A global machine learning based scoring function for protein structure prediction

Eshel Faraggi[*]

Department of Biochemistry and Molecular Biology,
Indiana University School of Medicine, Indianapolis,
Indiana 46202, USA ; Battelle Center for Mathematical Medicine,
Nationwide Children's Hospital, Columbus,
Ohio 43215, USA ; and Physics Division,
Research and Information Systems,
LLC, Carmel, Indiana, 46032, USA

Andrzej Kloczkowski[†]

Battelle Center for Mathematical Medicine,
Nationwide Children's Hospital, Columbus,
Ohio 43215, USA ; and Department of Pediatrics,
The Ohio State University, Columbus, Ohio 43215, USA

(Dated: October 3, 2013)

Abstract

We present a knowledge-based function to score protein decoys based on their similarity to native structure. A set of features is constructed to describe the structure and sequence of the entire protein chain. Furthermore, a qualitative relationship is established between the calculated features and the underlying electromagnetic interaction that dominates this scale. The features we use are associated with residue-residue distances, residue-solvent distances, pairwise knowledge-based potentials and a four-body potential. In addition we introduce a new target to be predicted, the fitness score, which measures the similarity of a model to the native structure. This new approach enables us to obtain information both from decoys and from native structures. It is also devoid of previous problems associated with knowledge-based potentials. These features were obtained for a large set of native and decoy structures and a back-propagating neural network was trained to predict the fitness score. Overall this new scoring potential proved to be superior to the knowledge-based scoring functions used as its inputs. In particular, in the latest CASP (CASP10) experiment our method was ranked third for all targets, and second for freely-modeled hard targets among about 200 groups for the top model predictions. Ours was the only method ranked in the top three for all targets and for hard targets. This shows that initial results from the novel approach are able to capture details that were missed by a broad spectrum of protein structure prediction approaches. Source codes and executable from this work are freely available at http://mathmed.org and http://mamiris.com/.

Keywords: Tertiary Protein Structure — Protein Knowledge Potentials — Protein Potential Energy — Protein Scoring Functions — Neural Network — Global Features

Short Title: Global protein scoring function

[*] Electronic address: efaraggi@gmail.com

[†] Correspondence to: Andrzej.Kloczkowski@nationwidechildrens.org

I. INTRODUCTION

Protein structure mediates the interactions that govern life. In general, protein tertiary structure prediction involves generating candidate structures (predicted models) and picking the closest to native among them. Various approaches have been designed to this selection problem with the ultimate goal of improving structure prediction. Since it is usually assumed that the native structure has the lowest energy scoring functions are commonly called potentials. They enable to distinguish between physically stable and unstable configurations of proteins and their partners and have numerous applications in medicine, biology, chemistry, and physics.

In general, a protein is a directed polypeptide chain of the 20 naturally occurring amino-acids. Critically, in-vivo it is immersed in a solution (water), or is influenced by other forms of boundary conditions. It is also well established that modification of these boundary conditions by changing the solution or its properties may involve a change in protein structure, dynamics and function. Hence, it is reasonable to assume that boundary effects will be critical in determining protein structure.

Four interactions are presently known to account for the observable universe. Out of these interactions three are out of range. The strong and weak nuclear forces deal with events on the subatomic scale and gravitation deals with events on, mostly, the extra-planetary scale. Thus, one interaction, electromagnetism, is left to account for much of the world as we know it, especially for the processes associated with biological systems. This electromagnetic interaction can be calculated in practice either by truncated multipole expansion or by numerical simulation of the discrete charge density. However, because of the amount of interactions in a given protein and the amounts of proteins to study, such fine detail calculations are very limited in scope. Instead a simplified representation of the system is utilized when designing a potential function. The aim of such functions is two fold: to determine the dynamics of proteins in molecular simulations, and to discriminate between decoys of native protein structures.

Two partly successful approaches to designing such a potential function have been reported. The so called physics-based potentials are based on some physical knowledge of the configuration of the system. Because of general structural features of the peptide chain, at the atomic level interactions are modeled as a force-field that contains additive contributions from spring-like potentials around equilibrium positions of bonds and angles, dihedral potentials, electrostatic and van der Waals potentials, and hydrogen bonds that effect propensity towards secondary structure formation [1]. Approaches that coarse grain the atomic model also exist, for example MARTINI [2, 3], UNRES [4–6], MSCG [7, 8], and others [9].

