface area (SASA) is maximized. Only residues separated by more than three residues along the sequence are considered in the calculation to reduce the bias due to sequence proximity. The total contact number of residue i was computed by summing up w_{ij} over the entire protein:

$$WCN_i = \sum_{j \in |j-i|>3}^{n} w_{ij} \quad \textbf{II-2}$$

where n is the length of the protein chain for computing monomeric WCN or the total number of residues in the protein for computing oligomeric WCN. All non-protein molecules were removed before computing WCNs. Non-protein molecules such as coenzymes, ligands, and internal waters play important roles for the function of membrane proteins. However, the biochemical identity of the interface between these molecules and membrane proteins requires detailed analysis and is beyond the scope of this study.

II-2.3 Computation of relative solvent accessibility

The relative solvent accessibility (RSA) of a residue was computed as the ratio between the absolute solvent accessibility (ASA) observed in the native structure and that in an extended tripeptide conformation (A-X-A). The ASA values were computed based on the oligomeric states provided by OPM using DSSP with a probe radius of 1.4 Å (Kabsch and Sander, 1983) as with previous studies (Pollastri *et al.*, 2002, Ahmad *et al.*, 2003, Chang *et al.*, 2008, Petersen *et al.*, 2009). No further exploration on probe sizes was conducted because it has been shown that probe size has little or no effect on the performance of RSA predictors (Illergard *et al.*, 2010). The ASA

value of each amino acid type in an extended tripeptide conformation was adopted from a similar study (Ahmad *et al.*, 2003).

II-2.4 Computation of feature vectors

The multiple sequence alignment (MSA) for each protein sequence in the data set was obtained by searching the UniRef50 (Suzek *et al.*, 2007) non-redundant sequence database with PSI-BLAST for five iterations (Altschul *et al.*, 1997). The E-value inclusion threshold was set to 0.01. Floating point-valued position-specific scoring matrix (PSSM, see Figure A-1 in APPENDICES for an example of PSSM) were generated from PSI-BLAST checkpoint files using the source code (chkparse.c) adapted from PSIPRED (McGuffin *et al.*, 2000). Floating point-valued PSSM was preferred over integer-valued PSSM as the former provides higher precision. PSSM is an $L \times 20$ matrix where L denotes sequence length. For each sequence position i, there are 20 entries, each corresponding to the score of one of the 20 naturally occurring amino acid. The BLAST probability profile (BPP) for amino acid j at sequence position i was computed by transforming each PSSM entry m_{ij} using the following equation:

$$p_{ij} = \frac{10^{\frac{m_{ij}}{10}}}{\sum_{j}^{20} 10^{\frac{m_{ij}}{10}}} \qquad \textbf{II-3}$$

where j runs from 1 to 20. The variance-based conservation index (CI) CI_i is one of the commonly used conservation indices and is defined by the following formula:

$$CI_i = \sqrt{\sum_{j}^{20} (p_{ij} - p_j)^2} \qquad \textbf{II-4}$$

where the summation is carried out over 20 amino acids, p_{ij} is the BLAST probability of amino acid j at position i such that $\sum_{j}^{20} p_{ij} = 1$, and p_j is the average BLAST probability of amino acid j and is defined as $\frac{1}{L}\sum_{i}^{L} p_{ij}$. The amino acid type at each sequence position is encoded by a vector with 20 binary entries (or 20 bits). When considering a window size of w centered at the residue whose WCN is to be predicted, the feature vector computed based on PSSM, BPP, or local

sequence composition (LSC) has a total of $w \times 20$ components. Whereas the feature vector computed based on CI has a total of $w \times 1$ components.

II-2.5 Training of dropout neural networks with back-propagation of errors

The support vector machine (SVM) algorithm has been applied to various bioinformatics tasks, especially solvent accessibility and WCN prediction (Yuan, 2005, Yuan *et al.*, 2006, Park *et al.*, 2007, Illergard *et al.*, 2010). It has the benefit of being less prone to overfitting than neural networks. Indeed, our preliminary test showed that neural networks trained without dropout (learning rate $\eta = 0.1$, momentum factor $\alpha = 0.1$, number of hidden layer neurons = 64, *and* number of epochs = 500) had a MAE (mean absolute error, see Eq. II-9 for its definition) of 2.70, whereas an optimized SVM (radial basis function kernel, $\gamma = 0.025$, cost = 0.1) had a MAE of 1.76. However, neural networks trained with dropout (learning rate $\eta = 0.1$, momentum factor $\alpha = 0.1$, number of hidden layer neurons = 64, number of epochs = 500, dropout rate in input layer = 0.05, and dropout rate in output layer = 0.5) had a MAE of 1.69. As dropout neural networks had a smaller MAE, we thus chose dropout neural networks as the learning algorithm in the current study.

The dropout neural networks trained in this study were fully connected three-layer feed-forward networks with a sigmoid activation function (Figure II-1(A)). The input layer contained one unit for each component in the feature vector. Inputs to the network are either local sequence information or evolutionary information derived from PSI-BLAST computed MSAs. The window size used for computing feature vectors was set to 15, an optimal value for WCN prediction found in our preliminary testing. The output layer was composed of a single node for residue-specific WCN or RSA. The hidden layer was composed of 64 neurons. A random of 5% of units in the input layer and 50% of neurons in the hidden layer were dropped during each presentation of each training case. The networks were trained with resilient back-propagation of errors (Rumelhart *et al.*, 1986) with the learning rate η being set to 0.1 and momentum factor α set to 0.1. Weights were updated after presentation of each residue to the network. A maximum of 2000 epochs were applied.

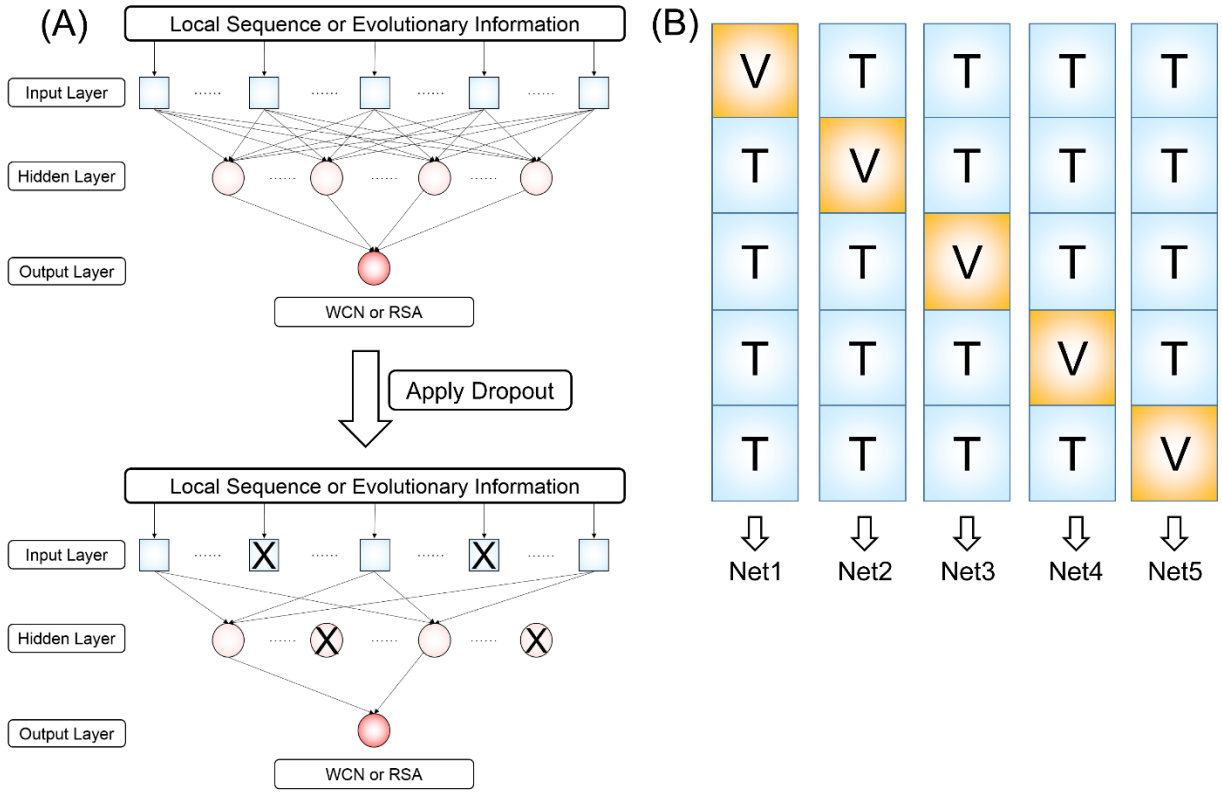


Figure II-1 Training of dropout neural networks with five-fold cross-validation

(A) Neural network architectures before and after applying dropout (neurons randomly dropped out are crossed); (B) five-fold cross-validation training protocol (T: training set, V: validation set).

II-2.6 Jackknife cross-validation

A relatively low sequence identity (25%) was used in the current study; however, such low sequence identity alone might not be sufficient to exclude homology among protein chains. In fact, substantial remote homology could still exist at this level placing HMPs in the same structural superfamily (Jaakkola *et al.*, 1999). Such remote homology between proteins in the training set and proteins in the validation set for testing the model can lead to an over-optimistic estimate of the performance for new folds. As a way of preventing such over-optimism, the data set was partitioned such that each OPM superfamily forms its own subset that contains all its members and no members from other OPM superfamilies. Cross-validation of the networks was done in a jackknife manner with respect to OPM superfamily. Of the 33 OPM superfamilies, one single OPM superfamily was withheld as the validation set for evaluating the neural networks, and a five-fold cross-validation protocol adopted for our transmembrane span and secondary structure

prediction algorithm (Leman *et al.*, 2013) was carried out on the remaining 32 superfamilies (Figure II-1(B)). This process was then repeated 33 times, with each of the 33 OPM superfamilies used exactly once as the validation set. Predictions for the 33 validation sets were combined to give the final estimate of the performance of the neural networks.

II-2.7 Performance measures

A set of performance measures were adopted to evaluate the performance of the neural networks. The primary measure was the Pearson correlation coefficient (PCC) between experimental and predicted WCNs and RSA. For a set of n data points (x_i, y_i), the PCC was computed as follows:

$$PCC = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \quad \text{II-5}$$

For comparing our results to that from previous studies, we incorporated the following measures that are commonly used to evaluate classifiers:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \quad \text{II-6}$$

$$FPR = \frac{FP}{TN + FP} \quad \text{II-7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{II-8}$$

where MCC is the Matthews correlation coefficient (Matthews, 1975), FPR is the false positive rate, TP is the number of correctly predicted buried residues, TN is the number of correctly predicted exposed residues, FP is the number of incorrectly predicted buried residues, and FN is the number of incorrectly predicted exposed residues. The real value WCN and RSA were transformed to binary states using the median as a cutoff such that the data set is equally partitioned. The mean absolute error (MAE) which is defined as the per-residue absolute difference between experimental and predicted WCN and RSA was used to evaluate prediction errors:

$$AE = \frac{\sum |v_{experimental} - v_{predicted}|}{n} \quad \text{II-9}$$

where v is either RSA or WCN, n is the total number of residues to be predicted. The summation is carried out over all predicted residues.

II-3 Results and Discussion

II-3.1 Statistics of the data set

The repository of HMPs with known structures has expanded tremendously in recent years. It was reported that the latest number of unique membrane protein structures deposited in the Protein Databank is 535 (http://blanco.biomol.uci.edu/mpstruc/) compared to ~150 back in 2005 (White, 2004). Curation of a dataset that is representative of the population is an essential step in producing a model with high predictive accuracy. We compared the data set used to train TMH-Expo to those used in two related works namely $ASAP_{mem}$ (Yuan *et al.*, 2006) and MPRAP (Illergard *et al.*, 2010). In terms of the size of data sets, the TMH-Expo data set consists of 71 HMPs (90 unique chains), significantly larger than the $ASAP_{mem}$ data set (also known as the Beuming-Weinstein or BW data set (Beuming and Weinstein, 2004)) which has 28 HMPs (59 unique chains). The MPRAP data set has 52 HMPs (80 unique chains). Interestingly, PISCES returned only 34 HMPs (60 unique chains) from the MPRAP data set using the same criteria applied to cull the TMH-Expo data set.

Table II-1 lists the frequency, mean WCN, as well as standard deviation of WCN for each amino acid residue type. Similar to observations made by Ulmschneider and coworkers (Ulmschneider and Sansom, 2001), residues with nonpolar side chain such as Ala, Phe, Ile, Leu, and Val are dominantly abundant. In addition, except in the case of Ala, their mean WCNs are not significantly higher than that of other amino acid residues. In fact, the mean WCNs for Phe, Ile, Leu, and Val are among the lowest, an expected observation given the fact that the membrane provides an environment that is more hydrophobic than the protein interior. On the other end, the mean WCNs for Ala, Cys, Gly, and Ser are among the highest, suggesting that on average helices enriched with these residues are more densely packed. In fact, Ala, Gly, and Ser are known to form the sequence motifs of the type AxxxA, GxxxG, and SxxxS that are believed to promote close helical packing (Russ and Engelman, 2000).

Table II-1 Summary of the TMH-Expo data set

Amino Acid Residue	Frequency	Mean WCN	Standard Deviation of WCN
A	1282	12.09	3.12
C	131	12.46	2.72
D	93	11.34	2.62
E	151	11.10	2.51
F	953	10.58	2.67
G	1008	12.76	3.15
H	134	11.31	2.36
I	1242	10.46	2.68
K	156	9.22	2.53
L	1938	10.59	2.63
M	437	11.65	2.54
N	204	11.84	2.88
P	329	10.79	3.23
Q	161	11.02	2.69
R	184	9.94	2.57
S	598	12.20	2.80
T	604	11.83	2.85
V	1256	10.91	2.88
W	323	10.08	2.62
Y	381	11.02	2.61

II-3.2 Relevance of input features

The performance of a data-trained machine learning method depends crucially on the judicious choice of the feature vector. For solvent accessibility prediction, feature vectors containing primary sequence information or evolutionary information have been tested (Ahmad *et al.*, 2003, Yuan, 2005, Park *et al.*, 2007, Chang *et al.*, 2008). Four feature vectors: CI, LSC, BPP, and PSSM were investigated in this study. CI, BPP, and PSSM can be considered as evolutionary information-containing feature vectors as they are derived based on MSA, whereas LSC contains purely primary sequence information. We initially examined the correlation coefficient of all features computed considering a window size of 41 (residues from $i - 20$ to $i + 20$, where i is the position of the residue of interest, inclusive) with WCNs. This resulted in 41×1, 41×20, 41×20, and 41×20 entries for feature vector of CI, LSC, BPP, and PSSM respectively (Figure A-1 in APPENDICES). Figure II-2 plots the correlation coefficients of entries in each feature vector with WCNs. For

sequence-based prediction, it is well known that the use of evolutionary information derived from MSA improves prediction performance. In fact, on average CI, BPP, and PSSM show stronger correlation with WCNs than local sequence composition does (compare Figure II-2(A), 2(C), 2(D) with 2(B)). It is also interesting to note that PSSM generally has more strongly correlated entries than either of the other two evolutionary information-containing feature vectors does (compare Figure II-2(D) with 2(A) and 2(C)).

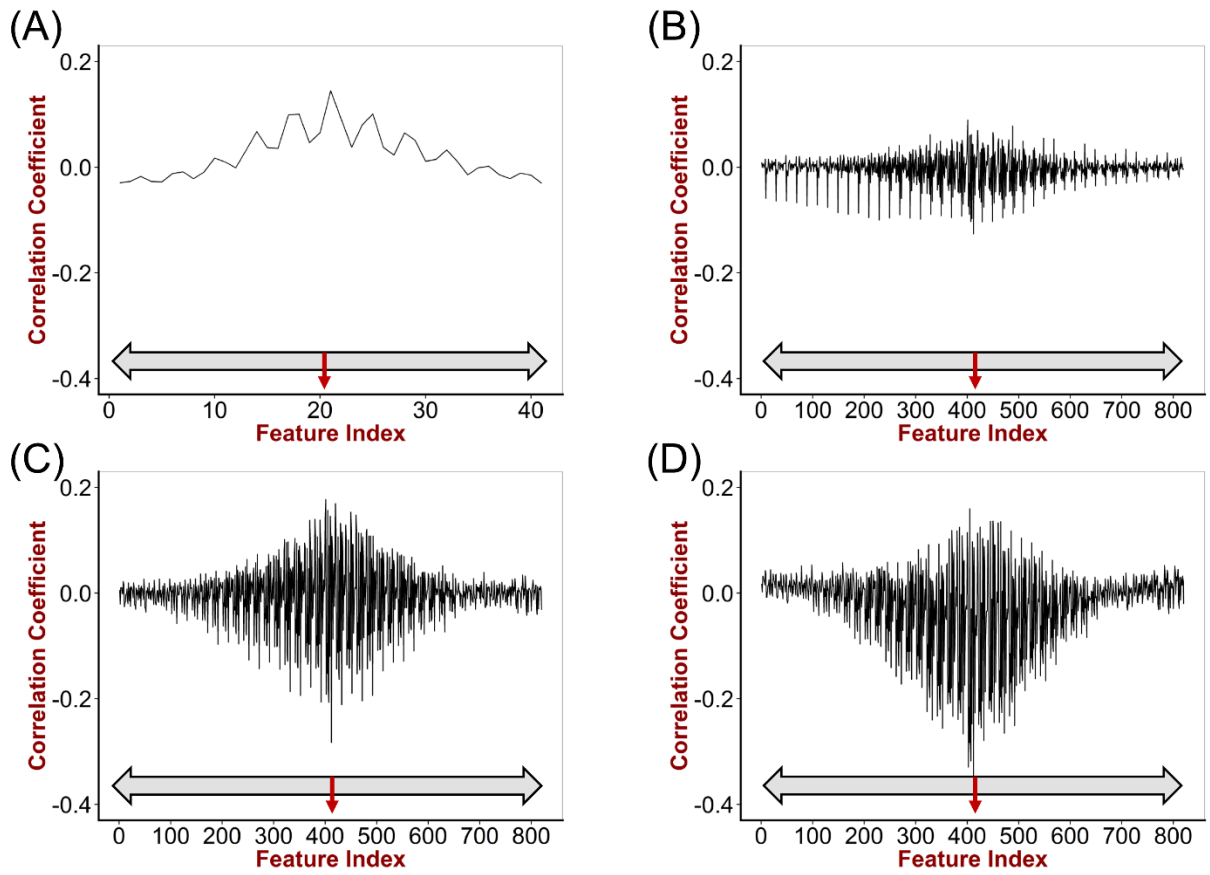


Figure II-2 Correlation of features with WCNs

(A) Correlation of entries in CI feature vector with WCNs; (B) Correlation of entries in LSC feature vector with WCNs; (C) Correlation of entries in BPP feature vector with WCNs; (D) Correlation of entries in PSSM feature vector with WCNs. Each entry in the feature vector is assigned a feature index sequentially such that it starts with 0 for the leftmost residue and ends with 40 (CI) or 820 (other feature vectors) for the rightmost residue (double-headed arrow bar). The red arrow from the arrow bar points to the central residue.

II-3.3 Choosing the optimal window size

One further observation made from Figure II-2 is that features computed from neighboring residues are substantially correlated with the WCN of the central residue and the correlation is dependent on sequence separation. Correlation coefficient decays gradually from very strong at the central residue to very weak at a separation of 15 or more residues. This suggests that there should be an optimal window size such that the signal-to-noise ratio is maximized. For solvent accessibility or WCN prediction, window sizes of 7 (Ahmad *et al.*, 2003), 9 (Illergard *et al.*, 2010), 11 (Ma and Wang, 2015), 15 (Yuan, 2005, Park *et al.*, 2007), 17 (Lai *et al.*, 2013), and 21 (Kinjo *et al.*, 2005) have been used in previous studies. These window sizes are either arbitrarily chosen or obtained by optimization over a relatively short range. We tested a wide spectrum of window sizes ranging from 1 to 41 with a step size of 2. The input feature vector was the PSSM and the architecture of the networks was kept the same across all window sizes.

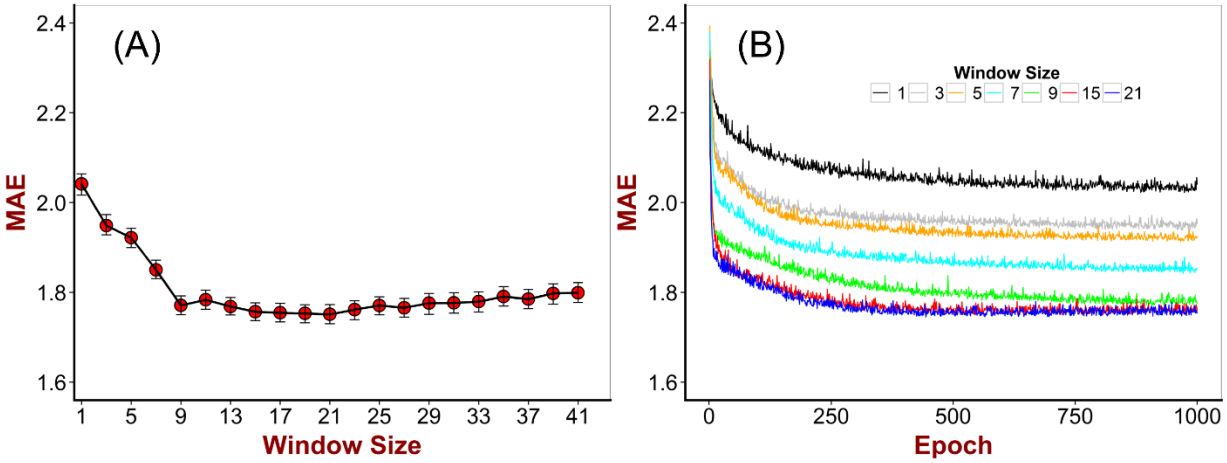


Figure II-3 Effect of window size on the performance of the neural networks

(A) Final MAE on validation sets averaged over cross-validated neural networks; (B) MAEs averaged over cross-validated neural networks as the neural networks were being iteratively trained.

Figure II-3(A) shows the effect of windows sizes on the performance of the neural networks. As window size increases from 1 to 9, the MAE decreases drastically from above 2.0 to below 1.8, a trend similar to the observation made by Park *et al.* (Park *et al.*, 2007). As the window size increases from 9 to 15, the MAE follows a decreasing trend that is slight but noticeable. The MAE rises gradually as the window size is further extended to beyond 21. Interestingly, the MAEs for

window sizes from 15 to 21 remain essentially identical. It was previously proposed that the identities of the residues lying just above (i + 4) and below (i – 4) the target residue on the same helix face are most indicative of the burial status of the central residue (Park *et al.*, 2007). However, our observation suggests that including up to 7 neighboring residues on either side of the central residue consistently improves the performance of the neural network. The fact that the MAE reaches its lowest value when the window size is 15 is especially intriguing given that heptad repeat is one of the signature patterns in helix-helix interactions (Walters and DeGrado, 2006). In fact, Adamian *et al.* developed a highly accurate method for predicting helix-lipid interfaces using heptad motifs as a structural template to assign helical faces of each helical residue (Adamian and Liang, 2006). However, whether the optimal window size arises from heptad repeat needs further investigation.

II-3.4 Dropout prevents overfitting and improves performance

Neuronal dropout is a technique developed for addressing the overfitting problem in neural networks where a large number of parameters are optimized. The key idea is to randomly drop neurons along with their connections from the neural network for each presentation of each training case (Figure II-1(A)) (Srivastava *et al.*, 2014). With this training feature, smaller networks are sampled from an exponential number of networks. Dropout also prevents hidden neurons from co-adapting too much, forcing each hidden neuron to build a relatively independent mapping from feature space onto output space. At test time, a single network without dropout whose weights are multiplied by the probability applied to drop neurons is used (Srivastava *et al.*, 2014). It has been demonstrated that dropout reduces overfitting and improves performance of neural networks on classification tasks in speech recognition and handwritten digit classification (Krizhevsky *et al.*, 2012, Deng *et al.*, 2013, Srivastava *et al.*, 2014).

In order to confirm that dropout reduces overfitting and improves the performance of neural networks for WCN prediction, we compared performances of networks trained with and without dropout. As shown in Figure II-4, compared to the performance of networks trained with dropout, the performance of networks trained without dropout is drastically worse. The MAE for networks trained with dropout converges to a value below 1.8 after ~500 epochs of training, whereas the MAE for networks trained without dropout reaches its lowest value at slightly above 1.8 after only a few epochs of training before it increases almost logarithmically. This observation mirrors the

result obtained from applying dropout to speech and image recognition (Srivastava *et al.*, 2014), confirming that overfitting of the networks for WCN prediction was prevented and performance was improved by using dropout.

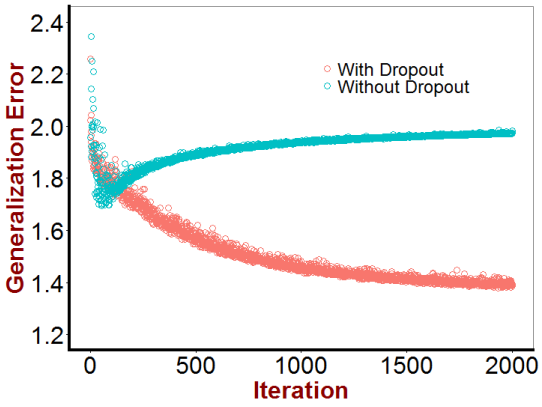


Figure II-4 MAE on validation sets for neural networks trained with or without dropout as learning progresses

II-3.5 Performances of the networks on polytopic HMPs

In light of the investigation on the effects of window sizes, we first examined the performance of the networks for polytopic HMPs using each of the feature vectors separately considering a window size of 15. The performance measures of the networks were averaged over the validation sets. Table 2 summarizes our findings. When using CI as the feature vector, only a moderate PCC of 0.23 was achieved. Switching from CI to LSC increased the performance from PCC = 0.30 to PCC = 0.41. Consistent with the previous conclusion that entries in PSSM generally show stronger correlation with WCNs, the networks achieved a significantly higher PCC (0.69) with PSSM. It is interesting to note that BP gave worse performance (PCC = 0.65) than PSSM despite the fact that it is derived from PSSM. The result of MAE mirrors the observation made on PCC with lower MAE corresponding to higher PCC.

Traditionally, prediction of WCN is treated as a classification problem in which a residue is categorized as either exposed or buried. It is also interesting to see the performance of the current method regarding classification of residue burial status. For computing accuracy and MCC for polytopic HMPs, the median WCN 11.44 in the subset of polytopic HMPs was used as the cutoff. The cutoff was set in this way so that the data set is class-balanced (the number of exposed residues equals that of buried residues) and the accuracy of a classifier that assigns all residues to one particular class is at most 50%. As shown in Table II-2, both accuracy and MCC follow the trend

found in the previous section in the sense that PSSM gives the highest accuracy (75.8%) and MCC (0.52) whereas CI gives the lowest. The final networks were trained with dropout, using PSSM with a window size of 15 as input feature vector. All results and discussions in the rest of the paper refer to the final networks.

Table II-2 Summary of performance measures for WCN prediction

Feature Vector	Size	MAE		PCC		Accuracy (%)		MCC	
		P*	B*	P	B	P	B	P	B
CI	15 × 1	2.33	2.63	0.23	0.33	58.4	50.1	0.18	0.00
LSC	15 × 20	2.18	2.79	0.41	0.15	63.9	50.1	0.28	0.00
BPP	15 × 20	1.79	2.62	0.65	0.28	73.1	51.5	0.47	0.05
PSSM	15 × 20	1.68	2.51	0.69	0.38	75.8	54.2	0.52	0.13

* P: polytopic, B: bitopic

II-3.6 WCNs for bitopic HMPs are difficult to predict

By comparing the performance of the networks on polytopic HMPs to that on bitopic HMPs, we observed that the performance on bitopic HMPs are substantially worse (Table II-2 and Figure A-2 in APPENDICES). The MAE on bitopic HMPs is considerable higher than that on polytopic HMPs (2.51 versus 1.68). The PCC, accuracy, and MCC (using a cutoff of 8.50, which is the median WCN for bitopic HMPs) on bitopic HMPs are significantly lower than those on polytopic HMPs. In fact, 11 out of 12 protein chains with MAE greater than 2.5 are bitopic (Table A-2 in APPENDICES). The reason why WCNs for bitopic HMPs are more difficult to predict is still unclear. One potential explanation could be that the distribution of WCNs for bitopic HPMs is significantly different from that for polytopic HMPs (Figure II-5). Using relative conservation analysis, Zviling *et al.* recently proposed that bitopic HPMs have various interaction modes (Zviling *et al.*, 2007). If this is the case, the interaction modes for bitopic HMPs observed in the data set might only represent one of multiple possible modes (e.g. the buried face of the helix of a bitopic HMP in one complex might be instead the exposed face when being part of another complex). Therefore, WCNs for bitopic HMPs computed based on complex structures observed in the current data set might be biased.

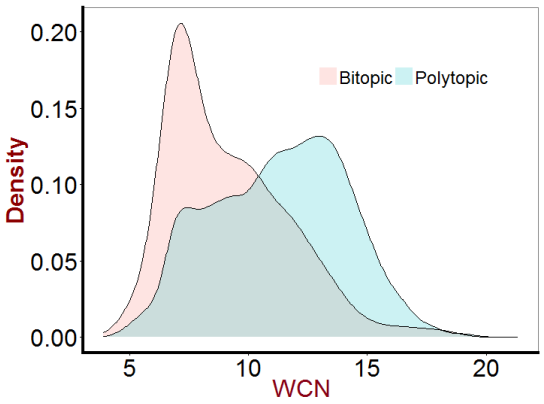


Figure II-5 Distribution of WCNs of bitopic and polytopic HMPs

II-3.7 WCNs for highly exposed or buried TMHs are difficult to predict

One reason why the distribution of WCNs for bitopic HMPs is drastically different from that for polytopic HMPs is that most bitopic HMPs are docked to the surface of large HMP complexes, leading to fewer interacting TMHs then a TMH at the center of a large HMP. In fact, out of the 20 bitopic HMPs in the data set, 17 are localized on the surface of a HMP complex. The fact that the WCN for bitopic HMPs are difficult to predict poses an interesting question: is it a general feature that WCNs for TMHs with fewer interacting TMHs are difficult to predict? In order to answer this question, we computed the MAE for each TMH. We also binned TMHs into groups according their average WCN, assuming that average WCN is a scaled indicator of the number of interacting TMHs. Figure 6 shows that TMHs with very few interacting partners have an increased group-averaged MAE. Interestingly, Figure II-6 also shows that completely buried TMHs have the highest group-averaged MAE.

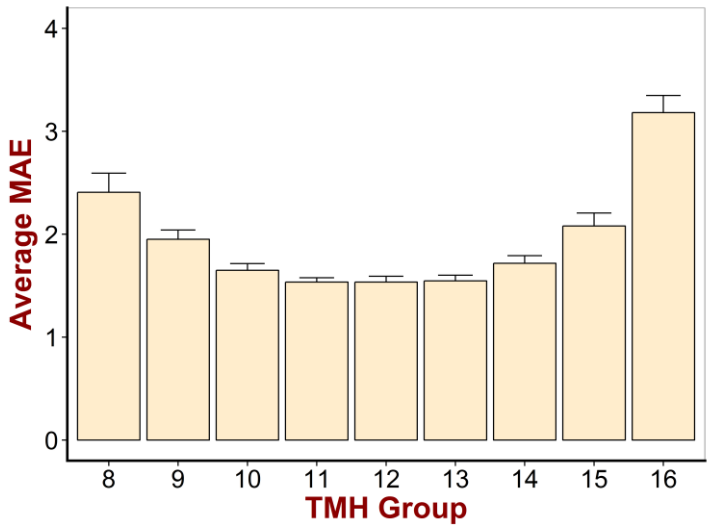


Figure II-6 Group-averaged MAEs for TMHs grouped according to their average WCNs

The x-axis denotes average WCN of a TMH group. For instance, 10 means the group of TMHs that have average WCN between 9 and 10.

II-3.8 WCNs of extremely exposed or buried residues are difficult to predict

In addition to the overall performance, the distribution of MAE was analyzed. The positive skewness of the unimodal density curve for the distribution MAE (Figure II-7(A)) indicates that the model was able to accurately predict WCN for most residues. In fact, 53.5% residues were predicted with an absolute error of less than 1.5, 66.6% residues were predicted with an absolute error of less than 2. Knowing whether the performance of the networks differs for different ranges of WCN is helpful as it indicates how reliable the result is when interpreting a prediction. We grouped residues using the same grouping scheme applied in the previous section and computed the group-averaged MAEs. Similar to the situation with TMHs, Figure II-7(B) shows that MAE is higher toward either end of the residue groups than in the middle. This relationship implies that WCNs for residues in the most buried groups (highest WCN) or the most exposed groups (lowest WCN) are the most difficult to predict.

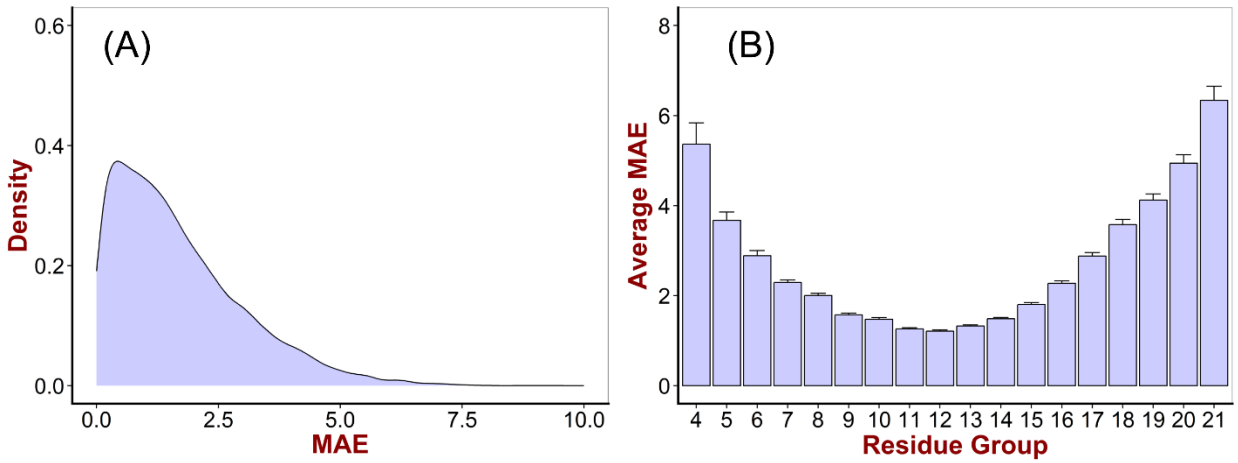


Figure II-7 Group-averaged MAEs for residues grouped according to their WCNs

The x-axis denotes the WCN of a residue group. For instance, 10 means the group of residues that have WCN between 9 and 10.

II-3.9 Amino acid bias in prediction error

In order to examine whether there are amino acid types for which the WCN is more difficult to predict, we computed the amino acid residue-specific MAEs. Figure II-8(A) shows the MAE for each amino acid type. In general, amino acids with charged side chains (Lys, Glu, His, Asp and Arg) have lower MAEs than those with uncharged side chains. This is likely because of the fact that these charged residues are functionally important and are often employed by membrane proteins to bind ligands (Illergard *et al.*, 2011), thus having similar buried states. In fact, the standard deviations of WCNs of these charged residues are among the lowest (Table II-1 and ref. (Adamian and Liang, 2001)). The MAEs for Pro, Ala, and Gly are among the highest and are significantly higher than those of the other residues. Prolines introduces kinks or π-bulges to TMHs (Senes *et al.*, 2004). Alanines and glycines form the sequence motifs of the type AxxxA, GxxxG that are believed to promote close helical packing (Russ and Engelman, 2000). These residues have a highly variable exposure environment as indicated by the high standard deviations of the WCNs (Table II-1 and ref. (Adamian and Liang, 2001)). The correlation between MAEs and the standard deviation of WCNs of amino acid types is 0.84 (Figure II-8b), suggesting that increased variability of exposure is an important determining factor for reduced prediction quality.

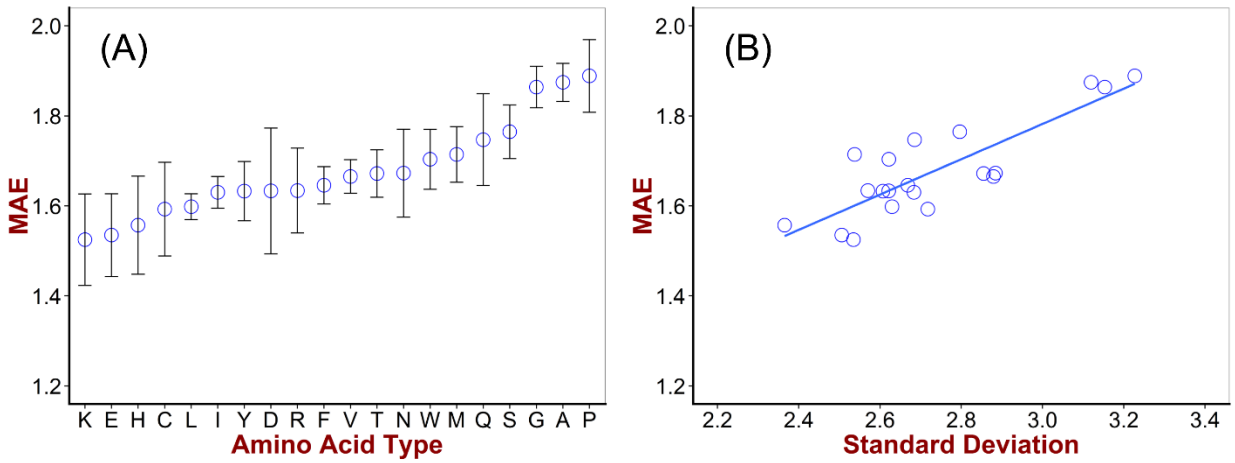


Figure II-8 Amino acid type-specific MAEs and the dependence of MAE on standard deviation of WCNs

(A) amino acid type-specific MAEs; (B) dependence of MAE on standard deviation of WCNs.

II-3.10 Predicted WCNs reveal exposure pattern

An important application of WCN predictors is that they can be incorporated into scoring functions for evaluating *de novo* predicted or homology-modeled 3D protein structures. However, the possibility of this application depends on the fact that predicted WCNs accurately reflect the exposure pattern of a protein. For illustrative purposes, we mapped the experimental and predicted WCNs onto the native structure for two protein chains (3tlwA, 4buoA). 3tlwA is one of the five subunits of the GLIC homopentameric ligand-gated ion channel (Tiefenbrunn *et al.*, 2011) and is among the cases for which the networks achieved the lowest MAE and highest PCC (Figure II-9(A)). 4buoA is a structure of the thermostable agonist-bound G-protein-coupled receptor neurotensin receptor 1 (Egloff *et al.*, 2014) for which the networks also achieved good prediction (Figure II-9(D)). Comparing Figure II-9(B) with (C) and (E) with (F) shows that WCNs predicted by the networks correctly reflect exposure patterns for membrane-facing as well as buried TMHs. The two-phases of membrane-facing TMHs are differentiated by the alternating nature of predicted WCNs. Thus, predicted contacts can be used to eliminate incorrectly predicted 3D structure models where buried TMHs are placed facing the membrane or vice versa.

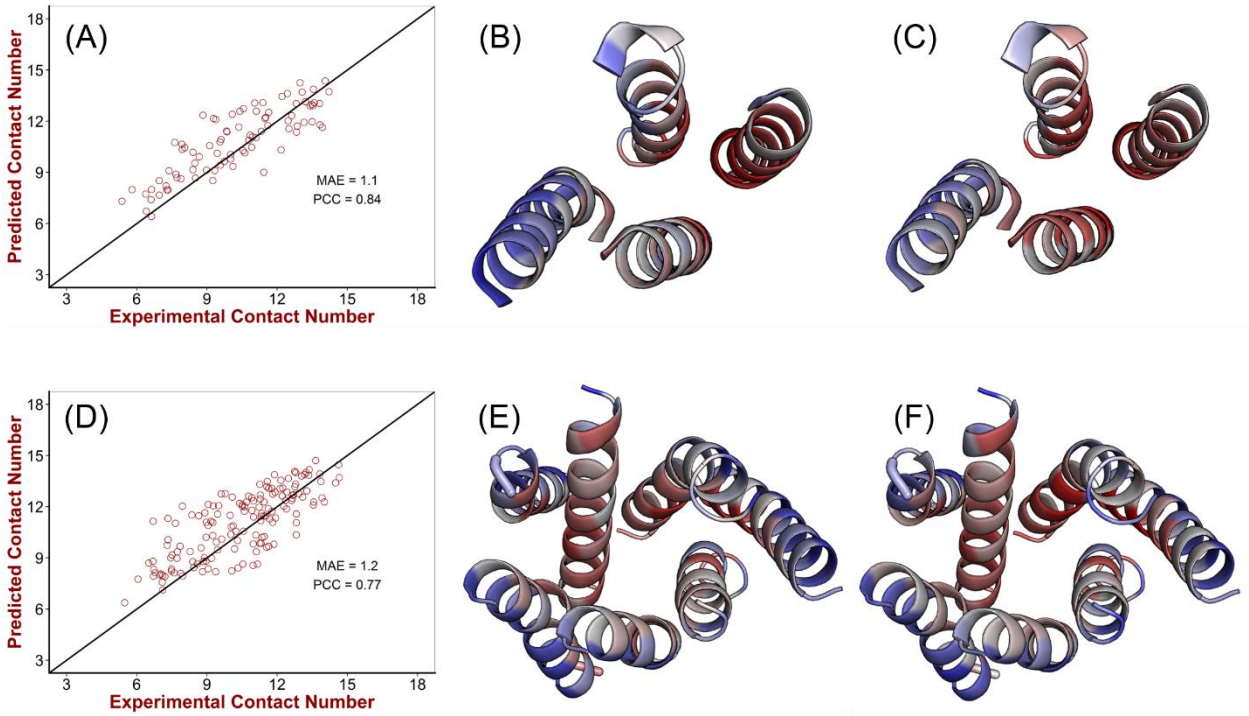


Figure II-9 Predicted WCNs reveal exposure pattern of TMHs

(A) correlation between experimental and predicted WCNs for 3tlwA; (B) mapping of experimental WCNs onto the crystal structure of 3tlwA; (C) mapping of predicted WCN to the crystal structure of 3tlwA; (D) correlation between experimental and predicted WCNs for 4buoA; (E) mapping of experimental WCNs onto the crystal structure of 4buoA; (F) mapping of predicted WCN to the crystal structure of 4buoA. Color scheme: as WCN increases, color changes gradually from blue to red. Only TMHs are shown.

II-3.11 Predicting membrane protein-membrane protein interface

Oligomerization is an essential mechanism by which many membrane proteins function (Kawano *et al.*, 2013). In fact, 49 out of 71 HMPs in the TMH-Expo data set are oligomers. Interaction between membrane protein and membrane protein is a research area that has gained increasing attention from the biochemical community (Miller *et al.*, 2005, Babu *et al.*, 2012). Given a monomer HMP with known structure, it desires to identify interface-forming residues with a reasonable accuracy. As experimental WCNs were calculated from structures where all transmembrane subunits are included, we hypothesized that predicted WCNs be generally higher for interface residues than for non-interface lipid-exposed residues. If our hypothesis proved correct, then interface-forming TMHs can be identified. For evaluating the performance of TMH-Expo on identifying interface residues, we defined a residue as an interface residue if $WCN_o - WCN_m \geq 1$, where WCN_o is the WCN in oligomeric state and WCN_m is that in monomeric state.

A residue is predicted as interface residue if $WCN_p - WCN_m \geq 1$, where WCN_p is the predicted WCN. The cutoff value 1 was chosen to reduce the chance of including residues on the protein core-buried face of a TMH as interface residues. 16.3% residues in the data set satisfied this definition. For classifying interface residues (Table A-3 in APPENDICES), TMH-Expo achieved an overall accuracy of 68.6%, and a sensitivity of 76.8%, significantly better than the performance reported in a similar study (Illergard *et al.*, 2010). One should be aware of the high FPR of TMH-Expo (33.0%), a complication that could be accounted for by the fact that the oligomeric state of many HMPs is not unambiguously defined (Duarte *et al.*, 2013).

As an example of predicting membrane protein-membrane protein interface residues, we investigated the performance of TMH-Expo for the subunit (4al0A) of the homotrimeric microsomal prostaglandin E2 synthase (Sjogren *et al.*, 2013). 4al0A has a similar FPR (32.1%) to the overall FPR of TMH-Expo. As shown in Table II-3, out of the 85 TMH residues, 66 were correctly classified, giving an overall accuracy of 77.7%. Among these 32 interface residues, 30 were identified, giving a sensitivity of 93.8 %. To visualize the prediction, we highlighted interface residues identified with experimental WCNs (Figure II-10(A)) and those identified with predicted WCNs (Figure II-10(B)) on the native structure. Despite the high FPR, most false positives can be reasonably eliminated if we only consider residues on the exposed face of a TMH.