For protein systems such approaches can provide a description of protein dynamics on a wide range of time scales [1]. There are two problems though. Even at this level of simplification such simulations are prohibitive in terms of the computational time requirements. More importantly, while such simulations have been successful for various problems of protein dynamics, it has not been established that they can generally fold proteins starting from an arbitrary initial conformation, nor keep them in their naturally occurring states.

The other approach, termed knowledge based potentials, are generally based on some phenomenological learning using known biological structural and evolutionary data. That is, information existent in current databases is used to train a potential function. Sometimes an additional simplification is used. Instead of a fully atomic representation of amino acids a coarse-grained representation is used, where each residue is represented by one or more points. For example by the positions of the $C\alpha$ atom, the $C\beta$ atom, and the center of the side chain. In this work, we dress a knowledge-based scoring function with a little physics and use a machine learner to optimize it.

Critically important for any machine learning is the presentation of the data in a machine usable form. We shall discuss later how we do this using general features of electromagnetic potentials.

The central theme of protein knowledge based potentials has been to assume the existence of functions f(A,B) for interactions between residue types A and B such that the potential energy of a protein is the sum over all pairs. Furthermore, the Boltzmann distribution was assumed to describe the relation between the energy of interactions of pairs of residues A and B and probabilities of finding them being in contact in structural databases [10–12] and this was used to parameterize pairwise scoring functions. However, it has been already shown that the pairwise potentials approach was less than satisfactory, to describe proteins due to the dense packing of residues in protein cores [13–16]. The unsatisfactory nature of two-body parameterization of the energy potentials is also demonstrated by improvements due to addition of multi-body potentials [11, 17–20]. In simple terms, the two body approach is unsatisfactory because the abundance of a particular pair of residues at a given distance may be determined by other residues in the protein and does not necessarily reflect any physical interaction between that particular pair of residues. Knowledge based potentials are also sometimes criticized [16] because of sampling across many proteins without imposing a standard environment. We suggest that sampling across many different sequences (proteins) to derive potential energy is preferable over sampling over a single sequence. On the other hand the non-standard environment in protein structure determination is of serious debilitating nature. Nevertheless we have to use information that is currently available hoping that in the future a standardized approach or at least the reporting of the environment will be established. Despite these misgivings numerous successes of knowledge based potentials suggest that the effect of a non-constant environment does not impair their usefulness.

As suggested recently [16, 21] knowledge based potentials can be rendered physically meaningful under certain conditions. Critically it is the account of many-body effects that proves most beneficial [16]. Here we shall start from the basic ideas and progressively build upon them. The proper phrasing of the Boltzmann criterion here is that for a sample of protein states, the probability of any given configuration is asymptotically proportional to the relative amount of the configuration in the sampled states. And, that the energy difference between two states can be obtained by taking the logarithm of the ratio between the probabilities of being in these states. The key issue to realize is that these probabilities are dependent on a collection of interactions across the entire protein chain. Hence, knowledge-based potentials that take into account a bigger part of the protein will be better at capturing the physics of the problem. Methods that take into account the whole protein will be best. These global considerations are necessary to derive potential energies in a meaningful way.

II. MATERIALS AND METHODS

How do we model interactions on the entire protein scale? One possible way is to use global features. That is, use features that are calculated over the whole protein. For this we need to sum or integrate over many interactions. As argued at the outset, electromagnetism governs the phenomena on this scale. Thus, the terms appearing in such a sum would be proportional to inverse distances, their cubes and so forth with increasing odd integers describing higher moments. Furthermore, we can assume that the terms appearing in the sum would contain parameters dependent on the types of residue pairs and possibly their environment. These parameters can depend, for example, on the charge and electric dipole moment of the residues, etc. Hence, we can imagine that if we take as features such partial sums, partitioned for example by pair types, we can use a machine learner to approximate the unknown parameters, which are generally dependent on the physical characteristics of the system.