Table II-3 Performance of interface residue identification of TMH-Expo on 4al0A

		Predicted		
		Interface	Non-interface	Total
Experimental	Interface	30	2	32
	Non-interface	17	36	53
	Total	47	38	85

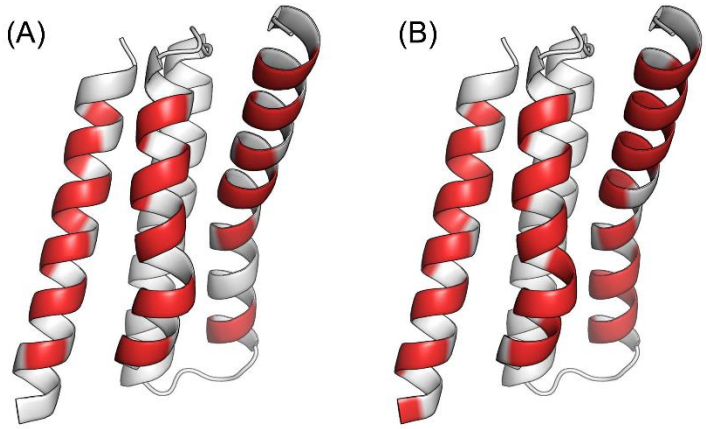


Figure II-10 Predicted WCNs reveal interface-forming residues of 4al0A

(A) Mapping of interface residues (colored in red) identified with experimental WCNs onto the crystal structure of 4al0A; (B) mapping of predicted interface residues (colored in red) onto the crystal structure of 4al0A.

II-3.12 Comparison with other WCN predictors

To the best of our knowledge, TMH-Expo is the first attempt that has been made to predict WCNs for membrane proteins. Therefore, a direct comparison of TMH-Expo with any of the other existing methods is not possible. To give an approximate sense of the performance of TMH-Expo, we compared TMH-Expo with two notable WCN predictors developed for soluble proteins. Using linear regression analysis, Kinjo et al. developed a real-valued WCN predictor with a PCC of 0.63 (Kinjo *et al.*, 2005), outperformed by TMH-Expo. Yuan developed a support vector regression-based predictor with a PCC of 0.70 (Yuan, 2005), slightly better than TMH-Expo. However, it should be noted that the performance of Yuan's method might be favorably biased since the data set was not split in a way such that proteins in the same superfamily stay within the same subset. In addition, the structural repository of soluble proteins is significantly bigger than that of HMPs, making the training set for soluble proteins more informative.

We also trained neural network models for RSA prediction using PSSM as feature vector and the same training parameters as with training networks for WCN prediction. For RSA prediction, TMH-Expo achieved a PCC of 0.58 for polytopic HMPs. Since both accuracy and MCC are dependent on the cutoff value applied, it is rather arbitrary to make comparisons based on these two performance measures. We therefore approximately (since the data set employed in different study varies) compared our method to predictors for which PCC was reported. Yuan *et al*. developed a support vector regression-based predictor termed $ASAP_{mem}$ with a PCC of 0.66 for

TM helical residues (Yuan *et al.*, 2006). A random forest-based method recently reported by Wang et al. achieved a PCC of 0.68 (Wang *et al.*, 2012). Although these two methods reportedly have better performance than TMH-Expo on RSA prediction, it should be pointed out that the cross-validation scheme employed in these studies might favorably biased the performance. In fact, using the same cross-validation scheme, Illergård et al. trained a RSA predictor MPRAP which achieved the same PCC as TMH-Expo (Illergard *et al.*, 2010).

II-4 Limitations and future directions

In the current implementation of the algorithm, total WCN is computed by summing over contributions made by residues inside a sphere centered at the C_β atom of the residue of interest. The contribution is assigned to each residue in a distance-dependent way such that close neighbors have an increased weight when compared to distant neighbors. This approach mirrors the distance-dependence of van der Waals and electrostatic interactions and is superior to the use of a single cut-off distance. Shortcomings of the current implementation include that the spatial distribution of neighboring residues is not taken into account (Durham *et al.*, 2009).

Another limitation comes from the coarse-grained C_β representation of the side chains in which size and bulkiness of side chains is ignored. While representing side chain atoms as a single 'superatom' improves computational efficiency and is necessary in early stages of de novo 3D structure prediction, it could result in loss of important structural information and lead to biased estimate of WCNs. For instance, residues with a bulky side chain have longer C_β-C_β distances than small residues. Thus, the average WCNs for bulky residues might be underestimated (Gimpelev *et al.*, 2004). When the information about the spatial distribution of neighboring residues is needed, a computationally slightly more demanding quantity called "neighbor vector" could be employed (Durham *et al.*, 2009). The neighbor vector is a vector associated with each residue whose direction and magnitude not only depend on the number of neighboring residues but also on the spatial distribution.

Similar to predicted lipid exposure, which has been leveraged to improve the inference of the rotational angles of TMHs (Lai *et al.*, 2013), we expect that accurately predicted WCNs would also improve the accuracy of predicted TMH-TMH packing interactions.

II-5 Conclusion

We have developed a dropout neural network-based WCN and RSA predictor TMH-Expo for HMPs. TMH-Expo is the first work that reports WCN prediction for HMPs. Trained on an expanded non-redundant data set of HMPs with five-fold cross-validation, TMH-Expo achieved an unprecedented PCC of 0.69 between experimental and predicted WCNs. We have also shown that the training was benefitted from using neuronal dropout. With neuronal dropout, overfitting was significantly reduced, and the performance was improved. Detailed examination of MAEs and PCCs indicated that it is generally easy to predict WCNs for polytopic HMPs than for bitopic HMPs. Mapping of predicted WCNs onto structure demonstrated that WCNs predicted by TMH-Expo reflect exposure patterns of TMHs and reveal interface-forming TMHs. This reinforces the idea of incorporating predicted WCNs for predicting helix-helix packing and protein-protein docking. *De novo* protein folding and protein-protein docking studies leveraging WCNs predicted by TMH-Expo are currently ongoing.

II-6 Supporting Information

HMP chains included in the TMH-Expo data set, an illustration of feature vectors, summary of 12 poorly predicted protein chains, distribution of prediction quality for of bitopic and polytopic HMPs, performance of TMH-Expo on identifying interface residues. This material is available in APPENDICES.

II-7 Notes

TMH-expo has been integrated into the Biochemical Library (BCL) software suite that is being actively developed. It is also available via a webserver at http://www.meilerlab.org/index.php/servers. The BCL software suite is available at http://www.meilerlab.org/bclcommons under academic and business site licenses. The BCL source code is published under the BCL license and is available at http://www.meilerlab.org/bclcommons.

II-8 Abbreviations

HMP, helical membrane protein; TMH, transmembrane helix; BCL, biochemical library; PSSM, position-specific scoring matrix; PCC, Pearson's correlation coefficient; MCC, Matthew's correlation coefficient; MAE, mean absolute error; MSA, multiple sequence alignment; RSA, relative solvent accessibility; ASA, absolute solvent accessibility.

III. IMPROVING PREDICTION OF HELIX–HELIX PACKING IN MEMBRANE PROTEINS USING PREDICTED CONTACT NUMBERS AS RESTRAINTS

This chapter has been published in (Li *et al.*, 2017a).

III-1 Introduction

Helical membrane proteins (HMPs) are essential components of a living cell. They play crucial roles in orchestrating the interactions of the cell with its environment; for example, by mediating cellular signaling, regulating ion gradients, and facilitating the transfer of molecules across the cell membrane. It was estimated that 20–30% of genes in most genomes encode HMPs (Krogh *et al.*, 2001). About 50% of therapeutics on the market target HMPs (Overington *et al.*, 2006). The availability of a three-dimensional (3D) structure of a HMP not only improves our understanding of how the protein works at the atomic level (Li *et al.*, 2012) but also facilitates the development of new therapeutics (Xiong *et al.*, 2011, Zhan *et al.*, 2011, Li *et al.*, 2014). Despite great progress in experimental techniques for determining HMP structures, only ~2 % structures in the protein databank are HMPs (Weiner *et al.*, 2013), highlighting the fact that HMP structure characterization is still a challenge. Further, experimental data for HMPs are often of limited resolution, requiring computational methods to elucidate atomic-level details. Similarly, not all biologically relevant conformations of HMPs – which tend to be very flexible – can be studied experimentally. Likewise, accurate computational methods for HMP structure prediction are a complement to existing experimental techniques to enable HMP structure determination from limited experimental data (Weiner *et al.*, 2014, Fischer *et al.*, 2015).

A commonly used computational approach for predicting protein tertiary structure is comparative modeling. However, a sequence identity of at least 25% between target and template proteins is recommended to give reliable models (Cavasotto and Phatak, 2009). Because the fold of most HMPs are unknown and it was estimated that comparative modeling covers at most 10% of HMPs (Hopf *et al.*, 2012), a few *de novo* methods have been developed, such as Rosetta-Membrane (Yarov-Yarovoy *et al.*, 2006) and BCL::MP-Fold (Weiner *et al.*, 2013). Rosetta-Membrane assembles models helix-by-helix starting from a helix near the middle of the protein (Yarov-Yarovoy *et al.*, 2006). For HMPs with ~150 residues or less, Rosetta-Membrane achieved RMSD100 (root-mean-square distance normalized to a sequence of 100 residues) values of < 4 Å to experimental structures. However, the prediction accuracy with respect to helix rotation around

the main axis was either not evaluated or very poor (Weiner *et al.*, 2013). BCL::MP–Fold uses secondary structure element (SSE) pools and inserts helices across the membrane to build complete models. It achieved RMSD100 values to the experimental structure in the range of 3 to 8 Å for most benchmark HMPs (Weiner *et al.*, 2013). For models assembled by BCL::MP-Fold, even though TMHs are predicted to span the membrane with the correct topology, ~40% were reported to contain helices with incorrect rotation (Weiner *et al.*, 2013). For example, contact-forming, buried residues are sometimes rotated toward the membrane. For HMP models to be useful in applications such as structure-based drug design, accurate modeling of helix rotation is essential.

One approach to improving the accuracy of *de novo* tertiary structure prediction is to incorporate restraints (Barth *et al.*, 2009). These restraints may be experimental, such as NMR chemical shifts (Weiner *et al.*, 2014) and electron-paramagnetic resonance (EPR) accessibilities (Fischer *et al.*, 2015), or computational, such as predicted residue–residue contacts (Barth *et al.*, 2009, Marks *et al.*, 2011, Hopf *et al.*, 2012, Nugent and Jones, 2012, Kosciolek and Jones, 2014). For example, Fischer *et al.* recently showed that using either experimental or simulated EPR accessibility increases the likelihood of sampling native-like HMP folds and improves the accuracy of predicting helix rotations (Fischer *et al.*, 2015). Residue–residue contacts derived from experiments or accurate computational predictions also provide substantial guiding information for sampling. For instance, Evfold_membrane developed by Hopf *et al.* enables *de novo* prediction of tertiary structures of 25 HMPs by incorporating amino acid covariation extracted from evolutionary sequence record (Hopf *et al.*, 2012).

Residue weighted contact number (WCN) is a real-valued quantity that measures the degree of local packing of a residue within the protein tertiary structure. The WCN of a given residue was originally computed by applying a clear distance cutoff and considering indiscriminately residues within the cutoff (Nishikawa and Ooi, 1980, Nishikawa and Ooi, 1986). Later improvements incorporated various distance-dependent weighting schemes to account for the distance-dependent nature of residue–residue interactions (Kinjo *et al.*, 2005, Lin *et al.*, 2008, Durham *et al.*, 2009). WCNs have been used to derive protein dynamic properties such as B-factor profile (Lin *et al.*, 2008). Studies have also shown that WCN is the main structural determinant of site-specific substitution rates of proteins (Echave *et al.*, 2016). Although it has been suggested that WCNs

could help in tertiary structure prediction, to our knowledge, no studies on tertiary structure prediction have explicitly incorporated WCNs.

The WCNs of interfacial TMHs (peripheral TMHs of a helical bundle) follow a signature periodic trend (Figure III-1(A)). Importantly, the WCN signature of a TMH is tightly coupled to its rotation: even a small perturbation of the helix rotation will disrupt the WCN signature (Figure III-1(B)). Hence, the WCN signature of a TMH should give a strong constraint over its rotation. However, experimental WCNs are not available until the tertiary structure of the protein is determined. Very recently, we developed a dropout neural network-based method, BCL::TMH-Expo, specifically for predicting WCNs for HMPs (Li *et al.*, 2016). WCNs predicted by BCL::TMH-Expo correlate well with WCNs computed from experimental structures and mirror exposure patterns of TMHs (Li *et al.*, 2016). In this study, WCNs predicted by BCL::TMH-Expo were incorporated into the empirical scoring function of BCL::MP-Fold in the form of restraints to improve prediction of helix–helix packing. We tested this method on a set of 15 benchmark HMPs that span a wide range of fold complexity.

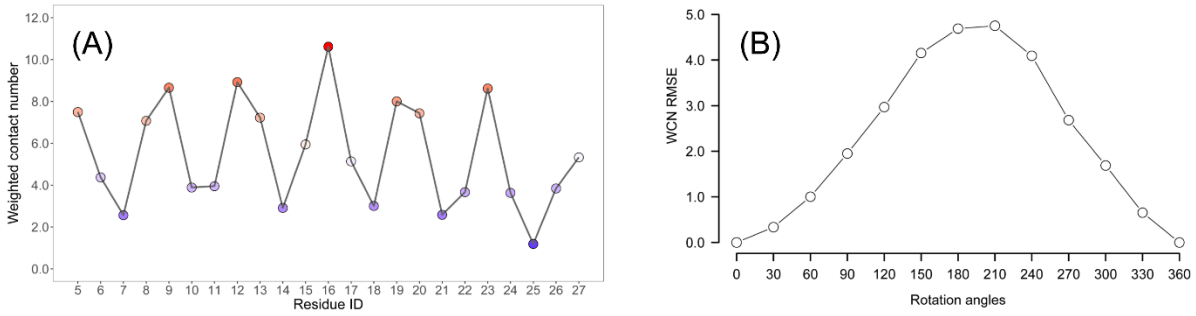


Figure III-1 An example of WCN signature of a TMH and its tight coupling to the rotation about the helix normal

(A) The native WCN signature of the first transmembrane helix of the bacteriorhodopsin (PDB ID: 1m0l); (B) The root-mean-square error of the WCN signature of the helix to that of the native signature after each rotation.

III-2 Materials and Methods

III-2.1 Benchmark set

A set of 15 multi-spanning HMP subunits were carefully selected to assess whether using WCN restraints can improve the prediction of helix–helix packing. This set consists of HMP subunits that are both structurally and functionally diverse (Table III-1). Pairwise sequence identity is 30% or less. Sequence length ranges from 156 to 467 residues. The number of TMHs ranges from 4 to 10. As a measure of the size of transmembrane domains, the number of TMH residues were also computed for each target. None of these HMPs was used in the training set of BCL::TMH-Expo or had a sequence identity of more than 30% to any of the HMPs in the training set of BCL::TMH-Expo. This benchmark set contains diverse folds ranging from simplistic four-helix bundles and 7-TM receptors, up to proteins with 10 TMHs or helices in reentrant regions. Six of these HMPs are homo-oligomers. Due to the complexity of folding oligomers, we limited the scope of the present investigation to consider only a single subunit of each oligomer.

Table III-1 Summary of the benchmark set

PDB ID	Structure Method	Resolution	Length	TMH	TMH Residue	PCC	MAE	Oligomeric State
1OED	EM	4.0	227	4	104	0.35	2.23	Homopentamer
1OKC	X-ray	2.2	292	6	214	0.39	2.37	Monomer
1PV6	X-ray	3.5	189	6	163	0.62	1.66	Monomer
1PY6	X-ray	1.8	249	7	177	0.72	1.29	Monomer
1U19	X-ray	2.2	348	7	173	0.58	1.63	Monomer
2BL2	X-ray	2.1	156	4	119	0.65	2.42	Homo 10-mer
2K73	NMR	NA	164	4	99	0.45	1.78	Monomer
2O9G	X-ray	1.9	234	6	166	0.69	1.74	Homotetramer
2Y01	X-ray	2.6	315	7	185	0.76	1.45	Monomer
3M71	X-ray	1.2	314	10	242	0.85	1.33	Homotrimer
3QAP	X-ray	1.9	239	7	168	0.69	1.35	Monomer
3UG9	X-ray	2.3	333	7	194	0.45	1.75	Homodimer
3UON	X-ray	3.0	467	7	183	0.66	1.60	Monomer
4A2N	X-ray	3.4	194	5	123	0.58	1.67	Monomer
4O6Y	X-ray	1.7	230	6	156	0.58	1.55	Homodimer
Mean			265	6.6	164	0.60	1.72	

PCC: Pearson correlation coefficient; MAE: mean absolute error; EM: electron microscopy; X-ray: x-ray diffraction; NMR: nuclear magnetic resonance; NA: not applicable

III-2.2 Computation of experimental and predicted WCNs

The details of the algorithm for computing WCNs from experimental strictures can be found in two previous studies (Durham *et al.*, 2009, Li *et al.*, 2016). Briefly, the experimental WCN of residue *i* was computed as a weighted sum of contacts contributed by residues over the entire protein:

$$WCN_{\mathrm{i}} = \sum_{\mathrm{j}\in|\mathrm{j}-\mathrm{i}|>3}^{\mathrm{n}} w_{\mathrm{ij}} \qquad \textbf{III-1}$$

where w_{ij} is the contribution made by residue j and is assigned in a distance-dependent manner such that short-range contacting residues have higher contribution than long-range contacting ones. Residues whose C_β atom is within 4.0 Å to the C_β atom of the residue of interest are assigned a contribution of 1.0; those with a distance longer than 11.4 Å are assigned a contribution of 0. Any residue 4-11.4 Å is assigned a contribution between 0.0 and 1.0 according to a smooth transition function (see Eq. II-1) (Durham *et al.*, 2009). Only residues separated by more than three residues along the sequence were considered in the calculation to reduce the bias due to sequence proximity and local secondary structure. Experimental WCNs were calculated based on structures retrieved from the OPM (Orientations of Proteins in Membranes) database (Lomize *et al.*, 2006).

Although a relatively low sequence identity (30%) was maintained while compiling a list of benchmark protein chains to reduce the homology between the modeling benchmark set and the training set for BCL::TMH-Expo, such level of sequence identity alone may not be sufficient to exclude homology among protein chains. In fact, substantial remote homology could still exist at this level placing HMPs in the same structural superfamily (Jaakkola *et al.*, 1999). Such remote homology between proteins in the training set and proteins in the modeling benchmark set can lead to an optimistic estimate of the performance for new folds. As a way of preventing such optimism, the original training set for BCL::TMH-Expo was partitioned such that each SCOP superfamily (Murzin *et al.*, 1995) forms its own subset that contains all its members and no members from other SCOP superfamilies. Predicted WCN of each residue of a modeling benchmark protein was then obtained through a specific variant of the neural network-based WCN predictor BCL::TMH-Expo. This variant was trained using all remaining proteins after excluding the subset of proteins that share the same SCOP superfamily as the modeling benchmark protein from the original

training set of BCL::TMH-Expo. For example, for predicting the WCNs for 3UON, all proteins that are in the same SCOP superfamily as 3UON were removed from the original training proteins of BCL::TMH-Expo and a neural network was trained using the remaining proteins. The WCNs were then predicted using this retrained neural network. This strategy was applied to each protein in the modeling benchmark set. Note that BCL::TMH-Expo method is a dropout neural network-based algorithm that predicts WCNs for HMPs. It uses the position-specific scoring matrix (PSSM) (Gribskov *et al.*, 1987) derived from multiple sequence alignment (MSA) by PSI-BLAST (Altschul *et al.*, 1997) as predictive features and outputs residue-specific WCN. The MSA for each protein chain in the benchmark set was obtained by searching the UniRef50 (Suzek *et al.*, 2007) non-redundant sequence database with PSI-BLAST for five iterations (Altschul *et al.*, 1997). The E-value inclusion threshold was set to 10^{-2}. Floating point-valued PSSM was generated from PSI-BLAST checkpoint files using the source code (chkparse.c) adapted from PSIPRED (McGuffin *et al.*, 2000). Predicted WCN was obtained by feeding the floating point-valued PSSM to BCL::TMH-Expo.

III-2.3 Incorporating WCNs as restraints in *de novo* structure prediction

The *de novo* membrane protein structure prediction algorithm BCL::MP-Fold (Weiner *et al.*, 2013) developed by adapting the original algorithm BCL::Fold (Karakas *et al.*, 2012) for membrane proteins was used to assemble 3D models. BCL::MP-Fold assembles 3D models by drawing TMHs from a pool of predicted TMHs. TMH pools were created from predictions made by the combined membrane association and secondary structure predictor BCL::MASP (Jeffrey L Mendenhall and Meiler, 2014). A Monte Carlo minimizer with Metropolis criterion (Metropolis *et al.*, 1953) was used to sample models with low energy. To use WCNs to guide sampling of helix-helix packing, a WCN-based penalty score was added to the knowledge-based scoring function of BCL::MP-Fold:

$$Score = \sum_i w_i \times S_i + w_p \times Penalty \qquad \textbf{III-2}$$

where S_i represents each of the individual knowledge-based potentials previously derived and w_i is the associated weight. These potentials have been detailed in prior studies (Woetzel *et al.*, 2012, Weiner *et al.*, 2013). The restraint scoring term was defined using the following formula:

$$enalty = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\delta_i^2} \qquad \textbf{III-3}$$

where n is the number of residues in the assembled structural model, δ is the difference between the WCN used as restraint and the WCN calculated from the assembled structural model. w_p is the corresponding weight of the penalty. An optimal balance between the knowledge-based potentials and the penalty score is critical for correcting helix rotation while sampling native-like folds. If the weight for the restraint penalty is too low, its capacity of correcting helix rotation is reduced, if the weight is too high, it dominants other scoring terms. An empirical approach, in which a range of w_p values were systematically tested in preliminary sampling, was used to determine a near-optimal weight. Finally, five thousand models were assembled for each target in the benchmark set. The procedure for generating 3D models is summarized in Figure III-2.

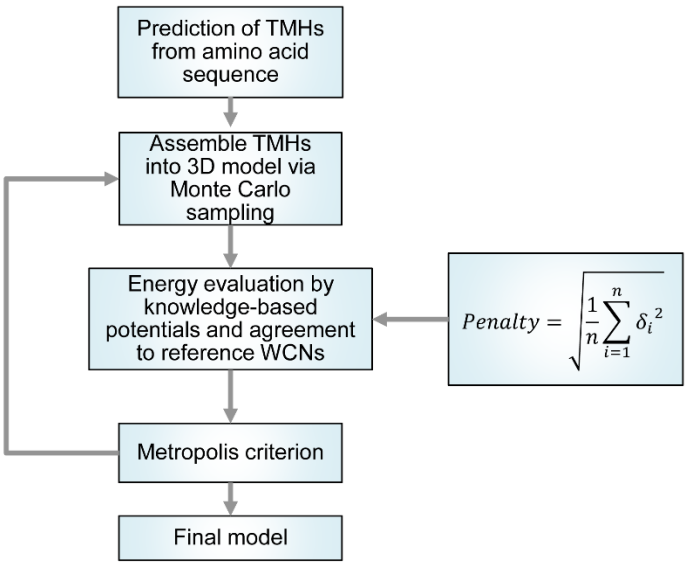


Figure III-2 Protocol for assembling 3D models. BCL::MP-Fold predicts the tertiary structure of a HMP by assembling predicted TMHs in the 3D space

In the first step, the TMHs are predicted using the neural network-based membrane association and secondary structure prediction algorithm BCL::MASP. Predicted TMHs are assembled into a 3D model, and perturbed using a Monte Carlo sampling algorithm. The energy of the model after each perturbation is evaluated by knowledge-based potentials and agreement to WCN restraints. The perturbation is subjected to the Metropolis criterion and is either accepted or rejected depending on the difference between the energies before and after the perturbation. This process is repeated

for a specific number of iterations or until the maximum number of 2000 iterations without energy improvement is reached.

III-2.4 Metrics for measuring of model quality

Root-mean-square distance (RMSD) gives a useful impression of the similarity between two structures if there is only a slight difference between their conformations. Unfortunately, a small perturbation in just one part of the protein (for instance, off position of a short loop) can lead to a large RMSD and it would seem that one structure substantially differs from the other. In order to address this issue, several quality measures have been introduced among which RMSD100 (Carugo and Pongor, 2001) is commonly used. RMSD100 is a normalized, sequence length-independent version of RMSD calculated using:

$$RMSD100 = \frac{RMSD}{1 + \ln\sqrt{\frac{n}{100}}} \qquad \textbf{III-4}$$

where *n* is the number of residues superimposed. Using RMSD100 as an indicator of structural variability reduces the influence of the intuition that larger proteins are more likely to differ from one another (Carugo and Pongor, 2001). In this study, RMSD100 was computed over the C_α atoms of all TMH residues.

A metric called contact recovery (CR), defined as the percentage of native contacts recovered in the assembled 3D model, was used to measure the accuracy of helix rotations in our previous study (Weiner *et al.*, 2013). However, the previous definition does not account for false positive contacts (FPC), which may be prevalent in 3D models assembled in a globular shape when the real shape of the protein is extended or rod-like and it has helices or strands that are somewhat "detached" from its main domain. In such cases, these "detached" secondary structure fragments could potentially be packed against the main domain of the protein by the folding algorithm, and thus, making a substantial fraction of FPCs. Thus, we redefined CR as the F1-score. Being the harmonic mean of precision and recall, the F1-score accounts for FPCs by weighting precision and recall equally:

$$ontact\ Recovery = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \textbf{III-5}$$

where

$$Precision = \frac{TPC}{TPC + FPC} \quad \textbf{III-6}$$

and

$$Recall = \frac{TPC}{TPC + FNC} \quad \textbf{III-7}$$

TPC (true positive contacts) denotes the number of contacts observed in the experimental structure that are correctly predicted in the assembled model and FNC (false negative contacts) is the number of contacts in the experimental structure that are missed in the assembled model. Two residues are considered in contact if they are separated along the sequence by at least 12 residues and the distance between their C_β atoms is within 8 Å. CR reaches its best value at 100% and worst at 0%.

III-2.5 Computation of enrichment

The enrichment was used to measure how capable a scoring function is to select the most accurate models from a pool of models. To calculate enrichment, models of a given set S are sorted by their CR values. The top 10% of the models with the highest CR values are put into the set *T* (true) and the rest of the models are put into the set *F* (false). The models in S are then sorted by their evaluated score. The top 10% of models with the lowest score are put into the set *P* (positive) and the rest are put into the set *N* (negative). The intersection of sets *T* and *P* are models that are correctly identified by the scoring function and referred to as *TP* (true positives). The intersection of sets *F* and *P* are models that are incorrectly identified by the scoring function and are referred to as *FP* (false positives). The enrichment value is then computed using the following formula:

$$Enrichment = \frac{TP}{TP + FP} \Big/ \frac{P}{P + N} \quad \textbf{III-8}$$

Intuitively, $\frac{P}{P+N}$ represents that probability of obtaining a native-like model when choosing a model from *S* at random, whereas $\frac{TP}{TP+FP}$ represents the probability of obtaining a native-like model when choosing from a set of models below an energy cutoff. By our experimental design, $\frac{P}{P+N}$ has a constant value of 0.1, and therefore, the maximum enrichment value that can be achieved is 10.

III-3 Results and Discussion

III-3.1 Predicting WCNs for HMPs in the benchmark set

Table III-1 shows the Pearson correlation coefficient (PCC) between experimental and predicted WCNs as well as the mean absolute error (MAE) of predicted WCNs for each target in the modeling benchmark set. The average PCC and the average MAE over the modeling benchmark set were 0.60 and 1.72 respectively. Notably, the WCNs for three proteins, namely 1PY6, 2Y01, and 3M71, were predicted with a PCC > 0.70. Whereas, for 1OED, 1OKC, 3UG9, and 4A2N, the PCCs were below 0.50. Factors affecting the accuracy of WCN prediction include oligomeric state, whether the protein chain is bitopic, and other factors that had been discussed previously in detail (Li *et al.*, 2016). To illustrate the agreement between experimental and predicted WCNs and visualize the predictions, the experimental and predicted WCNs of 1PY6 were plotted and mapped onto its experimental structure. As shown in Figure III-3(A), the predicted WCNs of 1PY6 are in close agreement with experimental WCNs, particularly in transmembrane regions (vertical gray bars). As expected, predicted WCNs generally distinguish between the exposed and buried faces of helices (Figure III-3(B) and (C)). We, therefore, reasoned that the native rotation of helices can be confined by forcing them to satisfy predicted WCN, thus improving the prediction of helix–helix packing.

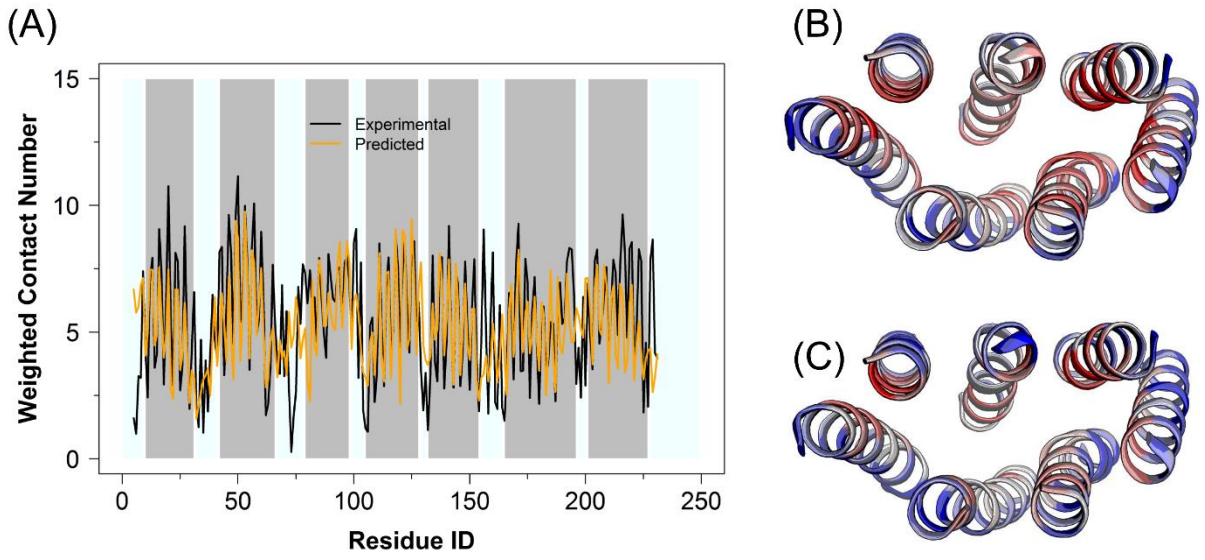


Figure III-3 Agreement between experimental and predicted WCNs of 1PY6

(A) Experimental and predicted WCNs plotted against residues sequence positions, (B) Experimental WCNs mapped onto structure; (C) Predicted WCNs mapped onto structure. Color scheme in (B) and (C): as WCN increases, color changes gradually from blue to red. Only TMHs are shown.

III-3.2 Incorporation of WCNs significantly improved CR

The following three CR-based parameters were compared among the three simulation groups (E: with experimental WCNs, P: with predicted WCNs, N: without WCNs):

β_{CR}: the highest CR achieved,

μ_{CR}: the average of the 10 highest CR values,

π_{20}: the percentage of models with a CR greater than 20%,

β_{CR} and μ_{CR} measure how accurate the best-assembled models can be, whereas π_{20} measures how often an accurate model can be sampled.

As summarized in Table III-2, model quality is generally improved using WCNs as restraints. Specifically, μ_{CR} is improved for all targets when models were assembled using predicted WCNs as restraints, and π_{20} is improved for all but two targets (1OKC and 2O9G). β_{CR} is improved for all targets except 4O6Y and by an average amount of 8.07% and μ_{CR} is improved by an average amount of 8.04% compared to folding without WCN restraints. A substantial increase in μ_{CR} (>5%) is seen for 10 of the 15 targets, with 4 of the targets (1PY6, 2K73, 3QAP, and 4A2N)

showing over 10% of improvement. By using WCN restraints, not only the best models are more accurate, but the probability that accurate models are sampled is also increased. For example, comparison of π_{20} among groups shows that π_{20} is increased by 6.75% on average when folded with predicted WCNs compared to folded without WCNs. It is worth noting that for 3 targets (2Y01, 3M71, and 3UON), models with CR greater than 20% were not sampled ($\pi_{20} = 0$) without WCN restraints but sampled with noticeable frequency with predicted WCNs as restraints. Experimental WCNs further improve CR, for example, β_{CR} is improved by an average amount of 17.78% when using experimental WCNs as restraints. In summary, both experimental and predicted WCNs enable strongly significant improvements in CR of folded protein models ($p < 0.01$, paired t-test).

Table III-2 Summary of contact recovery

	β_{CR} (%)			μ_{CR} (%)			Relative Improvement in μ_{CR} (%)	π_{20} (%)		
Target	**E**	**P**	**N**	**E**	**P**	**N**	$\frac{\mu_{CR(P)} - \mu_{CR(N)}}{\mu_{CR(N)}} \times 100$	**E**	**P**	**N**
1OED	**73.10**	**38.02**	**28.93**	**70.59**	**32.88**	**23.47**	**40.09**	**51.77**	**5.17**	**0.33**
1OKC	*18.34*	*10.78*	*9.96*	*14.81*	*9.14*	*8.17*	*11.87*	*0.00*	*0.00*	*0.00*
1PV6	**31.55**	**30.02**	**21.90**	**26.93**	**23.09**	**17.06**	**35.35**	**1.37**	**0.35**	**0.03**
1PY6	**54.65**	**41.86**	**22.31**	**44.45**	**35.56**	**20.08**	**77.09**	**13.01**	**10.31**	**0.11**
1U19	**30.28**	**25.31**	**20.46**	**26.98**	**23.59**	**16.57**	**42.37**	**2.43**	**1.87**	**0.04**
2BL2	**68.40**	**59.29**	**54.50**	**66.78**	**55.22**	**49.63**	**11.26**	**76.26**	**50.98**	**29.80**
2K73	**59.49**	**49.33**	**30.70**	**57.04**	**44.13**	**27.82**	**58.63**	**72.04**	**33.58**	**1.45**
2O9G	*14.65*	*14.15*	*11.67*	*11.47*	*11.92*	*10.76*	*10.78*	*0.00*	*0.00*	*0.00*
2Y01	*36.15*	*21.97*	*19.42*	*30.60*	*20.70*	*17.30*	*19.65*	*1.94*	*0.19*	*0.00*
3M71	**23.46**	**23.58**	**17.14**	**21.77**	**20.35**	**14.42**	**41.12**	**0.54**	**0.20**	**0.00**
3QAP	**48.24**	**43.64**	**26.16**	**39.67**	**39.48**	**22.24**	**77.52**	**15.63**	**10.86**	**0.32**
3UG9	**38.38**	**35.90**	**24.16**	**35.98**	**30.08**	**20.66**	**45.60**	**14.71**	**6.77**	**0.14**
3UON	*32.37*	*21.81*	*19.16*	*25.93*	*20.02*	*16.01*	*25.05*	*1.11*	*0.09*	*0.00*
4A2N	**49.64**	**42.45**	**27.24**	**41.56**	**38.93**	**24.65**	**57.93**	**11.49**	**11.26**	**0.48**
4O6Y	*57.08*	*31.92*	*35.29*	*46.25*	*29.25*	*24.94*	*17.28*	*8.09*	*2.84*	*0.47*
Mean	42.39	32.67	24.60	37.39	28.96	20.92	38.11	18.03	8.96	2.21

E: contact numbers computed using experimental structure; P: contact numbers predicted by neural network; N: no contact numbers; μ_{CR} improved by 5% or more (**bold**) and less than 5% (*italic*) when folded with predicted WCNs.

III-3.3 Accurate prediction of WCNs is not sufficient for improving prediction of TMH potations

Though a consistent improvement in CR is observed (Table III-2) when folded with predicted WCNs as restraints, the improvement is not as substantial as with experimental WCNs. In fact, the higher the PCC of WCN prediction is, the closer the μ_{CR} obtained with predicted WCNs ($\mu_{CR(P)}$) is to that obtained with experimental WCNs ($\mu_{CR(E)}$). This relationship is illustrated by a scatter plot (Figure III-4(A)) of the PCCs of WCN prediction and the values of $\frac{\mu_{CR(E)}-\mu_{CR(P)}}{\mu_{CR(E)}}$, which measures the relative difference between $\mu_{CR(E)}$ and $\mu_{CR(P)}$. And the correlation shows that there is still the need to improve the accuracy of WCN prediction if one is to make the best of using WCNs as restraints.

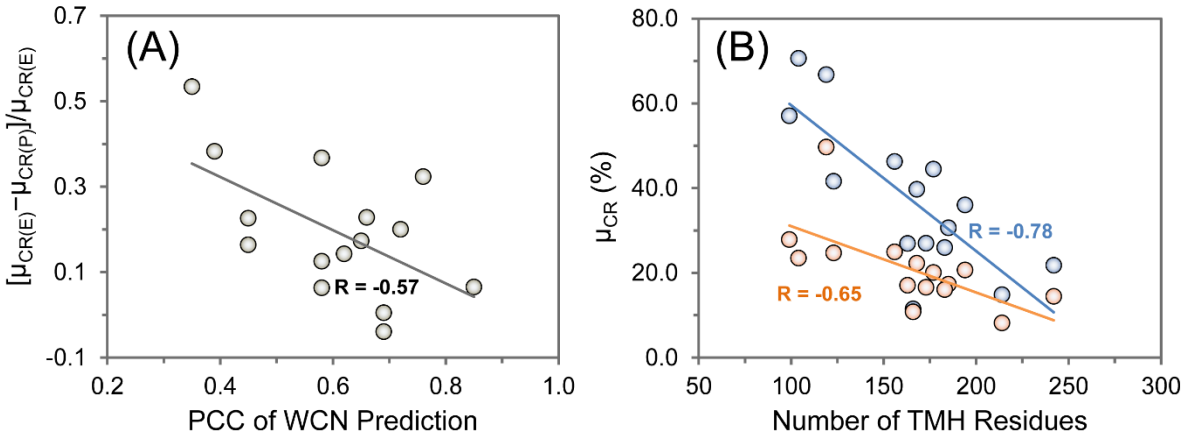


Figure III-4 Improvement in CR is determined by multiple factors

(A) Negative correlation between PCCs of WCN prediction and relative differences between $\mu_{CR(E)}$ and $\mu_{CR(P)}$. $\mu_{CR(P)}$ values obtained with better WCN predictions is closer to $\mu_{CR(E)}$than those obtained with poorer WCN predictions. (B) μCR is negatively correlated with number of TMH residues. Orange dots indicate $\mu_{CR(N)}$values and blue dots indicate $\mu_{CR(E)}$values.

Intuitively, one might also expect that more accurate prediction of WCNs leads to larger relative improvements in CR relative to folding without WCNs. However, the correlation between PCCs and the values of $\frac{\mu_{CR(P)}-\mu_{CR(N)}}{\mu_{CR(N)}}$, which measures the relative improvement in $\mu_{CR(P)}$ compared to $\mu_{CR(N)}$, is only very weak (0.28). For instance, $\mu_{CR(P)}$ is improved by 58.63% relative to $\mu_{CR(N)}$ for 2K73 although the accuracy of WCN prediction for it is low (PCC: 0.45). Whereas for 2BL2

and 2O9G for which WCN predictions are comparably accurate (PCCs are 0.65 and 0.69 respectively), $\mu_{CR(P)}$ is improved by only 11.26% and 10.78% relative to $\mu_{CR(N)}$ respectively. This suggests that other factors besides accurate WCN prediction affect improvement in CR.

One intuitive factor is the size of proteins. In fact, as the size of transmembrane domain (measured by the number of TMH residues) increases it becomes more difficult to predict the correct rotation of helices. To illustrate this, the values of $\mu_{\mathrm{CR(E)}}$ and $\mu_{\mathrm{CR(N)}}$ are plotted against number of TMH residues. As shown in Figure III-4(B), μ_{CR} is negatively correlated with number of TMH residues (R = -0.78 for $\mu_{\mathrm{CR(E)}}$ and -0.65 for $\mu_{CR(N)}$). In addition to this negative correlation, improvement in μ_{CR} also becomes less substantial as transmembrane domain becomes larger. This is reflected on the fact that the gap between the two fitted lines shrinks as TMH residues increases. It is also worth noting that $\mu_{CR(N)}$ is below 20% for 7 out of 11 targets with more than 150 TMH residues, whereas $\mu_{CR(E)}$ is above 20% for all but two targets (1OKC and 2OG9).

Another factor is that some proteins might just represent easy cases whereas others difficult cases for the BCL::MP-Fold algorithm no matter whether WCN restraints are incorporated or not. For easy cases, on the one hand, BCL::MP-Fold samples models with high CR even without WCN restraints and for them it is difficult to improve substantially upon such a high CR with the current level of accuracy of WCN prediction. For example, the membrane rotor of the V-type ATPase 2BL2 whose subunit adopts a four-helical bundle fold (Murata *et al.*, 2005) can be considered an easy case for BCL::MP-Fold. As mentioned previously, its $\mu_{CR(N)}$ is as high as 49.63% even without WCN restraints and the relative improvement in CR in terms of μ_{CR} is a comparably low value of 11.26%. For difficult cases on, the other hand, BCL::MP-Fold is not able to sample models with comparably high CR even experimental WCN restraints partially due these proteins' intrinsic topological complexity. In this modeling benchmark set, 1OKC and 2O9G represent such cases as all TMHs of 1OKC are kinked and 2O9G has two helices located in reentrant regions (Pebay-Peyroula *et al.*, 2003, Savage and Stroud, 2007).

III-3.4 RMSD100 is improved by using WCNs as restraints

While the primary motivation to introduce WCNs as restraints was to improve prediction of helix rotation, an improvement in RMSD100 was also expected after helix rotations are improved. To

verify this, we compared the following parameters among the three simulation groups (E: with experimental WCN, P: with predicted WCN, N: without WCN):

$\beta_{RMSD100}$: the lowest RMSD100 achieved,

$\mu_{RMSD100}$: the average of the 10 lowest RMSD100 values,

π_5: the percentage of models with an RMSD100 value lower than 5 Å.

When using experimental WCNs as restraints, $\beta_{RMSD100}$ was decreased for all targets and $\mu_{RMSD100}$ was decreased for all but 2O9G (Table III-3). As discussed in the previous section, the subunit of the tetrameric aquaporin 2O9G is a special case in that it has two reentrant helices sitting on top of each other (Savage and Stroud, 2007). When using predicted WCNs as restraints, a decrease in $\beta_{RMSD100}$ is seen for 13 targets and a decrease in $\mu_{RMSD100}$ is seen for 12 cases. In terms of $\mu_{RMSD100}$, a decrease of 0.5 Å or more is achieved for 4 targets and the most substantial improvement is a 0.79 Å decrease for 4A2N. Use of predicted WCNs yields smaller, albeit still statistically significant ($p < 0.05$, paired t-test), improvements to $\mu_{RMSD100}$. It is also interesting to note that models with RMSD100 within 5 Å to experimental structures were assembled with noticeable frequencies for three targets (1PV6, 1U19, and 3UON) when using predicted WCNs as restraints, whereas no such models were assembled without WCNs as restraints.

Table III-3 Summary of RMSD100

	$\beta_{RMSD100}$ (Å)			$\mu_{RMSD100}$ (Å)			π_5 (%)		
Target	**E**	**P**	**N**	**E**	**P**	**N**	**E**	**P**	**N**
1OED	*1.88*	*3.69*	*3.70*	*2.08*	*3.85*	*3.89*	*13.47*	*3.80*	*4.28*
1OKC	10.93	11.73	11.75	11.85	12.25	12.05	0	0	0
1PV6	**4.34**	**4.14**	**5.09**	**4.92**	**4.72**	**5.49**	**0.16**	**0.16**	**0**
1PY6	*3.13*	*3.40*	*4.20*	*3.99*	*4.38*	*4.70*	*0.63*	*0.35*	*0.22*
1U19	*3.83*	*4.44*	*5.10*	*5.17*	*5.42*	*5.80*	*0.07*	*0.04*	*0*
2BL2	*2.14*	*2.36*	*2.77*	*2.25*	*2.84*	*2.86*	*11.99*	*8.14*	*13.67*
2K73	*3.01*	*3.59*	*3.82*	*3.06*	*3.72*	*4.03*	*32.44*	*10.76*	*7.07*
2O9G	10.42	12.21	11.41	12.60	12.72	12.41	0	0	0
2Y01	*4.94*	*5.06*	*5.26*	*5.21*	*5.46*	*5.76*	*0.04*	*0*	*0*
3M71	*5.63*	*5.75*	*5.94*	*6.05*	*6.26*	*6.36*	*0*	*0*	*0*
3QAP	*3.33*	*3.89*	*4.26*	*4.25*	*4.50*	*4.65*	*0.56*	*0.39*	*0.3*
3UG9	**3.36**	**3.24**	**4.57**	**3.76**	**4.19**	**4.83**	**1.54**	**0.77**	**0.28**
3UON	**3.70**	**4.94**	**5.30**	**5.17**	**5.30**	**5.81**	**0.13**	**0.02**	**0**
4A2N	**3.51**	**3.56**	**4.30**	**3.94**	**3.79**	**4.58**	**1.28**	**1.53**	**0.55**
4O6Y	2.71	4.21	3.59	3.36	4.90	4.04	1.04	0.07	0.45
Mean	4.46	5.08	5.40	5.18	5.62	5.82	4.22	1.74	1.79

E: contact numbers computed using experimental structure; P: contact numbers predicted by neural network; N: no contact numbers; $\mu_{RMSD100}$ improved by 0.5 Å or more (**bold**), 0.0–0.5 Å (*italic*), and no improvement (normal) when folded with predicted CNs.

III-3.5 Helix rotation accuracy is improved by using predicted WCN as restraints

To visualize the refinement of helix rotation in models with good accuracy at the fold level, experimental WCNs were mapped onto the experimental structure and 3D models with the lowest RMSD100 values. Helices with incorrect rotation would have buried residues exposed and exposed residues buried, thus by coloring buried and exposed residues differentially, incorrectly rotated helices in models can be readily identified. 1PY6 was selected, in part because its fold was generally predicted correctly even without the WCN restraints. The CR values of the 1PY6 models with the lowest RMSD100 values are 41.86% and 7.29%, respectively when folded with predicted WCNs and without WCNs. Without WCN restraints, the buried face of TMH4 and that of TMH6 were modeled to be exposed. This can be readily seen by comparing the rotation of their buried face with that in the experimental structure (Figure III-5(A) and (C)). The incorrect rotation of these two helices disrupted many native contacts between the buried residues of TMH4 and TMH6 (exemplified by red spheres), and likewise, leading to a significantly lower CR. With WCN restraints, the rotations of TMH4 and TMH6 were consistent with the experimental structure (Figure III-5(B)).

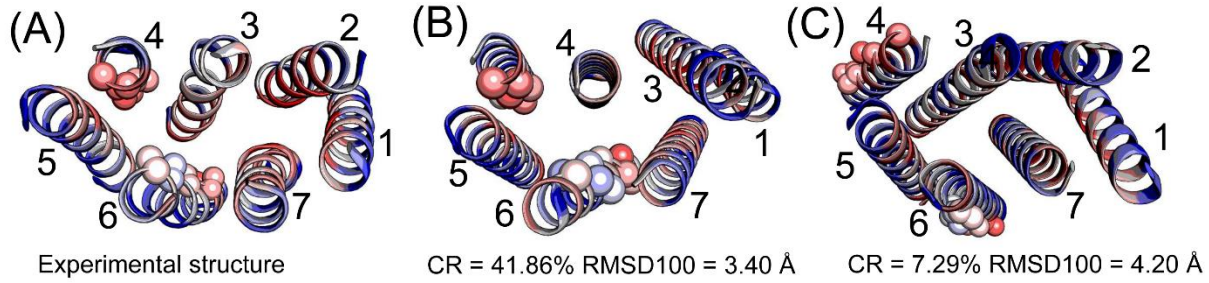


Figure III-5 Experimental CNs mapped onto experimental structures and folded models

(A) experimental structure; (B) model with lowest RMSD100 folded with predicted WCNs as restraints; (C) model with lowest RMSD100 folded without WCN restraints. Color scheme: gradient from blue – fully exposed, red – fully buried. Only TMHs are shown for clarity. Spheres stand for C_α atoms of buried residues of helices 4 and 6 in the experimental structure.

III-3.6 Increased ability of the scoring function at selecting accurate models

When folded without WCN restraints, the average enrichment value over the benchmark set was 1.12. Using predicted WCNs as restraints, enrichment was increased for 14 targets and the average

enrichment was improved to 1.64 (Table III-4). Paired t-test showed that enrichment is improved with statistical significance when folding with predicted WCNs ($p < 0.01$). Indeed, enrichment exceeded 1.50 when folding with predicted WCNs for 8 targets, vs. only 3 targets when folding without WCN restraints. Enrichment was improved even further by using experimental WCNs as restraints. For example, the average enrichment was increased to 1.92 and 13 targets had enrichment greater than or equal to 1.50. Due to the intrinsic inaccuracy of the scoring function in the approximation to the potential energy surface, it should be admitted that these relatively low enrichment values are indicative of a difficulty in selecting the most accurate models of the BCL::MP-Fold algorithm (Fischer *et al.*, 2016). Nevertheless, the statistically significant improvement in enrichment indicates that WCN restraints provide the scoring function with critical information about residue burial, often corresponding to mis-rotated helices.

Table III-4 Enrichment achieved with and without WCN restraints

	Enrichment		
Target	**E**	**P**	**N**
1OED	2.79	1.76	0.88
1OKC	0.51	0.97	0.71
1PV6	1.60	1.27	0.98
1PY6	1.95	1.97	1.16
1U19	1.61	1.29	1.05
2BL2	2.22	2.19	0.43
2K73	2.40	2.43	1.36
2O9G	2.22	1.05	1.08
2Y01	1.50	1.48	1.10
3M71	2.17	1.69	1.45
3QAP	2.20	1.75	1.60
3UG9	2.57	1.67	1.25
3UON	1.43	1.41	0.86
4A2N	1.61	1.89	0.80
4O6Y	1.97	1.80	2.14
Mean	1.92	1.64	1.12

E: contact numbers computed using experimental structure; P: contact numbers predicted by neural network; N: no contact numbers.

III-4 Limitations and Future Directions

Incorporating the burial status of residues has been shown to improve *de novo* structure prediction for soluble proteins (Simons *et al.*, 1997, Durham *et al.*, 2009, Karakas *et al.*, 2012). It is thought

that the benefit of incorporating burial status in *de novo* structure prediction is even larger for HMPs (Adamian and Liang, 2006, Park *et al.*, 2007) because distinguishing buried from exposed residues in the apolar membrane environment is more challenging for non-specific scoring functions. Our results indicate that explicit incorporation of WCN restraints into the BCL::MP-Fold algorithm significantly improves the prediction of TMH rotations and increases the accuracy of helix–helix packing.

Our results also show that using experimental WCNs as restrainted results in significantly more improvement in modeling performance than using predicted WCNs. This suggests that the performance of the WCN predictor BCL::TMH-Expo is an important factor in the BCL::MP-Fold algorithm for HMPs, especially for simple folds such as 1OED. Although using predicted WCNs improved modeling outcomes for most targets, we found that accurate prediction of WCNs does not guarantee a substantial improvement in CR or RMSD100 for every target. For example, only marginal improvement in CR was seen for 2O9G (Table III-2) even though its WCNs were predicted with high PCC (Table III-1) and using predicted WCNs did not improve RMSD100 for 1OKC or 2O9G (Table III-3).

1OKC and 2O9G represent intrinsically difficult targets for BCL::MP-Fold and probably for other methods too. The mitochondrial ADP/ATP carrier (1OKC) has its three odd-numbered TMHs kinked substantially by the presence of prolines (Pebay-Peyroula *et al.*, 2003), whereas the aquaporin (2O9G) contains two reentrant regions (Savage and Stroud, 2007). Tertiary structure prediction for them was either not benchmarked by methods such as Rosetta-Membrane (Yarov-Yarovoy *et al.*, 2006) or Evfold_membrane (Hopf *et al.*, 2012), or proved to be poor with BCL::MP-Fold. BCL::MP-Fold was not able to sample models remotely similar to their experimental structure. The best RMSD100 values for both are > 10 Å (Table III-3). BCL::MP-Fold does not typically accurately represent bent helices. It starts with an idealized, perfectly straight, pool of TMHs. While there are bending moves during the Monte Carlo sampling that bend the TMHs, the current algorithm does not adequately capture the kinks and bends that are commonly seen in native TMHs. This limitation can be overcome with increased probabilities for the bending Monte Carlo moves or more sophisticated bending moves that perturb several ϕ/ψ angles simultaneously by fitting to observed TMH fragments.

III-5 Conclusions

WCN is a key property of amino acid residues that indicate their local packing density. We have demonstrated that explicitly incorporating WCNs as restraints into the membrane protein structure prediction algorithm, BCL::MP-Fold, significantly improved prediction of helix–helix packing. Specifically, WCN restraints helped sample more accurate helix rotation and fold, and improved the ability of the scoring function to select native-like models. The relative improvement from using WCN restraints was often greatest for proteins with relatively simple folds, though improvements in contact recovery were observed across all proteins in the benchmark set when using predicted WCNs. More accurate contact number predictors and structure sampling algorithms that can sample the correct fold of large proteins will be critical to future development of *de novo* tertiary structure prediction for HMPs.

III-6 Software Availability

BCL::MP-Fold has been integrated into the Biochemical Library (BCL) software suite that is being actively developed. It is available at http://www.meilerlab.org/bclcommons under academic and business site licenses. The BCL source code is published under the BCL license and is available at http://www.meilerlab.org/bclcommons. Contact numbers can be readily predicted for novel HMPs using BCL::TMH-Expo via its webserver: http://www.meilerlab.org/servers/tmh_expo.

IV. INTERFACES ACROSS ALPHA-HELICAL TRANSMEMBRANE PROTEINS: CHARACTERIZATION, PREDICTION, AND IMPACT FOR DOCKING

This chapter will be submitted for publication under the same title.

IV-1 Introduction

Alpha-helical transmembrane proteins play essential roles in signal transduction, substance transport, and energy circulation among other critical cellular processes. It was estimated that about one quarter of the human genome encode alpha-helical transmembrane proteins (Fagerberg *et al.*, 2010). Frequently, these transmembrane proteins do not function as monomers but undergo concerted interactions to form either homo-oligomers or interacting with other transmembrane proteins to form hetero-oligomers (Miller *et al.*, 2005, Daley, 2008, Babu *et al.*, 2012).

Proteins can either form stable, obligate oligomers via permanent protein-protein interactions (PPIs) or non-obligate oligomers via transient PPIs. In an obligate PPI, the protomers are not found as stable structures on their own in vivo. Such oligomers are generally also functionally obligate. In contrast, protomers in transient interactions can exist independently and oligomers of this kind are usually involved in processes such as cellular signaling and receptor-ligand binding (Nooren and Thornton, 2003a). The properties of protein-protein interfaces between globular proteins have been extensively characterized in terms of size, amino acid composition, physicochemical texture, conservation, as well as coevolution of residue pairs between the interacting proteins. These properties usually differ for those oligomers that are transient versus those that are obligate (Perkins *et al.*, 2010). In general, interfaces of transiently interacting proteins are smaller in size than obligate interfaces and have amino acid compositions that are usually not drastically different from the rest of the protein surface (Jones and Thornton, 1996, Jones and Thornton, 1997a, Jones and Thornton, 1997b, Lo Conte *et al.*, 1999, Nooren and Thornton, 2003b, Ansari and Helms, 2005, Dey *et al.*, 2010). It was also found that the interface is usually more conserved than the rest of the protein surface (Caffrey *et al.*, 2004, Mintseris and Weng, 2005, Yan *et al.*, 2008), although residues in the interfaces of obligate PPIs tend to be more conserved and exhibit much stronger coevolution with their interacting partners than those in the interfaces of transient PPIs (Mintseris and Weng, 2005).

These observations have been essential to our understanding of the biochemistry and biophysics of PPIs between globular proteins. In contrast, little is known about the characteristics of the interfaces between alpha-helical transmembrane proteins. Here, we analyzed a non-redundant set of alpha-helical transmembrane protein oligomers whose structures have been experimentally determined to a high resolution in order to answer the following open and intriguing questions about membrane protein interfaces: 1) is membrane-protein interface physicochemically distinguishable from the rest of the protein surface? 2) are residues in the interfaces more conserved than the rest of the protein surface? 3) are there detectable coevolutionary signals across the interface? and 4) how do the interfaces in obligate oligomers differ from those in transient ones?

We found that in alpha-helical transmembrane protein oligomers, while the aqueous part of the interface exhibits similar characteristics to interfaces between globular proteins, the intramembranous part of the interface is not significantly different from the rest of the surface in terms of amino acid composition or hydrophobicity. However, on average, the interface is significantly more conserved than the rest of the surface both in the aqueous and intramembranous parts, and residue pairs that are in physical contact in the interface of homo-oligomers correlate more strongly than pairs not in contact. Based on our findings, we also developed a neural network-based method that predicts weighted contact numbers (WCNs) of surface residues from evolutionary information. We showed that interface residues can be accurately identified based on their predicted WCNs. Inspired by our previous study in which residue WCNs were effectively used as restraints to improve *de novo* tertiary structure prediction for alpha-helical membrane proteins (Li *et al.*, 2017a), we implemented an algorithm which leverages the high discriminatory power of a WCN-based penalty score for accurate docking of membrane proteins.

IV-2 Materials and Methods

IV-2.1 Data set

A set of multi-spanning alpha-helical transmembrane proteins whose structures have been determined to a resolution of 2.5 Å or better and an R-free value of 0.3 or better was extracted from the OPM (orientations of proteins in membranes) database (Lomize *et al.*, 2006). The data set was further refined by using the PISCES server (Wang and Dunbrack, 2003) to reduce redundancy such that pairwise sequence identity between protein chains is no more than 25%.

Proteins whose structures were not determined by X-ray crystallography were excluded from consideration. Classification of an oligomer as obligate or transient was carried out by inspection of the experimental structure. Assignment of oligomeric state was based on evidence found in the relevant literature. In summary, the data set consists of 36 obligate and 8 transient oligomers (Table A-4 in APPENDICES). In terms of oligomeric states, the dataset consists of 16 homodimers, 12 homotrimers, 4 homotetramers, 2 homopentamers, 2 homodecamers, 1 heterodimer, 4 heterotrimers, 2 heterotetramers, and 1 heteropentamer (the chimera channelrhodopsin 3ug9A was removed).

IV-2.2 Defining interface residues

Relative solvent accessibility of each residue was calculated with NACCESS (Hubbard and Thornton, 1993). A residue was categorized as core residue if it had < 5% relative solvent accessibility and as surface residue otherwise. Interface residues were defined as those surface residues that lost > 5% relative solvent accessibility upon oligomerization.

IV-2.3 Site-specific rate of evolution

Site-specific rate of evolution was estimated using the Rate4Site method (Pupko *et al.*, 2002). The multiple sequence alignment (MSA) of homologs to each monomer was obtained by running HHblits against the Uniprot20 sequence database (Remmert *et al.*, 2012) with minimum coverage of query sequence (sequence of the monomer) set to 75%, maximum sequence identity to query sequence set to 90%, maximum pairwise sequence identity set to 90%, and E-value cutoff for inclusion in result alignment set to 0.00001.

IV-2.4 Mutual information

We model a sequence position as a discrete random variable X, which takes on one of an alphabet of 20 possible letters $A_X = (A, C, D, \cdots, W, Y)$ with probabilities $(p_A, p_C, p_D, \cdots, p_W, p_Y)$, with $P(X = x) = p_x, p_x \geq 0 \; and \; \sum_{x \in A_X} p_x = 1$. The alphabet A_X contains one letter for each amino acid. Alignment gap was not considered because it introduces spuriously high conservation for alignment columns containing a high percentage of gaps. p_x is estimated by the relative frequency

f_x of amino acid residue x at the column of a MSA. f_x is adjusted by a pseudocount parameter $\lambda = 1$,

$$f_x = \frac{1}{20\lambda + M}\left[\lambda + \sum_{i=1}^{M} \delta(x, X_i)\right] \quad \textbf{IV-1}$$

Given two MSA columns X and Y, the mutual information $I(X,Y)$ between them is defined as

$$I(X,Y) = \sum_{i=1}^{20}\sum_{j=1}^{20} p_{ij}(x_i, y_j)\, log\, \frac{p_{ij}(x_i, y_j)}{p_i(x_i)p_j(y_j)} \quad \textbf{IV-2}$$

$I(X,Y)$ is a general measure of association between two random variables X and Y (two alignment columns in a MSA), it equals zero if and only if X and Y are independent, and it is positive otherwise. Intuitively, $I(X,Y)$ can be thought of as the average reduction in uncertainty about X that results from knowing the value of Y, and vice versa.

To correct for bias in $I(X,Y)$ that may arise from phylogeny (Wollenberg and Atchley, 2000), entropy of the MSA columns (Fodor and Aldrich, 2004), or background noise (Dunn *et al.*, 2008), and test for significance, we implemented a permutation test in which, for each pair of columns, one column was selected and permuted 200 times. This procedure was described in (Cline *et al.*, 2002) in detail. Briefly, a mutual information was calculated for the pair after each permutation. If the number of permutations for which the mutual information is greater than that of the original column pair is ≥ 2 (1% of the total number of permutations), we reject the hypothesis that the column pair is correlated, and its mutual information is set to 0. Otherwise, the bias-corrected mutual information for the pair was computed by subtracting the average mutual information of the 200 permutations from that of the original column pair.

IV-2.5 Training a neural network for predicting WCN

We trained a neural network for predicting residue WCN from amino acid sequence. The target WCNs used in the training were computed from the structure of oligomers in the data set, using the algorithm previously described (Li *et al.*, 2016). Each residue was numerically characterized by rate of evolution of the sequence site and entries in a window of size 15 from the position-

specific scoring matrix (PSSM) computed from a MSA of homologs of the protein family as previously described (Li *et al.*, 2017b). The output layer of the neural network consists of a single node for the residue-specific WCN. The hidden layer consists of 64 neurons. 5% of units in the input layer and 50% of neurons in the hidden layer were randomly silenced during each presentation of each training case. Connection weights were iteratively tuned with resilient back-propagation of errors with the learning rate η set to 0.1 and momentum factor α set to 0.1. The accuracy of WCN prediction was assessed by the Pearson correlation coefficient and the mean absolute error between predicted WCNs and target WCNs computed from the structure of oligomers.

IV-2.6 Predicting interface residues

Note that the WCN of a residue may be different depending on whether it is computed based on the structure of the protomer or the oligomer. To make the distinction straightforward, we refer to WCNs computed from protomers as protomeric WCNs and those computed from oligomers as oligomeric WCNs. For predicting interface residues, it is reasonable to assume that an experimental structure of the protomer is available, and as such, protomeric WCNs can be easily computed from the protomer structure. A surface residue is predicted to be an interface residue if its predicted oligomeric WCN is larger than its experimental protomeric WCN, which is computed based on the protomer structure, by at least a tunable threshold.

The performance of the neural network on predicting interface residues was assessed by the area under than Receiver Operating Characteristic (ROC) curve, or AUC (Hanley and Mcneil, 1982). The AUC was estimated through cross-validation where the data set was partitioned into subsets such that proteins from the same SCOP superfamily (Murzin *et al.*, 1995) were placed in the same subset. Each subset was used exactly once as the testing set for evaluating the performance of the neural network trained on the remaining of the data set. Effectively, a value of AUC was computed from each testing subset. The final estimate of the AUC was computed as the mean of all the AUCs.

IV-2.7 Membrane protein docking

One of the factors that makes modeling membrane protein complexes a more tractable problem than modeling soluble protein complexes is that the membrane imposes a smaller conformational

space that needs to be searched through. We designed an algorithm called BCL::MP-Dock for predicting structure of membrane protein oligomers given the individual tertiary structures of their subunits. The input structures to BCL:: MP-Dock are a structure of the "receptor" and a structure of the "ligand", both oriented in the membrane where the z-axis is aligned with the membrane normal using the PPM server (Lomize *et al.*, 2012) separately. Generation of docking candidates begins with a random rotation of the ligand about the z-axis and translation of the ligand on the membrane to create a glancing contact with the receptor. The ligand is then randomly moved with respect to the receptor using Monte Carlo search {Karakas, 2012 #708}. Translation along the z-axis is limited to no more than 5.4 Å and the step size of tilt angle from the z-axis is limited to no more than 0.05 radians. The baseline scoring function of BCL::MP-Dock consists of a clashing term against residue clashes, a residue pair contact potential term for interface interaction, and a radius of gyration term that favors dense packing between the two docking partners. BCL::MP-Dock was designed to be able to use various experimental and computational information about the interaction between docking partners. In the current study, we tested the idea of using predicted interface residues and their predicted WCNs as restraints for scoring docking solutions on a set of 16 alpha-helical transmembrane protein oligomers (

Table **IV-1**).

IV-2.8 Computation of enrichment

The enrichment was used to measure how capable a scoring function is to select the most accurate docking solutions from a pool of solutions. To calculate enrichment, a given set S of docking solutions are sorted by their RMSD100 values. The top 10% of the solutions with the highest RMSD100 values are put into the set *T* (true) and the rest of the solutions are put into the set *F* (false). The solutions in S are then sorted by their evaluated score. The top 10% of solutions with the lowest score are put into the set *P* (positive) and the rest are put into the set *N* (negative). The intersection of sets *T* and *P* are solutions that are correctly identified by the scoring function and referred to as *TP* (true positives). The intersection of sets *F* and *P* are solutions that are incorrectly identified by the scoring function and are referred to as *FP* (false positives). The enrichment value is then computed using the following formula:

$$Enrichment = \frac{TP}{TP + FP} / \frac{P}{P + N} \quad \textbf{IV-3}$$

Intuitively, $\frac{P}{P+N}$ represents that probability of obtaining a native-like model when choosing a model from *S* at random, whereas $\frac{TP}{TP+FP}$ represents the probability of obtaining a native-like model when choosing from a set of models below an energy cutoff. By our experimental design, $\frac{P}{P+N}$ has a constant value of 0.1, and therefore, the maximum enrichment value that can be achieved is 10.

Table IV-1 Summary of the benchmark set of transmembrane protein complexes

Protein ID	Resolution (Å)	Oligomeric state	Obligate	Protein name	TMH
1q16_CF	1.9	Homodimer	No	Nitrate reductase A	5
2a65_AB	1.7	Homodimer	Yes	LEUTAA	12
2bs2_CF	1.8	Homodimer	No	Quinol fumarate reductase	5
2nq2_AB	2.4	Homodimer	Yes	Putative metal-chelate type ABC transporter	10
2vpz_CG	2.4	Homodimer	No	Polysulfide reductase from Thermus thermophilus	8
2z73_AB	2.5	Homodimer	No	Squid rhodopsin	7
3odu_AB	2.5	Homodimer	No	CXCR4 chemokine receptor	7
3puw_FG	2.3	Heterodimer	Yes	MBP-Maltose transporter transmembrane subunits	8
4a01_AB	2.4	Homodimer	Yes	H1-Translocating Pyrophosphatase	16
4jkv_AB	2.5	Homodimer	No	Human smoothened receptor	7
4mrs_AB	2.4	Homodimer	Yes	Bacterial Atm1-family ABC transporter	6
4o6m_AB	1.9	Homodimer	Yes	CDP-alcohol phosphotransferase	6
4o6y_AB	1.7	Homodimer	Yes	Cytochrome b561	6
4qnd_AB	1.7	Homodimer	Yes	Bacterial homologue of SWEET transporters	3
4rng_AC	2.4	Homodimer	Yes	Bacterial homologue of SWEET transporters (sequence identity with 4qnd_AB < 25%)	3
4u9n_AB	2.2	Homodimer	Yes	Mg(2+) channel MgtE	5

Protein IDs are formatted as four-letter PDB ID followed by the chain IDs of the two subunits comprising the oligomer

IV-3 Results

Each protomer is divided into three disjoint regions: protein core, interface, and non-interface surface with no overlapping residues. And each region is further divided into two parts: aqueous part and intramembranous part based on the atomic coordinates and membrane thickness of each oligomer provided by OPM.

IV-3.1 Amino acid composition and interface propensities

Figure IV-1(A) compares the residue compositions of protein cores, interfaces, and non-interface surfaces in the aqueoust part and Figure IV-1(B) compares those in the membrane. The comparisons show that in the aqueous part, protein cores have the highest frequencies of hydrophobic residues (e.g. Cys, Phe, Ile, Leu, Met, and Val), whereas surfaces have the highest frequencies of hydrophilic residues (e.g. Asp, Glu, His, Lys, Asn, Pro, Ser, and Trp), similar to the amino acid compositions of solube protein complexes (Yan *et al.*, 2008). In the membrane, it is expected that hydrophobic residues will prefer to locate on non-interface surfaces over interfaces and protein cores due to the hydrophobic nature of lipid tails. However, this is only partially true according to the statistics from our dataset. Figure IV-1(B) shows that non-interface surfaces have the highest frequencies of some hydrophilic residues (e.g. Ile, Leu, Val) as well as some hydropholic residues (e.g. His, Lys, Arg, and Trp), whereas some hydrophilic residues have the highest frequencies in protein cores (e.g. Glu, Asn, Pro, Gln, Ser, and Thr). It is also intertesing that Gly and Ala, which have the lowest volume, have the highest frequencies in protein cores both in the aqueous part and in the membrane. This is likely because Gly and Ala, which usually occur in GxxxG and AxxxA type of motifs, may facilitate more efficient helix packing (Senes *et al.*, 2000, Senes *et al.*, 2004).

For the purpose of predicting interface residues, it is worth looking into whether amino acid type distribution in the interface differs from that on the rest of the surface, and estimating the interface propensity of each amino acid type. In fact, the amino acid type distribution in the interface is significantly different from that on the rest of the surface for both the aqueous part ($p = 1.7 \times 10^{-11}$, χ^2 test) and the intramembranous part ($p = 0.011$, χ^2 test). Figure IV-2 shows the interface prosensity of each residue type. In the aqueous part, residue types that prefer interface over the rest of the surface are Phe, Ile, Leu, Met, Val, and Tyr. In the membrane, residue types that prefer interface over the rest of the surface are Ala, Asp, Met, Asn, Pro, Gln, Thr, and Tyr.

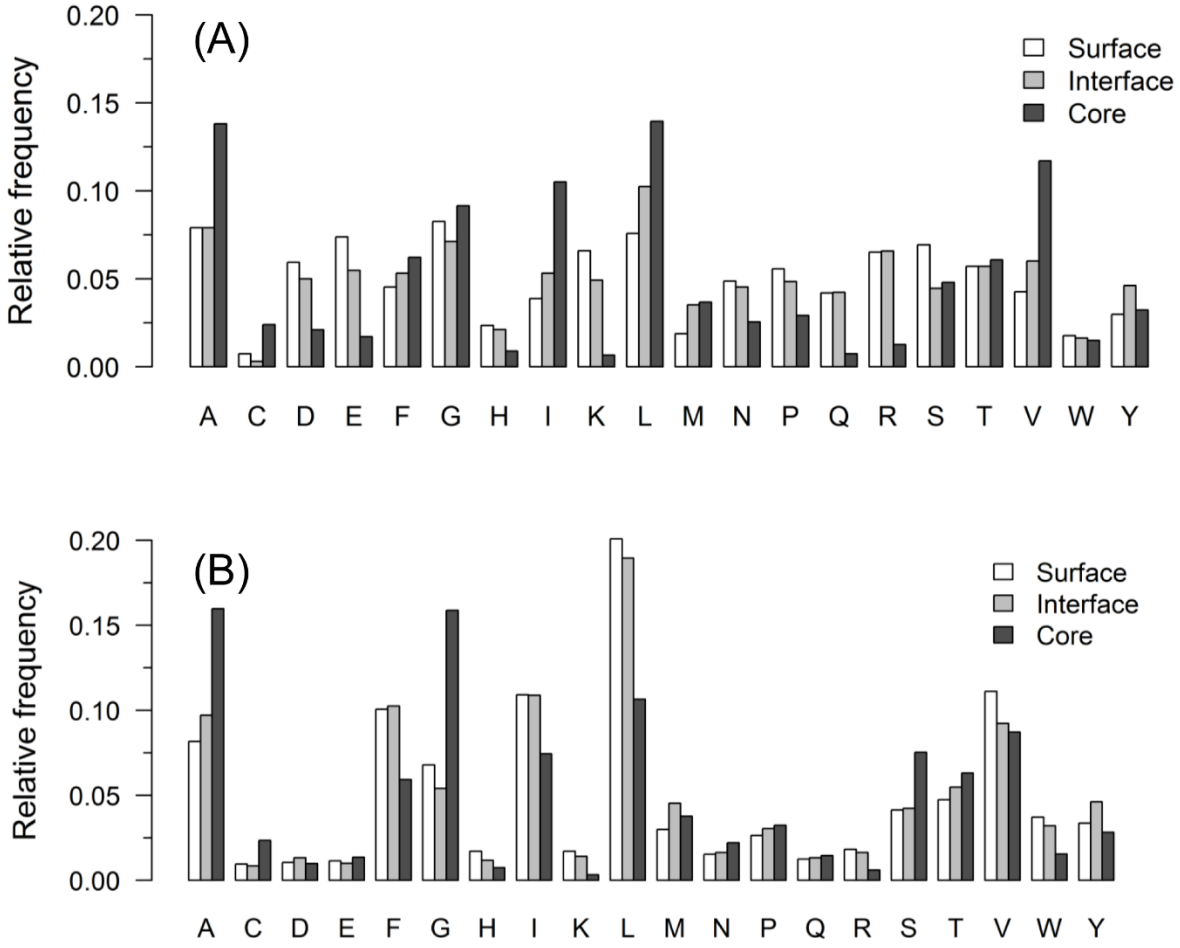


Figure IV-1 Amino acid composition of the core, interface, and the rest of the surface

(A) Amino acid composition in the aqueous part; (B) amino acid composition in the intramembranous part.

Figure IV-2 Interface versus surface propensity in the aqueous part and the intramembranous part for each amino acid type

The interface propensity of an amino acid type A, IP_A, was calculated according to $IP_A = -\log {f_{A,i}}/{f_{A,s}}$, where $f_{A,i}$ is its fraction in the interface and $f_{A,s}$ is its fraction on the surface as a whole. A negative propensity value indicates the amino acid prefers interface.

IV-3.2 Hydrophobicity

Since transmembrane proteins experience two drastically different environments (lipid bilayer and aqueous) at the same time, we analyzed the effect of these two environments on the hydrophobicity of the interface, noninterface surface region, and the protein core separately. In the aqueous part, the interface is significantly more hydrophobic than noninterface surface region, and the protein core is significantly more hydrophobic than the interface (Figure IV-3(A)), consistent with the observation made on soluble proteins (Bordner and Abagyan, 2005). In the lipid bilayer, the protein core is significantly less hydrophobic than the noninterface surface region ($p = 0.0001$, paired t-test) (Figure IV-3(B)), due to the nonpolar nature of lipid tails. The protein core is also less hydrophobic than the interface, although the evidence for this is weaker ($p = 0.011$, paired t-test). To our surprise, no significant difference in average hydrophobicity is found between the interface and the rest of the surface ($p = 0.16$, paired t-test). It is likely that this is partially because protomers of transient membrane protein complexes, when they are not part of a complex, need their interface hydrophobic enough to make favorable contact with lipid tails. In fact, in obligate complexes the interface is less hydrophobic than the rest of the surface, although it is only marginally significant ($p = 0.048$, paired t-test) (Figure IV-3(C)).

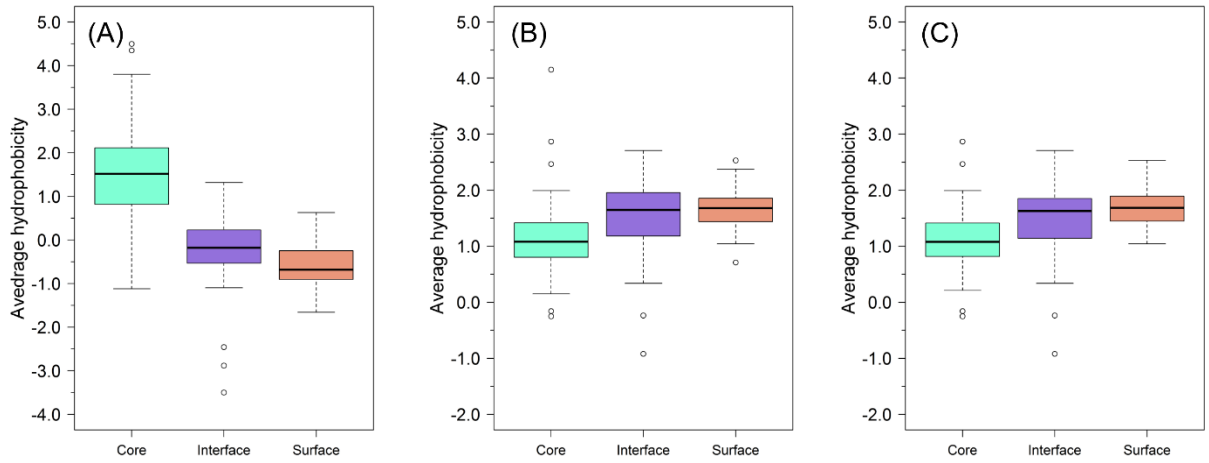


Figure IV-3 Distribution of average hydrophobicity of core, interface, and noninterface surface of alpha-helical transmembrane proteins

(A) Distribution of average hydrophobicity in the aqueous part; (B) distribution of average hydrophobicity in the intramembranous part; (C) distribution of average hydrophobicity for obligate oligomers.

IV-3.3 The interface is more conserved than the rest of the surface in obligate oligomers

The selective pressure is higher on the interface than on the rest of the surface because proteins interact by making specific contacts in the interface whereas, the interaction with lipid tails or solvent is usually nonspecific. Thus, it is expected that the interface is more conserved than the rest of the surface. To confirm this, we computed the average rate of evolution of the interface and that of the rest of the surface for each oligomer in the data set. A lower average rate of evolution indicates a stronger conservation. While the average rate of evolution of the interface is significantly lower than that of the rest of the surface in both the aqueous region ($p = 0.00033$, paired t-test) (Figure IV-4(A)) and the intramembranous region ($p = 0.00056$, paired t-test) (Figure IV-4(D)) overall, it is of more interest to look into the difference between obligate and transient oligomers. Our analysis shows that, in obligate oligomers, the interface is significantly more conserved than the rest of the surface in both the aqueous region ($p = 4.7 \times 10^{-5}$, paired t-test) (Figure IV-4(B)) and the intramembranous region ($p = 6.3 \times 10^{-5}$, paired t-test) (Figure IV-4(E)). In contrast, in transient oligomers, the interface is not seen to be conserved compared to the rest of the surface in neither region (Figure IV-4(C) and (F)). This confirms that, in general, the selection pressure on obligate complexes is stronger than on transient complexes (Nooren and Thornton, 2003a, Mintseris and Weng, 2005).

We also investigated whether interface residues in the lipid bilayer are more conserved than interface residues in the aqueous part. On average, the average rate of evolution of interface residues in the lipid bilayer is significantly lower than that of interface residues in the aqueous part ($p = 0.0046$, paired t-test). However, it is worth noting that for many interfaces (12 out of 44 cases in our data set), the average rate of evolution is lower in the aqueous part than that in the membrane. This suggests that, for some membrane proteins, the extramembrane domain may be critical to protein-protein interactions.

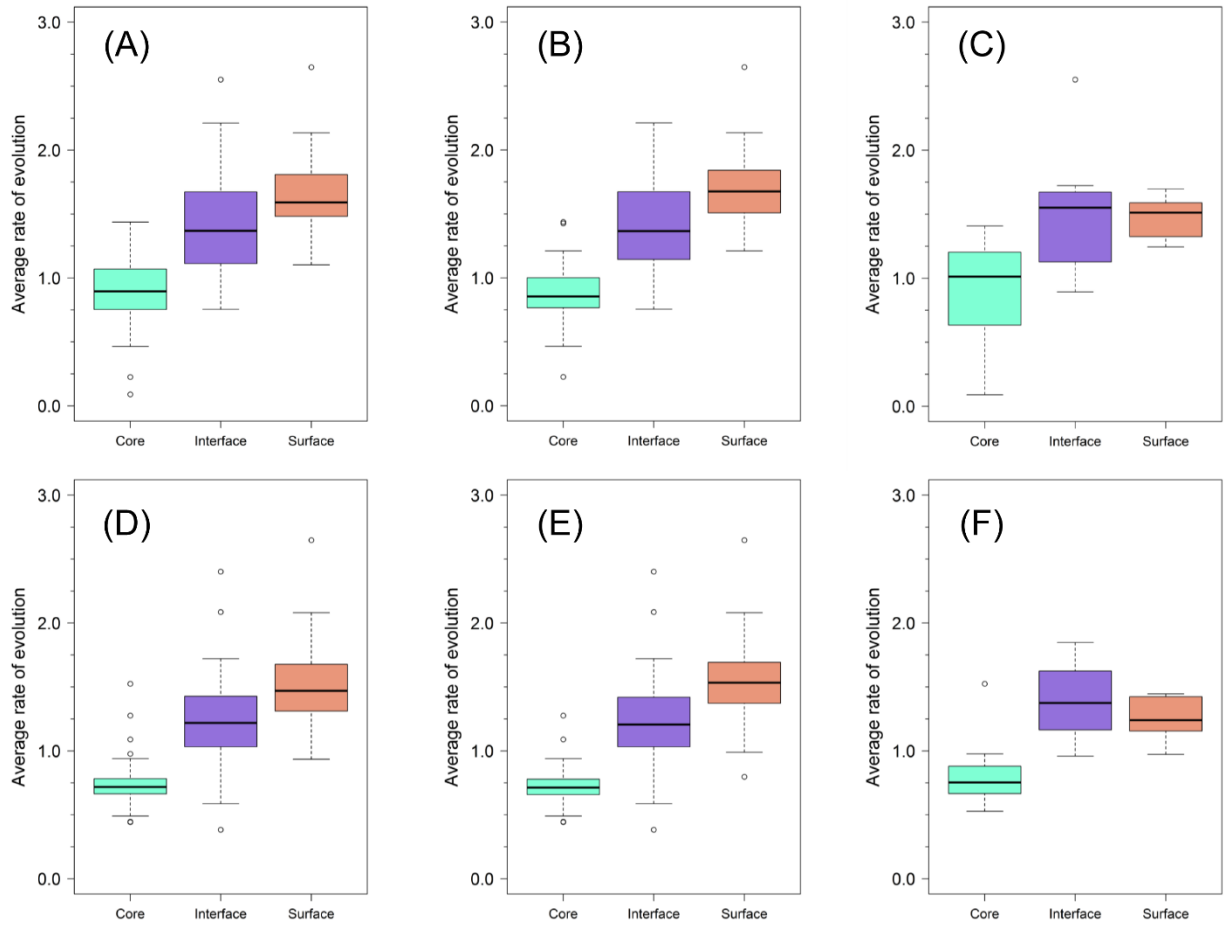


Figure IV-4 Distribution of average rate of evolution of core, interface, and noninterface surface of alpha-helical transmembrane proteins

Distribution of average rate of evolution for the aqueous region taking (A) obligate and transient oligomers together, (B) obligate oligomers only, (C) transient oligomers only; and for the intramembranous region taking (D) obligate and transient oligomers together, (E) obligate oligomers only, (F) transient oligomers only.

IV-3.4 Contacting interface residue pairs show stronger correlation than non-contacting pairs

Previously studies have shown that, in globular proteins, correlated positions have a tendency to be spatially closer in the inter-domain interface (Pazos *et al.*, 1997). However, it is not clear whether residue pairs in the interface of membrane protein complexes are correlated. Figure IV-5 compares the distribution of mutual information between pairs of residues in contact with that of the mutual information between pairs of residues not in contact. Note that a residue i in one subunit forming the interface is considered in physical contact with a residue j in the other subunit if the

separation between any heavy atom of i and j is ≤ 4.0 Å. A clear shift of residue pairs that are in contact toward higher mutual information is observed ($p < 10^{-8}$, Mann-Whitney test). This shift suggests that, in the interface of membrane protein complexes, residue pairs that are in physical contact tend to correlate stronger than those that are not in contact, and that, while a small number of important contacting pairs have high levels of mutual information, there are many more pairs with lower levels of mutual information that may also contribute to the stabilization of membrane protein complex.

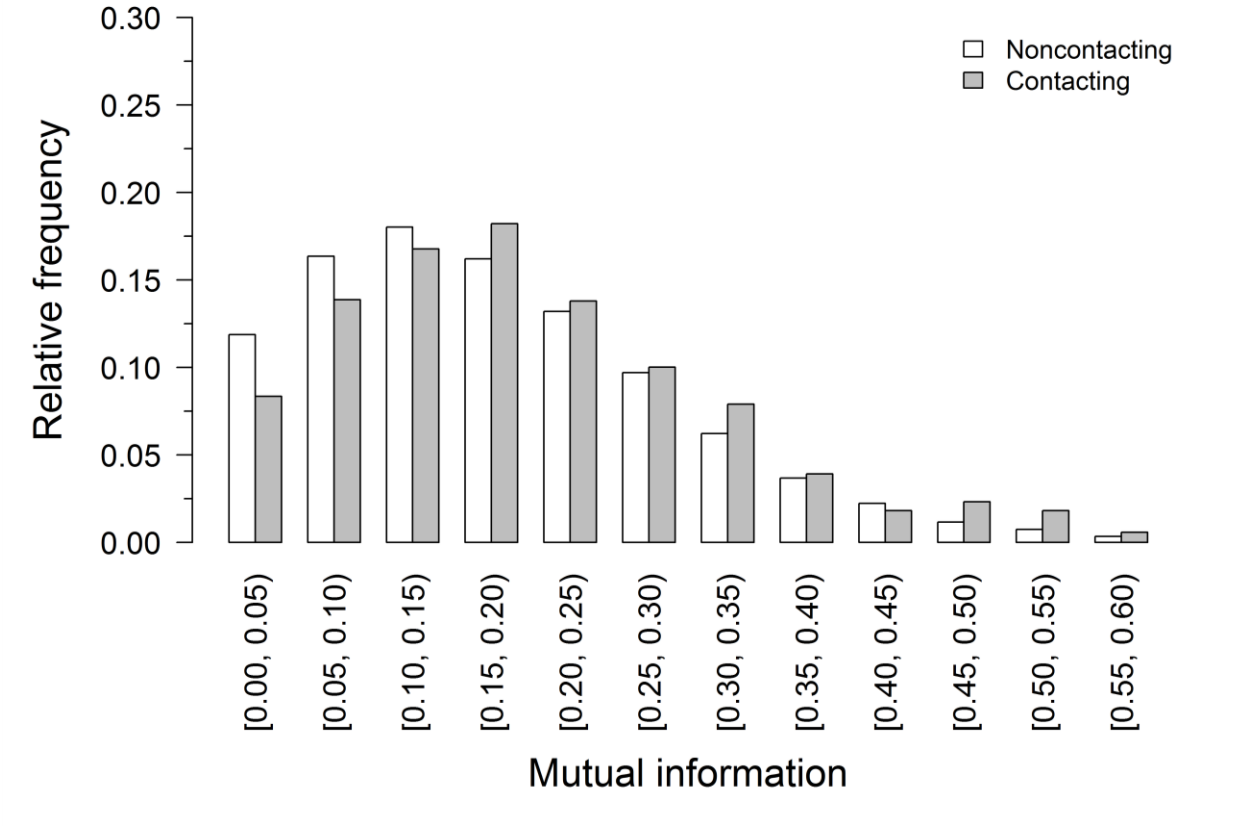


Figure IV-5. Bar diagrams comparing the distribution of the mutual information between pairs of residues in contact with that of the mutual information between pairs of residues not in contact

IV-3.5 Predicting interface residues in the membrane

We evaluate the performance of the trained neural networks in distinguishing residues in the interface from those on the rest of the surface, because the assumption in the current study is that an experimental structure for the protomer is available and it is straightforward to tell which

residues are buried by computing residue relative solvent accessible surface areas. A score for a residue to be in the interface is computed by subtracting its protomeric WCN from its predicted WCN, this score is termed Δ_{WCN}. It is considered that the higher the score the more that the residue is buried in the interface.

Figure IV-6(A) shows the ROC curve plotted using cross-validated predictions from a neural network trained to distinguish interface residues from residues on the rest of the surface for both the aqueous part and the intramembranous part. The area under this curve (AUC) is 0.72. Because the environment of the aqueous part differs drastically from the membrane, we expect that a neural network trained specifically using intramembranous residues will perform better than one trained using an agglomerated dataset of aqueous and intramembranous residues. Figure IV-6(B) shows the ROC curve plotted using cross-validated predictions from a neural network trained to distinguish interface residues from residues on the rest of the surface for the intramembranous part. In fact, the AUC is 0.75, higher than the neural network trained using an agglomerated dataset. This AUC value is also comparable to a previous random forest-based protein-protein binding sites predictor for membrane proteins (Bordner, 2009).