In Table I we give the protein descriptive features used as inputs. res2res refers to the partial energy sums between pairs of residues A and B. We distinguish between interaction AB and BA depending on their order in the sequence. For a given pair of types A and B, we consider interactions as $...A...B...$, with "..." referring to any or no point along the sequence. For a given protein we denote by N_{AB} the number of such interactions and let d^i_{AB}, $i = 1, .., N_{AB}$, denote the distances between the $C\alpha$ of these residue pairs. We collect the terms $\sum_i (d^i_{AB}))^{-1}$ and $\sum_i (d^i_{AB}))^{-3}$ normalized by the cube root of the length of the sequence of the protein. We also collect N_{AB}, normalized by the length. A final z-score normalization was applied to all features. This means that for a given feature (partial sum) we subtract the average over all proteins in our PDB dataset (discussed below) and divide by the standard deviation over that set. One extra residue type was used for all unknown types. In total there are $3 * 21 * 21 = 1323$ features for this feature type. A possible improvement would be to use the same approach but with atom types instead of residue types. Additionally, including the minimum distance between residue/atom types across the protein, defined as the distance between the closest two atoms for residues, can result in an improved assessment of the similarity of a given structure to the native one associated with that sequence.

So far we have considered residue-residue interactions, typ233w.norm contains information about the distance of individual atoms to the solvent. This depth is a critical piece of information since much of the nature of proteins is an outcome of interactions associated with the solvent. We distinguish between atoms associated with different residue types and calculate as before the partial sums with respect to the inverse distance and its cube. We use a total of 233 heavy atom-residue types with 234th reserved for unknown atoms. We have used a recursive program through PDB files to come up with this list of types, allowing new types to be defined on the fly when first encountered in a PDB file. The list of types accompanies this publication as the supplementary file ‘aa.atom.types.233’. In essence this allows us to avoid placing entries with an ‘unknown’ attribute. We use a stand alone version of the DEPTH program [22] to calculate the distance to the solvent. Distances are defined by the shortest paths to a water molecule in a solvated protein. In this case, for each atom-residue distance partial sum (feature) we also calculate the average standard deviation of the distances to the solvent appearing in the sum for that feature. All these quantities were normalized as before. A significant improvement to this feature would result from collecting distances into several bins and assigning separate sums in each individual bin to separate features.

In addition to these global input features we also use three global features that are publicly available and are easy to use. 4bod, dfire2, and rwplus refer to the four-body [23–25], DFIRE2 [10, 11], and RWPlus [12] potentials respectively. While 4bod is designed to capture some multi-body effects, DFIRE2 and RWPlus are two-body knowledge-based potentials. They are used here though parameterized only for entire proteins. Hence they can be viewed as capturing global information regarding the two-body decomposition of a protein. For these potentials we calculate the z-score of their raw values. We also normalize the raw values by the length of the protein and calculate that z-score and the normalized value scaled to be in the $[-1, 1]$ region. For the four-body potential we also use the number of atoms for normalization.

These are the global features we chose to describe the protein. We are now faced with determining the parameters associated with each of our features. We could, and it might be an interesting study, reduce the amount of our global features and try to ascertain some parameterization of them through statistical analysis of currently available protein structures. Unfortunately, using such global features would reduce the amount of information to parameterize the potential. Nevertheless initial testing of these ideas were performed first, by using simple additive non-weighted parameterization which gave a robust signal. This was followed by training the neural network on a subset of the PDB [26] with a sequence of more than 20 residues (29381 structures), assigning the target value of the protein potential to be -1 for representing the native structure. These first simple steps showed enough promise to continue this study further.

At this point it is instructive to consider what we would like to get. Our main aim is to develop a scoring function that for given protein sequence and given set of structural models of protein, will score them according to their similarity to the native structure associated with the given sequence. For this purpose it is just as important to parameterize the scoring function on non-native models. If we do this we obtain a new dimension which we can use to greatly expand our statistics. This is the innovative approach we have taken.

For a given non-native structure we calculate the structural similarity of it to the native structure corresponding to the same sequence. We use the TM-score [27, 28] to measure similarity between native and model structures (decoys). The TM-score values were renormalized to be fitness scores as $\text{FS} = 2(0.5-\text{TM})$, with TM denoting the TM-score and FS the fitness score. This conforms to our previous definition of -1 as the score of a perfect match to native structure. To obtain as realistic as possible set of protein models we used server models from the Critical Assessment of protein Structure Prediction (CASP) [29] rounds 5 through 9 (94717 structures) to further train and test our protein potential. Models from earlier CASPs were discarded because of differences in their cataloging. In addition several other freely available databases of protein decoys were collected. These include the ModPipe Decoys [30] (168,632 decoys for 6,877 native protein chain structures) and decoys from Decoys 'R' Us [31] (multiple decoys set). However they were not used in training because of the abundance of CASP 5 thorough CASP 9 models, and because of computational difficulties associated with abundance of data and shortage of dedicated resources. The set of CASP server models was used in combination with the set of native PDB structures mentioned earlier.