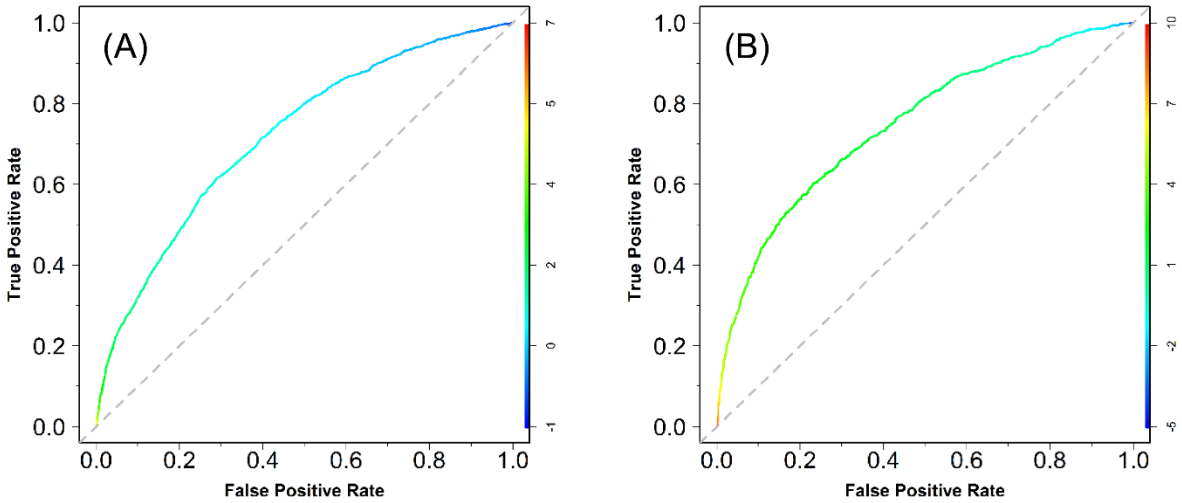


Figure IV-6 Receiver-operating characteristic (ROC) curves

(A) The ROC curve for the neural network-based interface residue classification model trained using interface and surface residues in both the aqueous and the intramembranous regions. The area under this curve is 0.72. (B) The ROC curve for the neural network-based interface residue classification model trained using interface and surface residues in the intramembranous region only. The area under this curve is 0.75. Both curves are colorized according

to classification thresholds and the scale on the right vertical axis indicates the mapping between colors and classification thresholds ΔWCN.

IV-3.6 Docking membrane proteins using predicted WCNs as restraints

The BCL::MP-Dock algorithm was evaluated on a set consisting of 1 heterodimer and 15 homodimer structures. The results for global searches from fully randomized starting positions are detailed in Table IV-2. Of the 15 homodimer cases, BCL::MP-Dock is able to reconstruct the structure of the complex for 14 cases where the best RMSD100 of the docked ligand subunit to its native structure is less than 2.5 Å (2.5 Å is the threshold of resolution used in creating the dataset). The two cases where the best RMSD100 of the docked ligand is greater than 2.5 Å are the bacterial Atm1-family ABC transporter (PDB ID: 4mrs_AB) and the heterodimeric transmembrane subunits of the maltose ABC transporter (PDB ID: 3puw_FG). To confirm the effectiveness of the sampling scheme of BCL::MP-Dock, we also computed the means of RMSD100 of the docked ligand of the top 1% models ranked by RMSD100. Again, except for 3puw_FG and 4mrs_AB, the mean RMSD100 is either less than or only slightly greater than 3.0 Å.

Given that BCL::MP-Dock is able to sample the native structure of most complexes in the testing set, the next question one would then ask is: how effective is the scoring function at identifying the docked ligands that are top-ranked by RMSD100? To answer this question, we computed the means of RMSD100 of the docked ligand of the 1% models ranked by the baseline scoring function. Comparing to the means of RMSD100 of the top 1% models ranked by RMSD100, it can be seen that the baseline scoring function is remotely effective in only two cases (2a65_AB, 2vpz_CG, and 4a01_AB). The effectiveness of a scoring function can also be evaluated by computing its enrichment (see Computation of enrichment). As shown in *Table IV-2*, while the enrichment of the baseline scoring function is greater than 1.0 for 11 cases, it is greater than 2.0 for only 5 cases. These results indicate that the baseline scoring function is not able to identify docked ligands that are top-ranked by RMSD100 in most cases.

Inspired by our previous study which demonstrates that residue weighted contact numbers (WCN) can be effectively used as restraints to improve *de novo* tertiary structure prediction for alpha-helical membrane proteins (Li *et al.*, 2017a), we hypothesized that a scoring term that assigns a penalty to a docked models according the magnitude of the deviation the WCN of interface residues from ANN-predicted WCN will be highly effective:

$$estraint\ score = \sqrt{\frac{1}{n}\sum_{\substack{i \in predicted \\ interface \\ residues}}^{n} (WCN_i - WCN_i^p)^2} \qquad \textbf{IV-4}$$

where WCN_i is the WCN of interface residue i computed based on the docked model, WCN_i^p is the WCN of interface residue i predicted by the neural network. We tested three Δ_{WCN} (ANN-predicted oligomeric WCN – true protomeric WCN) thresholds, namely Δ_{WCN} = 1, 2, or 3 (τ_1, τ_2, and τ_3 in Table IV-2), for classifying interface residues. The predicted WCNs of the resulting interfaces residues of each docking partners were used as restraints. Note that, to avoid potential overestimation of the effectiveness of this approach, the WCNs used as restraints in docking the partners of each testing complex are predicted by a neural network trained with a dataset containing no members from the SCOP superfamily of the testing complex.

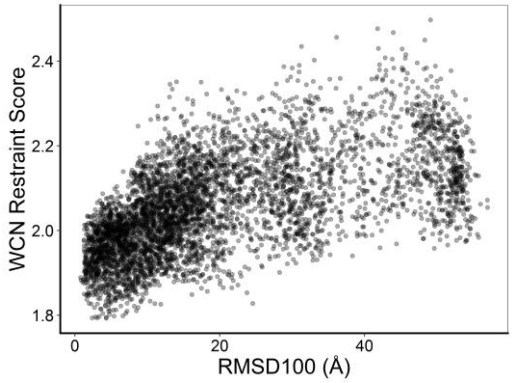
As shown in Table IV-2, with classification thresholds τ_1, τ_2, and τ_3, the enrichment of the WCN-based penalty score is improved in 13, 11, and 11 cases, respectively, compared to the enrichment of the baseline scoring function. More importantly, the number of cases where the enrichment is greater than 2.0 is increased to 10, 10, and 8, respectively. The cases where the enrichment of the penalty score is worse than that of the baseline scoring function or less than 2.0 is likely due to poor classification of interface residues or poor accuracy of WCN prediction. For example, when using classification threshold τ_1, the cases where the enrichment of the penalty score is less than 2.0 has an average TPR, PPV, and MAE of 0.71, 0.20, and 2.09, whereas the those for cases where the enrichment is greater than 2.0 are 0.80, 0.44, and 1.75, respectively. Similar results can be obtained with other two classification thresholds. To demonstrate the effectiveness of the penalty score, we plot the funnel-shaped score v.s. rmsd relationship and compare the best docked model in the 1% models ranked by this score to the native complex structure for cases where using classification threshold τ_1 achieved an enrichment > 2.0 (Figure IV-7).

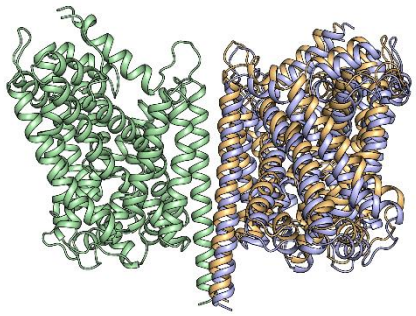
Table IV-2 Summary of the global docking of transmembrane proteins using predicted interface residue WCN as restraints

Oligomer ID	No. of predicted interface residues			TPR			PPV			MAE			Best RMSD100 (Å)				Mean RMSD100 of top 1% models ranked by RMSD100 (Å)				Mean RMSD100 of top 1% models ranked by score (Å)				Enrichment			
	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	τ_1	τ_2	τ_3	β	τ_1	τ_2	τ_3	β	τ_1	τ_2	τ_3	β	τ_1	τ_2	τ_3	β	τ_1	τ_2	τ_3
1q16_CF	46	19	4	0.67	0.33	0.00	0.17	0.21	0.00	2.09	2.59	3.71	1.1	1.1	0.6	0.8	2.2	2.1	2.0	2.3	27.5	25.7	30.7	34.8	0.6	0.0	0.1	0.2
2a65_AB	65	27	12	0.64	0.57	0.43	0.14	0.30	0.50	1.74	2.26	2.41	0.7	0.8	0.9	0.5	1.4	1.4	1.5	1.4	5.3	5.9	4.8	4.3	3.0	**3.1**	**3.7**	**3.6**
2bs2_CF	54	26	8	0.67	0.33	0.00	0.22	0.23	0.00	1.81	2.43	3.47	1.2	1.3	1.4	1.4	2.9	2.9	3.4	3.3	28.0	33.3	37.9	43.0	0.6	0.6	0.4	0.0
2nq2_AB	81	48	30	0.92	0.80	0.76	0.28	0.42	0.63	1.89	2.25	2.21	0.7	1.1	1.1	0.9	2.9	2.2	2.1	1.8	40.0	5.1	4.6	4.1	0.4	**6.6**	**7.0**	**6.8**
2vpz_CG	52	27	14	0.61	0.39	0.17	0.21	0.26	0.21	2.18	2.85	3.47	0.7	0.9	0.9	0.8	2.0	1.9	1.8	1.8	10.5	24.2	29.0	38.7	2.8	0.8	1.0	0.2
2z73_AB	66	36	12	0.69	0.38	0.15	0.14	0.14	0.17	2.16	2.70	3.16	1.9	2.6	1.8	1.5	4.2	4.2	4.1	4.2	26.1	28.7	30.3	18.4	0.7	**0.8**	0.5	**1.0**
3odu_AB	81	57	24	0.86	0.57	0.14	0.07	0.07	0.04	2.50	2.95	3.88	0.9	1.2	0.9	1.7	3.5	4.1	4.4	4.9	35.2	30.0	33.1	35.2	1.1	**1.3**	0.9	**1.2**
3puw_FG	123	93	61	0.93	0.81	0.64	0.45	0.52	0.62	1.99	2.26	2.44	3.2	1.0	4.6	4.7	6.9	7.4	7.2	7.6	25.7	13.4	10.9	14.0	1.8	**3.5**	**3.9**	**4.0**
4a01_AB	50	25	14	0.72	0.58	0.39	0.52	0.84	1.00	1.46	1.68	1.34	0.6	0.7	1.4	1.2	2.4	2.6	2.7	2.4	5.4	4.3	4.0	3.8	4.1	**8.0**	**7.0**	**6.3**
4jkv_AB	58	30	15	0.70	0.39	0.22	0.28	0.30	0.33	2.14	2.79	3.26	0.9	0.5	1.0	1.2	2.4	2.1	2.4	2.5	19.8	8.0	15.9	18.6	2.2	**4.4**	**2.5**	0.6
4mrs_AB	55	29	11	0.75	0.54	0.21	0.38	0.52	0.55	1.77	2.05	2.12	2.9	9.5	1.1	9.1	16.4	15.8	16.2	16.3	29.1	26.5	26.0	24.0	0.5	**1.0**	**0.9**	**1.4**
4o6m_AB	24	9	4	0.55	0.23	0.09	0.50	0.56	0.50	1.46	2.16	2.59	0.6	0.9	1.1	1.1	2.3	2.2	2.2	2.3	12.4	3.4	7.5	14.1	2.1	**4.8**	**3.8**	1.2
4o6y_AB	56	30	14	0.95	0.67	0.48	0.36	0.47	0.71	1.77	2.21	2.30	1.1	0.8	0.7	0.7	2.0	1.8	2.0	2.0	21.4	5.6	7.1	5.9	1.6	**5.6**	**5.4**	**6.2**
4qnd_AB	30	21	17	0.81	0.65	0.58	0.70	0.81	0.88	1.70	1.93	1.88	1.7	0.9	1.5	1.3	3.0	2.9	3.1	3.0	12.7	5.3	7.0	5.8	1.8	**4.6**	**4.6**	**4.2**
4rng_AC	53	42	23	0.87	0.83	0.70	0.49	0.60	0.91	1.49	1.62	1.42	1.0	1.2	1.5	0.6	3.0	2.9	3.1	2.6	21.8	4.1	4.4	3.5	1.5	**7.6**	**7.2**	**5.3**
4u9n_AB	72	55	36	0.90	0.80	0.60	0.63	0.73	0.83	1.86	2.00	2.27	0.7	1.2	1.2	1.9	3.1	2.8	3.1	3.0	17.3	4.9	5.8	5.7	1.4	**6.2**	**5.4**	**4.6**

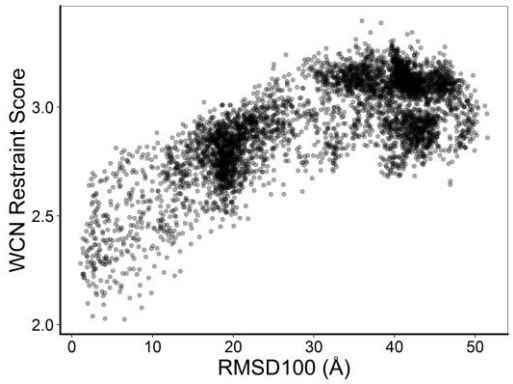
Note: TPR stands for true positive rate; PPV stands for positive predictive value; and MAE stands for mean absolute error; τ_1 indicates predicting residues whose Δ_{WCN} is greater than 1.0 to be in the interface; τ_2 indicates for predicting residues whose Δ_{WCN} is greater than 2.0 to be in the interface; and τ_3 indicates for predicting residues whose Δ_{WCN} is greater than 3.0 to be in the interface; β indicates docking using the baseline scoring function, which consists of a clashing term against residue clashes in docking candidates, a residue pair contact potential term for interface interaction, and a radius of gyration term that favors dense packing between the two docking partners.

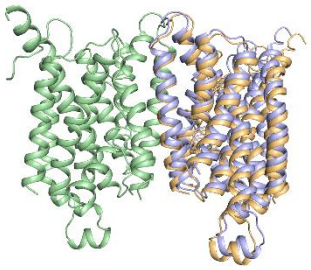
2a65_AB (1.6 Å)



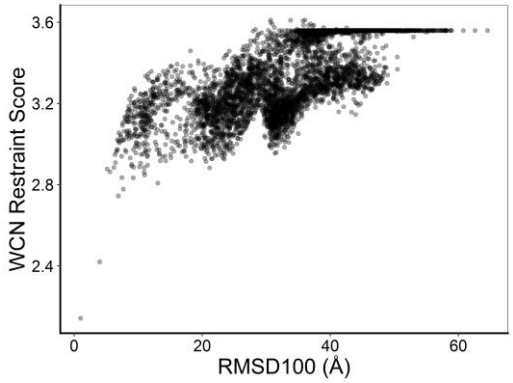


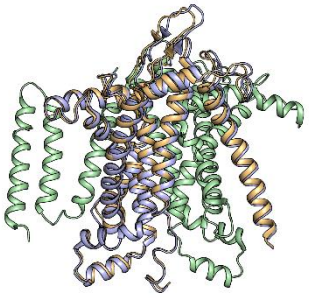
2nq2_AB (1.3 Å)



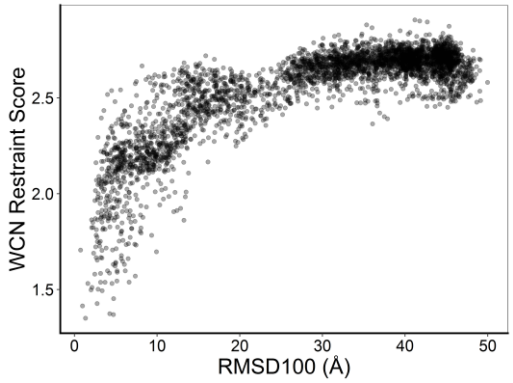


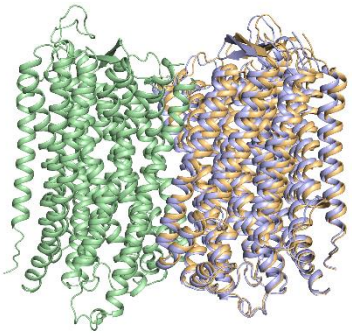
3puw_FG (1.0 Å)



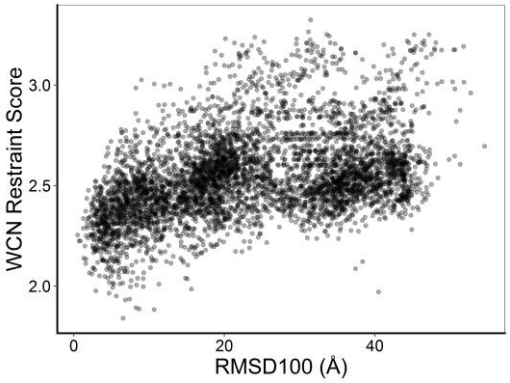


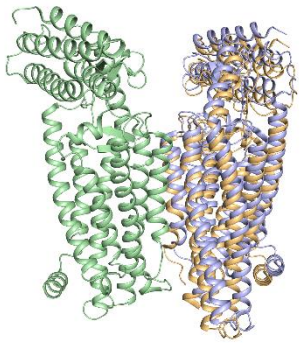
4a01_AB (0.7 Å)



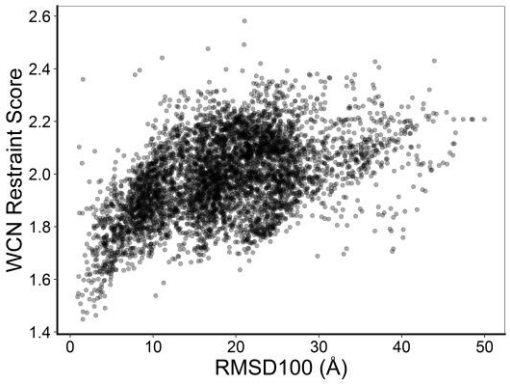


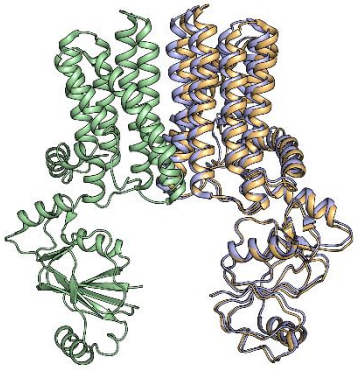
4jkv_AB (2.1 Å)



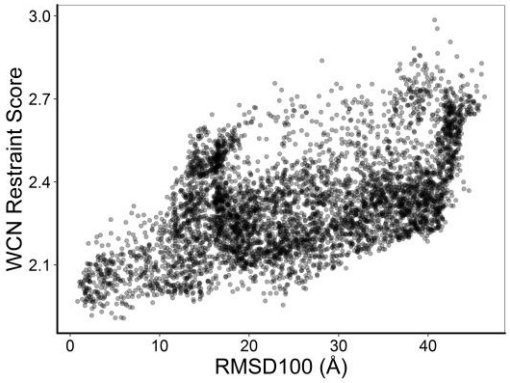


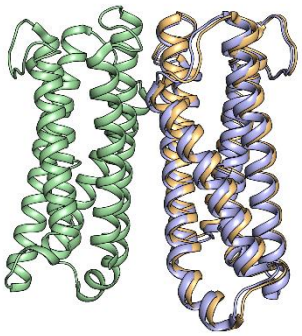
4o6m_AB (0.9 Å)



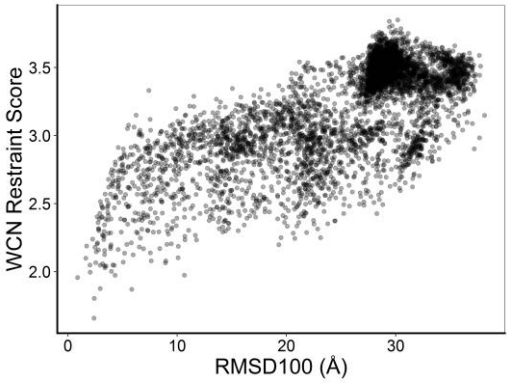


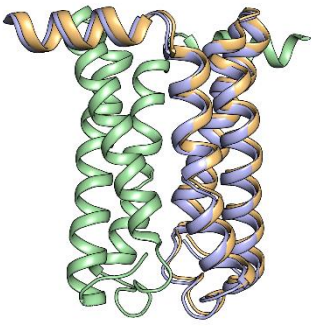
4o6y_AB (0.8 Å)





4qnd_AB (0.9 Å)





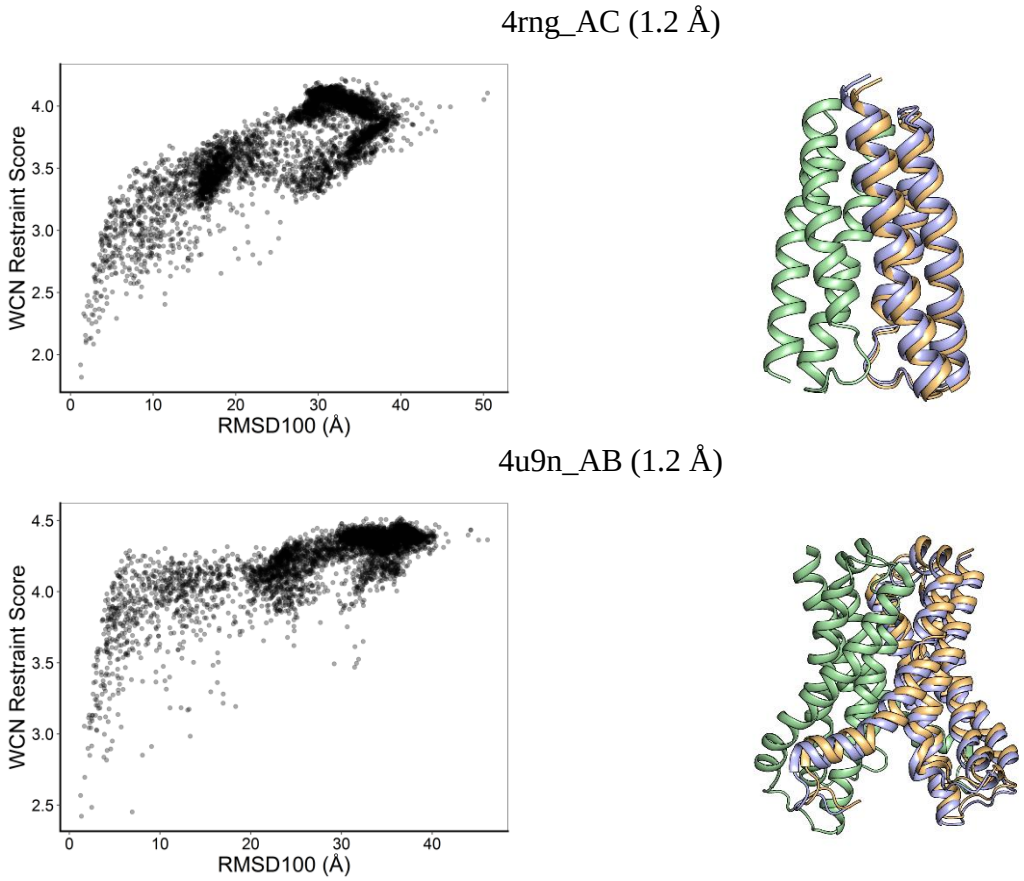


Figure IV-7 Examples of oligomers for which WCNs helped identify the correct docking solutions

Plots on WCN restraint score v.s. RMSD100 relationship, and comparison of the best docked model in the 1% models ranked by WCN restraint score to the native oligomer structure (pale green: native receptor subunit, pale blue: native ligand subunit, light orange: docked ligand subunit) for cases where the enrichment of the WCN restraint score is higher than 2 using ΔWCN = 1 as the threshold for classifying interface residue. The values in parentheses are the RMSD100 of the models shown.

IV-4 Discussion

While the properties of interfaces in both transient and obligate oligomers of globular proteins have been well characterized, little is known about the characteristics of interfaces in oligomers of transmembrane proteins. In the current work, we compiled a nonredundant dataset of oligomers of alpha-helical membrane proteins whose structure have been determined to high resolution and studied the properties of the interfaces in terms of hydrophobicity, amino acid composition, interface propensity of amino acids, evolutionary conservation, and correlation of amino acid pairs in the interface. We found that the aqueous part of the interface in oligomers of alpha-helical membrane proteins has similar properties to the interface in oligomers of globular proteins (Jones

and Thornton, 1997a, Nooren and Thornton, 2003b, Bordner and Abagyan, 2005, Yan *et al.*, 2008, Acuner Ozbabacan *et al.*, 2011). Within the membrane, while the average hydrophobicity of the interface was not found to be statistically different from that of the rest of the surface, the interface is significantly more conserved than the rest of the surface. We also observed that contacting residue pairs across the interface tend to correlate more strongly than non-contacting residue pairs.

Based on the observation that the interface is significantly more conserved than the rest of the surface in the membrane, we adapted our previously developed neural network-based method (Li *et al.*, 2016) for distinguishing interface residues from residues on the rest of the surface. This classification was based on the weighted contact numbers of surface residues predicted by the neural network. While the performance of this method is comparable to the Random Forest-based binary classifier developed by Bordner (Bordner, 2009), its strength lies in the fact that it also predicts residues' real-valued weighted contact number. Residues' WCN not only is an effective restraint for improving the fraction of native contacts in predicted structural models for *de novo* prediction of tertiary structures of alpha-helical membrane proteins (Li *et al.*, 2017a), as we have shown in the current study, it can also be a powerful score for selecting native-like docking candidates of membrane protein complexes.

The sampling problem in docking transmembrane proteins is inherently more tractable than that in docking globular proteins. This is mainly because the membrane imposes a strong constraint on the rotational degree of freedom with respect to the membrane normal and the translational degree of freedom along the membrane normal. Despite this simplification in sampling, docking transmembrane proteins is still a challenging problem in terms of scoring, particularly for docking algorithms where the scoring function primarily concerns about physicochemical complementarity. Another complication is that scoring functions replying on shape complementarity would not perform well neither because of the typical cylindrical shape of alpha-helical membrane proteins. In contrast, using restraints derived from evolutionary analysis may represent an effective approach to narrowing down the set of viable docking candidates. In fact, the effectiveness of using predicted coevolving residue pairs to single out the right docking solution has been demonstrated in docking globular proteins (Pazos *et al.*, 1997, Madaoui and Guerois, 2008, Weigt *et al.*, 2009, Ovchinnikov *et al.*, 2014, Ovchinnikov *et al.*, 2015, Uguzzoni

et al., 2017) and some isolated cases of transmembrane proteins (Ovchinnikov *et al.*, 2014, Ovchinnikov *et al.*, 2015).

While coevolutionary analysis has the benefit of pinpointing specific interacting residue pairs, it requires construction of paired MSA of interacting partners, and computationally, this problem becomes difficult by itself due to the existence of paralogs (Burger and van Nimwegen, 2008, Bitbol *et al.*, 2016, Gueudre *et al.*, 2016). We have taken a different approach where the first step is to predict WCNs of surface residues from sequence and to identify potential interface residues in each interaction partner using a neural network trained to map from evolutionary information to residue WCN. The second step is to use the identified interface residues as "sticky" points and their predicted WCNs as restraints for ranking docking solutions. This approach is computationally more efficient and more tractable as it eliminates the necessity of creating pairedMSAs. We have also demonstrated the effectiveness of this approach using a benchmark set of 16 alpha-helical transmembrane protein oligomers.

We also note that the effectiveness of our approach depends on the accuracy of the prediction of WCNs and the identification of interface residues, although it is robust to the extent that the enrichment of the WCN restraint score will be greater than 1.0 when either the PPV of classifying interface residues is above 0.25 or the MAE of WCN prediction is below 2.0 (Table IV-2). However, when the PPV drops below 0.25 and the MAE goes above 2.0, the WCN score starts to act against identifying correct docking solutions. This issue stands out especially in some transient oligomers where structural or functional constraint on the interface may be too weak to leave detectable evolutionary information on interface residues or incidental oligomers where the interface is neither structurally nor functionally relevant (Nooren and Thornton, 2003a). Nevertheless, the current study provides essential statistics about some properties of the interface in alpha-helical transmembrane protein oligomers and also presents a relatively effective and robust approach for docking transmembrane proteins. The statistics may give insights into the development of methods that are more accurate in predicting interface residues and the docking approach may be a valuable tool for constructing structural models for transmembrane protein oligomers.

V. PREDICTING THE FUNCTIONAL IMPACT OF KCNQ1 VARIANTS OF UNKNOWN SIGNIFICANCE

This chapter has been published under (Li *et al.*, 2017b).

V-1 Introduction

Congenital long QT syndrome (LQTS) is a heart rhythm disorder that affects ~ 1 in 2,500 births (Schwartz *et al.*, 2009). It predisposes children and young adults to a type of ventricular tachycardia (torsades de pointes) and sudden cardiac death (Goldenberg and Moss, 2008). LQTS is associated with pathogenic variants in several genes that lead to dysfunctional cardiac ion channels. Among the 16 known LQTS-associated genes, *KCNQ1* variants account for ~ 30-35% of all LQTS cases. *KCNQ1* encodes the α-subunit of the voltage-gated K^+ channel KCNQ1 (also known as $K_V7.1$) that regulates the slow delayed rectifier current (I_{Ks}), a major driver of cardiac repolarization (Barhanin *et al.*, 1996). Loss of KCNQ1 function leads to diminished or dysfunctional I_{Ks}, impaired myocardial repolarization and LQTS (Schwartz *et al.*, 2013).

An emerging standard-of-care for LQTS employs clinical genetic testing to identify LQTS-associated variants (Schwartz *et al.*, 2013). Established genotype-phenotype relations should be factored into the assessment of the risk of sudden cardiac death and the selection of appropriate therapeutic interventions (Giudicessi and Ackerman, 2013). However, variants of unknown significance (VUS) for which there is inadequate evidence to classify as being pathogenic are common findings (Ackerman, 2015). This issue is further confounded by the presence of background genetic "noise" (the frequency of genetic variations of a particular gene in a healthy population) and variants with incomplete penetrance (MacArthur and Tyler-Smith, 2010, Giudicessi and Ackerman, 2013, Ackerman, 2015). Variant interpretation is bound to present an increasingly daunting challenge in the era of next-generation sequencing (MacArthur and Tyler-Smith, 2010, Cooper and Shendure, 2011, Katsanis and Katsanis, 2013).

Ideally, except for certain well-established disease-causing variants, positive LQTS genetic testing results should be evaluated by physiologically relevant experimental functional assays, but experimental characterization remains labor-intensive and costly to scale (Bhuiyan, 2012, Katsanis and Katsanis, 2013). Under such constraints, computational methods, which are usually machine learning-based, represent a common predictive approach (Ng and Henikoff, 2006, Cooper and

Shendure, 2011, Richards *et al.*, 2015). However, hardly any computational methods are sufficiently accurate for clinical use related to channelopathies or other genetic disorders (Tchernitchko *et al.*, 2004, Ohanian *et al.*, 2012). Most existing computational methods have been trained on datasets pulled from online databases that have not been subjected to rigorous functional validation (Richards *et al.*, 2015). These datasets may be significantly contaminated with erroneous annotations and thereby provide machine-learning algorithms with misleading information (Care *et al.*, 2007, Richards *et al.*, 2015). Further, a potentially even more crucial issue is that current methods intermingle two related but separate questions: whether a given variant causes functional impact at the molecular level and, if so, whether that functional effect will be manifested at the organismal level. Making such distinctions is important when delivering predictions because dysfunction at molecular level does not necessarily equate to organismal deleteriousness (MacArthur and Tyler-Smith, 2010, Cooper and Shendure, 2011).

In this study, we sought to develop a protein-specific algorithm capable of accurately predicting functional consequences of KCNQ1 variants. We first curated a set of functionally validated KCNQ1 variants. We then trained a neural network-based, KCNQ1-specific genotype−channel function relationship predictor Q1VarPred. In contrast to genome-wide methods, whose performances have suffered from dataset contamination and heterogeneity and do not differentiate between functional impact and organismal deleteriousness when delivering predictions, Q1VarPred was trained on the functionally validated dataset to predict molecular functional impact.

V-2 Materials and Methods

V-2.1 Dataset and criteria for annotating functional impact

KCNQ1 variants and their associated electrophysiological (EP) effects in the dataset for this study were collected from the literature (Table A-5 in APPENDICES). We only considered data from experiments where the auxiliary subunit KCNE1 was also expressed. Each variant was annotated in terms of functional impact based on two experimental parameters (peak current relative to the wild-type and change in voltage of half-maximal activation V½). Specifically, a variant was defined as "Normal" if 1) $75\% \leq peak\ current \leq 125\%$ and 2) there was ≤10 mV depolarization or hyperpolarization shift in V½. "Mild Loss of Function" was defined as 1) $25\% < peak\ current < 75\%$ or 2) 10-20 mV depolarization shift in V½. "Severe Loss of Function" was

defined as 1) $peak\ current < 25\%$ or 2) >20 mV depolarization shift in V½. "Severe Gain of Function" was defined as 1) >150% peak current or 2) 120 to 150 % peak current and >15 mV hyperpolarization shift in V½. Clinical classification (case variant versus control) of each variant was sourced from previous large-scale clinical studies (Kapa *et al.*, 2009, Giudicessi *et al.*, 2012), or EP studies that reported such information. Case variants were identified in patient cohort whereas control variants were found in healthy cohort. In addition, in accordance to the recent ACMG/AMP standards and guidelines for the interpretation of sequence variants (Richards *et al.*, 2015), variants with a minor allele frequency of > 1 / 2500 (LQTS prevalence) in the general population were removed. For training the binary classification model Q1VarPred, loss-of-function and gain-of-function variants were grouped together as dysfunctional and a mild loss-of-function variant was either labeled as dysfunctional if its peak current was < 50%, or normal otherwise. The common variant G643S was classified as having normal function (Modell and Lehmann, 2006).

V-2.2 Neural network architecture and training

The neural network in the present study was a fully connected three-layer feed-forward network with a sigmoid transfer function. The input layer consists of two nodes, one for each predictive feature. The output layer consists of a single neuron, which outputs a numerical prediction of the functional impact of a given variant on the scale of 0 to 1 with 1 being complete dysfunction. A hidden layer with 3 neurons was chosen considering the fact that the "dropout" technique (Srivastava *et al.*, 2014) was adopted to prevent the neural network from overfitting, a phenomenon in which the learned model is excessively complex (e.g. too many model parameters relative to the number of observations for training) and is poorly generalizable. However, we also tested hidden layers with up to 8 neurons, the results of which showed that the size of the hidden layer did not affect the performance of the neural network in a significant way (Table A-6 in APPENDICES). The neural network was trained on numeric encoding of variant functional labels (1 for complete dysfunction 0 for normal), with back-propagation of errors. The learning rate was set to 0.05 and momentum was set to 0.8. Weights were updated after each presentation of a variant to the network and a constant weight decay of 0.02 was applied to reduce model flexibility.

V-2.3 Predictive features

We used two features to characterize an amino acid substitution, namely rate of evolution, which quantifies the conservation of the sequence position where the substitution has occurred, and position-specific scoring matrix (PSSM)-based perturbation, which measures the radicalness of the substitution itself. These two features were chosen, before the dataset was inspected, based on the rationale that a very conserved position may tolerate less radical substitutions while a variable position may not tolerate more radical substitutions, as—for example—observed in a systematic mutation study of bacteriophage T4 lysozyme (Rennell *et al.*, 1991). We confirmed that these features are critical among a limited number of features tested (Table A-7 in APPENDICES). Details on how these two features were computed can be found in Computation of predictive features in APPENDICES.

V-2.4 Performance metrics

The performance of the learned neural network model and other evaluated methods were quantified using the following metrics: true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), accuracy, Matthew's correlation coefficient (MCC),(Matthews, 1975) and area under the receiver operating characteristic (ROC) curve (AUC). Note that the first six metrics can be computed only after all variants are classified at a specific threshold. Using the notation of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), these metrics are defined as:

$$TPR = \frac{TP}{TP + FN} \quad \textbf{V-1}$$

$$TNR = \frac{TN}{TN + FP} \quad \textbf{V-2}$$

$$PPV = \frac{TP}{TP + FP} \quad \textbf{V-3}$$

$$NPV = \frac{TN}{TN + FN} \quad \textbf{V-4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{V-5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN)(TP + FP)(TP + FN)(TN + FP)}} \quad \textbf{V-6}$$

respectively. A TP is a dysfunctional variant classified as dysfunctional and TN is a normal variant classified as normal. MCC measures the correlation between predicted and observed binary classifications, with a value between -1 and 1. A MCC of 1 means perfect classification, a value of 0 means no better than random classification, and -1 indicates a completely reversed classification. As MCC is unaffected by class size, it is a particularly useful measure of classification quality when the two classes are of very different sizes.(Matthews, 1975) Computation of all performance metrics was accomplished using the ROCR package (Sing *et al.*, 2005) implemented in the R programming environment (R Development Core Team, 2015).

V-2.5 Estimating generalization ability

The generalization ability of a learned model is defined as its performance in predicting new variants that are not used for training. A model with higher generalization ability is favored over ones with lower generalization ability. A common practice to estimate a model's generalization ability is through a procedure called "k-fold cross validation" where the dataset is randomly divided into k equally-sized mutually exclusive subsets. The model is trained on $k - 1$ subsets (collectively known as the training set) and its generalization ability is estimated on the remaining one subset (test set). Specifically, after the model is trained, a threshold is determined which maximizes the MCC on the training set, the same threshold is then used for computing the performance metrics on the test set. This process is repeated k times each using a different one of the k subsets as the test set and the remaining $k - 1$ subsets as the training set. Every time a model is trained, its performance metrics are computed on the test set. In a k-fold cross-validation, the generalization ability is estimated as the average of performance metrics over k test sets. Because the number of ways a dataset can be split into k subsets is enormous, it is desirable to repeat the random splitting p times to reduce artifacts. In the current study, we chose $k = 3$ and $p = 200$, similar to a previous study (Smith *et al.*, 2014). The splitting was stratified such that the class proportions of the training set and the test set are as close to that of the whole dataset as possible.

To ensure the consistency of comparison, the performance metrics of all evaluated methods were estimated using the exact same data. This means that every time the dataset was randomly split into 3 subsets, these subsets were used for calculating the performance metrics of all methods. The variability in performance metrics associated with random splitting of dataset is presented in Table A-8 in APPENDICES.

V-3 Results

V-3.1 Functional studies do not always agree with clinical testing

We compiled a total of 107 functionally characterized KCNQ1 variants (Table A-5 in APPENDICES). Two important observations were made on this dataset. First, not all case variants (variants identified in LQTS patient cohort, a total of 99 in our dataset) are severely dysfunctional. Per our scheme of functional annotation (see Dataset and criteria for annotating functional impact), 6 out of 99 case variants are functionally normal and 8 out of 99 cause only mild loss of function. Interestingly, these two fractions roughly agree with the previous estimate that ~10% case variants may be false positives (Kapa *et al.*, 2009). On the other hand, a few variants identified in presumed healthy controls are severely dysfunctional (for example, V110I and A300T). A300T, which occurs within the pore helix of the channel was shown to cause a massive reduction of I_{Ks} and hyperpolarization of the voltage of half-activation of the channel both with and without the presence of the wild-type subunit (Bianchi *et al.*, 2000). The V110I variant showed significant reduction in I_{Ks} and depolarization of voltage of half-maximal activation when expressed in the absence of the wild-type subunit (Cordeiro *et al.*, 2010). This analysis reinforces the argument that translating protein dysfunction at the molecular level to clinical manifestation and also attributing clinical manifestation to protein dysfunction both need to be carried out with caution (Giudicessi and Ackerman, 2013).

V-3.2 Position-specific rate of evolution reflects functionally-critical subdomains

The importance of a sequence site for protein structure or function can often be inferred from its conservation over evolution. We computed the position-specific rate of evolution for the entire sequence as well as the mean rate of evolution for each of the 24 subdomains of KCNQ1 (see Computation of predictive features in APPENDICES). A lower rate of evolution indicates higher conservation.

Overall, the N-terminal domain (NTD) and C-terminal domain (CTD) are generally less conserved than subdomains within the channel domain (CD), as shown in Figure V-1. The average rates of evolution for the NTD and CTD are 3.2 and 2.5, respectively, whereas the average rate of evolution in the CD is 1.0. Within the CD, six subdomains have a mean rate of evolution below 1.0 (S4, S4-S5, S5, pore-helix, pore-loop, and S6). As expected, the pore-helix (residues 299-312) and pore-loop (residues 313-322) of the channel are the most conserved subdomains, with mean rates of evolution of only 0.38 and 0.41, respectively. This correlates with the critical role played by these components in achieving high ion selectivity for K^+ over Na^+ ions (Doyle *et al.*, 1998). The S4 segment of the CD, which harbors basic residues for sensing and responding to changes in membrane potential (Choe, 2002), has a mean rate of evolution of 0.61. The S4-S5 linker, which is believed to be responsible for transferring the conformational changes in the voltage sensor domain to the pore (Labro *et al.*, 2011) and serve as binding sites for phosphatidylinositol-4,5-bisphosphate (PIP2) to modulate the deactivation rate of the channel (Taylor and Sanders, 2016), has a mean rate of evolution of 0.92. The S2-S3 linker, proposed in a recent study to also bind PIP2 (Chen *et al.*, 2015), is only moderately conserved. Interestingly, although most subdomains of the CD exhibit a low mean rate of evolution, two subdomains namely the S1-S2 linker and the S5-Pore linker, show substantially higher mean rates of evolution (2.5 and 1.9, respectively) than the rest of the CD.

As the CTD has been shown to have four helices designated A-D (Wiener *et al.*, 2008), we computed the mean rate of evolution of each of these helices and their linkers to see if any of these subdomains are conserved. Our analysis shows that only helices A, B, and C have a mean rate of evolution < 1.0, whereas the mean rate of evolution of helix D is substantially higher (1.9). This observation agrees with the functional role of helices A and B in binding calmodulin (CaM) and the critical role of helix C in tetramerization of the intracellular C-terminal domain (Wiener *et al.*, 2008, Sachyani *et al.*, 2014). The juxtramembrane subdomain S6-A, with a mean rate of evolution of 0.88, as well as the B-C linker, considered extremely conserved according to its mean rate of evolution (0.24), have yet to be shown to play any particular functional role.

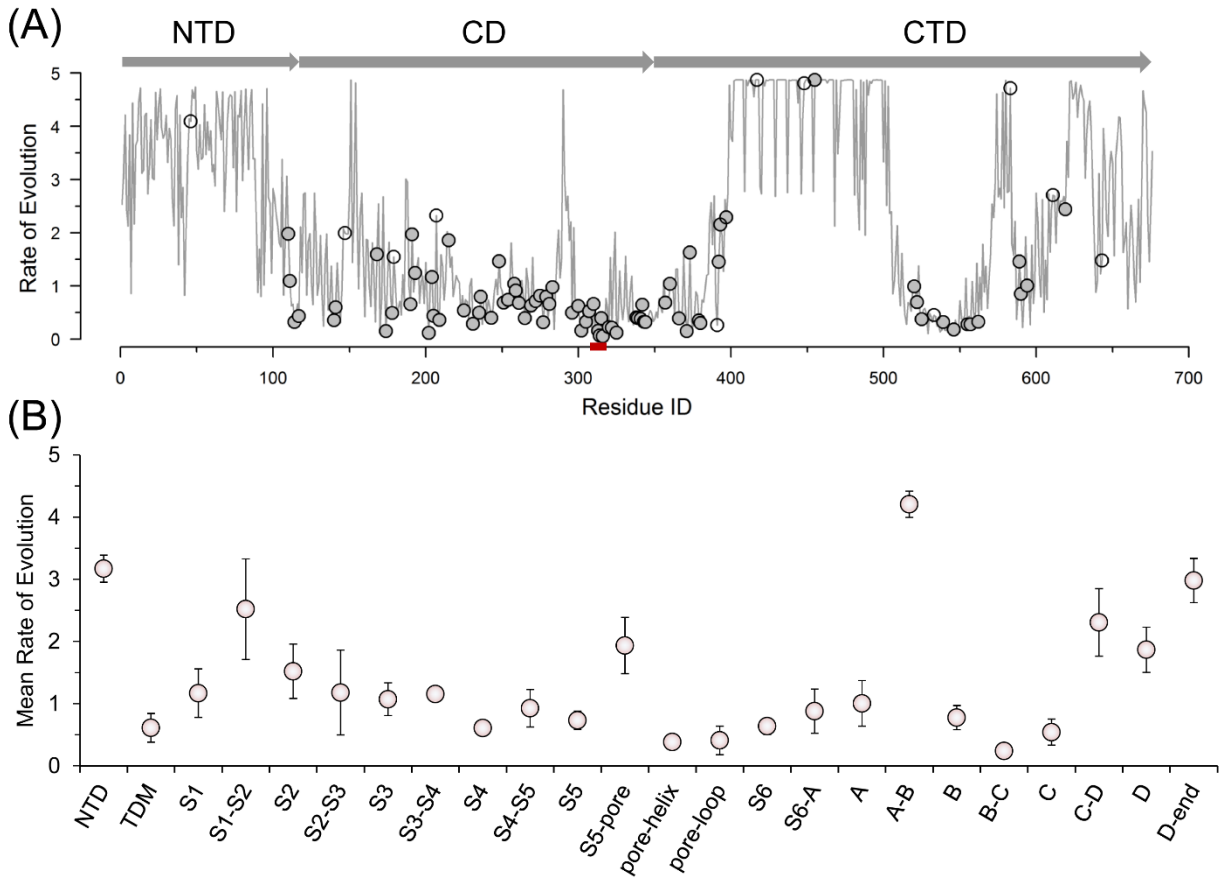


Figure V-1 Analysis on the evolutionary variability of the KCNQ1 sequence

(A) Position-specific rate of evolution. Shaded arrow bars on the top indicate the sequence range of NTD, CD, and CTD respectively. The small red bar on the horizontal axis highlights the "selectivity filter" TIGYG. Closed circles represent dysfunctional variants and open circles represent normal variants. (B) mean rates of evolution for structurally distinct subdomains of NTD, CD, and CTD. Note that the trafficking determinant motif (TDM), which resides within the NTD, is singled out for its distinct functional role. Error bars indicate the 95% confidence intervals (under Student-t distribution) for the mean rate of evolution.

V-3.3 Dysfunctional variants are enriched in selected subdomains

Results from a recent study suggested that the probability of pathogenicity of a KCNQ1 variant depends in part on the topological location of the variant (Giudicessi *et al.*, 2012). However, in the previous study the protein was only divided into three topological domains namely NTD, CD, and CTD. We mapped all variants in our dataset onto the curve of position-specific rates of evolution (Figure V-1(A)). We observed that dysfunctional variants preferentially occur at positions with low rate of evolution, especially within a selected set of subdomains.

In fact, 95.7% (90 / 94) dysfunctional variants occur at positions where the rate of evolution is under 2. In contrast, 61.5% (8 / 13) of normal variants occur at positions with rates of evolution above 2. The five normal variants that occur at positions with a rate of evolution under 2 are: Q147R, G179S, T391I, R533W, and G643S. Interestingly, Q147R, G179S, T391I, and G643S are chemically conserved, as judged by their Grantham distances:(Grantham, 1974) Q→R (68), G→S (56), T→I (89). Nevertheless, this clear segregation of the functional impact of variants with respect to position-specific rate of evolution indicates that the rate of evolution of a sequence site pre-selected as one of the predictive features is indeed a strong predictor on whether variants occurring at the site will be dysfunctional or not.

In addition, we also computed the enrichment of dysfunctional variants for each subdomain, to confirm that such variants are indeed localized within a selected set of subdomains (Calculation of enrichment of dysfunctional variants and Table A-9 in APPENDICES). An enrichment of > 1.0 indicates that the corresponding subdomain has higher than random chance of harboring dysfunctional variants. As shown in Figure V-2, subdomains with higher than random chance for dysfunctional variants are: S0, S2-S3 linker, S3, S4, S4-S5, S5, pore-helix, pore-loop, S6, S6-A, B-C, and C. In particular, S0, S3, S4-S5 linker, S5, pore-loop, and S6-A each have an enrichment ≥ 3. As discussed in the previous section, these subdomains are highly conserved.

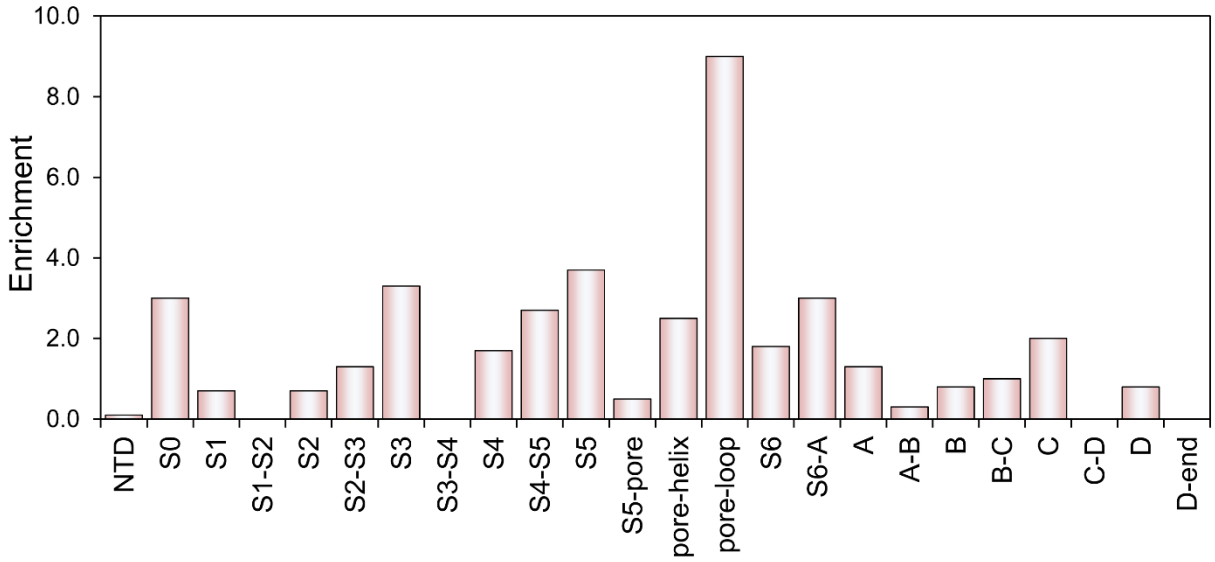


Figure V-2 Bar graph of subdomain-specific enrichment of dysfunctional variants

This bar graph plots subdomain-specific enrichment of dysfunctional variants, showing that dysfunctional variants are enriched in a selected set of subdomains (S0, S3, S4-S5, S5, pore-helix, pore-loop, S6-A, see Table S5 for the residue ranges these subdomains correspond to). One needs to keep in mind that due to the sparsity of functionally characterized variants, the estimates of enrichments are likely to be biased.

V-3.4 Q1VarPred: a KCNQ1-specific predictor

A schematic representation of the architecture of Q1VarPred is shown in Figure V-3(A). Figure V-3(B) shows a visualization of the Q1VarPred model of the relationship between predictive features (rate of evolution and PSSM-based perturbation) and the prediction about functional impact (impact score 0 – most likely normal, 1 – most likely dysfunctional). The contour surface indicates that the impact score has a sharper dependence on the rate of evolution than it does on PSSM-based perturbation. In particular, variants at conserved positions (rate of evolution close to 0) are very likely to be dysfunctional (impact score > 0.5) even if the perturbation is very small. An example of such variants is the dysfunctional V307L whose impact was predicted to be 0.68. The estimated rate of evolution of this position is 0.52, whereas the perturbation introduced by substituting Val for Leu at this position is considerably small (3.7). Similarly, variants at evolutionarily tolerated positions (rate of evolution > 3.0 for example) tend to be normal even if the perturbation is very large (for example, R583H). However, the impact score does rise along with increasing magnitude of perturbation, which is particularly important for predicting the impact at positions exhibiting intermediate rates of evolution.

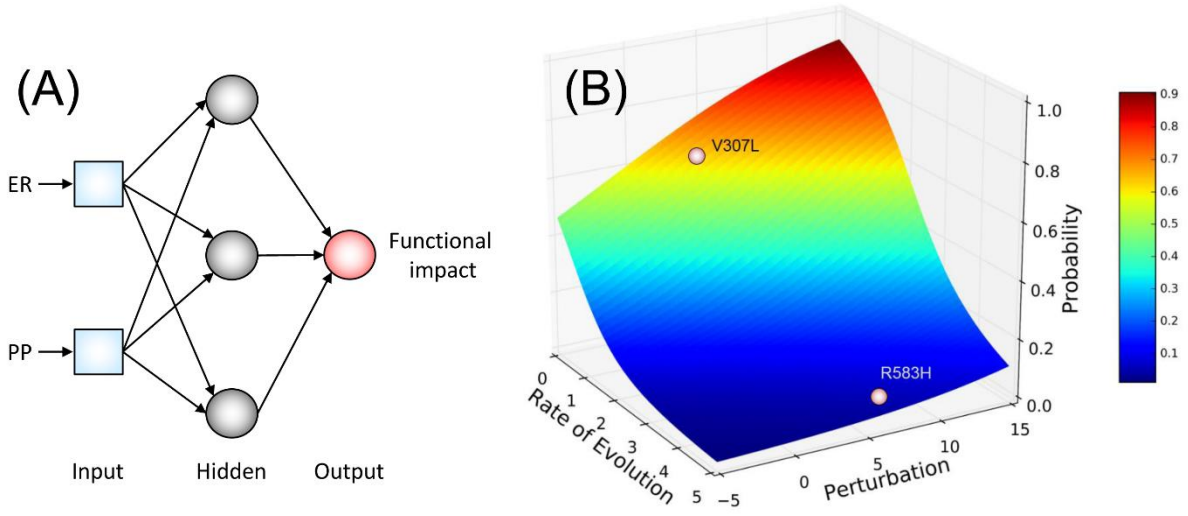


Figure V-3 The neural network architecture and a visualization of Q1VarPred

(A) A schematic representation of the architecture of Q1VarPred. The input layer is composed of two predictive features: rate of evolution (ER) and perturbation derived from PSSM (PP). The hidden layer has three neurons and the output layer has one neuron that computes the final predicted functional impact. (B) A visualization of the Q1VarPred-mapped mathematical relationship between predictive features (rate of evolution and perturbation) and functional impact. The vertical axis is functional impact on the scale of 0 to 1 with 1 being complete dysfunction.

V-3.5 Comparing Q1VarPred with other methods

We employed a procedure called "repeated cross-validation" (Smith *et al.*, 2014) to estimate the generalization ability of Q1VarPred and other methods (see Estimating generalization ability). Seven commonly used genome-wide methods: PhD-SNP, Polyphen-2, PredictSNP, PROVEAN, SIFT, SNAP, and SNPs&GO and one potassium channel-specific method called KvSNP were examined (Methods and Table S6 in Data Supplement). Table V-1 shows that all performance metrics rank Q1VarPred the best, except for NPV and TPR. In general, AUC and MCC are considered the most robust metrics for evaluating classifiers. AUC is independent of user-chosen and therefore possibly biased thresholds. MCC has the advantage to consider all four numbers (TP, TN, FP, FN) and provides a much more balanced evaluation than TPR or TNR individually (Baldi *et al.*, 2000). In terms of AUC, Q1VarPred > PROVEAN > PhD-SNP > SNPs&GO > SIFT > KvSNP > PredictSNP > PolyPhen-2 > SNAP. This is similar to the findings of Leong *et al.* (Leong *et al.*, 2015) except that PolyPhen-2 was shown to rank between PROVEAN and SNP&GO, and PhD-SNP and KvSNP were not evaluated in Leong *et al.* Methods that perform better than Q1VarPred in TPR, do so at a cost of a very low TNR, i.e. the threshold is chosen to minimize the loss of true positives at the cost of predicting many false positives. In some disease conditions, a high fraction of false positives might be acceptable. However, in LQTS and related channelopathies, the cost of false positives is as drastic as that of false negatives (Ackerman, 2015). It is also worth noting that while KvSNP is gene-specific, our evaluation shows that its performance is worse than most genome-wide methods on this dataset. The primary cause of the inflation in KvSNP's claimed performance is probably its convolution of dataset preparation and feature selection, where 85.5% of "neutral variants" were generated from variable sequence positions and later several sequence conservation-based features were selected as predictive features (Stead *et al.*, 2011).

Table V-1 Comparison of Q1VarPred with other methods

Method	Mean performance metric							
	AUC	MCC	PPV	NPV	Accuracy	TPR+TNR	TPR	TNR
Q1VarPred	0.884	0.581	0.968	0.537	0.881	1.680	0.895	0.785
KvSNP	0.662	0.313	0.922	0.344	0.832	1.255	0.887	0.438
PhD-SNP	0.727	0.386	0.941	0.390	0.820	1.453	0.850	0.603
PolyPhen-2	0.636	0.340	0.912	0.547	0.866	1.272	0.939	0.333
PredictSNP	0.652	0.355	0.918	0.459	0.850	1.303	0.912	0.391
PROVEAN	0.770	0.510	0.949	0.537	0.869	1.536	0.902	0.634
SIFT	0.680	0.360	0.927	0.503	0.861	1.364	0.921	0.443
SNAP	0.542	0.101	0.895	0.158	0.771	1.085	0.844	0.241
SNPs&GO	0.697	0.307	0.939	0.296	0.767	1.384	0.792	0.592

V-4 Discussion

V-4.1 From functional impact to clinical disease diagnosis

The goal of our study was to create a highly tailored computational method to predict functional impact. However, translating evidence on functional impact to clinical disease diagnosis is far from trivial. First, every computational method has a certain degree of accuracy and reliability, and those of genome-wide methods are particularly limited. In fact, this is one of the primary motivations of the present study. Second, variants that are dysfunctional at the molecular level may not have clinical manifestation. For example, the A300T variant, which was confirmed experimentally to be severely dysfunctional (Bianchi *et al.*, 2000), was later identified in a cohort considered to be clinically normal.(Kapa *et al.*, 2009) Such dysfunctional variants may have been rescued by compensating genetic variations. Third, trying to predict the clinical outcome without considering the mode of inheritance of LQTS may be problematic. The mode of inheritance is a key factor when determining the clinical relevance of a genotype for LQTS. For example, four variants in our dataset (R231H, W305S, A525T, and R594Q) were functionally normal when expressed in combination with the wild-type channel but were severely dysfunctional in the absence of the wild-type. W305S was identified in members of two consanguineous families with the recessive JLN syndrome (Neyroud *et al.*, 1998) and A525T was suspected to cause the recessive form of RW syndrome (Larsen *et al.*, 1999). Moreover, a functionally normal variant may have compound genetic variations within the same gene or other genes that may obviate or, alternatively, contribute to the clinical phenotype (Westenskow *et al.*, 2004). In light of these

considerations, Q1VarPred was intended for judicious use by researchers or clinicians in conjunction with complementary clinical and genetic evidence to assess the disease susceptibility caused by KCNQ1 variants.

V-4.2 Unexpected conserved subdomains in the C-terminal domain

Figure V-4 shows the topological distributions of position-specific rate of evolution and subdomain-specific enrichment of dysfunctional variants. In our analysis of the rate of evolution in the CTD, we found a few topological subdomains with conserved mean rate of evolution (Figure V-1(B)), predicting important functional or structural roles. Two subdomains, the S6-A linker and the B-C linker, were shown to have a surprisingly low mean rate of evolution (0.88 and 0.24, respectively). While S6-A has an estimated enrichment of dysfunctional variants of 3.0, that of the B-C linker is unexpectedly low (1.0) (Figure V-2 and Table A-9 in APPENDICES). The low enrichment of the B-C linker is likely biased because of the sparsity of functionally validated variants (e.g. only three functionally validated variants are located in the B-C linker). In fact, another six variants (Table A-11 in APPENDICES) found in this subdomain have been deposited in ClinVar (Landrum *et al.*, 2014). However, they were not included into our dataset as we were not able to find literature describing their functional validation. The enrichment of the B-C linker is likely to increase when larger datasets of functionally validated variants become available for estimating enrichments. More importantly, there seems to be a lack of study documenting the functional roles the S6-A linker and the B-C linker. Nevertheless, based on their low rate of evolution, we alert investigators about the potential high functional impact of variants found in these two subdomains.

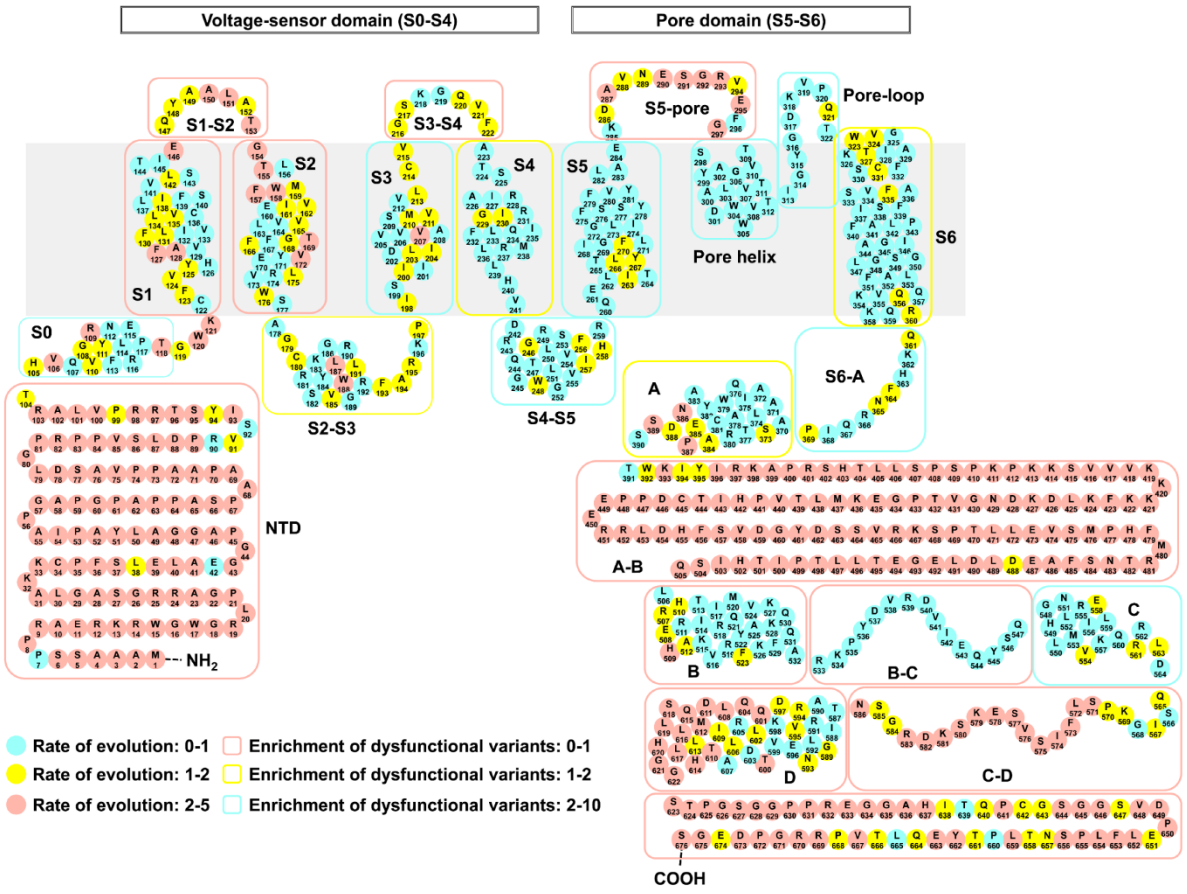


Figure V-4 A "global" view of the topological distribution of rate of evolution and enrichment of dysfunctional variants

V-4.3 The machine learning model

Ideally, a machine learning algorithm should produce a learned model that is accurate at predicting new observations and, at the same time, simple enough to allow straightforward interpretation. In general, linear models are easier to interpret, while nonlinear models are more powerful in cases where classes are not linearly separable. We chose a neural network, which generally is considered to be a nonlinear model, for the present study to leverage our extensive experience with neural networks and an established library for feature engineering and model building (Butkiewicz *et al.*, 2013, Leman *et al.*, 2013, Li *et al.*, 2016, Mendenhall and Meiler, 2016, Li *et al.*, 2017a). Admittedly, a logistic regression model performed only slightly worse (AUC = 0.855) and a linear discriminant classifier performed comparably (AUC = 0.870). However, given the complexity in the mechanisms behind KCNQ1 dysfunction, we expect that the "true" decision boundary between normal and dysfunctional variants is complex. As additional experimental data become available,

the advantage of neural networks for prediction over linear models is likely to become more substantial.

V-4.4 Factors contributing to the improved performance of Q1VarPred

Q1VarPred offers improved overall performance in predicting functional impact of variants on a KCNQ1-specific basis compared to the other evaluated tools (Table V-1). Although most tools allow for predictions for a wide range of proteins, the fact that each method applies a single threshold to classify variants on all proteins may be partially responsible for their weaker overall performance on KCNQ1 variants. Additionally, recent work has shown that contemporary variant–phenotype and variant–stability prediction algorithms are substantially worse at predicting outcomes for membrane proteins, such as KCNQ1, than for water soluble proteins (Kroncke *et al.*, 2016).

The observed higher performance of Q1VarPred may also be attributed to better predictive features. Many methods use MSA-derived position-specific conservation scores as predictive feature, presumably based on the assumptions that the functional importance of a given position dictates how conserved this position is and, conversely, that the degree of conservation indicates the functional importance of this position. While this latter assumption is often valid, position-specific conservation scores computed directly from MSA without considering the evolutionary history of the aligned protein family may be biased because of unevenly sampled sequence space. Numerous position-specific quantitative conservation scores have been proposed over the years (Valdar, 2002) and all evaluated methods except the meta-predictor PredictSNP use as position-specific conservation measures of some sort derived from MSA as predictive features. However, none of these methods consider the topology and branch lengths of phylogenetic trees as the method used in the current study does (Methods in Data Supplement). Thus, these conservation measures may lead to less accurate estimations of rate of evolution.

The other predictive feature used in Q1VarPred is the perturbation derived in the context of a PSSM. This feature measures how much less likely it is for the variant to occur at a sequence position relative to the wild-type. The higher the perturbation the less likely for the variant to replace the wild type residue at a specific position. While the position-specific rate of evolution presumably is a strong predictor of functional impact, it only indicates how likely it is that the wild-type amino acid at this sequence position changes. It does not, however, tell how likely it is

that the wild-type amino acid is changed to one particular amino acid type over the others. In other words, the perturbation adds additional information by complementing position-specific rates of evolution with what the actual variants are.

V-5 Limitations and future direction

The primary limitation of the current study is the size of the dataset. Although a substantial amount of effort was spent by many labs to experimentally characterize the 107 variants treated in this study, the dataset used in this study is still very small, relative to that used to train other contemporary variant-effect predictors. As a result, we were limited from selecting a set of most relevant features in a systematic, algorithmic manner. Thus, it is very likely that we missed some very informative sequence-based features. When larger datasets become available, Q1VarPred can be re-trained and new predictive features can be tested. In addition, our estimation of enrichment of dysfunctional variants for each subdomain is also likely to be biased due to this data sparsity. Even though the enrichment values correlate well with average rates of evolution and our analysis shows that functionally important subdomains tend to be more enriched with dysfunctional variants, there is currently not enough data available to demonstrate that such relationship for KCNQ1 is statistically significant.

Recent investigations into machine learning have shown that training neural networks on multiple traits/outcomes per training example can improve performance (Qi *et al.*, 2012, Heffernan *et al.*, 2015). Specifically, the advantages of simultaneously training a neural network to predict multiple outcome variables (disease severity, electrophysiological parameters, etc.) may enable a more accurate prediction of phenotype traits as well. Previous work aimed at predicting secondary structure and membrane burial for residues has suggested that neural networks trained to predict multiple outcomes are particularly beneficial when the dataset size is especially small (Leman *et al.*, 2013). This suggests that such neural networks may be particularly suitable to leverage the diverse experimental parameters available for LQTS variants and phenotypes.

The method developed in this study is modular in the sense that one possible future direction is to combine this method with other predictors—such as estimation of the impact of genetic variations on protein stability, to come up with predictions that are both more reliable and that also suggest mechanisms underlying variation-induced gain or loss of function.

V-6 Data and Software Availability

The curated data set is included in APPENDICES and designated as Table . The dataset for training Q1VarPred is provided as a spreadsheet in Supplemental Materials. Q1VarPred was developed under the framework of the Biochemical Library (available at http://www.meilerlab.org/bclcommons) and is made publicly available as a web server at http://meilerlab.org/q1varpred.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

VI-1 Contributions

De novo tertiary structure prediction for proteins is still an unsolved problem. This is partially because the size of the conformational space of proteins is intrinsically large and functions approximating the free energy landscape of the conformational space of large biomolecules is far from sufficiently accurate. An approach developed by my colleagues to improving the efficiency of sampling and the power of scoring is to assemble secondary structure elements predicted by machine learning-based methods into 3D models (see I-5.5 BCL::Fold) and further using sparse experimental data as restraints to reduce conformational space and improve the likelihood of identifying native-like models by the energy function (see I-4 Improving sampling and scoring with restraints).

In the first part of this work (chapters II, III, and IV), a novel approach was developed where machine learning models were trained to predict structural properties of amino acid residues from sequence, which were in turn encoded as restraints into the energy function to improve tertiary and quaternary structure prediction for transmembrane proteins and their complexes. Specifically, in chapter II a neural network method was trained with the "dropout" regularization technique to predict WCNs for HMPs. This method is named TMH-Expo, and to the best of our knowledge, it is the first published method for predicting WCNs for HMPs. Trained on an expanded non-redundant data set of HMPs with the "jackknife" cross-validation technique, TMH-Expo achieved an unprecedented Pearson correlation coefficient of 0.69 between experimental and predicted WCNs. A web server at http://meilerlab.org/servers/tmh_expo was also set up for public access to TMH-Expo. In chapter III, WCNs predicted by TMH-Expo were explicitly incorporated as restraints into the membrane protein structure prediction algorithm, BCL::MP-Fold, and were tested whether they will help tertiary structure prediction for HMPs. It was demonstrated that WCN restraints helped sample more accurate helix rotation angles (e.g. increased fraction of native contacts) and fold and improved the ability of the scoring function to select native-like models. In chapter IV, a novel algorithm for docking HMPs, named BCL::MP-Dock, was described and benchmarked. It was shown that for all of the 15 test homodimers, BCL::MP-Dock is able to reconstruct a model of the complex in which the RMSD100 of the docked ligand subunit to its native structure is less than 3.0 Å. It was also shown that the ability of BCL::MP-Dock to identify

native-like models can be substantially improved if the algorithm is supplied with predicted interface residues and their predicted WCNs as restraints

In the second part (chapter V), a detailed quantitative analysis on the sequence conservation patterns of subdomains of KCNQ1 and the distribution of pathogenic variants was conducted based on a "high-quality" set of 107 functionally characterized KCNQ1 variants curated from the literature. It was found that conserved subdomains generally are critical for channel function and are enriched with dysfunctional variants. Using this experimentally validated dataset, a neural network, designated Q1VarPred, was trained specifically for predicting the functional impact of KCNQ1 variants of unknown significance. The estimated predictive performance of Q1VarPred in terms of Matthew's correlation coefficient and area under the receiver operating characteristic curve were 0.581 and 0.884, respectively, superior to the performance of eight previous methods tested in parallel. Q1VarPred was made publicly available as a web server at http://meilerlab.org/q1varpred. Although a plethora of tools are available for making pathogenicity predictions over a genome-wide scale, previous tools fail to perform in a robust manner when applied to KCNQ1. The contrasting and favorable results for Q1VarPred suggests a promising approach, where a machine learning algorithm is tailored to a specific protein target and trained with a functionally validated dataset to calibrate informatics tools.

The success of the novel approach developed in the first part of this work calls for some reflection. First, why can WCNs be accurately predicted from sequence? WCN is a structural feature that indicates how densely packed a residue is within the context of protein tertiary structure. In light of the fact that protein interior (densely packed region) is generally more conserved than the surface (loosely packed region) and the evidence that WCN is the main determinant of site-specific rate of evolution (Echave *et al.*, 2016), accurate prediction of WCNs from multiple alignment of protein family sequences is expected. Second, why does incorporating WCNs predicted from sequence as restraints give better sampling and scoring than a knowledge-based potential derived from a database of structures did? Statistics have shown that in membrane proteins, most amino acid types, hydrophobic ones in particular, have a wide range of preferred WCNs, say from 4 to 12 (Weiner *et al.*, 2013). Thus, a knowledge-based potential would score equally favorably no matter whether these amino acids are buried or exposed while in really in some positions these amino acids are buried and in others exposed. Using WCNs predicted from

sequence as restraints solves this problem by explicitly teaching the algorithm the expected WCN at each position. Even for positions of the same amino acid type, the expected WCNs predicted from sequence by a machine-learning method may differ considerably depending on how structurally constrained these positions are.

VI-2 Limitations and future directions

BCL::Fold (and BCL::MP-Fold as well) assembles secondary structure elements (SSEs), namely α-helices and β-strands, which are geometrically pure. Individual residues are represented by their backbone and C_β atoms only ($H_{\alpha 2}$ for glycine), but side-chains are not explicitly modeled. This representation scheme was employed in part to make room for computational speed. However, it has some major limitations. First, under this representation scheme all residues are essentially identical from a physicochemical point of view, the algorithm relies solely on the associated knowledge-based potential function, which by itself has limited accuracy (Thomas and Dill, 1996). Second, it may result in physically unrealistic models. For example, two residues not clashing as judged according to the separation between their C_β atoms may have substantial clashes if their side-chains were present. Third, representing SSEs in a geometrically pure fashion without a compensating efficient scheme for sampling the backbone flexibility within individual SSEs produces unreasonably "rigid" structural models. To address these limitations would require modifications to the current representation scheme. For example, it would be physically more sensible to reduce the side-chain to a pseudo-atom whose location is taken as the average location of the side-chain atoms than simply to the C_β atom. Such a representation together with a knowledge-based potential derived from updated statistics based on side-chain centroids should partially resolve the first two limitations. To resolve the inefficiency of sampling backbone flexibility, one potential approach to test is to use ensembles of probable conformations extracted from protein structure database for individual SSEs as starting pools for the algorithm to sample from.

In training TMH-expo for WCN prediction, it was found that while in most cases high WCN is correlated with high sequence site conservation, this is not generally true (unpublished result, APPENDICES Figure A-3). This observation may partially explain why in some cases, especially where site conservation is not a strong predictor of WCN, our method TMH-Expo failed to give reasonably accurate prediction (say PCC > 0.5 and MAE < 2.0) of WCNs. The ultimate root cause

why WCNs may not be strongly correlated with site conservation is itself an interesting topic to study. One potential explanation could be that some proteins may be under positive selection where site variability is required to explore amino acid substitutions that produce variants with higher fitness. Another possibility is that there may only be a specific subset of residues whose WCNs are strongly correlated with their site conservation and only this specific subset of residues is truly essential for maintaining the protein's structural integrity. There may exist a substantial fraction of residues that are not structural constrained even though their apparent WCNs as calculated from available experimental structures are high. Another critical aspect of protein tertiary structure is that proteins are dynamic entities, they often undergo substantial conformational change when they are performing their biological function. In light of this conformational flexibility, each residue should have a range of allowed WCNs, meaning that the WCN of each residue is less likely to be a fixed number.

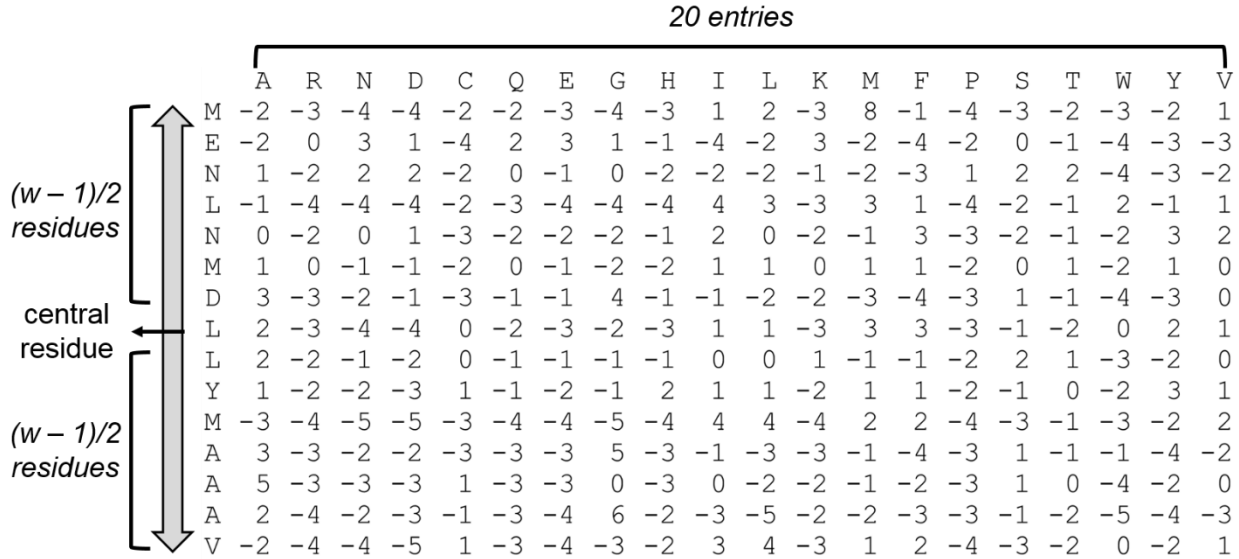
In developing the neural network-based method for predicting interface residues and their WCNs, it was hypothesized that interface residues are either functionally or structurally more constrained (e.g. lower rate of evolution) than residues on the rest of the surface. The higher degree of conservation of interface residues would allow them to be identified and their WCNs to be accurately predicted. Surprisingly, the interface is significantly more conserved than the rest of the surface in only 26 out of 44 cases from the data set (APPENDICES Figure A-4). There are 4 cases where the average rate of evolution of the interface is higher than that of the rest of the surface, indicating that these interfaces might not be the most biologically relevant and there may be other patches on the surface that are structurally or functionally more important. In the remaining 14 cases, the distribution of rate of evolution of interface residues is not easily distinguishable from that of residues on the rest of the surface. It is still an open question as to what this observation implies.

Interpretation of the functional impact of amino acid substitutions in proteins is still a challenging problem. In the second half of this work, a neural network named Q1VarPred was trained to classify the functional impact (normal versus dysfunctional) of variants of unknown significance of the KCNQ1 potassium channel. While this method demonstrated superior performance on KCNQ1 when compared with eight other methods, it's applicability may be limited due to its nature of being a classifier. Ideally, one would not only like to have a

computational method that does perfect classification, but also produces prediction values correlated with phenotype severity and gives testable hypotheses about the potential mechanisms by which a dysfunctional variant affect the protein. In the future, it is desirable to have updated versions of Q1VarPred that are able to make real-valued prediction of how the physiological parameters change upon amino acid substitution and enable an interpretation of such changes at the structural level.

APPENDICES

Accurate prediction of contact numbers for multi-spanning helical membrane proteins



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
M	-2	-3	-4	-4	-2	-2	-3	-4	-3	1	2	-3	8	-1	-4	-3	-2	-3	-2	1
E	-2	0	3	1	-4	2	3	1	-1	-4	-2	3	-2	-4	-2	0	-1	-4	-3	-3
N	1	-2	2	2	-2	0	-1	0	-2	-2	-2	-1	-2	-3	1	2	2	-4	-3	-2
L	-1	-4	-4	-4	-2	-3	-4	-4	-4	4	3	-3	3	1	-4	-2	-1	2	-1	1
N	0	-2	0	1	-3	-2	-2	-2	-1	2	0	-2	-1	3	-3	-2	-1	-2	3	2
M	1	0	-1	-1	-2	0	-1	-2	-2	1	1	0	1	1	-2	0	1	-2	1	0
D	3	-3	-2	-1	-3	-1	-1	4	-1	-1	-2	-2	-3	-4	-3	1	-1	-4	-3	0
L	2	-3	-4	-4	0	-2	-3	-2	-3	1	1	-3	3	3	-3	-1	-2	0	2	1
L	2	-2	-1	-2	0	-1	-1	-1	-1	0	0	1	-1	-1	-2	2	1	-3	-2	0
Y	1	-2	-2	-3	1	-1	-2	-1	2	1	1	-2	1	1	-2	-1	0	-2	3	1
M	-3	-4	-5	-5	-3	-4	-4	-5	-4	4	4	-4	2	2	-4	-3	-1	-3	-2	2
A	3	-3	-2	-2	-3	-3	-3	5	-3	-1	-3	-3	-1	-4	-3	1	-1	-1	-4	-2
A	5	-3	-3	-3	1	-3	-3	0	-3	0	-2	-2	-1	-2	-3	1	0	-4	-2	0
A	2	-4	-2	-3	-1	-3	-4	6	-2	-3	-5	-2	-2	-3	-3	-1	-2	-5	-4	-3
V	-2	-4	-4	-5	1	-3	-4	-3	-2	3	4	-3	1	2	-4	-3	-2	0	-2	1

Figure A-1 Illustration of feature vectors

PSSM, BPP, or LSC is an l × 20 matrix where l denotes the length a protein sequence. For each sequence position i, there are 20 entries with each corresponding to the score (PSSM) or probability (BPP) of one of the 20 naturally occurring amino acid to occur at position i. In the case of LSC, each position is encoded by a vector of 20 binary entries (bits). When considering w nearest residues on either side of the residue whose contact number is to be predicted (central residue), the feature vector computed based on PSSM, BPP, or LSC has a total of w × 20 entries. Whereas the feature vector computed based on CI has a total of w × 1 entries. Shown in this figure is an original integer-valued PSSM obtained from PSI-BLAST search (Altschul *et al.*, 1997).

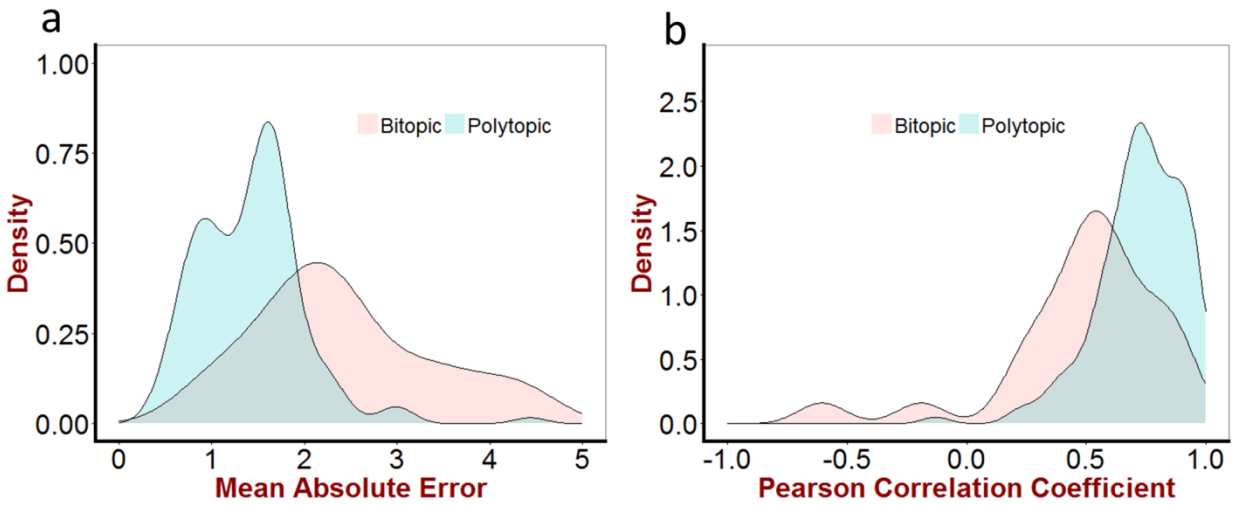


Figure A-2 Distributions of MAE and PCC for bipotic and polytopic HPMs

(a) Distribution of MAE; (b) Distribution of PCC.

Table A-1 HMP chains in the TMH-Expo data set

1kqgB	3ar4A	3tuiA
1kqgC	3ayfA	3tx3A
1m57D	3b9zA	3w4tA
1orsC	3cn5A	3wguB
1qlbC	3cx5C	3ze3D
1u7cA	3cx5D	4a01A
2fyuK	3cx5E	4al0A
2h88C	3cx5H	4bbjA
2h88D	3cx5I	4buoA
2nq2A	3d31C	4bwzA
2q72A	3detA	4cadC
2vpyC	3egwC	4dveA
2wdqC	3gi8C	4dx5A
2wdqD	3ob6A	4ezcA
2wjnH	3oufA	4g7vS
2wjnM	3p4pC	4gc0A
2wswA	3p4pD	4gycB
2xq2A	3qe7A	4huqS
2zxeG	3rfrA	4huqT
2zy9A	3rfrB	4hyjA
3ag3A	3rfrC	4jrzA
3ag3B	3rlbA	4k1cA
3ag3C	3rlfF	4kppA
3ag3D	3rvyA	4kt0F
3ag3G	3s3wA	4ky0A
3ag3I	3s8gA	4kytB
3ag3J	3s8gB	4lp8A
3ag3K	3tdsA	4n6hA
3ag3L	3tijA	4n7wA
3ag3M	3tlwA	4njpA

Table A-2 Summary of 12 poorly predicted protein chains

Protein Chain	MAE	PCC	Bitopic
4kytB	2.52	0.35	Bitopic
3ag3G	2.59	0.28	Bitopic
3cx5E	2.69	0.30	Bitopic
4gycB	2.82	0.14	Polytopic
3ag3M	2.85	0.41	Bitopic
4kt0F	3.11	0.56	Bitopic
2zxeG	3.29	0.61	Bitopic
3cx5I	3.72	0.74	Bitopic
3ag3K	3.91	0.45	Bitopic
3ag3I	4.02	0.36	Bitopic
2fyuK	4.10	0.63	Bitopic
1m57D	4.41	-0.21	Bitopic

Table A-3 Performance of TMH-Expo on Identifying Interface Residues

		Predicted		
		Interface	Non-interface	Total
Experimental	Interface	1444	437	1881
	Non-interface	3191	6493	9684
	Total	4635	6930	11565

Table A-4 Alpha-helical transmembrane protein chains that form the oligomers in the data set

Protein chain	Resolution (Å)	Oligomeric state	Obligate
1m56A	2.3	ht4	1
1m56C	2.3	ht4	1
1q16C	1.9	hm2	0
1u7gA	1.4	hm3	1
1yq3C	2.2	ht4	0
1yq3D	2.2	ht4	0
2a65A	1.7	hm2	1
2bl2A	2.1	hm10	1
2bs2C	1.8	hm2	0
2j8cM	1.9	ht3	1
2nq2A	2.4	hm2	1
2qtsA	1.9	hm3	1
2uuhA	2.2	hm3	1
2vpzC	2.4	hm2	0
2w2eA	1.2	hm4	1
2wswA	2.3	hm3	1
2yevC	2.4	ht3	1
2z73A	2.5	hm2	0
2zxeA	2.4	ht3	1
3b9yA	1.9	hm3	1
3c02A	2.1	hm4	1
3cx5C	1.9	ht5	1
3k3fA	2.3	hm3	1
3klyA	2.1	hm5	1
3m73A	1.2	hm3	1
3oduA	2.5	hm2	0
3oufA	1.6	hm4	1
3puwF	2.3	ht2	1
3s8gA	1.8	ht3	1
3spcA	2.5	hm4	1
3tijA	2.4	hm3	1
4a01A	2.4	hm2	1
4bpmA	2.1	hm3	1
4d2eA	2.3	hm3	1
4dx5A	1.9	hm3	1
4f4sA	1.9	hm10	1
4jkvA	2.5	hm2	0
4mrsA	2.4	hm2	1
4o6mA	1.9	hm2	1
4o6yA	1.7	hm2	1
4qndA	1.7	hm2	1
4rngC	2.4	hm2	1
4u9nA	2.2	hm2	1
4wd8A	2.3	hm5	1

hm2: homodimer, hm3: homotrimer, hm4: homotetramer, hm5: homopentamer, hm10: homodecamer, ht2: heterodimer, ht3: heterotrimer, ht4: heterotetramer, ht5: heteropentamer; Obligate: 1 (yes), 0 (no).