We are now faced with the key issue of the parameterization of the model. A common approach to parameterization is to multiply each feature by a parameter optimized according to some criterion. For example, in 2-body knowledge based protein potentials the distribution of residue pairs in the PDB is used to determine the values of such parameters. Indeed our initial studies were along those lines. After further thought, our approach evolved around the model system we just presented including the extra fitness score target. Our hypothesis was that it may be useful to use the architecture of a neural network to achieve the required parameterization. If we use the features we described as inputs, then multiplication by a parameter is the first input layer of the neural network. Model-to-native fitness is used for training the output layer of the network, and we are guaranteed that we can approximate any function by virtue of the mathematical nature of a neural network.

A two layer neural network with the input features described above plus an additional bias neuron was set up for this problem using the package GENN (GEneral Neural Network) [32].

The first hidden layer had 51 neurons and the second had 31. Each with an additional bias neuron. The output layer is a single neuron for the protein fitness to native score (FS). In this way the input is the protein sequence and model structure, as given for example in a PDB format, and the output is a measure for the similarity of the protein to a native structure, or, in the case of multi-structured proteins, to the first realization reported in the PDB. As mentioned previously once we include non-native models there is boundless statistics and certainly more than computational resources at our disposal can handle. We found that about thirty thousand structures is all we can do in one batch and we average over several training realizations with different initial conditions. For each we randomize the list of proteins and select 30% to compose the over-fit-protection set while the remaining 70% is used for on-line training. Over-fit testing is done after each training epoch. For testing purposes the newer CASP results are excluded as will be described below.

III. RESULTS

Performance of the methods described are analyzed in two ways using data from the last two CASP experiments. We note that our choice of CASP as a testing ground was due to it being a well-established and natural form of analysis of progress in development of protein structure prediction methodologies. Hence our testing procedure closely resembles a real scenario. For the first test (Test 1), models up to CASP7 were used for training and over-fit protection. These are all model up to and including CASP ID T0386, a total of 38254 models. The rest of available CASP models (CASP8 and CASP9 56463 models) are used as an independent test set. For the second test (Test 2), all models up to CASP8 (T0129-T0514, 65509 models) were used for the training and over-fit protection sets. In this case the 29208 models of CASP9 were used for testing. In our first round of developing the server native chain structures as given by CASP were also used in training. We term this server Seder1. We soon discovered that this may be a liability for the CASP experiment since the networks seemed to work better for native structures but were less successful in ranking non-native models. To overcome this liability a second round of training was performed with only predicted server models. In the second round the initial conditions of the neural network weights were taken as the optimized weights of the first round. We term this server Seder2.

In Tables II and III we give the results we got from these tests. Test 1 and 2 refer to previously described cases. All scores are calculated for each CASP target separately and then averaged over all targets in the test set. Here we use the following accuracy parameters. The correlation between fitness score and potential score is a measure of the accuracy for a range of structure realizations and should equal 1.0 for a perfect prediction according to our criteria. Another measure of accuracy is ability to pick structures most resembling the native structure when ranked according to the predicted protein potential. In this respect we report the average TM-score for the top 1 and the top 5 structures, and the average ranking position of the structure most closely resembling the native one. When ranking we distinguish between cases where a native structure is or is not included in the structures to be ranked. We also give the results using the values of the external scoring potentials for comparison. We see that there is very little difference between DFire2 and RWplus, except that RWplus seems to rank better. In our testing the 4body potential produced worse correlation but ranked higher the native structure. This did not translate however to a better TM-score for the top 1 or top 5 models. Overall Seder1 and Seder2 got better results than all external models across all accuracy parameters. Most significantly Seder improved the average TM-score of the top 1 structure by about 10% and significantly more if native structures are in the models set.