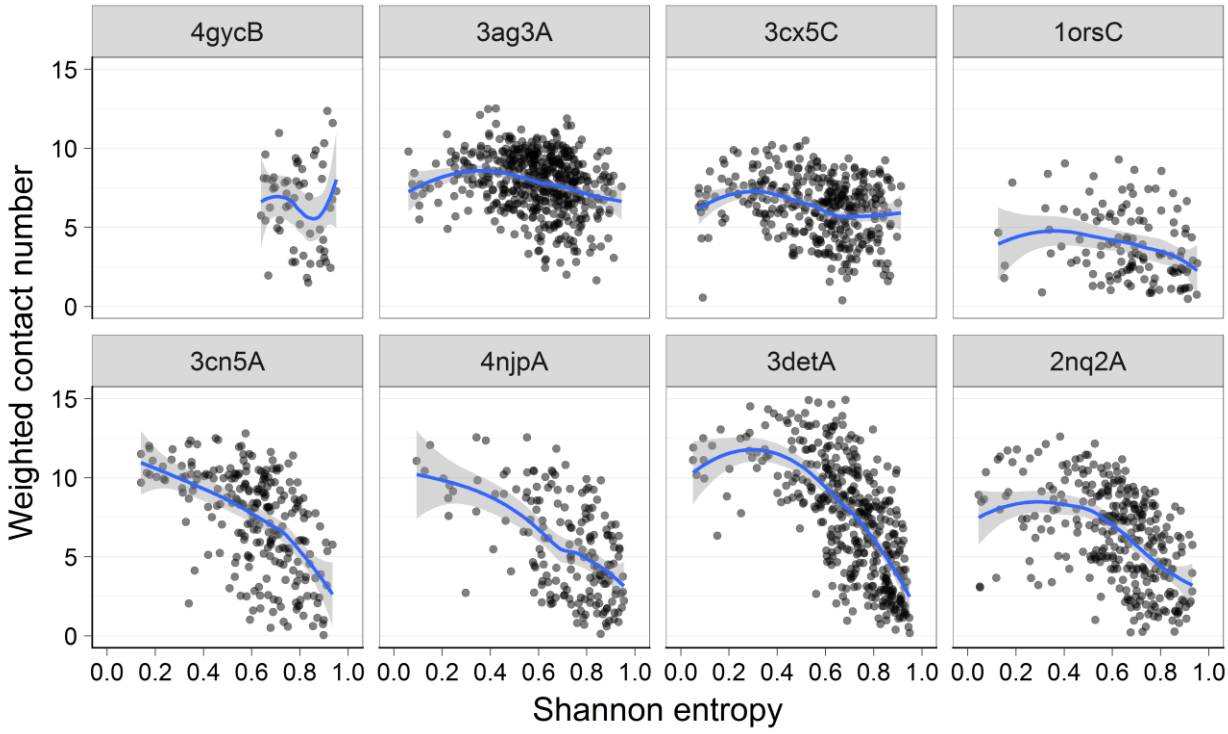


Figure A-3 Illustration of the weak and strong correlation between site variability and WCN

Figure A-4 Distribution of site-specific rate of evolution of interface residues and non-interface surface residues

Hypotheses were tested using the Mann-Whitney test. The null hypothesis H_0 was that the mean rate of evolution of interface residues is not lower than that of the non-interface residues.

Predicting the functional impact of KCNQ1 variants of unknown significance

Computation of predictive features

Position-specific rate of evolution was estimated using the Rate4Site method (Pupko *et al.*, 2002). While rates of evolution are commonly measured as the number of substitutions per sequence position per year (Lanfear *et al.*, 2014), it should be noted that the rate estimated by Rate4Site is relative to the average evolutionary rate across all positions and hence is unitless. The input multiple sequence alignment (MSA) of KCNQ1 homologs to Rate4Site was obtained by running HHblits against the Uniprot20 sequence database (Remmert *et al.*, 2012), with minimum coverage of master sequence (KCNQ1 wild-type sequence) set to 25%, minimum sequence identity to master sequence set to 15%, maximum pairwise sequence identity set to 90%, and E-value cutoff

for inclusion in result alignment set to 0.001. The total number of aligned sequences was limited to 300 as our testing showed that Rate4Site suffered from underflow problems when larger numbers of sequences were used.

For characterizing the severity of amino acid substitutions at a position, it is important to conduct the assessment in the context of MSA where the perturbation resulting from amino acid substitution can be quantified from the perspective of protein evolution. We derived this perturbation from the position-specific scoring matrix (PSSM, Figure S1) obtained by searching the NCBI non-redundant sequence database (Pruitt *et al.*, 2007) with PSI-BLAST (Altschul *et al.*, 1997) for four iterations. The E-value inclusion threshold was set to 0.00001. For a protein of length L, a PSSM is a $L \times 20$ matrix containing log ratios of the estimated frequency of each of the 20 amino acids to occur at each position relative to the expected frequency of the wild-type amino acid in a random sequence. If P_A is the probability for amino acid A to occupy a position and P_A^0 is its background probability, then the PSSM entry for A at this position equals $\lambda \ln \frac{P_A}{P_A^0}$, where λ is a scaling factor built in PSI-BLAST.(Altschul *et al.*, 1997) One might recognize that this formula resembles the equation for calculating Gibbs free energy change for a chemical reaction ($\Delta G = -RT \ln K$). Similar in spirit to free energy perturbation, we define the perturbation introduced by amino acid substitution from A to B in the context of MSA as: $\lambda \left(\ln \frac{P_A}{P_A^0} - \ln \frac{P_B}{P_B^0} \right)$. Intuitively, the more substantial the perturbation the less likely it is for a variation to occur without a functional or structural impact.

Tested genome-wide tools

Seven genome-wide prediction tools: PhD-SNP (Capriotti *et al.*, 2006), PolyPhen-2 (Adzhubei *et al.*, 2010), PredictSNP (Bendl *et al.*, 2014), PROVEAN (Choi *et al.*, 2012), SIFT (Ng and Henikoff, 2001), SNAP (Bromberg and Rost, 2007), and SNPs&GO (Calabrese *et al.*, 2009) and a potassium channel-specific method KvSNP (Stead *et al.*, 2011) were tested for their ability to predict functionality of KCNQ1 variants. PhD-SNP, PolyPhen-2, PredictSNP, SIFT, and SNAP were recently shown to have an overall Matthew's correlation coefficient (MCC) > 0.35 and an overall area under the receiver-operating characteristics curve (AUC) > 0.70 on a fully independent test set consisting of variants from multiple genes (Bendl *et al.*, 2014). PROVEAN and SNPs&GO were shown to have high accuracy to classify LQTS gene variants. These selected tools differ in

the machine learning algorithms with which they were trained and in the required input features. A summary of these tools is presented in Table A-10.

Calculation of enrichment of dysfunctional variants

Based on a homology model of the homotetrameric transmembrane channel domain (Smith *et al.*, 2007), and a structural study of the C-terminal domain of KCNQ1 (Wiener *et al.*, 2008), we mapped the sequence of KCNQ1 into 24 topologically distinct regions and assigned each variant to the region within which it is located (Table A-9). The enrichment of dysfunctional variants for a region is computed as the ratio of observed number of dysfunctional variants (O_v) to the number of dysfunctional variants that would otherwise be observed if each sequence position were equally likely to raise dysfunctional variants, denoted as E_v. E_v can be easily obtained with

$$E_v = \frac{L_s}{L_p} \times N_v$$

where L_s and L_p are the length of the segment and the protein, respectively, and N_v is the total number of dysfunctional variants in the data set.

Table A-5 Functionally characterized KCNQ1 variants curated from the literature.

Residue ID	Wild	Variant	Clinical	I_{ks} ratio	$V_{1/2}$ (mV)	Activation τ ratio	Deactivation τ ratio	Surface	Annotation	Label	Reference
46	A	T	Case	100 %	0	0.6			Normal	Normal	(Yang *et al.*, 2009)
110	V	I	Control	40%	30			Normal	Severe LOF	Dysfunctional	(Cordeiro *et al.*, 2010)
111	Y	C	Case	0%				Absent	Severe LOF	Dysfunctional	(Dahimene *et al.*, 2006)
114	L	P	Case	0%				Absent	Severe LOF	Dysfunctional	(Dahimene *et al.*, 2006)
117	P	L	Case	0%				Impaired	Severe LOF	Dysfunctional	(Dahimene *et al.*, 2006)
140	S	G	Case	150 %					Severe GOF	Dysfunctional	(Campbell *et al.*, 2013)
141	V	M	Case	300 %	0				Severe GOF	Dysfunctional	(Hong *et al.*, 2005)

147	Q	R	Case	60%	0				Mild LOF	Normal	(Lundby *et al.*, 2007)
168	G	R	Case	5%					Severe LOF	Dysfunctional	(Westenskow *et al.*, 2004)
174	R	C	Case	47%	17	1			Mild LOF	Dysfunctional	(Matavel *et al.*, 2010)
178	A	T	Case	41%	45	1.68	0.86	Impaired	Severe LOF	Dysfunctional	(Harmer *et al.*, 2014)
179	G	S	Control	54%	-12				Mild LOF	Normal	(Westenskow *et al.*, 2004)
190	R	Q	Case	0%					Severe LOF	Dysfunctional	(Chouabe *et al.*, 2000)
191	L	P	Case	22%	0			Impaired	Severe LOF	Dysfunctional	(Pan *et al.*, 2009)
193	F	L	Case	80%	0	1.83			Severe LOF	Dysfunctional	(Yamaguchi *et al.*, 2003)
202	D	E	Case	11%	54.6	1	0.33		Severe LOF	Dysfunctional	(Eldstrom *et al.*, 2010)
202	D	H	Case	41%	16.6	0.83	0.26	Normal	Severe LOF	Dysfunctional	(Eldstrom *et al.*, 2010)
202	D	N	Case	20%	23.8	0.55	0.09	Normal	Severe LOF	Dysfunctional	(Eldstrom *et al.*, 2010)
204	I	F	Case	23%	53.3	7.25	0.43	Normal	Severe LOF	Dysfunctional	(Eldstrom *et al.*, 2010)
204	I	M	Case	34%	36.1	1.16	0.65	Normal	Severe LOF	Dysfunctional	(Eldstrom *et al.*, 2010)
204	I	N	Case		32.9	2.47	0.7		Severe LOF	Dysfunctional	(Eldstrom *et al.*, 2010)
205	V	M	Case	36%	20	1.48	0.42		Severe LOF	Dysfunctional	(Eldstrom *et al.*, 2015)
207	V	M	Control	93%	7.1	1.4	1.2		Normal	Normal	(Eldstrom *et al.*, 2010)

209	S	F	Case	35%	-48.7					Severe LOF	Dysfunctional	(Eldstrom *et al.*, 2010)
209	S	P	Case	200 %	-42.4			5.7		Severe GOF	Dysfunctional	(Das *et al.*, 2009)
215	V	M	Case	41%	20.2					Severe LOF	Dysfunctional	(Eldstrom *et al.*, 2010)
225	S	L	Case	10%	11				Normal	Severe LOF	Dysfunctional	(Bianchi *et al.*, 2000)
231	R	C	Case	5%						Severe LOF	Dysfunctional	(Itoh *et al.*, 2009)
231	R	H	Case	15%	40					Severe LOF	Dysfunctional	(Itoh *et al.*, 2009)
235	I	N	Case	10%						Severe LOF	Dysfunctional	(Bartos *et al.*, 2014)
236	L	R	Case	0%	54				Impaired	Severe LOF	Dysfunctional	(Steffensen *et al.*, 2015)
243	R	C	Case	12%	67	1				Severe LOF	Dysfunctional	(Matavel *et al.*, 2010)
243	R	H	Case	13%					Normal	Severe LOF	Dysfunctional	(Huang *et al.*, 2001)
248	W	R	Case	0%						Severe LOF	Dysfunctional	(Franqueza *et al.*, 1999)
251	L	P	Case	0%					Normal	Severe LOF	Dysfunctional	(Deschenes *et al.*, 2003)
254	V	M	Case	7%	41.5					Severe LOF	Dysfunctional	(Wedekind *et al.*, 2004)
258	H	R	Case	5%	-44	0.5		2.5	Impaired	Severe LOF	Dysfunctional	(Labrom *et al.*, 2010)
259	R	C	Case	30%	10					Severe LOF	Dysfunctional	(Kubota *et al.*, 2000)
259	R	H	Case	200 %	1			1.7	Normal	Severe GOF	Dysfunctional	(Wu *et al.*, 2015)
261	E	D	Case	9%						Severe LOF	Dysfunctional	(Huang *et al.*, 2001)

261	E	K	Case	5%					Severe LOF	Dysfunctional	(Franqueza *et al.*, 1999)
265	T	I	Case	100 %	8	2			Severe LOF	Dysfunctional	(Yang *et al.*, 2009)
269	G	D	Case	0%					Severe LOF	Dysfunctional	(Chouabe *et al.*, 1997)
269	G	S	Case	15%	70.7	1	0.4	Impaired	Severe LOF	Dysfunctional	(Wu *et al.*, 2014)
272	G	V	Case	34%	10				Severe LOF	Dysfunctional	(Oka *et al.*, 2010)
275	F	S	Case	34%	27	1.5	2	Impaired	Severe LOF	Dysfunctional	(Li *et al.*, 2009b)
277	S	L	Case	0%	-8.7				Severe LOF	Dysfunctional	(Aidery *et al.*, 2011)
279	F	I	Case	150 %	-25	0.42	1	Normal	Severe GOF	Dysfunctional	(Moreno *et al.*, 2015)
281	Y	C	Case	0%				Normal	Severe LOF	Dysfunctional	(Bianchi *et al.*, 2000)
283	A	T	Case	20%	9				Severe LOF	Dysfunctional	(Crotti *et al.*, 2013)
296	F	S	Case	12%	-10				Severe LOF	Dysfunctional	(Yang *et al.*, 2009)
300	A	T	Control	15%	-19			Normal	Severe LOF	Dysfunctional	(Bianchi *et al.*, 2000)
302	A	V	Case	5%					Severe LOF	Dysfunctional	(Yang *et al.*, 2009)
305	W	S	Case	0%					Severe LOF	Dysfunctional	(Chouabe *et al.*, 1997)
307	V	L	Case	130 %	-18	0.52			Severe GOF	Dysfunctional	(Bellocq *et al.*, 2004)
310	V	I	Case	0%	60				Severe LOF	Dysfunctional	(Westenskow *et al.*, 2004)
313	I	K	Case	0%					Severe LOF	Dysfunctional	(Ikrar *et al.*, 2009)

314	G	S	Case	12%					Severe LOF	Dysfunctional	(Li *et al.*, 2009a)
315	Y	C	Case	0%				Normal	Severe LOF	Dysfunctional	(Bianchi *et al.*, 2000)
315	Y	S	Case	0%					Severe LOF	Dysfunctional	(Chouabe *et al.*, 1997)
316	G	E	Case	18%	0				Severe LOF	Dysfunctional	(Yang *et al.*, 2009)
320	P	A	Case	0%					Severe LOF	Dysfunctional	(Thomas *et al.*, 2010)
320	P	H	Case	0%					Severe LOF	Dysfunctional	(Thomas *et al.*, 2010)
322	T	A	Case	0%				Impaired	Severe LOF	Dysfunctional	(Burgess *et al.*, 2012)
322	T	M	Case	0%				Impaired	Severe LOF	Dysfunctional	(Burgess *et al.*, 2012)
325	G	R	Case	0%					Severe LOF	Dysfunctional	(Aidery *et al.*, 2012)
338	S	F	Case	5%	12			Normal	Severe LOF	Dysfunctional	(Hoosien *et al.*, 2013)
339	F	S	Case	4%	1			Normal	Severe LOF	Dysfunctional	(Hoosien *et al.*, 2013)
341	A	V	Case	6%	60	5.59	0.29	Normal	Severe LOF	Dysfunctional	(Heijman *et al.*, 2012)
342	L	F	Case	0%					Severe LOF	Dysfunctional	(Chouabe *et al.*, 1997)
343	P	S	Case	0%					Severe LOF	Dysfunctional	(Zehelein *et al.*, 2004)
344	A	V	Case	100 %	40				Severe LOF	Dysfunctional	(Siebrands *et al.*, 2006)
357	Q	R	Case	27%	20	3	1	Impaired	Severe LOF	Dysfunctional	(Boulet *et al.*, 2006)
360	R	G	Case	20%					Severe LOF	Dysfunctional	(Yang *et al.*, 2009)

366	R	P	Case	0%	24.1					Severe LOF	Dysfunctional	(Shamgar *et al.*, 2006)
366	R	Q	Case	22%	29	1				Severe LOF	Dysfunctional	(Matavel *et al.*, 2010)
366	R	W	Case	30%	39.2				Impaired	Severe LOF	Dysfunctional	(Shamgar *et al.*, 2006)
371	A	T	Case	0%	21.9					Severe LOF	Dysfunctional	(Shamgar *et al.*, 2006)
373	S	P	Case	5%	37.9				Impaired	Severe LOF	Dysfunctional	(Shamgar *et al.*, 2006)
379	W	R	Case	0%					Impaired	Severe LOF	Dysfunctional	(Steffensen *et al.*, 2015)
380	R	S	Case	33%	0				Normal	Mild LOF	Dysfunctional	(Li *et al.*, 2013)
391	T	I	Case	85%	0					Normal	Normal	(Westenskow *et al.*, 2004)
392	W	R	Case	0%	28.3					Severe LOF	Dysfunctional	(Shamgar *et al.*, 2006)
393	K	M	Case	33%	0				Normal	Mild LOF	Dysfunctional	(Li *et al.*, 2013)
393	K	N	Control	100 %	13.3					Normal	Normal	(Shamgar *et al.*, 2006)
397	R	Q	Control	90%	0				Impaired	Normal	Normal	(Xiong *et al.*, 2015)
397	R	W	Control	40%	0	1		1	Normal	Mild LOF	Dysfunctional	(Li *et al.*, 2013)
417	V	M	Case	100 %	0	1				Normal	Normal	(Wedekind *et al.*, 2004)
448	P	R	Control	120 %	0					Normal	Normal	(Westenskow *et al.*, 2004)
455	H	Y	Case	43%	0	0.6				Mild LOF	Dysfunctional	(Yang *et al.*, 2009)
520	M	R	Case	0%					Absent	Severe LOF	Dysfunctional	(Schmitt *et al.*, 2007)

522	Y	S	Case	10%	7			Impaired	Severe LOF	Dysfunctional	(Steffensen *et al.*, 2015)
525	A	T	Case	36%	22	1.34	1.08	Impaired	Severe LOF	Dysfunctional	(Harmer *et al.*, 2014)
533	R	W	Case	72%	13.9	1			Normal	Normal	(Chouabe *et al.*, 2000)
539	R	W	Case	17%	33.9	1	0.41		Severe LOF	Dysfunctional	(Chouabe *et al.*, 2000)
546	S	L	Case	25%	50.7	1.3	0.81	Normal	Severe LOF	Dysfunctional	(Dvir *et al.*, 2014)
555	R	C	Case	25%	60				Severe LOF	Dysfunctional	(Chouabe *et al.*, 1997)
555	R	H	Case	12%	50	1.1	0.72	Normal	Severe LOF	Dysfunctional	(Aromolaran *et al.*, 2014)
557	K	E	Case	0%				Normal	Severe LOF	Dysfunctional	(Spatjens *et al.*, 2014)
562	R	M	Case	43%	43.3	1.55	1.07	Normal	Severe LOF	Dysfunctional	(Dvir *et al.*, 2014)
583	R	H	Case	100 %	0				Normal	Normal	(Detta, 2010)
589	G	D	Case	15%	33			Impaired	Severe LOF	Dysfunctional	(Piippo *et al.*, 2001)
590	A	T	Case	45%	10			Normal	Mild LOF	Dysfunctional	(Kinoshita *et al.*, 2014)
594	R	Q	Case	5%	60				Severe LOF	Dysfunctional	(Westenskow *et al.*, 2004)
611	D	Y	Case	100 %	0				Normal	Normal	(Yamaguchi *et al.*, 2005)
619	L	M	Case	1%				Normal	Severe LOF	Dysfunctional	(Aromolaran *et al.*, 2014)
643	G	S	Control	35%	1.1	1	0.72		Mild LOF	Normal	(Kubota *et al.*, 2001)

Table A-6 Performance of the neural network model with varied sizes of hidden layer.

# hidden neurons	MCC	AUC
1	0.568	0.882
2	0.567	0.881
3	0.572	0.884
4	0.562	0.883
5	0.581	0.881
6	0.584	0.886
7	0.581	0.885
8	0.559	0.880

Table A-7 Information gain of a set of tested predictive features.

Feature	Information gain	Threshold maximizes information gain
Rate of evolution	0.22	1.46
PSSM perturbation	0.18	5.89
Change in hydrophobicity	0.035	0.01
Predicted residue packing density (Li *et al.*, 2016)	0.024	11.96
Grantham score (Grantham, 1974)	0.020	103
Change in charge	0.018	NA
Change in SASA*	0.017	29.91

**SASA: solvent accessible surface area*

Table A-8 Summary of the median and interquartile interval [Q1, Q3] of each performance metric.

Method	Medians and [Q1, Q3] intervals of performance metrics						
	AUC	MCC	PPV	NPV	Accuracy	TPR	TNR
Q1VarPred	0.884 [0.876, 0.890]	0.584 [0.560, 0.608]	0.967 [0.966, 0.968]	0.533 [0.502, 0.565]	0.889 [0.871, 0.890]	0.905 [0.885, 0.906]	0.783 [0.767, 0.783]
KvSNP	0.669 [0.577, 0.753]	0.306 [0.213, 0.462]	0.926 [0.900, 0.938]	0.333 [0.250, 0.429]	0.829 [0.800, 0.865]	0.903 [0.839, 0.935]	0.500 [0.250, 0.600]
PhD-SNP	0.726 [0.653, 0.794]	0.369 [0.293, 0.494]	0.935 [0.913, 0.963]	0.364 [0.273, 0.500]	0.829 [0.771, 0.865]	0.871 [0.774, 0.935]	0.600 [0.500, 0.750]
PolyPhen-2	0.625 [0.593, 0.718]	0.372 [0.298, 0.477]	0.912 [0.899, 0.935]	0.500 [0.333, 0.667]	0.886 [0.857, 0.914]	0.968 [0.935, 1.000]	0.250 [0.250, 0.500]
PredictSNP	0.653 [0.593, 0.718]	0.306 [0.211, 0.470]	0.912 [0.906, 0.936]	0.500 [0.333, 0.600]	0.865 [0.838, 0.892]	0.935 [0.903, 0.968]	0.400 [0.250, 0.500]
PROVEAN	0.788 [0.722, 0.810]	0.556 [0.468, 0.576]	0.956 [0.938, 0.957]	0.557 [0.500, 0.593]	0.896 [0.880, 0.899]	0.926 [0.925, 0.936]	0.683 [0.579, 0.700]
SIFT	0.684 [0.593, 0.786]	0.435 [0.313, 0.532]	0.926 [0.900, 0.962]	0.500 [0.333, 0.600]	0.865 [0.838, 0.886]	0.935 [0.875, 0.969]	0.500 [0.250, 0.750]
SNAP	0.512 [0.484, 0.605]	0.170 [0.089, 0.255]	0.886 [0.875, 0.909]	0.167 [0.000, 0.222]	0.800 [0.714, 0.865]	0.875 [0.750, 0.969]	0.200 [0.000, 0.400]
SNPs&GO	0.706 [0.638, 0.762]	0.326 [0.232, 0.405]	0.933 [0.920, 0.960]	0.286 [0.250, 0.333]	0.771 [0.730, 0.829]	0.806 [0.742, 0.871]	0.600 [0.500, 0.750]

Table A-9 Topological subdomains of KCNQ1 and the enrichment of dysfunctional variants within each region.

Subdomain	Range	Length	Observed number of variants	Expected number of variants	Enrichment
NTD	1-110+118-121	114	1	16	0.1
S0	111-117	7	3	1	3.0
S1	122-146	25	2	3	0.7
S1-S2	147-153	7	0	1	0.0

S2	154-177	24	2	3	0.7
S2-S3	178-197	20	4	3	1.3
S3	198-215	18	10	3	3.3
S3-S4	216-222	7	0	1	0.0
S4	223-241	19	5	3	1.7
S4-S5	242-259	18	8	3	2.7
S5	260-284	25	11	3	3.7
S5-pore	285-298	14	1	2	0.5
pore-helix	299-312	14	5	2	2.5
pore-loop	313-322	10	9	1	9.0
S6	323-360	38	9	5	1.8
S6-A	361-369	9	3	1	3.0
A	370-389	20	4	3	1.3
A-B	390-506	117	4	16	0.3
B	507-532	26	3	4	0.8
B-C	533-547	15	2	2	1.0
C	548-562	15	4	2	2.0
C-D	563-587	25	0	3	0.0
D	588-622	35	4	5	0.8
D-end	623-676	54	0	8	0.0

NTD: N-terminal domain

Table A-10 Summary of methods evaluated in this study.

Tool	Algorithm	Link	Reference
KvSNP	Fast random forest	http://www.bioinformatics.leeds.ac.uk/KvDB/KvSNP.html	(Stead *et al.*, 2011)
PhD-SNP	Support vector machine	http://snps.biofold.org/phd-snp/phd-snp.html	(Capriotti *et al.*, 2006)
PolyPhen-2	Naïve Bayes classification	http://genetics.bwh.harvard.edu/pph2/bgi.shtml	(Adzhubei *et al.*, 2010)
PredictSNP	Metaserver	http://loschmidt.chemi.muni.cz/predictsnp1/	(Bendl *et al.*, 2014)
PROVEAN	Sequence conservation	http://provean.jcvi.org/seq_submit.php	(Choi *et al.*, 2012)
SIFT	Sequence conservation	http://siftdna.org/www/SIFT_pid_subst_all_submit.html	(Ng and Henikoff, 2001)
SNAP	Neural networks	https://rostlab.org/services/snap2web/	(Bromberg and Rost, 2007)
SNPs&GO	Support vector machine	http://snps.biofold.org/snps-and-go/snps-and-go.html	(Calabrese *et al.*, 2009)

Table A-11 Six other variants in the B-C linker deposited in the ClinVar database as of June 2017.

Variant	Clinical significance	Review status
R539Q	Uncertain significance	Criteria provided, single submitter
V541I	Uncertain significance	Criteria provided, multiple submitters, no conflicts
E543K	Not provided	No assertion provided

Q544L	Uncertain significance	Criteria provided, single submitter
S546L	Pathogenic/likely pathogenic, not provided	Criteria provided, multiple submitters, no conflicts
Q547R	Not provided	No assertion provided

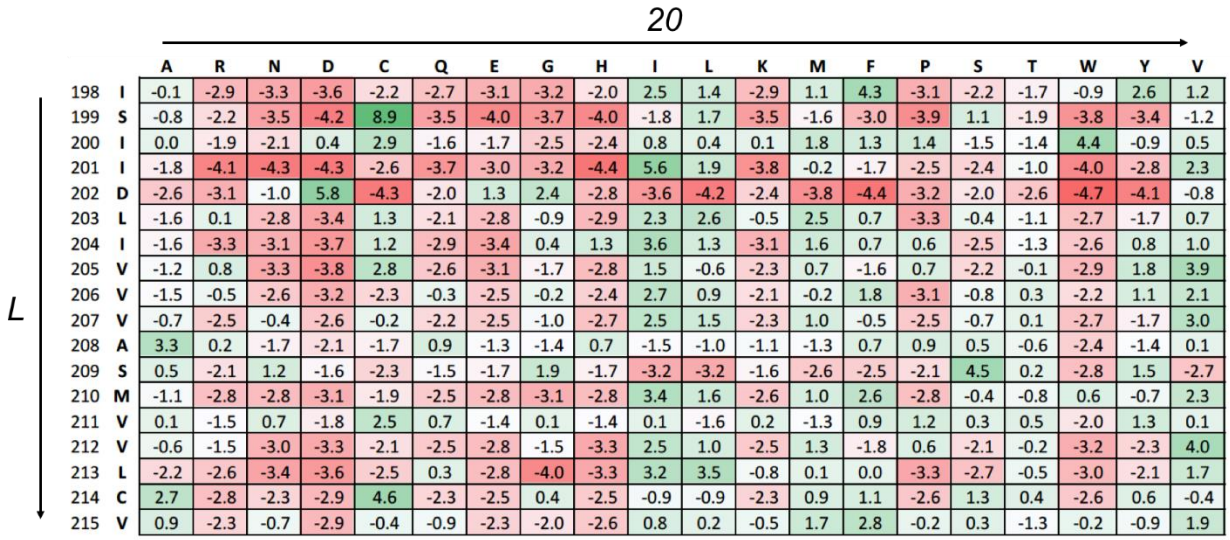


Figure A-5 An illustration of position-specific scoring matrix (PSSM)

For a protein of length L, a PSSM is a $L \times 20$ matrix containing log ratios of the estimated frequency of each of the 20 amino acids to occur at each position relative to the expected frequency of the wild-type amino acid in a random sequence. If P_A is the probability for amino acid A to occupy a position and P_A^0 is its background probability, then the PSSM entry for A at this position equals $\lambda \ln \frac{P_A}{P_A^0}$, where λ is a scaling factor built in PSI-BLAST (Altschul *et al.*, 1997).

A structural model for the glutamate A2 (GluA2) receptor and its cornichon 3 (CNIH3) auxiliary subunit

The AMPA (GluA1-GluA4) receptors are a subfamily of ionotropic glutamate receptors that mediate fast excitation within and between brain regions. The mammalian cornichon family was recently discovered to be cognate binders of AMPA receptors (Schwenk *et al.*, 2009). The family member CNIH3 was shown to slow down the deactivation and desensitization of activated GluA2 receptor (Schwenk *et al.*, 2009). However, how this regulation occurs at the atomic level is not clear for two primary reasons. First, a model of the structure of CNIH3 is unavailable. Second, crystalizing the GluA2/CNIH3 complex is a formidable experimental endeavor. Here, we present structural models for CNIH3 and GluA2/CNIH3 complex, predicted by using the BCL::MP-Fold

de novo structure prediction algorithm (Weiner *et al.*, 2013) and the ROSETTA protein-protein docking method (Gray *et al.*, 2003).

Tertiary structure models of CNIH3 were created using the BCL::MP–Fold protocol (see III-2.3 Incorporating WCNs as restraints in de novo structure prediction for details). 5000 models were constructed. The models were clustered and the top 50 models were submitted to side-chain and loop modeling using the ROSETTA *de novo* protein structure prediction algorithm (Simons *et al.*, 1997). The final models were selected according to ROSETTA full-atom score. The final selected model of CNIH3 was docked to a crystal structure of GluA2 using the ROSETTA protein-protein docking method (Gray *et al.*, 2003). 5000 docking solutions were created and then clustered. Candidate models were selected from the cluster centers of the top-ranked clusters.

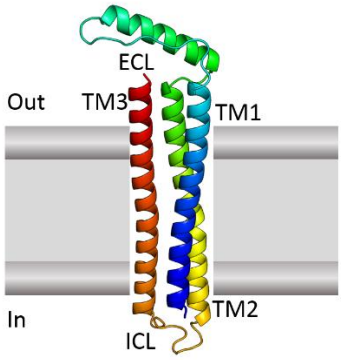


Figure A-6 One of the top-ranked de novo model for CNIH3

TM: transmembrane, ECL: extracellular loop, ICL: intracellular loop.

One of the top-scoring models is shown in Figure . In this model, CNIH3 adopts an antiparallel three-helical bundle fold with a counterclockwise arrangement when viewed from outside of the plasma membrane. The extracellular loop adopts a 16-residue long amphipathic helix that might be involved in binding to GluA2. This model is consistent with predictions on the topology of CNIH3 (Roth *et al.*, 1995, Diaz, 2010). As a preliminary investigation, we also docked this model to a crystal structure of a homotetrameric GluA2 (Sobolevsky *et al.*, 2009). Shown in Figure is one of the models of GluA2/CNIH3 complex with the lowest interface energy. The ROSETTA interface energy measures the strength of binding between docking partners, the lower the energy the stronger the binding. This model suggests three interaction clusters: salt-bridges between Arg^{51}, Arg^{55}, and Arg^{59} in the ECL of CNIH3 and Glu^{678}, Pro^{679} in the ligand-binding domain of GluA2; cation–pi interaction between Tyr^{72} at the extracellular tip of TM2 of CNIH-3 and Lys^{505},

Lys^{695}, and Lys^{697} in the ligand-binding domain of GluA2; hydrophobic packing between Val^{95} Leu^{98}, Phe^{99} at the intracellular tip of TM2 of CNIH3 and Val^{538}, Phe^{541}, and Leu^{542} at the intracellular tip of TM1 of GluA2. The salt-bridge interaction and hydrophobic packing are consistent with findings from peptide-array studies (Shanks *et al.*, 2014). The cation–pi interaction needs further experimental verification.

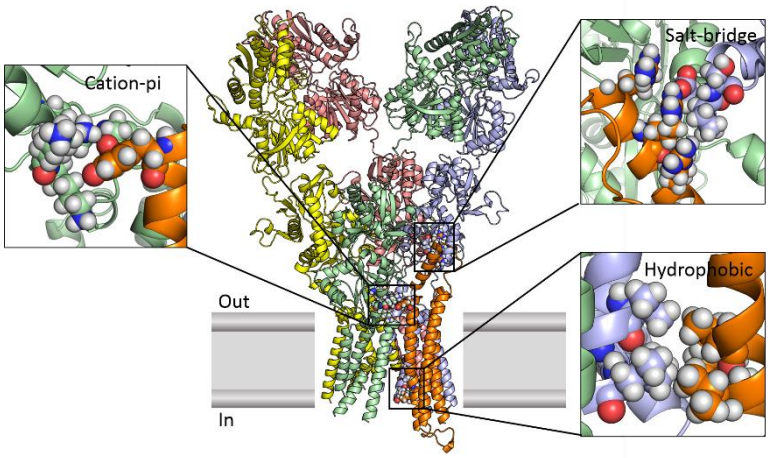


Figure A-7 One of the top-ranked models of GluA2/CNIH-3 complex predicted by protein-protein docking

This model suggests three interaction clusters two of which are consistent with peptide-array studies (see text for details).

REFERENCES

Abeysinghe, S., Ju, T., Baker, M.L. & Chiu, W., 2008. Shape modeling and matching in identifying 3D protein structures. *Computer-Aided Design,* 40**,** 708-720.

Ackerman, M.J., 2015. Genetic purgatory and the cardiac channelopathies: Exposing the variants of uncertain/unknown significance issue. *Heart Rhythm,* 12**,** 2325-31.

Acuner Ozbabacan, S.E., Engin, H.B., Gursoy, A. & Keskin, O., 2011. Transient protein-protein interactions. *Protein Eng Des Sel,* 24**,** 635-48.

Adamian, L. & Liang, J., 2001. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J Mol Biol,* 311**,** 891-907.

Adamian, L. & Liang, J., 2006. Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Struct Biol,* 6**,** 13.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. & Sunyaev, S.R., 2010. A method and server for predicting damaging missense mutations. *Nat Methods,* 7**,** 248-9.

Ahmad, S., Gromiha, M.M. & Sarai, A., 2003. Real value prediction of solvent accessibility from amino acid sequence. *Proteins,* 50**,** 629-35.

Aidery, P., Kisselbach, J., Schweizer, P.A., Becker, R., Katus, H.A. & Thomas, D., 2011. Biophysical properties of mutant KCNQ1 S277L channels linked to hereditary long QT syndrome with phenotypic variability. *Biochim Biophys Acta,* 1812**,** 488-94.

Aidery, P., Kisselbach, J., Schweizer, P.A., Becker, R., Katus, H.A. & Thomas, D., 2012. Impaired ion channel function related to a common KCNQ1 mutation - implications for risk stratification in long QT syndrome 1. *Gene,* 511**,** 26-33.

Alexander, N., Bortolus, M., Al-Mestarihi, A., Mchaourab, H. & Meiler, J., 2008. De novo high-resolution protein structure determination from sparse spin-labeling EPR data. *Structure,* 16**,** 181-95.

Altenbach, C., Froncisz, W., Hemker, R., Mchaourab, H. & Hubbell, W.L., 2005. Accessibility of nitroxide side chains: absolute Heisenberg exchange rates from power saturation EPR. *Biophys J,* 89**,** 2103-12.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25**,** 3389-402.

Amir, E.D., Kalisman, N. & Keasar, C., 2008. Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. *Proteins,* 72**,** 62-73.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. *Science,* 181**,** 223-30.

Ansari, S. & Helms, V., 2005. Statistical analysis of predominantly transient protein-protein interfaces. *Proteins,* 61**,** 344-55.

Aromolaran, A.S., Subramanyam, P., Chang, D.D., Kobertz, W.R. & Colecraft, H.M., 2014. LQT1 mutations in KCNQ1 C-terminus assembly domain suppress IKs using different mechanisms. *Cardiovasc Res,* 104**,** 501-11.

Ashkenazy, H. & Kliger, Y., 2010. Reducing phylogenetic bias in correlated mutation analysis. *Protein Eng Des Sel,* 23**,** 321-6.

Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B.D., Burston, H.E., Vizeacoumar, F.J., Snider, J., Phanse, S., Fong, V., Tam, Y.Y., Davey, M., Hnatshak, O., Bajaj, N., Chandran, S., Punna, T., Christopolous, C., Wong, V., Yu, A., Zhong, G., Li, J., Stagljar, I., Conibear, E., Wodak, S.J., Emili, A. & Greenblatt, J.F., 2012. Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. *Nature,* 489**,** 585-9.

Back, J.W., De Jong, L., Muijsers, A.O. & De Koster, C.G., 2003. Chemical cross-linking and mass spectrometry for protein structural modeling. *J Mol Biol,* 331**,** 303-13.

Bahar, I. & Jernigan, R.L., 1997. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol,* 266**,** 195-214.

Baker, D., 2000. A surprising simplicity to protein folding. *Nature,* 405**,** 39-42.

Baker, D. & Sali, A., 2001. Protein structure prediction and structural genomics. *Science,* 294**,** 93-6.

Baker, M.L., Ju, T. & Chiu, W., 2007. Identification of secondary structure elements in intermediate-resolution density maps. *Structure,* 15**,** 7-19.

Baker, M.L., Yu, Z., Chiu, W. & Bajaj, C., 2006. Automated segmentation of molecular subunits in electron cryomicroscopy density maps. *J Struct Biol,* 156**,** 432-41.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A. & Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics,* 16**,** 412-24.

Baldwin, R.L., 1989. How does protein folding get started? *Trends in Biochemical Sciences,* 14**,** 291-294.

Barducci, A., Bonomi, M. & Parrinello, M., 2011. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science,* 1**,** 826-843.

Barhanin, J., Lesage, F., Guillemare, E., Fink, M., Lazdunski, M. & Romey, G., 1996. K(V)LQT1 and lsK (minK) proteins associate to form the I(Ks) cardiac potassium current. *Nature,* 384**,** 78-80.

Baron, R., De Vries, A.H., Hunenberger, P.H. & Van Gunsteren, W.F., 2006a. Comparison of atomic-level and coarse-grained models for liquid hydrocarbons from molecular dynamics configurational entropy estimates. *J Phys Chem B,* 110**,** 8464-73.

Baron, R., De Vries, A.H., Hunenberger, P.H. & Van Gunsteren, W.F., 2006b. Configurational entropies of lipids in pure and mixed bilayers from atomic-level and coarse-grained molecular dynamics simulations. *J Phys Chem B,* 110**,** 15602-14.

Baron, R., Trzesniak, D., De Vries, A.H., Elsener, A., Marrink, S.J. & Van Gunsteren, W.F., 2007. Comparison of thermodynamic properties of coarse-grained and atomic-level simulation models. *Chemphyschem,* 8**,** 452-61.

Barth, P., Wallner, B. & Baker, D., 2009. Prediction of membrane protein structures with complex topologies using limited constraints. *Proc Natl Acad Sci U S A,* 106**,** 1409-14.

Bartlett, A.I. & Radford, S.E., 2009. An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nat Struct Mol Biol,* 16**,** 582-8.

Bartos, D.C., Giudicessi, J.R., Tester, D.J., Ackerman, M.J., Ohno, S., Horie, M., Gollob, M.H., Burgess, D.E. & Delisle, B.P., 2014. A KCNQ1 mutation contributes to the concealed type 1 long QT phenotype by limiting the Kv7.1 channel conformational changes associated with protein kinase A phosphorylation. *Heart Rhythm,* 11**,** 459-68.

Battey, J.N., Kopp, J., Bordoli, L., Read, R.J., Clarke, N.D. & Schwede, T., 2007. Automated server predictions in CASP7. *Proteins,* 69 Suppl 8**,** 68-82.

Battiste, J.L. & Wagner, G., 2000. Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. *Biochemistry,* 39**,** 5355-65.

Beauchamp, K.A., Mcgibbon, R., Lin, Y.S. & Pande, V.S., 2012. Simple few-state models reveal hidden complexity in protein folding. *Proc Natl Acad Sci U S A,* 109**,** 17807-13.

Behler, J., 2016. Perspective: Machine learning potentials for atomistic simulations. *J Chem Phys,* 145**,** 170901.

Bellocq, C., Van Ginneken, A.C., Bezzina, C.R., Alders, M., Escande, D., Mannens, M.M., Baro, I. & Wilde, A.A., 2004. Mutation in the KCNQ1 gene leading to the short QT-interval syndrome. *Circulation,* 109**,** 2394-7.

Ben-Naim, A., 1997. Statistical potentials extracted from protein structures: Are these meaningful potentials? *The Journal of Chemical Physics,* 107**,** 3698-3706.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E.D., Zendulka, J., Brezovsky, J. & Damborsky, J., 2014. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput Biol,* 10**,** e1003440.

Bernardi, R.C., Melo, M.C. & Schulten, K., 2015. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim Biophys Acta,* 1850**,** 872-7.

Best, R.B., Zhu, X., Shim, J., Lopes, P.E., Mittal, J., Feig, M. & Mackerell, A.D., Jr., 2012. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J Chem Theory Comput,* 8**,** 3257-3273.

Betancourt, M.R. & Skolnick, J., 2004. Local propensities and statistical potentials of backbone dihedral angles in proteins. *J Mol Biol,* 342**,** 635-49.

Betancourt, M.R. & Thirumalai, D., 1999. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci,* 8**,** 361-9.

Beuming, T. & Weinstein, H., 2004. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics,* 20**,** 1822-35.

Bhuiyan, Z.A., 2012. Silent mutation in long QT syndrome: pathogenicity prediction by computer simulation. *Heart Rhythm,* 9**,** 283-4.

Bianchi, L., Priori, S.G., Napolitano, C., Surewicz, K.A., Dennis, A.T., Memmi, M., Schwartz, P.J. & Brown, A.M., 2000. Mechanisms of I(Ks) suppression in LQT1 mutants. *Am J Physiol Heart Circ Physiol,* 279**,** H3003-11.

Bihnstein, E., And Ohi, Melanie., 2015. Cryo-Electron Microscopy and the Amazing Race to Atomic Resolution. *Biochemistry,* 54**,** 3133-3141.

Bitbol, A.F., Dwyer, R.S., Colwell, L.J. & Wingreen, N.S., 2016. Inferring interaction partners from protein sequences. *Proc Natl Acad Sci U S A,* 113**,** 12180-12185.

Bjorkholm, P., Daniluk, P., Kryshtafovych, A., Fidelis, K., Andersson, R. & Hvidsten, T.R., 2009. Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. *Bioinformatics,* 25**,** 1264-70.

Bonneau, R. & Baker, D., 2001. Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct,* 30**,** 173-89.

Bonneau, R., Ruczinski, I., Tsai, J. & Baker, D., 2002. Contact order and ab initio protein structure prediction. *Protein Sci,* 11**,** 1937-44.

Borbat, P.P., Mchaourab, H.S. & Freed, J.H., 2002. Protein Structure Determination Using Long-Distance Constraints from Double-Quantum Coherence ESR: Study of T4 Lysozyme. *Journal of the American Chemical Society,* 124**,** 5304-5314.

Bordner, A.J., 2009. Predicting protein-protein binding sites in membrane proteins. *BMC Bioinformatics,* 10**,** 312.

Bordner, A.J. & Abagyan, R., 2005. Statistical analysis and prediction of protein-protein interfaces. *Proteins,* 60**,** 353-66.

Bošković, B. & Brest, J., 2016. Genetic algorithm with advanced mechanisms applied to the protein structure prediction in a hydrophobic-polar model and cubic lattice. *Applied Soft Computing,* 45**,** 61-70.

Boulet, I.R., Raes, A.L., Ottschytsch, N. & Snyders, D.J., 2006. Functional effects of a KCNQ1 mutation associated with the long QT syndrome. *Cardiovasc Res,* 70**,** 466-74.

Bowers, K.J., Chow, E., Xu, H., Dror, R.O., Eastwood, M.P., Gregersen, B.A., Klepeis, J.L., Kolossvary, I., Moraes, M.A., Sacerdoti, F.D., Salmon, J.K., Shan, Y. & Shaw, D.E., 2006. Scalable algorithms for molecular dynamics simulations on commodity clusters. *Proceedings of the 2006 ACM/IEEE conference on Supercomputing.* Tampa, Florida: ACM, 84.

Bowers, P.M., 2000. De novo protein structure determination using sparse NMR data. *Journal of Biomolecular NMR,* 18**,** 311-318.

Bowie, J., Luthy, R. & Eisenberg, D., 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* 253**,** 164-170.

Bowman, G.R., Ensign, D.L. & Pande, V.S., 2010. Enhanced modeling via network theory: Adaptive sampling of Markov state models. *J Chem Theory Comput,* 6**,** 787-94.