IV. DISCUSSION

We also analyze the prediction accuracy that can be derived from each feature type separately. We do this for the human targets (all groups) of the CASP10 dataset of submitted server predictions. This resembles the hardest part of a realistic test of protein structure prediction. All training on the weights presented here was done before the CASP10 experiment begun. To test the contribution from each input feature we turn it off (zero its values), predict, and test the accuracy using correlations and mean absolute error (MAE) in percent. In Table IV we give these accuracy parameters in each case removing the specified input feature. This means that the worse the results the more important is the contribution to the prediction from the removed feature. We see that by far the most significant input feature is the typ233w which describes the distance of the residues to the surface (or the solvent). Removing this feature results in about a 50% drop in correlation and an increase of almost 100% in MAE. The next input feature is res2res which describes the distance between residues. Removing this feature results in a decrease of about 10% in the correlations. In an interesting twist, removing res2res results in the lowest MAE. We note however that our final aim is to compare different structures and while the MAE accounts for the accuracy of individual predictions the correlations is better suited to evaluate the relationships between different predictions. This observation deserves further study. The three last features, which are various other protein scoring functions, contribute each a few percent to the accuracy. Overall we find that not including natives in the training set produces better results when considering predictions for models only (Seder2 versus Seder1).

We had an opportunity to test blindly the performance of the method presented here in the CASP10 [29] experiment as the "Kloczkowski_Lab" group. For each target, the set of top 150 server models was downloaded from the CASP server, run through Seder2 for a predicted fitness score and ranked accordingly. No human intervention was carried out. Top five models were submitted in order, but since our methodology was developed specifically for the recognition of the best model, our approach excelled in top model (Model 1) category. The official group performance ranking over all targets [33] show that the method presented above was ranked third, with only the Zhang group and Zhang server ranked higher. The Zhang group of the University of Michigan is currently the leading group in protein structure prediction having won several recent CASPs, hence it is encouraging that our initial approach did so well.

For many proteins in the CASP experiment significantly close homologs with solved structure can be found and building structural models for these cases is relatively easier. Those proteins with no structurally resolved sequentially similar proteins are generally known as hard targets. For those hard targets the Kloczkowski Lab ranked second in the official ranking [33]. The top group in this category was ProQ2 group, however ProQ2 ranked eighth for all models. Other state-of-the-art quality assessment methods and best human predictors with their official rankings in CASP10 for (all models/hard models) are: ProQ2 (8/1), ProQ2clust (30/39), Zhang (1/4), keasar (3/17), Pcomb (5/6), Pcons (6/10), BAKER (14/37), MULTICOM (4/11), Mufold (12/49). Under this convention our method ranked as (3/2) and we find it highly promising that our automated approach was the only one to be highly ranked for both hard and all targets and was successful even against top human prediction groups in template-based and free modeling categories.

V. CONCLUSIONS

We have presented a new approach for designing a function to discriminate between protein decoys based on their fitness to native structure. At its base this discrimination is based on a set of input features that are used to describe the structure and the sequence of the protein chain. For the current work these are associated with residue-solvent distances (residue depth), residue-residue distances, pairwise knowledge potentials and a four-body potential. We found that the distances between the residue delineated atoms of the protein and the solvent is the most important input feature. These features were calculated for a large set of both native and decoy structures and a back-propagating neural network was used to train for predicting the fitness score of a model to the native structure. Overall this new scoring potential proved to be superior to the knowledge based scoring functions used as its inputs and was also ranked among the top performing groups in the recent CASP10 experiment. Possible improvements of the scoring function presented here include refinement to atom types in residue-residue interactions and refinement of the interactions based on distance bins. Possible extensions of this work include a protein interface scoring function for selection of native interfaces of interacting proteins.

One of the interesting possible extensions of this work would be to incorporate other approaches of protein representations and of cataloging its internal interactions. This may lead to other physically meaningful partitioning and parameterization of the interactions and improvements in identification of native structures. In this respect, simplified representations of a protein [4–6, 34–43] pose significant promise. Furthermore, a simplification route can also be taken through re-concentration on the four or so atomic building blocks of a protein and parameterizing the interactions between them. Such an approach may be most effective if one does not disregarded the hydrogens, as is sometimes done, for they influence much of the properties of proteins [44, 45]. Use of evolutionary knowledge and sequence alignment information [46–48] will also likely contribute due to the strong relationship between sequence similarity and structure conservation. Probably some combination [49] of all these approaches will prove superior to any one of them individually for the identification of naturally occurring protein structures.