Bowman, G.R., Voelz, V.A. & Pande, V.S., 2011. Taming the complexity of protein folding. *Curr Opin Struct Biol,* 21**,** 4-11.

Bradley, P., Chivian, D., Meiler, J., Misura, K.M., Rohl, C.A., Schief, W.R., Wedemeyer, W.J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C.E. & Baker, D., 2003. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins,* 53 Suppl 6**,** 457-68.

Bradley, P., Malmstrom, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D.E., Meiler, J., Misura, K.M. & Baker, D., 2005a. Free modeling with Rosetta in CASP6. *Proteins,* 61 Suppl 7**,** 128-34.

Bradley, P., Misura, K.M. & Baker, D., 2005b. Toward high-resolution de novo structure prediction for small proteins. *Science,* 309**,** 1868-71.

Bromberg, Y. & Rost, B., 2007. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res,* 35**,** 3823-35.

Brooks, B.R., Brooks, C.L., 3rd, Mackerell, A.D., Jr., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M. & Karplus, M., 2009. CHARMM: the biomolecular simulation program. *J Comput Chem,* 30**,** 1545-614.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. & Karplus, M., 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry,* 4**,** 187-217.

Brooks, C.L., 3rd, Onuchic, J.N. & Wales, D.J., 2001. Statistical thermodynamics. Taking a walk on a landscape. *Science,* 293**,** 612-3.

Bryngelson, J.D., Onuchic, J.N., Socci, N.D. & Wolynes, P.G., 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins,* 21**,** 167-95.

Burger, L. & Van Nimwegen, E., 2008. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol,* 4**,** 165.

Burger, L. & Van Nimwegen, E., 2010. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol,* 6**,** e1000633.

Burger V, B.I., Chennubhotla C., 2011. A Hierarchical Elastic Network Model for Unsupervised EM Density Map Segmentation. 2.

Burgess, D.E., Bartos, D.C., Reloj, A.R., Campbell, K.S., Johnson, J.N., Tester, D.J., Ackerman, M.J., Fressart, V., Denjoy, I., Guicheney, P., Moss, A.J., Ohno, S., Horie, M. & Delisle, B.P., 2012. High-risk long QT syndrome mutations in the Kv7.1 (KCNQ1) pore disrupt the molecular basis for rapid K(+) permeation. *Biochemistry,* 51**,** 9076-85.

Butkiewicz, M., Lowe, E.W., Jr., Mueller, R., Mendenhall, J.L., Teixeira, P.L., Weaver, C.D. & Meiler, J., 2013. Benchmarking ligand-based virtual High-Throughput Screening with the PubChem database. *Molecules,* 18**,** 735-56.

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J. & Huang, E.S., 2004. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci,* 13**,** 190-202.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L. & Casadio, R., 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat,* 30**,** 1237-44.

Campbell, C.M., Campbell, J.D., Thompson, C.H., Galimberti, E.S., Darbar, D., Vanoye, C.G. & George, A.L., Jr., 2013. Selective targeting of gain-of-function KCNQ1 mutations predisposing to atrial fibrillation. *Circ Arrhythm Electrophysiol,* 6**,** 960-6.

Capriotti, E., Calabrese, R. & Casadio, R., 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics,* 22**,** 2729-34.

Care, M.A., Needham, C.J., Bulpitt, A.J. & Westhead, D.R., 2007. Deleterious SNP prediction: be mindful of your training data! *Bioinformatics,* 23**,** 664-72.

Carugo, O. & Pongor, S., 2001. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci,* 10**,** 1470-3.

Cavasotto, C.N. & Phatak, S.S., 2009. Homology modeling in drug discovery: current trends and applications. *Drug Discov Today,* 14**,** 676-83.

Chan, H.S. & Dill, K.A., 1993. The Protein Folding Problem. *Physics Today,* 46**,** 24-32.

Chang, D.T., Huang, H.Y., Syu, Y.T. & Wu, C.P., 2008. Real value prediction of protein solvent accessibility using enhanced PSSM features. *BMC Bioinformatics,* 9 Suppl 12**,** S12.

Chen, L., Zhang, Q., Qiu, Y., Li, Z., Chen, Z., Jiang, H., Li, Y. & Yang, H., 2015. Migration of PIP2 lipids on voltage-gated potassium channel surface influences channel deactivation. *Sci Rep,* 5**,** 15079.

Chen, M., Baldwin, P.R., Ludtke, S.J. & Baker, M.L., 2016. De Novo modeling in cryo-EM density maps with Pathwalking. *J Struct Biol,* 196**,** 289-298.

Chen, P. & Li, J., 2010. Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers. *BMC Struct Biol,* 10 Suppl 1**,** S2.

Cheng, J. & Baldi, P., 2007. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics,* 8**,** 113.

Choe, S., 2002. Potassium channel structures. *Nat Rev Neurosci,* 3**,** 115-21.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P., 2012. Predicting the functional effect of amino acid substitutions and indels. *PLoS One,* 7**,** e46688.

Chothia, C. & Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J,* 5**,** 823-6.

Chou, K.-C. & Shen, H.-B., 2009. FoldRate: A web-server for predicting protein folding rates from primary sequence. *The Open Bioinformatics Journal,* 3**,** 31-50.

Chouabe, C., Neyroud, N., Guicheney, P., Lazdunski, M., Romey, G. & Barhanin, J., 1997. Properties of KvLQT1 K+ channel mutations in Romano-Ward and Jervell and Lange-Nielsen inherited cardiac arrhythmias. *EMBO J,* 16**,** 5472-9.

Chouabe, C., Neyroud, N., Richard, P., Denjoy, I., Hainque, B., Romey, G., Drici, M.D., Guicheney, P. & Barhanin, J., 2000. Novel mutations in KvLQT1 that affect Iks activation through interactions with Isk. *Cardiovasc Res,* 45**,** 971-80.

Chung, H.S., Piana-Agostinetti, S., Shaw, D.E. & Eaton, W.A., 2015. Structural origin of slow diffusion in protein folding. *Science,* 349**,** 1504-10.

Cline, M.S., Karplus, K., Lathrop, R.H., Smith, T.F., Rogers, R.G., Jr. & Haussler, D., 2002. Information-theoretic dissection of pairwise contact potentials. *Proteins,* 49**,** 7-14.

Cooper, G.M. & Shendure, J., 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet,* 12**,** 628-40.

Cordeiro, J.M., Perez, G.J., Schmitt, N., Pfeiffer, R., Nesterenko, V.V., Burashnikov, E., Veltmann, C., Borggrefe, M., Wolpert, C., Schimpf, R. & Antzelevitch, C., 2010. Overlapping LQT1 and LQT2 phenotype in a patient with long QT syndrome associated with loss-of-function variations in KCNQ1 and KCNH2. *Can J Physiol Pharmacol,* 88**,** 1181-90.

Crotti, L., Tester, D.J., White, W.M., Bartos, D.C., Insolia, R., Besana, A., Kunic, J.D., Will, M.L., Velasco, E.J., Bair, J.J., Ghidoni, A., Cetin, I., Van Dyke, D.L., Wick, M.J., Brost, B., Delisle, B.P., Facchinetti, F., George, A.L., Schwartz, P.J. & Ackerman, M.J., 2013. Long QT syndrome-associated mutations in intrauterine fetal death. *JAMA,* 309**,** 1473-82.

Cui, Y., Chen, R.S. & Wong, W.H., 1998. Protein folding simulation with genetic algorithm and supersecondary structure constraints. *Proteins,* 31**,** 247-57.

Custodio, F.L., Barbosa, H.J.C. & Dardenne, L.E., 2004. Investigation of the three-dimensional lattice HP protein folding model using a genetic algorithm. *Genetics and Molecular Biology,* 27**,** 611-615.

Custodio, F.L., Barbosa, H.J.C. & Dardenne, L.E., 2014. A multiple minima genetic algorithm for protein structure prediction. *Applied Soft Computing,* 15**,** 88-99.

Daggett, V. & Fersht, A., 2003a. The present view of the mechanism of protein folding. *Nat Rev Mol Cell Biol,* 4**,** 497-502.

Daggett, V. & Fersht, A.R., 2003b. Is there a unifying mechanism for protein folding? *Trends in Biochemical Sciences,* 28**,** 18-25.

Dahimene, S., Alcolea, S., Naud, P., Jourdon, P., Escande, D., Brasseur, R., Thomas, A., Baro, I. & Merot, J., 2006. The N-terminal juxtamembranous domain of KCNQ1 is critical for channel surface expression - Implications in the Romano-Ward LQT1 syndrome. *Circulation Research,* 99**,** 1076-1083.

Daley, D.O., 2008. The assembly of membrane proteins into complexes. *Current Opinion in Structural Biology,* 18**,** 420-424.

Dandekar, T. & Argos, P., 1992. Potential of genetic algorithms in protein folding and protein engineering simulations. *Protein Eng,* 5**,** 637-45.

Das, S., Makino, S., Melman, Y.F., Shea, M.A., Goyal, S.B., Rosenzweig, A., Macrae, C.A. & Ellinor, P.T., 2009. Mutation in the S3 segment of KCNQ1 results in familial lone atrial fibrillation. *Heart Rhythm,* 6**,** 1146-53.

De Juan, D., Pazos, F. & Valencia, A., 2013. Emerging methods in protein co-evolution. *Nat Rev Genet,* 14**,** 249-61.

Dehouck, Y., Gilis, D. & Rooman, M., 2006. A new generation of statistical potentials for proteins. *Biophys J,* 90**,** 4010-7.

Dekker, J.P. & Boekema, E.J., 2005. Supramolecular organization of thylakoid membrane proteins in green plants. *Biochim Biophys Acta,* 1706**,** 12-39.

Deluca, S., Dorr, B. & Meiler, J., 2011. Design of native-like proteins through an exposure-dependent environment potential. *Biochemistry,* 50**,** 8521-8.

Deng, L., Hinton, G. & Kingsbury, B., Year. New types of deep neural network learning for speech recognition and related applications: An overviewed.^eds. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*IEEE, 8599-8603.

Deschenes, D., Acharfi, S., Pouliot, V., Hegele, R., Krahn, A., Daleau, P. & Chahine, M., 2003. Biophysical characteristics of a new mutation on the KCNQ1 potassium channel (L251P) causing long QT syndrome. *Can J Physiol Pharmacol,* 81**,** 129-34.

Detta, N., 2010. Molecular Basis of Cardiac Arrhythmias: Genetics of Natural Variants and Electrophysiological Investigation of Mutant Proteins. University of Napoli Federico II.

Dey, S., Pal, A., Chakrabarti, P. & Janin, J., 2010. The Subunit Interfaces of Weakly Associated Homodimeric Proteins. *Journal of Molecular Biology,* 398**,** 146-160.

Diaz, E., 2010. Regulation of AMPA receptors by transmembrane accessory proteins. *European Journal of Neuroscience,* 32**,** 261-268.

Dill, K.A., 1999. Polymer principles and protein folding. *Protein Sci,* 8**,** 1166-80.

Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D. & Chan, H.S., 1995. Principles of protein folding--a perspective from simple exact models. *Protein Sci,* 4**,** 561-602.

Dill, K.A. & Chan, H.S., 1997. From Levinthal to pathways to funnels. *Nat Struct Biol,* 4**,** 10-9.

Dill, K.A. & Maccallum, J.L., 2012. The protein-folding problem, 50 years on. *Science,* 338**,** 1042-6.

Dill, K.A., Ozkan, S.B., Shell, M.S. & Weikl, T.R., 2008. The protein folding problem. *Annu Rev Biophys,* 37**,** 289-316.

Dobson, C.M. & Karplus, M., 1999. The fundamentals of protein folding: bringing together theory and experiment. *Curr Opin Struct Biol,* 9**,** 92-101.

Dobson, C.M., Sali, A. & Karplus, M., 1998. Protein folding: A perspective from theory and experiment. *Angewandte Chemie-International Edition,* 37**,** 868-893.

Doyle, D.A., Morais Cabral, J., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T. & Mackinnon, R., 1998. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science,* 280**,** 69-77.

Dror, R.O., Mildorf, T.J., Hilger, D., Manglik, A., Borhani, D.W., Arlow, D.H., Philippsen, A., Villanueva, N., Yang, Z., Lerch, M.T., Hubbell, W.L., Kobilka, B.K., Sunahara, R.K. & Shaw, D.E., 2015. SIGNAL TRANSDUCTION. Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science,* 348**,** 1361-5.

Dror, R.O., Pan, A.C., Arlow, D.H., Borhani, D.W., Maragakis, P., Shan, Y., Xu, H. & Shaw, D.E., 2011. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc Natl Acad Sci U S A,* 108**,** 13118-23.

Duan, Y. & Kollman, P.A., 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science,* 282**,** 740-4.

Duarte, J.M., Biyani, N., Baskaran, K. & Capitani, G., 2013. An analysis of oligomerization interfaces in transmembrane proteins. *BMC Struct Biol,* 13**,** 21.

Dunbrack, R.L., Jr. & Karplus, M., 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol,* 230**,** 543-74.

Dunn, S.D., Wahl, L.M. & Gloor, G.B., 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics,* 24**,** 333-340.

Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R. & Meiler, J., 2009. Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *J Mol Model,* 15**,** 1093-108.

Durrant, J.D. & Mccammon, J.A., 2011. Molecular dynamics simulations and drug discovery. *BMC Biol,* 9**,** 71.

Dvir, M., Strulovich, R., Sachyani, D., Ben-Tal Cohen, I., Haitin, Y., Dessauer, C., Pongs, O., Kass, R., Hirsch, J.A. & Attali, B., 2014. Long QT mutations at the interface between KCNQ1 helix C and KCNE1 disrupt I(KS) regulation by PKA and PIP(2). *J Cell Sci,* 127**,** 3943-55.

Echave, J., Spielman, S.J. & Wilke, C.O., 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet,* 17**,** 109-21.

Egloff, P., Hillenbrand, M., Klenk, C., Batyuk, A., Heine, P., Balada, S., Schlinkmann, K.M., Scott, D.J., Schutz, M. & Pluckthun, A., 2014. Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in Escherichia coli. *Proc Natl Acad Sci U S A,* 111**,** E655-62.

Ekeberg, M., Hartonen, T. & Aurell, E., 2014. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics,* 276**,** 341-356.

Eldstrom, J., Wang, Z., Werry, D., Wong, N. & Fedida, D., 2015. Microscopic mechanisms for long QT syndrome type 1 revealed by single-channel analysis of I(Ks) with S3 domain mutations in KCNQ1. *Heart Rhythm,* 12**,** 386-94.

Eldstrom, J., Xu, H., Werry, D., Kang, C., Loewen, M.E., Degenhardt, A., Sanatani, S., Tibbits, G.F., Sanders, C. & Fedida, D., 2010. Mechanistic basis for LQT1 caused by S3 mutations in the KCNQ1 subunit of IKs. *J Gen Physiol,* 135**,** 433-48.

Englander, S.W. & Mayne, L., 2014. The nature of protein folding pathways. *Proc Natl Acad Sci U S A,* 111**,** 15873-80.

Englander, S.W., Mayne, L. & Krishna, M.M., 2007. Protein folding and misfolding: mechanism and principles. *Q Rev Biophys,* 40**,** 287-326.

Eyal, E., Najmanovich, R., Mcconkey, B.J., Edelman, M. & Sobolev, V., 2004. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J Comput Chem,* 25**,** 712-24.

Fagerberg, L., Jonasson, K., Von Heijne, G., Uhlen, M. & Berglund, L., 2010. Prediction of the human membrane proteome. *Proteomics,* 10**,** 1141-1149.

Fang, Q. & Shortle, D., 2005. A consistent set of statistical potentials for quantifying local side-chain and backbone interactions. *Proteins,* 60**,** 90-6.

Farahbakhsh, Z.T., Altenbach, C. & Hubbell, W.L., 1992. SPIN LABELED CYSTEINES AS SENSORS FOR PROTEIN-LIPID INTERACTION AND CONFORMATION IN RHODOPSIN. *Photochemistry and Photobiology,* 56**,** 1019-1033.

Fariselli, P. & Casadio, R., 1999. A neural network based predictor of residue contacts in proteins. *Protein Engineering,* 12**,** 15-21.

Fariselli, P. & Casadio, R., 2000. Prediction of the number of residue contacts in proteins. *Proc Int Conf Intell Syst Mol Biol,* 8**,** 146-51.

Fariselli, P., Olmea, O., Valencia, A. & Casadio, R., 2001. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins: Structure, Function, and Bioinformatics,* 45**,** 157-162.

Fersht, A.R., 1997. Nucleation mechanisms in protein folding. *Curr Opin Struct Biol,* 7**,** 3-9.

Finkelstein, A.V., Badretdinov, A. & Gutin, A.M., 1995. Why do protein architectures have Boltzmann-like statistics? *Proteins,* 23**,** 142-50.

Fischer, A.W., Alexander, N.S., Woetzel, N., Karakas, M., Weiner, B.E. & Meiler, J., 2015. BCL::MP-fold: Membrane protein structure prediction guided by EPR restraints. *Proteins,* 83**,** 1947-62.

Fischer, A.W., Heinze, S., Putnam, D.K., Li, B., Pino, J.C., Xia, Y., Lopez, C.F. & Meiler, J., 2016. CASP11--An Evaluation of a Modular BCL::Fold-Based Protein Structure Prediction Pipeline. *PLoS One,* 11**,** e0152517.

Fiser, A., Feig, M., Brooks, C.L., 3rd & Sali, A., 2002. Evolution and physics in comparative protein structure modeling. *Acc Chem Res,* 35**,** 413-21.

Fodor, A.A. & Aldrich, R.W., 2004. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins,* 56**,** 211-21.

Franqueza, L., Lin, M., Shen, J., Splawski, I., Keating, M.T. & Sanguinetti, M.C., 1999. Long QT syndrome-associated mutations in the S4-S5 linker of KvLQT1 potassium channels modify gating and interaction with minK subunits. *J Biol Chem,* 274**,** 21063-70.

Frauenfelder, H., Sligar, S.G. & Wolynes, P.G., 1991. The energy landscapes and motions of proteins. *Science,* 254**,** 1598-603.

Fukunishi, H., Watanabe, O. & Takada, S., 2002. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *Journal of Chemical Physics,* 116**,** 9058-9067.

Gelin, B.R. & Karplus, M., 1979. Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. *Biochemistry,* 18**,** 1256-68.

Gimpelev, M., Forrest, L.R., Murray, D. & Honig, B., 2004. Helical packing patterns in membrane and soluble proteins. *Biophys J,* 87**,** 4075-86.

Giudicessi, J.R. & Ackerman, M.J., 2013. Genetic testing in heritable cardiac arrhythmia syndromes: differentiating pathogenic mutations from background genetic noise. *Curr Opin Cardiol,* 28**,** 63-71.

Giudicessi, J.R., Kapplinger, J.D., Tester, D.J., Alders, M., Salisbury, B.A., Wilde, A.A. & Ackerman, M.J., 2012. Phylogenetic and physicochemical analyses enhance the classification of rare nonsynonymous single nucleotide variants in type 1 and 2 long-QT syndrome. *Circ Cardiovasc Genet,* 5**,** 519-28.

Göbel, U., Sander, C., Schneider, R. & Valencia, A., 1994. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics,* 18**,** 309-317.

Godzik, A., 1996. Knowledge-based potentials for protein folding: what can we learn from known protein structures? *Structure,* 4**,** 363-6.

Gohlke, H., Hendlich, M. & Klebe, G., 2000. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol,* 295**,** 337-56.

Goldberg, D.E., 1989. Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading, MA.

Goldenberg, I. & Moss, A.J., 2008. Long QT syndrome. *J Am Coll Cardiol,* 51**,** 2291-300.

Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. *Science,* 185**,** 862-4.

Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A. & Baker, D., 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology,* 331**,** 281-299.

Gribskov, M., Mclachlan, A.D. & Eisenberg, D., 1987. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A,* 84**,** 4355-8.

Gromiha, M.M., 2009. Multiple Contact Network Is a Key Determinant to Protein Folding Rates. *Journal of Chemical Information and Modeling,* 49**,** 1130-1135.

Gromiha, M.M. & Selvaraj, S., 2001. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. *Journal of Molecular Biology,* 310**,** 27-32.

Gromiha, M.M. & Selvaraj, S., 2004. Inter-residue interactions in protein folding and stability. *Prog Biophys Mol Biol,* 86**,** 235-77.

Gromiha, M.M., Thangakani, A.M. & Selvaraj, S., 2006. FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Research,* 34**,** W70-W74.

Gruebele, M., 2002. Protein folding: the free energy surface. *Curr Opin Struct Biol,* 12**,** 161-8.

Gueudre, T., Baldassi, C., Zamparo, M., Weigt, M. & Pagnani, A., 2016. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling

analysis. *Proceedings of the National Academy of Sciences of the United States of America,* 113**,** 12186-12191.

Gumbart, J., Wang, Y., Aksimentiev, A., Tajkhorshid, E. & Schulten, K., 2005. Molecular dynamics simulations of proteins in lipid bilayers. *Curr Opin Struct Biol,* 15**,** 423-31.

Guo, J. & Rao, N., 2011. Predicting protein folding rate from amino acid sequence. *J Bioinform Comput Biol,* 9**,** 1-13.

Hamelryck, T., Borg, M., Paluszewski, M., Paulsen, J., Frellsen, J., Andreetta, C., Boomsma, W., Bottaro, S. & Ferkinghoff-Borg, J., 2010. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS One,* 5**,** e13714.

Hanley, J.A. & Mcneil, B.J., 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (Roc) Curve. *Radiology,* 143**,** 29-36.

Hansmann, U.H.E., 1997. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters,* 281**,** 140-150.

Hardin, C., Pogorelov, T.V. & Luthey-Schulten, Z., 2002. Ab initio protein structure prediction. *Curr Opin Struct Biol,* 12**,** 176-81.

Harmer, S.C., Mohal, J.S., Royal, A.A., Mckenna, W.J., Lambiase, P.D. & Tinker, A., 2014. Cellular mechanisms underlying the increased disease severity seen for patients with long QT syndrome caused by compound mutations in KCNQ1. *Biochem J,* 462**,** 133-42.

Haste Andersen, P., Nielsen, M. & Lund, O., 2006. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci,* 15**,** 2558-67.

Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y. & Zhou, Y., 2015. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep,* 5**,** 11476.

Heijman, J., Spatjens, R.L., Seyen, S.R., Lentink, V., Kuijpers, H.J., Boulet, I.R., De Windt, L.J., David, M. & Volders, P.G., 2012. Dominant-negative control of cAMP-dependent IKs upregulation in human long-QT syndrome type 1. *Circ Res,* 110**,** 211-9.

Heinig, M. & Frishman, D., 2004. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res,* 32**,** W500-2.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M.J., 1990. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J Mol Biol,* 216**,** 167-80.

Hirst, S.J., Alexander, N., Mchaourab, H.S. & Meiler, J., 2011. RosettaEPR: An Integrated Tool for Protein Structure Determination from Sparse EPR Data. *Journal of structural biology,* 173**,** 506-514.

Hong, K., Piper, D.R., Diaz-Valdecantos, A., Brugada, J., Oliva, A., Burashnikov, E., Santos-De-Soto, J., Grueso-Montero, J., Diaz-Enfante, E., Brugada, P., Sachse, F., Sanguinetti, M.C. & Brugada, R., 2005. De novo KCNQ1 mutation responsible for atrial fibrillation and short QT syndrome in utero. *Cardiovasc Res,* 68**,** 433-40.

Hoosien, M., Ahearn, M.E., Myerburg, R.J., Pham, T.V., Miller, T.E., Smets, M.J., Baumbach-Reardon, L., Young, M.L., Farooq, A. & Bishopric, N.H., 2013. Dysfunctional potassium channel subunit interaction as a novel mechanism of long QT syndrome. *Heart Rhythm,* 10**,** 728-37.

Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C. & Marks, D.S., 2012. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell,* 149**,** 1607-21.

Hopf, T.A., Scharfe, C.P., Rodrigues, J.P., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M. & Marks, D.S., 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife,* 3.

Hoque, M.T., Chetty, M. & Sattar, A., 2009. Genetic Algorithm inAb Initio Protein Structure Prediction Using Low Resolution Model: A Review. *In* A.S. Sidhu & T.S. Dillon (eds.) *Biomedical Data and Applications.* Berlin, Heidelberg: Springer Berlin Heidelberg, 317-342.

Hu, X., Beratan, D.N. & Yang, W., 2009. A gradient-directed Monte Carlo method for global optimization in a discrete space: Application to protein sequence design and folding. *The Journal of Chemical Physics,* 131**,** 154117.

Huang, C., Yang, X. & He, Z., 2010. Protein folding simulations of 2D HP model by the genetic algorithm based on optimal secondary structures. *Comput Biol Chem,* 34**,** 137-42.

Huang, J. & Mackerell, A.D., Jr., 2013. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem,* 34**,** 2135-45.

Huang, L., Bitner-Glindzicz, M., Tranebjaerg, L. & Tinker, A., 2001. A spectrum of functional effects for disease causing mutations in the Jervell and Lange-Nielsen syndrome. *Cardiovasc Res,* 51**,** 670-80.

Huang, S.Y. & Zou, X., 2006a. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J Comput Chem,* 27**,** 1866-75.

Huang, S.Y. & Zou, X., 2006b. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J Comput Chem,* 27**,** 1876-82.

Hubbard, S.J. & Thornton, J.M., 1993. NACCESS. University College London.

Ikrar, T., Hanawa, H., Watanabe, H., Aizawa, Y., Ramadan, M.M., Chinushi, M., Horie, M. & Aizawa, Y., 2009. Evaluation of channel function after alteration of amino acid residues at the pore center of KCNQ1 channel. *Biochem Biophys Res Commun,* 378**,** 589-94.

Illergard, K., Ardell, D.H. & Elofsson, A., 2009. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins,* 77**,** 499-508.

Illergard, K., Callegari, S. & Elofsson, A., 2010. MPRAP: an accessibility predictor for a-helical transmembrane proteins that performs well inside and outside the membrane. *BMC Bioinformatics,* 11**,** 333.

Illergard, K., Kauko, A. & Elofsson, A., 2011. Why are polar residues within the membrane core evolutionary conserved? *Proteins,* 79**,** 79-91.

Ingram, J.R., Knockenhauer, K.E., Markus, B.M., Mandelbaum, J., Ramek, A., Shan, Y., Shaw, D.E., Schwartz, T.U., Ploegh, H.L. & Lourido, S., 2015. Allosteric activation of apicomplexan calcium-dependent protein kinases. *Proc Natl Acad Sci U S A,* 112**,** E4975-84.

Itoh, H., Sakaguchi, T., Ding, W.G., Watanabe, E., Watanabe, I., Nishio, Y., Makiyama, T., Ohno, S., Akao, M., Higashi, Y., Zenda, N., Kubota, T., Mori, C., Okajima, K., Haruna, T., Miyamoto, A., Kawamura, M., Ishida, K., Nagaoka, I., Oka, Y., Nakazawa, Y., Yao, T., Jo, H., Sugimoto, Y., Ashihara, T., Hayashi, H., Ito, M., Imoto, K., Matsuura, H. & Horie, M., 2009. Latent genetic backgrounds and molecular pathogenesis in drug-induced long-QT syndrome. *Circ Arrhythm Electrophysiol,* 2**,** 511-23.

Itoh, S.G., Damjanovic, A. & Brooks, B.R., 2011. pH replica-exchange method based on discrete protonation states. *Proteins-Structure Function and Bioinformatics,* 79**,** 3420-3436.

Ivankov, D.N. & Finkelstein, A.V., 2004. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc Natl Acad Sci U S A,* 101**,** 8942-4.

Jaakkola, T., Diekhans, M. & Haussler, D., 1999. Using the Fisher Kernel Method to Detect Remote Protein Homologies. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, 149-158.

Jacobsen, R.B., Sale, K.L., Ayson, M.J., Novak, P., Hong, J., Lane, P., Wood, N.L., Kruppa, G.H., Young, M.M. & Schoeniger, J.S., 2006. Structure and dynamics of dark-state bovine rhodopsin revealed by chemical cross-linking and high-resolution mass spectrometry. *Protein Science : A Publication of the Protein Society,* 15**,** 1303-1317.

Jauch, R., Yeo, H.C., Kolatkar, P.R. & Clarke, N.D., 2007. Assessment of CASP7 structure predictions for template free targets. *Proteins,* 69 Suppl 8**,** 57-67.

Jeffrey L Mendenhall & Meiler, J., 2014. Prediction of Transmembrane Proteins and Regions using Fourier Spectral Analysis and Advancements in Machine Learning. *SERMACS 2014.* Nashville, TN.

Jiang, F. & Wu, Y.D., 2014. Folding of fourteen small proteins with a residue-specific force field and replica-exchange molecular dynamics. *J Am Chem Soc,* 136**,** 9536-9.

Jiang, W., Baker, M.L., Ludtke, S.J. & Chiu, W., 2001. Bridging the information gap: computational tools for intermediate resolution structure interpretation1. *Journal of Molecular Biology,* 308**,** 1033-1044.

Jones, D.T., 1997a. Progress in protein structure prediction. *Curr Opin Struct Biol,* 7**,** 377-87.

Jones, D.T., 1997b. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins-Structure Function and Genetics***,** 185-191.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol,* 292**,** 195-202.

Jones, D.T., 2001. Predicting novel protein folds by using FRAGFOLD. *Proteins,* Suppl 5**,** 127-32.

Jones, D.T., Buchan, D.W.A., Cozzetto, D. & Pontil, M., 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics,* 28**,** 184-190.

Jones, D.T. & Mcguffin, L.J., 2003. Assembling novel protein folds from super-secondary structural fragments. *Proteins,* 53 Suppl 6**,** 480-5.

Jones, D.T., Singh, T., Kosciolek, T. & Tetchner, S., 2015. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics,* 31**,** 999-1006.

Jones, S. & Thornton, J.M., 1996. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A,* 93**,** 13-20.

Jones, S. & Thornton, J.M., 1997a. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol,* 272**,** 121-32.

Jones, S. & Thornton, J.M., 1997b. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol,* 272**,** 133-43.

Jorgensen, W.L. & Tirado-Rives, J., 1988. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc,* 110**,** 1657-66.

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

Kahsay, R.Y., Gao, G. & Liao, L., 2005. An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics,* 21**,** 1853-1858.

Kaján, L., Hopf, T.A., Kalaš, M., Marks, D.S. & Rost, B., 2014. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics,* 15**,** 1-6.

Kalkhof, S., Ihling, C., Mechtler, K. & Sinz, A., 2005. Chemical Cross-Linking and High-Performance Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for Protein Interaction Analysis: Application to a Calmodulin/Target Peptide Complex. *Analytical Chemistry,* 77**,** 495-503.

Kamisetty, H., Ovchinnikov, S. & Baker, D., 2013. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 15674-15679.

Kapa, S., Tester, D.J., Salisbury, B.A., Harris-Kerr, C., Pungliya, M.S., Alders, M., Wilde, A.A. & Ackerman, M.J., 2009. Genetic testing for long-QT syndrome: distinguishing pathogenic mutations from benign variants. *Circulation,* 120**,** 1752-60.

Karakas, M., Woetzel, N., Staritzbichler, R., Alexander, N., Weiner, B.E. & Meiler, J., 2012. BCL::Fold--de novo prediction of complex and large protein topologies by assembly of secondary structure elements. *PLoS One,* 7**,** e49240.

Karplus, M., 2011. Behind the folding funnel diagram. *Nature Chemical Biology,* 7**,** 401-404.

Karplus, M. & Kuriyan, J., 2005. Molecular dynamics and protein function. *Proc Natl Acad Sci U S A,* 102**,** 6679-85.

Karplus, M. & Mccammon, J.A., 2002. Molecular dynamics simulations of biomolecules. *Nat Struct Biol,* 9**,** 646-52.

Karplus, M. & Petsko, G.A., 1990. Molecular dynamics simulations in biology. *Nature,* 347**,** 631-9.

Karplus, M. & Weaver, D.L., 1994. Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci,* 3**,** 650-68.

Katsanis, S.H. & Katsanis, N., 2013. Molecular genetic testing and the future of clinical genomics. *Nat Rev Genet,* 14**,** 415-26.

Kawano, K., Yano, Y., Omae, K., Matsuzaki, S. & Matsuzaki, K., 2013. Stoichiometric analysis of oligomerization of membrane proteins on living cells using coiled-coil labeling and spectral imaging. *Anal Chem,* 85**,** 3454-61.

Kim, D.E., Blum, B., Bradley, P. & Baker, D., 2009. Sampling bottlenecks in de novo protein structure prediction. *J Mol Biol,* 393**,** 249-60.

Kim, D.E., Dimaio, F., Yu-Ruei Wang, R., Song, Y. & Baker, D., 2014. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins: Structure, Function, and Bioinformatics,* 82**,** 208-218.

Kim, P.S. & Baldwin, R.L., 1982. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu Rev Biochem,* 51**,** 459-89.

Kim, P.S. & Baldwin, R.L., 1990. Intermediates in the folding reactions of small proteins. *Annu Rev Biochem,* 59**,** 631-60.

Kim, T.R., Yang, J.S., Shin, S. & Lee, J., 2013. Statistical torsion angle potential energy functions for protein structure modeling: a bicubic interpolation approach. *Proteins,* 81**,** 1156-65.

Kinch, L., Yong Shi, S., Cong, Q., Cheng, H., Liao, Y. & Grishin, N.V., 2011. CASP9 assessment of free modeling target predictions. *Proteins,* 79 Suppl 10**,** 59-73.

Kinch, L.N., Li, W., Monastyrskyy, B., Kryshtafovych, A. & Grishin, N.V., 2016. Evaluation of free modeling targets in CASP11 and ROLL. *Proteins,* 84 Suppl 1**,** 51-66.

Kinjo, A.R., Horimoto, K. & Nishikawa, K., 2005. Predicting absolute contact numbers of native protein structure from amino acid sequence. *Proteins,* 58**,** 158-65.

Kinoshita, K., Komatsu, T., Nishide, K., Hata, Y., Hisajima, N., Takahashi, H., Kimoto, K., Aonuma, K., Tsushima, E., Tabata, T., Yoshida, T., Mori, H., Nishida, K., Yamaguchi, Y., Ichida, F., Fukurotani, K., Inoue, H. & Nishida, N., 2014. A590T mutation in KCNQ1 C-terminal helix D decreases IKs channel trafficking and function but not Yotiao interaction. *J Mol Cell Cardiol,* 72**,** 273-80.

Kirkpatrick, S., Gelatt, C.D., Jr. & Vecchi, M.P., 1983. Optimization by simulated annealing. *Science,* 220**,** 671-80.

Klare, J.P. & Steinhoff, H.J., 2009. Spin labeling EPR. *Photosynth Res,* 102**,** 377-90.

Klepeis, J.L., Lindorff-Larsen, K., Dror, R.O. & Shaw, D.E., 2009. Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol,* 19**,** 120-7.

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A.E. & Kolinski, A., 2016. Coarse-Grained Protein Models and Their Applications. *Chem Rev,* 116**,** 7898-936.

Kocher, J.P., Rooman, M.J. & Wodak, S.J., 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol,* 235**,** 1598-613.

Kong, Y. & Ma, J., 2003. A Structural-informatics Approach for Mining β-Sheets: Locating Sheets in Intermediate-resolution Density Maps. *Journal of Molecular Biology,* 332**,** 399-413.

Kong, Y., Zhang, X., Baker, T.S. & Ma, J., 2004. A Structural-informatics Approach for Tracing β-Sheets: Building Pseudo-C(α) Traces for β-Strands in Intermediate-resolution Density Maps. *Journal of molecular biology,* 339**,** 117-130.

Kortemme, T., Morozov, A.V. & Baker, D., 2003. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol,* 326**,** 1239-59.

Kosciolek, T. & Jones, D.T., 2014. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS One,* 9**,** e92197.

Kosciolek, T. & Jones, D.T., 2015. Accurate contact predictions using covariation techniques and machine learning. *Proteins: Structure, Function, and Bioinformatics***,** n/a-n/a.

Krizhevsky, A., Sutskever, I. & Hinton, G.E., Year. Imagenet classification with deep convolutional neural networksed.^eds. *Advances in neural information processing systems*, 1097-1105.

Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol,* 305**,** 567-80.

Kroncke, B.M., Duran, A.M., Mendenhall, J.L., Meiler, J., Blume, J.D. & Sanders, C.R., 2016. Documentation of an Imperative To Improve Methods for Predicting Membrane Protein Stability. *Biochemistry,* 55**,** 5002-9.

Kubota, T., Horie, M., Takano, M., Yoshida, H., Takenaka, K., Watanabe, E., Tsuchiya, T., Otani, H. & Sasayama, S., 2001. Evidence for a single nucleotide polymorphism in the KCNQ1 potassium channel that underlies susceptibility to life-threatening arrhythmias. *J Cardiovasc Electrophysiol,* 12**,** 1223-9.

Kubota, T., Shimizu, W., Kamakura, S. & Horie, M., 2000. Hypokalemia-induced long QT syndrome with an underlying novel missense mutation in S4-S5 linker of KCNQ1. *J Cardiovasc Electrophysiol,* 11**,** 1048-54.

Kuszewski, J., Gronenborn, A.M. & Clore, G.M., 1996. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci,* 5**,** 1067-80.

Labro, A.J., Boulet, I.R., Choveau, F.S., Mayeur, E., Bruyns, T., Loussouarn, G., Raes, A.L. & Snyders, D.J., 2011. The S4-S5 linker of KCNQ1 channels forms a structural scaffold with the S6 segment controlling gate closure. *J Biol Chem,* 286**,** 717-25.

Labro, A.J., Boulet, I.R., Timmermans, J.P., Ottschytsch, N. & Snyders, D.J., 2010. The rate-dependent biophysical properties of the LQT1 H258R mutant are counteracted by a dominant negative effect on channel trafficking. *J Mol Cell Cardiol,* 48**,** 1096-104.

Lai, J.S., Cheng, C.W., Lo, A., Sung, T.Y. & Hsu, W.L., 2013. Lipid exposure prediction enhances the inference of rotational angles of transmembrane helices. *BMC Bioinformatics,* 14**,** 304.

Laio, A. & Parrinello, M., 2002. Escaping free-energy minima. *Proc Natl Acad Sci U S A,* 99**,** 12562-6.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M. & Maglott, D.R., 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res,* 42**,** D980-5.

Lane, T.J., Bowman, G.R., Beauchamp, K., Voelz, V.A. & Pande, V.S., 2011. Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *J Am Chem Soc,* 133**,** 18413-9.

Lane, T.J., Shukla, D., Beauchamp, K.A. & Pande, V.S., 2013. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr Opin Struct Biol,* 23**,** 58-65.

Lanfear, R., Kokko, H. & Eyre-Walker, A., 2014. Population size and the rate of evolution. *Trends Ecol Evol,* 29**,** 33-41.

Larsen, L.A., Fosdal, I., Andersen, P.S., Kanters, J.K., Vuust, J., Wettrell, G. & Christiansen, M., 1999. Recessive Romano-Ward syndrome associated with compound heterozygosity for two mutations in the KVLQT1 gene. *Eur J Hum Genet,* 7**,** 724-8.

Lasker, K., Förster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., Beck, F., Aebersold, R., Sali, A. & Baumeister, W., 2012. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proceedings of the National Academy of Sciences of the United States of America,* 109**,** 1380-1387.

Latek, D., Ekonomiuk, D. & Kolinski, A., 2007. Protein structure prediction: Combining de novo modeling with sparse experimental data. *Journal of Computational Chemistry,* 28**,** 1668-1676.

Lau, K.F. & Dill, K.A., 1989. A Lattice Statistical-Mechanics Model of the Conformational and Sequence-Spaces of Proteins. *Macromolecules,* 22**,** 3986-3997.

Lazaridis, T. & Karplus, M., 1999. Effective energy function for proteins in solution. *Proteins-Structure Function and Bioinformatics,* 35**,** 133-152.

Lazaridis, T. & Karplus, M., 2000. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol,* 10**,** 139-45.

Lazaridis, T. & Karplus, M., 2003. Thermodynamics of protein folding: a microscopic view. *Biophys Chem,* 100**,** 367-95.

Leach, A.R., 2001. Molecular modelling: principles and applications. Pearson education.

Lee, B. & Richards, F.M., 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol,* 55**,** 379-400.

Lee, J., Wu, S. & Zhang, Y., 2009. Ab initio protein structure prediction. *From protein structure to function with bioinformatics.* Springer, 3-25.

Leman, J.K., Mueller, R., Karakas, M., Woetzel, N. & Meiler, J., 2013. Simultaneous prediction of protein secondary structure and transmembrane spans. *Proteins,* 81**,** 1127-40.

Leong, I.U.S., Stuckey, A., Lai, D., Skinner, J.R. & Love, D.R., 2015. Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations. *Bmc Medical Genetics,* 16.

Levinthal, C., 1968. Are There Pathways for Protein Folding. *Journal De Chimie Physique Et De Physico-Chimie Biologique,* 65**,** 44-+.

Levinthal, C., Year. How to fold graciouslyed.^eds. *Mossbauer spectroscopy in biological systems In Proceedings of a Meeting held at Allerton House,* Monticello, Illinois: University of Illinois Press, 22-24.

Li, B., Fooksa, M., Heinze, S. & Meiler, J., 2018. Finding the needle in the haystack: towards solving the protein-folding problem computationally. *Crit Rev Biochem Mol Biol,* 53**,** 1-28.

Li, B., Li, W., Du, P., Yu, K.Q. & Fu, W., 2012. Molecular insights into the D1R agonist and D2R/D3R antagonist effects of the natural product (-)-stepholidine: molecular modeling and dynamics simulations. *J Phys Chem B,* 116**,** 8121-30.

Li, B., Mendenhall, J., Nguyen, E.D., Weiner, B.E., Fischer, A.W. & Meiler, J., 2016. Accurate Prediction of Contact Numbers for Multi-Spanning Helical Membrane Proteins. *J Chem Inf Model,* 56**,** 423-34.

Li, B., Mendenhall, J., Nguyen, E.D., Weiner, B.E., Fischer, A.W. & Meiler, J., 2017a. Improving prediction of helix-helix packing in membrane proteins using predicted contact numbers as restraints. *Proteins,* 85**,** 1212-1221.

Li, B., Mendenhall, J.L., Kroncke, B.M., Taylor, K.C., Huang, H., Smith, D.K., Vanoye, C.G., Blume, J.D., George, A.L., Jr., Sanders, C.R. & Meiler, J., 2017b. Predicting the Functional Impact of KCNQ1 Variants of Unknown Significance. *Circ Cardiovasc Genet,* 10.

Li, B., Xu, L.L., Shen, Q., Gu, X.F. & Fu, W., 2014. Discovery of novel small-molecule Src kinase inhibitors via a kinase-focused druglikeness rule and structure-based virtual screening. *Molecular Simulation,* 40**,** 341-348.

Li, D.W. & Bruschweiler, R., 2010. NMR-based protein potentials. *Angew Chem Int Ed Engl,* 49**,** 6778-80.

Li, W., Du, R., Wang, Q.F., Tian, L., Yang, J.G. & Song, Z.F., 2009a. The G314S KCNQ1 mutation exerts a dominant-negative effect on expression of KCNQ1 channels in oocytes. *Biochem Biophys Res Commun,* 383**,** 206-9.

Li, W., Wang, Q.F., Du, R., Xu, Q.M., Ke, Q.M., Wang, B., Chen, X.L., Tian, L., Zhang, S.Y., Kang, C.L., Guan, S.M., Yang, J.G. & Song, Z.F., 2009b. Congenital long QT syndrome caused by the F275S KCNQ1 mutation: mechanism of impaired channel function. *Biochem Biophys Res Commun,* 380**,** 127-31.

Li, Y., Fang, Y. & Fang, J., 2011. Predicting residue–residue contacts using random forest models. *Bioinformatics,* 27**,** 3379-3384.

Li, Y., Gao, J.Y., Lu, Z.J., Mcfarland, K., Shi, J.Y., Bock, K., Cohen, I.S. & Cui, J.M., 2013. Intracellular ATP binding is required to activate the slowly activating K+ channel I-Ks. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 18922-18927.

Lifson, S. & Warshel, A., 1968. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *The Journal of Chemical Physics,* 49**,** 5116-5129.

Lin, C.P., Huang, S.W., Lai, Y.L., Yen, S.C., Shih, C.H., Lu, C.H., Huang, C.C. & Hwang, J.K., 2008. Deriving protein dynamical properties from weighted protein contact number. *Proteins,* 72**,** 929-35.

Lindahl, E. & Sansom, M.S., 2008. Membrane proteins: molecular dynamics simulations. *Curr Opin Struct Biol,* 18**,** 425-31.

Lindert, S., Alexander, N., Wotzel, N., Karakas, M., Stewart, P.L. & Meiler, J., 2012a. EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. *Structure,* 20**,** 464-78.

Lindert, S., Hofmann, T., Wotzel, N., Karakas, M., Stewart, P.L. & Meiler, J., 2012b. Ab initio protein modeling into CryoEM density maps using EM-Fold. *Biopolymers,* 97**,** 669-77.

Lindert, S., Staritzbichler, R., Wotzel, N., Karakas, M., Stewart, P.L. & Meiler, J., 2009a. EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. *Structure,* 17**,** 990-1003.

Lindert, S., Stewart, P.L. & Meiler, J., 2009b. Hybrid approaches: applying computational methods in cryo-electron microscopy. *Curr Opin Struct Biol,* 19**,** 218-25.

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M.P., Dror, R.O. & Shaw, D.E., 2012. Systematic validation of protein force fields against experimental data. *PLoS One,* 7**,** e32131.

Lindorff-Larsen, K., Piana, S., Dror, R.O. & Shaw, D.E., 2011. How fast-folding proteins fold. *Science,* 334**,** 517-20.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O. & Shaw, D.E., 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins,* 78**,** 1950-8.

Lippi, M. & Frasconi, P., 2009. Prediction of protein β-residue contacts by Markov logic networks with grounding-specific weights. *Bioinformatics,* 25**,** 2326-2333.

Lo Conte, L., Chothia, C. & Janin, J., 1999. The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology,* 285**,** 2177-2198.

Lomize, M.A., Lomize, A.L., Pogozheva, I.D. & Mosberg, H.I., 2006. OPM: orientations of proteins in membranes database. *Bioinformatics,* 22**,** 623-5.

Lomize, M.A., Pogozheva, I.D., Joo, H., Mosberg, H.I. & Lomize, A.L., 2012. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res,* 40**,** D370-6.

Lopes, P.E., Guvench, O. & Mackerell, A.D., Jr., 2015. Current status of protein force fields for molecular dynamics simulations. *Methods Mol Biol,* 1215**,** 47-71.

Lu, H. & Skolnick, J., 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins,* 44**,** 223-32.

Lundby, A., Ravn, L.S., Svendsen, J.H., Olesen, S.P. & Schmitt, N., 2007. KCNQ1 mutation Q147R is associated with atrial fibrillation and prolonged QT interval. *Heart Rhythm,* 4**,** 1532-41.

Ma, J. & Wang, S., 2015. AcconPred: Predicting Solvent Accessibility and Contact Number Simultaneously by a Multitask Learning Framework under the Conditional Neural Fields Model. *Biomed Res Int,* 2015**,** 678764.

Ma, J., Wang, S., Zhao, F. & Xu, J., 2013. Protein threading using context-specific alignment potential. *Bioinformatics,* 29**,** i257-i265.

Macarthur, D.G. & Tyler-Smith, C., 2010. Loss-of-function variants in the genomes of healthy humans. *Hum Mol Genet,* 19**,** R125-30.

Maccallum, R.M., 2004. Striped sheets and protein contact prediction. *Bioinformatics,* 20**,** i224-i231.

Mackerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-Mccarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M., 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B,* 102**,** 3586-616.

Mackerell, A.D., Jr., 2004. Empirical force fields for biological macromolecules: overview and issues. *J Comput Chem,* 25**,** 1584-604.

Mackerell, A.D., Jr., Feig, M. & Brooks, C.L., 3rd, 2004. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem,* 25**,** 1400-15.

Madaoui, H. & Guerois, R., 2008. Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proc Natl Acad Sci U S A,* 105**,** 7708-13.

Majek, P. & Elber, R., 2009. A coarse-grained potential for fold recognition and molecular dynamics simulations of proteins. *Proteins,* 76**,** 822-36.

Majumdar, I., Krishna, S.S. & Grishin, N.V., 2005. PALSSE: a program to delineate linear secondary structural elements from protein structures. *BMC Bioinformatics,* 6**,** 202.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R. & Sander, C., 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS One,* 6**,** e28766.

Marks, D.S., Hopf, T.A. & Sander, C., 2012. Protein structure prediction from sequence variation. *Nat Biotechnol,* 30**,** 1072-80.

Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P. & De Vries, A.H., 2007. The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B,* 111**,** 7812-24.

Marrink, S.J. & Tieleman, D.P., 2013. Perspective on the Martini model. *Chem Soc Rev,* 42**,** 6801-22.

Marsh, J.A. & Teichmann, S.A., 2011. Relative solvent accessible surface area predicts protein conformational changes upon binding. *Structure,* 19**,** 859-67.

Martins, J.M., Ramos, R.M., Pimenta, A.C. & Moreira, I.S., 2014. Solvent-accessible surface area: How well can be applied to hot-spot detection? *Proteins,* 82**,** 479-90.

Matavel, A., Medei, E. & Lopes, C.M., 2010. PKA and PKC partially rescue long QT type 1 phenotype by restoring channel-PIP2 interactions. *Channels (Austin),* 4**,** 3-11.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta,* 405**,** 442-51.

Maximova, T., Moffatt, R., Ma, B., Nussinov, R. & Shehu, A., 2016. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comput Biol,* 12**,** e1004619.

Mccammon, J.A., Gelin, B.R. & Karplus, M., 1977. Dynamics of folded proteins. *Nature,* 267**,** 585-90.

Mcguffin, L.J., Bryson, K. & Jones, D.T., 2000. The PSIPRED protein structure prediction server. *Bioinformatics,* 16**,** 404-5.

Mclachlan, A.D., 1971. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c551. *Journal of Molecular Biology,* 61**,** 409-424.

Melo, F. & Feytmans, E., 1997. Novel knowledge-based mean force potential at atomic level. *J Mol Biol,* 267**,** 207-22.

Mendenhall, J. & Meiler, J., 2016. Improving quantitative structure-activity relationship models using Artificial Neural Networks trained with dropout. *J Comput Aided Mol Des,* 30**,** 177-89.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E., 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics,* 21**,** 1087-1092.

Michel, M., Hayat, S., Skwark, M.J., Sander, C., Marks, D.S. & Elofsson, A., 2014. PconsFold: improved contact predictions improve protein models. *Bioinformatics,* 30**,** i482-i488.

Miller, J.P., Lo, R.S., Ben-Hur, A., Desmarais, C., Stagljar, I., Noble, W.S. & Fields, S., 2005. Large-scale identification of yeast integral membrane protein interactions. *Proc Natl Acad Sci U S A,* 102**,** 12123-8.

Mintseris, J. & Weng, Z.P., 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 10930-10935.

Miyazawa, S. & Jernigan, R.L., 1985. Estimation of Effective Interresidue Contact Energies from Protein Crystal-Structures - Quasi-Chemical Approximation. *Macromolecules,* 18**,** 534-552.

Miyazawa, S. & Jernigan, R.L., 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol,* 256**,** 623-44.

Modell, S.M. & Lehmann, M.H., 2006. The long QT syndrome family of cardiac ion channelopathies: a HuGE review. *Genet Med,* 8**,** 143-55.

Monastyrskyy, B., D'andrea, D., Fidelis, K., Tramontano, A. & Kryshtafovych, A., 2015. New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins: Structure, Function, and Bioinformatics***,** n/a-n/a.

Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R.G., Tieleman, D.P. & Marrink, S.J., 2008. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J Chem Theory Comput,* 4**,** 819-34.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T. & Weigt, M., 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A,* 108**,** E1293-301.

Moreno, C., Oliveras, A., De La Cruz, A., Bartolucci, C., Munoz, C., Salar, E., Gimeno, J.R., Severi, S., Comes, N., Felipe, A., Gonzalez, T., Lambiase, P. & Valenzuela, C., 2015. A new KCNQ1 mutation at the S5 segment that impairs its association with KCNE1 is responsible for short QT syndrome. *Cardiovasc Res,* 107**,** 613-23.

Mori, T., Miyashita, N., Im, W., Feig, M. & Sugita, Y., 2016. Molecular dynamics simulations of biological membranes and membrane proteins using enhanced conformational sampling algorithms. *Biochim Biophys Acta,* 1858**,** 1635-51.

Moult, J., 1997. Comparison of database potentials and molecular mechanics force fields. *Curr Opin Struct Biol,* 7**,** 194-9.

Moult, J., 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol,* 15**,** 285-9.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A., 2016. Critical assessment of methods of protein structure prediction (CASP) – progress and new directions in Round XI. *Proteins: Structure, Function, and Bioinformatics***,** n/a-n/a.

Moult, J., Pedersen, J.T., Judson, R. & Fidelis, K., 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins,* 23**,** ii-v.

Munteanu, C.R., Pimenta, A.C., Fernandez-Lozano, C., Melo, A., Cordeiro, M.N. & Moreira, I.S., 2015. Solvent accessible surface area-based hot-spot detection methods for protein-protein and protein-nucleic acid interfaces. *J Chem Inf Model,* 55**,** 1077-86.

Murata, T., Yamato, I., Kakinuma, Y., Leslie, A.G. & Walker, J.E., 2005. Structure of the rotor of the V-Type Na+-ATPase from Enterococcus hirae. *Science,* 308**,** 654-9.

Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol,* 247**,** 536-40.

Neher, E., 1994. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences,* 91**,** 98-102.

Neyroud, N., Denjoy, I., Donger, C., Gary, F., Villain, E., Leenhardt, A., Benali, K., Schwartz, K., Coumel, P. & Guicheney, P., 1998. Heterozygous mutation in the pore of potassium channel gene KvLQT1 causes an apparently normal phenotype in long QT syndrome. *Eur J Hum Genet,* 6**,** 129-33.

Ng, P.C. & Henikoff, S., 2001. Predicting deleterious amino acid substitutions. *Genome Res,* 11**,** 863-74.

Ng, P.C. & Henikoff, S., 2006. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet,* 7**,** 61-80.

Nguyen, H., Maier, J., Huang, H., Perrone, V. & Simmerling, C., 2014. Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent. *Journal of the American Chemical Society,* 136**,** 13959-13962.

Nishikawa, K. & Ooi, T., 1980. Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int J Pept Protein Res,* 16**,** 19-32.

Nishikawa, K. & Ooi, T., 1986. Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J Biochem,* 100**,** 1043-7.

Nooren, I.M.A. & Thornton, J.M., 2003a. Diversity of protein-protein interactions. *Embo Journal,* 22**,** 3486-3492.

Nooren, I.M.A. & Thornton, J.M., 2003b. Structural characterisation and functional significance of transient protein-protein interactions. *Journal of Molecular Biology,* 325**,** 991-1018.

Nugent, T. & Jones, D.T., 2009. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics,* 10**,** 1-11.

Nugent, T. & Jones, D.T., 2012. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci U S A,* 109**,** E1540-7.

Ohanian, M., Otway, R. & Fatkin, D., 2012. Heuristic Methods for Finding Pathogenic Variants in Gene Coding Sequences. *Journal of the American Heart Association,* 1.

Oka, Y., Itoh, H., Ding, W.G., Shimizu, W., Makiyama, T., Ohno, S., Nishio, Y., Sakaguchi, T., Miyamoto, A., Kawamura, M., Matsuura, H. & Horie, M., 2010. Atrioventricular block-induced Torsades de Pointes with clinical and molecular backgrounds similar to congenital long QT syndrome. *Circ J,* 74**,** 2562-71.

Okamoto, Y., 2004. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J Mol Graph Model,* 22**,** 425-39.

Olmea, O. & Valencia, A., 1997. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding and Design,* 2**,** S25-S32.

Onuchic, J.N., Luthey-Schulten, Z. & Wolynes, P.G., 1997. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem,* 48**,** 545-600.

Onuchic, J.N. & Wolynes, P.G., 2004. Theory of protein folding. *Curr Opin Struct Biol,* 14**,** 70-5.

Ouyang, Z. & Liang, J., 2008. Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Sci,* 17**,** 1256-63.

Ovchinnikov, S., Kamisetty, H. & Baker, D., 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife,* 3.

Ovchinnikov, S., Kinch, L., Park, H., Liao, Y.X., Pei, J.M., Kim, D.E., Kamisetty, H., Grishin, N.V. & Baker, D., 2015. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife,* 4.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.S., Pavlopoulos, G.A., Kim, D.E., Kamisetty, H., Kyrpides, N.C. & Baker, D., 2017. Protein structure determination using metagenome sequence data. *Science,* 355**,** 294-298.

Overington, J.P., Al-Lazikani, B. & Hopkins, A.L., 2006. How many drug targets are there? *Nat Rev Drug Discov,* 5**,** 993-6.

Paci, E., Lindorff-Larsen, K., Dobson, C.M., Karplus, M. & Vendruscolo, M., 2005. Transition state contact orders correlate with protein folding rates. *J Mol Biol,* 352**,** 495-500.

Pan, N., Sun, J., Lv, C., Li, H. & Ding, J., 2009. A hydrophobicity-dependent motif responsible for surface expression of cardiac potassium channel. *Cell Signal,* 21**,** 349-55.

Pande, V.S., Baker, I., Chapman, J., Elmer, S.P., Khaliq, S., Larson, S.M., Rhee, Y.M., Shirts, M.R., Snow, C.D., Sorin, E.J. & Zagrovic, B., 2003. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers,* 68**,** 91-109.

Pande, V.S., Beauchamp, K. & Bowman, G.R., 2010. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods,* 52**,** 99-105.

Park, B. & Levitt, M., 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol,* 258**,** 367-92.

Park, B.H., Huang, E.S. & Levitt, M., 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J Mol Biol,* 266**,** 831-46.

Park, Y., Hayat, S. & Helms, V., 2007. Prediction of the burial status of transmembrane residues of helical membrane proteins. *BMC Bioinformatics,* 8**,** 302.

Patapati, K.K. & Glykos, N.M., 2011. Three force fields' views of the 3(10) helix. *Biophys J,* 101**,** 1766-71.

Pazos, F., Helmercitterich, M., Ausiello, G. & Valencia, A., 1997. Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology,* 271**,** 511-523.

Pebay-Peyroula, E., Dahout-Gonzalez, C., Kahn, R., Trezeguet, V., Lauquin, G.J. & Brandolin, G., 2003. Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. *Nature,* 426**,** 39-44.

Pedersen, J.T. & Moult, J., 1996. Genetic algorithms for protein structure prediction. *Curr Opin Struct Biol,* 6**,** 227-31.

Periole, X., 2017. Interplay of G Protein-Coupled Receptors with the Membrane: Insights from Supra-Atomic Coarse Grain Molecular Dynamics Simulations. *Chem Rev,* 117**,** 156-185.

Perkins, J.R., Diboun, I., Dessailly, B.H., Lees, J.G. & Orengo, C., 2010. Transient protein-protein interactions: structural, functional, and network properties. *Structure,* 18**,** 1233-43.

Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M. & Lundegaard, C., 2009. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol,* 9**,** 51.

Phatak, M., Adamczak, R., Cao, B., Wagner, M. & Meller, J., 2011. Solvent and lipid accessibility prediction as a basis for model quality assessment in soluble and membrane proteins. *Curr Protein Pept Sci,* 12**,** 563-73.

Piana, S., Klepeis, J.L. & Shaw, D.E., 2014. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr Opin Struct Biol,* 24**,** 98-105.

Piana, S. & Laio, A., 2007. A bias-exchange approach to protein folding. *J Phys Chem B,* 111**,** 4553-9.

Piana, S., Lindorff-Larsen, K. & Shaw, D.E., 2012. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc Natl Acad Sci U S A,* 109**,** 17845-50.

Piana, S., Lindorff-Larsen, K. & Shaw, D.E., 2013. Atomic-level description of ubiquitin folding. *Proc Natl Acad Sci U S A,* 110**,** 5915-20.

Piippo, K., Swan, H., Pasternack, M., Chapman, H., Paavonen, K., Viitasalo, M., Toivonen, L. & Kontula, K., 2001. A founder mutation of the potassium channel KCNQ1 in long QT syndrome: implications for estimation of disease prevalence and molecular diagnostics. *J Am Coll Cardiol,* 37**,** 562-8.

Pillardy, J., Czaplewski, C., Liwo, A., Lee, J., Ripoll, D.R., Kazmierkiewicz, R., Oldziej, S., Wedemeyer, W.J., Gibson, K.D., Arnautova, Y.A., Saunders, J., Ye, Y.J. & Scheraga, H.A., 2001. Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc Natl Acad Sci U S A,* 98**,** 2329-33.

Pintilie, G.D., Zhang, J., Goddard, T.D., Chiu, W. & Gossard, D.C., 2010. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *Journal of structural biology,* 170**,** 427-438.

Pitera, J.W. & Swope, W., 2003. Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins. *Proc Natl Acad Sci U S A,* 100**,** 7587-92.

Plaxco, K.W., Simons, K.T. & Baker, D., 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol,* 277**,** 985-94.

Pollastri, G., Baldi, P., Fariselli, P. & Casadio, R., 2001. Improved prediction of the number of residue contacts in proteins by recurrent neural networks. *Bioinformatics,* 17 Suppl 1**,** S234-42.

Pollastri, G., Baldi, P., Fariselli, P. & Casadio, R., 2002. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins,* 47**,** 142-53.

Pollock, D.D. & Taylor, W.R., 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Engineering,* 10**,** 647-657.

Ponder, J.W. & Case, D.A., 2003. Force fields for protein simulations. *Adv Protein Chem,* 66**,** 27-85.

Poole, A.M. & Ranganathan, R., 2006. Knowledge-based potentials in protein design. *Curr Opin Struct Biol,* 16**,** 508-13.

Prinz, J.H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J.D., Schutte, C. & Noe, F., 2011. Markov models of molecular kinetics: generation and validation. *J Chem Phys,* 134**,** 174105.

Pruitt, K.D., Tatusova, T. & Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res,* 35**,** D61-5.

Ptitsyn, O., 1996. How molten is the molten globule? *Nature Structural Biology,* 3**,** 488-490.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F. & Ben-Tal, N., 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics,* 18 Suppl 1**,** S71-7.

Qi, Y., Oja, M., Weston, J. & Noble, W.S., 2012. A unified multitask architecture for predicting local protein properties. *PLoS One,* 7**,** e32235.

Qiu, J. & Elber, R., 2005. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins,* 61**,** 44-55.

R Development Core Team, 2015. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rabenstein, M.D. & Shin, Y.K., 1995. Determination of the distance between two spin labels attached to a macromolecule. *Proceedings of the National Academy of Sciences,* 92**,** 8239-8243.

Radford, S.E., 2000. Protein folding: progress made and promises ahead. *Trends in Biochemical Sciences,* 25**,** 611-618.

Ramakrishnan, C. & Ramachandran, G.N., 1965. Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophys J,* 5**,** 909-33.

Rapaport, D.C., 2004. *The art of molecular dynamics simulation*: Cambridge university press.

Rashid, M.A., Iqbal, S., Khatib, F., Hoque, M.T. & Sattar, A., 2016. Guided macro-mutation in a graded energy based genetic algorithm for protein structure prediction. *Comput Biol Chem,* 61**,** 162-77.

Remmert, M., Biegert, A., Hauser, A. & Söding, J., 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods,* 9**,** 173-5.

Rennell, D., Bouvier, S.E., Hardy, L.W. & Poteete, A.R., 1991. Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol,* 222**,** 67-88.

Reva, B.A., Finkelstein, A.V., Sanner, M.F. & Olson, A.J., 1997. Residue-residue mean-force potentials for protein structure recognition. *Protein Eng,* 10**,** 865-76.

Rhee, Y.M. & Pande, V.S., 2003. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys J,* 84**,** 775-86.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H.L. & Committee, A.L.Q.A., 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med,* 17**,** 405-24.

Robertson, M.J., Tirado-Rives, J. & Jorgensen, W.L., 2015. Improved Peptide and Protein Torsional Energetics with the OPLSAA Force Field. *J Chem Theory Comput,* 11**,** 3499-509.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S. & Baker, D., 2004. Protein structure prediction using rosetta. *Numerical Computer Methods, Pt D,* 383**,** 66-+.

Rooman, M. & Gilis, D., 1998. Different derivations of knowledge-based potentials and analysis of their robustness and context-dependent predictive power. *European Journal of Biochemistry,* 254**,** 135-143.

Rost, B. & Sander, C., 1993. Prediction of Protein Secondary Structure at Better Than 70-Percent Accuracy. *Journal of Molecular Biology,* 232**,** 584-599.

Rost, B., Schneider, R. & Sander, C., 1993. Progress in Protein-Structure Prediction. *Trends in Biochemical Sciences,* 18**,** 120-123.

Rosta, E. & Hummer, G., 2009. Error and efficiency of replica exchange molecular dynamics simulations. *J Chem Phys,* 131**,** 165102.

Roth, S., Neumansilberberg, F.S., Barcelo, G. & Schupbach, T., 1995. Cornichon and the Egf Receptor Signaling Process Are Necessary for Both Anterior-Posterior and Dorsal-Ventral Pattern-Formation in Drosophila. *Cell,* 81**,** 967-978.

Roy, A., Kucukural, A. & Zhang, Y., 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc,* 5**,** 725-38.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J., 1986. Learning representations by back-propagating errors. *Nature,* 323**,** 533-536.

Russ, W.P. & Engelman, D.M., 2000. The GxxxG motif: a framework for transmembrane helix-helix association. *J Mol Biol,* 296**,** 911-9.

Sachyani, D., Dvir, M., Strulovich, R., Tria, G., Tobelaim, W., Peretz, A., Pongs, O., Svergun, D., Attali, B. & Hirsch, J.A., 2014. Structural basis of a Kv7.1 potassium channel gating module: studies of the intracellular c-terminal domain in complex with calmodulin. *Structure,* 22**,** 1582-94.

Samudrala, R. & Moult, J., 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol,* 275**,** 895-916.

Savage, D.F. & Stroud, R.M., 2007. Structural basis of aquaporin inhibition by mercury. *J Mol Biol,* 368**,** 607-17.

Schmitt, N., Calloe, K., Nielsen, N.H., Buschmann, M., Speckmann, E.J., Schulze-Bahr, E. & Schwarz, M., 2007. The novel C-terminal KCNQ1 mutation M520R alters protein trafficking. *Biochem Biophys Res Commun,* 358**,** 304-10.

Schulze-Kremer, S., 2000. Genetic Algorithms and Protein Folding. *In* D.M. Webster (ed.) *Protein Structure Prediction: Methods and Protocols.* Totowa, NJ: Humana Press, 175-222.

Schwartz, P.J., Ackerman, M.J., George, A.L., Jr. & Wilde, A.A., 2013. Impact of genetics on the clinical management of channelopathies. *J Am Coll Cardiol,* 62**,** 169-80.

Schwartz, P.J., Stramba-Badiale, M., Crotti, L., Pedrazzini, M., Besana, A., Bosi, G., Gabbarini, F., Goulene, K., Insolia, R., Mannarino, S., Mosca, F., Nespoli, L., Rimini, A., Rosati, E.,

Salice, P. & Spazzolini, C., 2009. Prevalence of the congenital long-QT syndrome. *Circulation,* 120**,** 1761-7.

Schwenk, J., Harmel, N., Zolles, G., Bildl, W., Kulik, A., Heimrich, B., Chisaka, O., Jonas, P., Schulte, U., Fakler, B. & Klocker, N., 2009. Functional proteomics identify cornichon proteins as auxiliary subunits of AMPA receptors. *Science,* 323**,** 1313-9.

Senes, A., Engel, D.E. & Degrado, W.F., 2004. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr Opin Struct Biol,* 14**,** 465-79.

Senes, A., Gerstein, M. & Engelman, D.M., 2000. Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol,* 296**,** 921-36.

Shackelford, G. & Karplus, K., 2007. Contact prediction using mutual information and neural nets. *Proteins: Structure, Function, and Bioinformatics,* 69**,** 159-164.

Shamgar, L., Ma, L., Schmitt, N., Haitin, Y., Peretz, A., Wiener, R., Hirsch, J., Pongs, O. & Attali, B., 2006. Calmodulin is essential for cardiac IKS channel gating and assembly: impaired function in long-QT mutations. *Circ Res,* 98**,** 1055-63.

Shan, Y., Gnanasambandan, K., Ungureanu, D., Kim, E.T., Hammaren, H., Yamashita, K., Silvennoinen, O., Shaw, D.E. & Hubbard, S.R., 2014. Molecular basis for pseudokinase-dependent autoinhibition of JAK2 tyrosine kinase. *Nat Struct Mol Biol,* 21**,** 579-84.

Shan, Y., Kim, E.T., Eastwood, M.P., Dror, R.O., Seeliger, M.A. & Shaw, D.E., 2011. How does a drug molecule find its target binding site? *Journal of the American Chemical Society,* 133**,** 9181-9183.

Shanks, N.F., Cais, O., Maruo, T., Savas, J.N., Zaika, E.I., Azumaya, C.M., Yates, J.R., Greger, I. & Nakagawa, T., 2014. Molecular Dissection of the Interaction between the AMPA Receptor and Cornichon Homolog-3. *Journal of Neuroscience,* 34**,** 12104-12120.

Shaw, D.E., Deneroff, M.M., Dror, R.O., Kuskin, J.S., Larson, R.H., Salmon, J.K., Young, C., Batson, B., Bowers, K.J. & Chao, J.C., 2007. Anton, a special-purpose machine for molecular dynamics simulation. *ACM SIGARCH Computer Architecture News,* 35**,** 1-12.

Shaw, D.E., Grossman, J., Bank, J.A., Batson, B., Butts, J.A., Chao, J.C., Deneroff, M.M., Dror, R.O., Even, A. & Fenton, C.H., Year. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputered.^eds. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*IEEE Press, 41-53.

Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y. & Wriggers, W., 2010. Atomic-level characterization of the structural dynamics of proteins. *Science,* 330**,** 341-6.

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

Shortle, D., 2003. Propensities, probabilities, and the Boltzmann hypothesis. *Protein Sci,* 12**,** 1298-302.

Siebrands, C.C., Binder, S., Eckhoff, U., Schmitt, N. & Friederich, P., 2006. Long QT 1 mutation KCNQ1A344V increases local anesthetic sensitivity of the slowly activating delayed rectifier potassium current. *Anesthesiology,* 105**,** 511-20.

Simons, K.T., Kooperberg, C., Huang, E. & Baker, D., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol,* 268**,** 209-25.

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T., 2005. ROCR: visualizing classifier performance in R. *Bioinformatics,* 21**,** 3940-1.

Sinz, A., 2003. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *Journal of Mass Spectrometry,* 38**,** 1225-1237.

Sippl, M.J., 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol,* 213**,** 859-83.

Sippl, M.J., 1993. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des,* 7**,** 473-501.

Sippl, M.J., 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol,* 5**,** 229-35.

Sippl, M.J., 1996. Helmholtz free energy of peptide hydrogen bonds in proteins. *J Mol Biol,* 260**,** 644-8.

Sippl, M.J., Ortner, M., Jaritz, M., Lackner, P. & Flockner, H., 1996. Helmholtz free energies of atom pair interactions in proteins. *Fold Des,* 1**,** 289-98.

Sjogren, T., Nord, J., Ek, M., Johansson, P., Liu, G. & Geschwindner, S., 2013. Crystal structure of microsomal prostaglandin E2 synthase provides insight into diversity in the MAPEG superfamily. *Proc Natl Acad Sci U S A,* 110**,** 3806-11.

Skolnick, J., 2006. In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol,* 16**,** 166-71.

Skwark, M.J., Abdel-Rehim, A. & Elofsson, A., 2013. PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics,* 29**,** 1815-1816.

Skwark, M.J., Raimondi, D., Michel, M. & Elofsson, A., 2014. Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS Comput Biol,* 10**,** e1003889.

Smith, G.C., Seaman, S.R., Wood, A.M., Royston, P. & White, I.R., 2014. Correcting for optimistic prediction in small data sets. *Am J Epidemiol,* 180**,** 318-24.

Smith, J.A., Vanoye, C.G., George, A.L., Jr., Meiler, J. & Sanders, C.R., 2007. Structural models for the KCNQ1 voltage-gated potassium channel. *Biochemistry,* 46**,** 14141-52.

Sobolevsky, A.I., Rosconi, M.P. & Gouaux, E., 2009. X-ray structure, symmetry and mechanism of an AMPA-subtype glutamate receptor. *Nature,* 462**,** 745-U66.

Spatjens, R.L., Bebarova, M., Seyen, S.R., Lentink, V., Jongbloed, R.J., Arens, Y.H., Heijman, J. & Volders, P.G., 2014. Long-QT mutation p.K557E-Kv7.1: dominant-negative suppression of IKs, but preserved cAMP-dependent up-regulation. *Cardiovasc Res,* 104**,** 216-25.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15**,** 1929-1958.

Stead, L.F., Wood, I.C. & Westhead, D.R., 2011. KvSNP: accurately predicting the effect of genetic variants in voltage-gated potassium channels. *Bioinformatics,* 27**,** 2181-2186.

Steffensen, A.B., Refaat, M.M., David, J.P., Mujezinovic, A., Calloe, K., Wojciak, J., Nussbaum, R.L., Scheinman, M.M. & Schmitt, N., 2015. High incidence of functional ion-channel abnormalities in a consecutive Long QT cohort with novel missense genetic variants of unknown significance. *Sci Rep,* 5**,** 10009.

Stout, M., Bacardit, J., Hirst, J.D., Smith, R.E. & Krasnogor, N., 2008. Prediction of topological contacts in proteins using learning classifier systems. *Soft Computing,* 13**,** 245-258.

Sugita, Y. & Okamoto, Y., 1999. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters,* 314**,** 141-151.

Summa, C.M. & Levitt, M., 2007. Near-native structure refinement using in vacuo energy minimization. *Proc Natl Acad Sci U S A,* 104**,** 3177-82.

Summa, C.M., Levitt, M. & Degrado, W.F., 2005. An atomic environment potential for use in protein structure prediction. *J Mol Biol,* 352**,** 986-1001.

Suzek, B.E., Huang, H., Mcgarvey, P., Mazumder, R. & Wu, C.H., 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics,* 23**,** 1282-8.

Tai, C.H., Bai, H., Taylor, T.J. & Lee, B., 2014. Assessment of template-free modeling in CASP10 and ROLL. *Proteins,* 82 Suppl 2**,** 57-83.

Tanaka, S. & Scheraga, H.A., 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules,* 9**,** 945-950.

Taylor, K.C. & Sanders, C.R., 2016. Regulation of KCNQ/Kv7 family voltage-gated K+ channels by lipids. *Biochim Biophys Acta*.

Tchernitchko, D., Goossens, M. & Wajcman, H., 2004. In silico prediction of the deleterious effect of a mutation: Proceed with caution in clinical genetics. *Clinical Chemistry,* 50**,** 1974-1978.

Tegge, A.N., Wang, Z., Eickholt, J. & Cheng, J., 2009. NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Research,* 37**,** W515-W518.

Thomas, D., Khalil, M., Alter, M., Schweizer, P.A., Karle, C.A., Wimmer, A.B., Licka, M., Katus, H.A., Koenen, M., Ulmer, H.E. & Zehelein, J., 2010. Biophysical characterization of KCNQ1 P320 mutations linked to long QT syndrome 1. *J Mol Cell Cardiol,* 48**,** 230-7.

Thomas, P.D. & Dill, K.A., 1996. Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol,* 257**,** 457-69.

Tiefenbrunn, T., Liu, W., Chen, Y., Katritch, V., Stout, C.D., Fee, J.A. & Cherezov, V., 2011. High resolution structure of the ba3 cytochrome c oxidase from Thermus thermophilus in a lipidic environment. *PLoS One,* 6**,** e22348.

Tsallis, C. & Stariolo, D.A., 1996. Generalized simulated annealing. *Physica A,* 233**,** 395-406.

Uguzzoni, G., John Lovis, S., Oteri, F., Schug, A., Szurmant, H. & Weigt, M., 2017. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc Natl Acad Sci U S A,* 114**,** E2662-E2671.

Ulmschneider, M.B. & Sansom, M.S., 2001. Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta,* 1512**,** 1-14.

Unger, R., 2004. The genetic algorithm approach to protein structure prediction. *Applications of Evolutionary Computation in Chemistry,* 110**,** 153-175.

Unger, R. & Moult, J., 1993. Genetic algorithms for protein folding simulations. *J Mol Biol,* 231**,** 75-81.

Valdar, W.S., 2002. Scoring residue conservation. *Proteins,* 48**,** 227-41.

Valsson, O., Tiwary, P. & Parrinello, M., 2016. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu Rev Phys Chem,* 67**,** 159-84.

Van Gunsteren, W. & Berendsen, H., 1987. Groningen molecular simulation (GROMOS) library manual. *Biomos, Groningen,* 24**,** 13.

Van Gunsteren, W.F. & Berendsen, H.J.C., 1990. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angewandte Chemie International Edition in English,* 29**,** 992-1023.

Van Gunsteren, W.F., Daura, X. & Mark, A.E., 1998. GROMOS force field. *Encyclopedia of computational chemistry*.

Venters, R.A., Huang, C.-C., Farmer, B.T., Trolard, R., Spicer, L.D. & Fierke, C.A., 1995. High-level 2H/13C/15N labeling of proteins for NMR studies. *Journal of Biomolecular NMR,* 5**,** 339-344.

Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A., 2008. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics,* 24**,** 2928-2929.

Viklund, H. & Elofsson, A., 2008. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics,* 24**,** 1662-8.

Vitalis, A. & Pappu, R.V., 2009. Methods for Monte Carlo simulations of biomacromolecules. *Annu Rep Comput Chem,* 5**,** 49-76.

Voelz, V.A., Jager, M., Yao, S., Chen, Y., Zhu, L., Waldauer, S.A., Bowman, G.R., Friedrichs, M., Bakajin, O., Lapidus, L.J., Weiss, S. & Pande, V.S., 2012. Slow unfolded-state structuring in Acyl-CoA binding protein folding revealed by simulation and experiment. *J Am Chem Soc,* 134**,** 12565-77.

Wallner, B. & Elofsson, A., 2006. Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Science,* 15**,** 900-913.

Walters, R.F. & Degrado, W.F., 2006. Helix-packing motifs in membrane proteins. *Proc Natl Acad Sci U S A,* 103**,** 13658-63.

Wang, C., Xi, L., Li, S., Liu, H. & Yao, X., 2012. A sequence-based computational model for the prediction of the solvent accessible surface area for alpha-helix and beta-barrel transmembrane residues. *J Comput Chem,* 33**,** 11-7.

Wang, G. & Dunbrack, R.L., Jr., 2003. PISCES: a protein sequence culling server. *Bioinformatics,* 19**,** 1589-91.

Wang, R.Y.-R., Kudryashev, M., Li, X., Egelman, E.H., Basler, M., Cheng, Y., Baker, D. & Dimaio, F., 2015. De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nat Meth,* 12**,** 335-338.

Wang, W., Donini, O., Reyes, C.M. & Kollman, P.A., 2001. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu Rev Biophys Biomol Struct,* 30**,** 211-43.

Weber, J.K. & Pande, V.S., 2011. Characterization and rapid sampling of protein folding Markov state model topologies. *J Chem Theory Comput,* 7**,** 3405-3411.

Wedekind, H., Schwarz, M., Hauenschild, S., Djonlagic, H., Haverkamp, W., Breithardt, G., Wulfing, T., Pongs, O., Isbrandt, D. & Schulze-Bahr, E., 2004. Effective long-term control of cardiac events with beta-blockers in a family with a common LQT1 mutation. *Clin Genet,* 65**,** 233-41.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A. & Hwa, T., 2009. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 67-72.

Weiner, B.E., Alexander, N., Akin, L.R., Woetzel, N., Karakas, M. & Meiler, J., 2014. BCL::Fold--protein topology determination from limited NMR restraints. *Proteins,* 82**,** 587-95.

Weiner, B.E., Woetzel, N., Karakas, M., Alexander, N. & Meiler, J., 2013. BCL::MP-fold: folding membrane proteins through assembly of transmembrane helices. *Structure,* 21**,** 1107-17.

Weiner, P.K. & Kollman, P.A., 1981. Amber - Assisted Model-Building with Energy Refinement - a General Program for Modeling Molecules and Their Interactions. *Journal of Computational Chemistry,* 2**,** 287-303.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P., 1984. A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins. *Journal of the American Chemical Society,* 106**,** 765-784.

Westenskow, P., Splawski, I., Timothy, K.W., Keating, M.T. & Sanguinetti, M.C., 2004. Compound mutations: a common cause of severe long-QT syndrome. *Circulation,* 109**,** 1834-41.

Wetlaufer, D.B., 1973. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proceedings of the National Academy of Sciences of the United States of America,* 70**,** 697-701.

White, S.H., 2004. The progress of membrane protein structure determination. *Protein Sci,* 13**,** 1948-9.

Wiener, R., Haitin, Y., Shamgar, L., Fernández-Alonso, M.C., Martos, A., Chomsky-Hecht, O., Rivas, G., Attali, B. & Hirsch, J.A., 2008. The KCNQ1 (Kv7.1) COOH terminus, a multitiered scaffold for subunit assembly and protein interaction. *J Biol Chem,* 283**,** 5815-30.

Wodak, S.J., Protein Structure and Stability: Database-derived Potentials and Prediction. *Encyclopedia of Computational Chemistry*.

Woetzel, N., Karakas, M., Staritzbichler, R., Muller, R., Weiner, B.E. & Meiler, J., 2012. BCL::Score--knowledge based energy potentials for ranking protein models represented by idealized secondary structure elements. *PLoS One,* 7**,** e49242.

Woetzel, N., Lindert, S., Stewart, P.L. & Meiler, J., 2011. BCL::EM-Fit: rigid body fitting of atomic structures into density maps using geometric hashing and real space refinement. *J Struct Biol,* 175**,** 264-76.

Wollenberg, K.R. & Atchley, W.R., 2000. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of Sciences of the United States of America,* 97**,** 3288-3291.

Wolynes, P.G., 2015. Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie,* 119**,** 218-30.

Wu, J., Naiki, N., Ding, W.G., Ohno, S., Kato, K., Zang, W.J., Delisle, B.P., Matsuura, H. & Horie, M., 2014. A molecular mechanism for adrenergic-induced long QT syndrome. *J Am Coll Cardiol,* 63**,** 819-27.

Wu, S., Skolnick, J. & Zhang, Y., 2007. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol,* 5**,** 17.

Wu, S. & Zhang, Y., 2008. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics (Oxford, England),* 24**,** 924-931.

Wu, X.W., Hodoscek, M. & Brooks, B.R., 2012. Replica exchanging self-guided Langevin dynamics for efficient and accurate conformational sampling. *Journal of Chemical Physics,* 137.

Wu, Z.J., Huang, Y., Fu, Y.C., Zhao, X.J., Zhu, C., Zhang, Y., Xu, B., Zhu, Q.L. & Li, Y., 2015. Characterization of a Chinese KCNQ1 mutation (R259H) that shortens repolarization and causes short QT syndrome 2. *J Geriatr Cardiol,* 12**,** 394-401.

Xiangian Hu, H.H., David N. Beratan, Weitao Yang, 2010. A Gradient-Directed Monte Carlo Approach for Protein Design. *J. Comp. Chem.,* 31.

Xiong, Q., Cao, Q., Zhou, Q., Xie, J., Shen, Y., Wan, R., Yu, J., Yan, S., Marian, A.J. & Hong, K., 2015. Arrhythmogenic cardiomyopathy in a patient with a rare loss-of-function KCNQ1 mutation. *J Am Heart Assoc,* 4**,** e001526.

Xiong, Z.J., Du, P., Li, B., Xu, L.L., Zhen, X.C. & Fu, W., 2011. Discovery of a Novel 5-HT2A Inhibitor by Pharmacophore-based Virtual Screening. *Chemical Research in Chinese Universities,* 27**,** 655-660.

Xu, D. & Zhang, Y., 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins,* 80**,** 1715-35.

Xue, B., Faraggi, E. & Zhou, Y., 2009. Predicting residue-residue contact maps by a two-layer, integrated neural-network method. *Proteins,* 76**,** 176-183.

Yamaguchi, M., Shimizu, M., Ino, H., Terai, H., Hayashi, K., Kaneda, T., Mabuchi, H., Sumita, R., Oshima, T., Hoshi, N. & Higashida, H., 2005. Compound heterozygosity for mutations

Asp611-->Tyr in KCNQ1 and Asp609-->Gly in KCNH2 associated with severe long QT syndrome. *Clin Sci (Lond),* 108**,** 143-50.

Yamaguchi, M., Shimizu, M., Ino, H., Terai, H., Hayashi, K., Mabuchi, H., Hoshi, N. & Higashida, H., 2003. Clinical and electrophysiological characterization of a novel mutation (F193L) in the KCNQ1 gene associated with long QT syndrome. *Clin Sci (Lond),* 104**,** 377-82.

Yan, C., Wu, F., Jernigan, R.L., Dobbs, D. & Honavar, V., 2008. Characterization of protein-protein interfaces. *Protein J,* 27**,** 59-70.

Yan, R., Xu, D., Yang, J., Walker, S. & Zhang, Y., 2013. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Scientific Reports,* 3**,** 2619.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. & Zhang, Y., 2015. The I-TASSER Suite: protein structure and function prediction. *Nat Methods,* 12**,** 7-8.

Yang, J.S., Kim, J.H., Oh, S., Han, G., Lee, S. & Lee, J., 2012. STAP Refinement of the NMR database: a database of 2405 refined solution NMR structures. *Nucleic Acids Res,* 40**,** D525-30.

Yang, T., Chung, S.K., Zhang, W., Mullins, J.G., Mcculley, C.H., Crawford, J., Maccormick, J., Eddy, C.A., Shelling, A.N., French, J.K., Yang, P., Skinner, J.R., Roden, D.M. & Rees, M.I., 2009. Biophysical properties of 9 KCNQ1 mutations associated with long-QT syndrome. *Circ Arrhythm Electrophysiol,* 2**,** 417-26.

Yarov-Yarovoy, V., Schonbrun, J. & Baker, D., 2006. Multipass membrane protein structure prediction using Rosetta. *Proteins,* 62**,** 1010-25.

Young, M.M., Tang, N., Hempel, J.C., Oshiro, C.M., Taylor, E.W., Kuntz, I.D., Gibson, B.W. & Dollinger, G., 2000. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America,* 97**,** 5802-5806.

Yuan, Z., 2005. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics,* 6**,** 248.

Yuan, Z., Zhang, F., Davis, M.J., Boden, M. & Teasdale, R.D., 2006. Predicting the solvent accessibility of transmembrane residues from protein sequence. *J Proteome Res,* 5**,** 1063-70.

Zagrovic, B., Snow, C.D., Shirts, M.R. & Pande, V.S., 2002. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J Mol Biol,* 323**,** 927-37.

Zehelein, J., Thomas, D., Khalil, M., Wimmer, A.B., Koenen, M., Licka, M., Wu, K., Kiehn, J., Brockmeier, K., Kreye, V.A., Karle, C.A., Katus, H.A., Ulmer, H.E. & Schoels, W., 2004. Identification and characterisation of a novel KCNQ1 mutation in a family with Romano-Ward syndrome. *Biochim Biophys Acta,* 1690**,** 185-92.

Zhan, C., Li, B., Hu, L., Wei, X., Feng, L., Fu, W. & Lu, W., 2011. Micelle-based brain-targeted drug delivery enabled by a nicotine acetylcholine receptor ligand. *Angew Chem Int Ed Engl,* 50**,** 5482-5.

Zhang, C., Liu, S., Zhu, Q. & Zhou, Y., 2005. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J Med Chem,* 48**,** 2325-35.

Zhang, W., Yang, J., He, B., Walker, S.E., Zhang, H., Govindarajoo, B., Virtanen, J., Xue, Z., Shen, H.B. & Zhang, Y., 2016. Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11. *Proteins,* 84 Suppl 1**,** 76-86.

Zhang, X., Wang, T., Luo, H., Yang, J.Y., Deng, Y., Tang, J. & Yang, M.Q., 2010. 3D protein structure prediction with genetic tabu search algorithm. *BMC Syst Biol,* 4 Suppl 1**,** S6.

Zhang, Y., 2008. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol,* 18**,** 342-8.

Zhang, Y., 2009. I-TASSER: fully automated protein structure prediction in CASP8. *Proteins,* 77 Suppl 9**,** 100-13.

Zhang, Y., Kolinski, A. & Skolnick, J., 2003. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J,* 85**,** 1145-64.

Zhang, Y. & Skolnick, J., 2004. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophys J,* 87**,** 2647-55.

Zhou, H. & Zhou, Y., 2002a. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci,* 11**,** 2714-26.

Zhou, H.Y. & Zhou, Y.Q., 2002b. Folding rate prediction using total contact distance. *Biophysical Journal,* 82**,** 458-463.

Zviling, M., Kochva, U. & Arkin, I.T., 2007. How important are transmembrane helices of bitopic membrane proteins? *Biochim Biophys Acta,* 1768**,** 387-92.