Acknowledgments

This work was funded in part by the financial support of the National Institutes of Health through Grants R01GM072014 and R01GM073095, and the National Science Foundation through Grant NSF MCB 1071785. Both authors would like to thank the organizers of CASP10 conference in Gaeta, Italy, for inviting them to the conference and providing the registration for EF. EF would also like to thank Keith Dunker for hosting him.

[1] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652, 2002.

[2] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H de Vries. The martini force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.

[3] Luca Monticelli, Senthil K Kandasamy, Xavier Periole, Ronald G Larson, D Peter Tieleman, and Siewert-Jan Marrink. The martini coarse-grained force field: extension to proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834, 2008.

[4] Adam Liwo, Piotr Arłukowicz, Cezary Czaplewski, Stanisław Ołdziej, Jarosław Pillardy, and Harold A Scheraga. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: application to the unres force field. *Proceedings of the National Academy of Sciences*, 99(4):1937–1942, 2002.

[5] Adam Liwo, Mey Khalili, Cezary Czaplewski, Sebastian Kalinowski, Stanislaw Oldziej, Katarzyna Wachucik, and Harold A Scheraga. Modification and optimization of the united-residue (unres) potential energy function for canonical simulations. i. temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *The Journal of Physical Chemistry B*, 111(1):260–285, 2007.

[6] Yi He, Yi Xiao, Adam Liwo, and Harold A Scheraga. Exploring the parameter space of the coarse-grained unres force field by random search: Selecting a transferable medium-resolution force field. *Journal of computational chemistry*, 30(13):2127–2135, 2009.

[7] Sergei Izvekov and Gregory A Voth. A multiscale coarse-graining method for biomolecular systems. *The Journal of Physical Chemistry B*, 109(7):2469–2473, 2005.

[8] WG Noid, Pu Liu, Yanting Wang, Jhih-Wei Chu, Gary S Ayton, Sergei Izvekov, Hans C Andersen, and Gregory A Voth. The multiscale coarse-graining method. ii. numerical implementation for coarse-grained molecular models. *The Journal of chemical physics*, 128(24):244115–244115, 2008.

[9] Nathalie Basdevant, Daniel Borgis, and Tap Ha-Duong. A coarse-grained protein-protein potential derived from an all-atom force field. *The Journal of Physical Chemistry B*, 111(31):9390–9399, 2007.

[10] H. Zhou and Y. Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*, 11:2714–2726, 2002.

[11] Y. Yang and Y. Zhou. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Science*, 17:1212–1219, 2008.

[12] J. Zhang and Y. Zhang. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS one*, 5(10):e15386, 2010.

[13] Paul D Thomas and Ken A Dill. Statistical potentials extracted from protein structures: how accurate are they? *Journal of molecular biology*, 257(2):457–469, 1996.

[14] A. Ben-Naim. Statistical potentials extracted from protein structures: Are these meaningful potentials? *The Journal of chemical physics*, 107(9):3698–3706, 1997.

[15] John Moult. Comparison of database potentials and molecular mechanics force fields. *Current opinion in structural biology*, 7(2):194–199, 1997.

[16] JW Mullinax and WG Noid. Recovering physical potentials from a model protein databank. *Proceedings of the National Academy of Sciences*, 107(46):19867–19872, 2010.

[17] Adam Liwo, R Kaźmierkiewicz, Cezary Czaplewski, Malgorzata Groth, S Ołdziej, Ryszard J Wawak, Shelly Rackovsky, Matthew R Pincus, and Harold A Scheraga. United-residue force field for off-lattice protein-structure simulations: Iii. origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *Journal of computational chemistry*, 19(3):259–276, 1998.

[18] Adam Liwo, Cezary Czaplewski, Jarosław Pillardy, and Harold A Scheraga. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic inter-

actions in the united-residue force field. *The Journal of chemical physics*, 115:2323, 2001.

[19] Gary S Ayton, Will G Noid, and Gregory A Voth. Multiscale modeling of biomolecular systems: in serial and in parallel. *Current opinion in structural biology*, 17(2):192–198, 2007.

[20] E.A.D. Amir, N. Kalisman, and C. Keasar. Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. *Proteins: Structure, Function, and Bioinformatics*, 72(1):62–73, 2008.

[21] Marcos R Betancourt. Another look at the conditions for the extraction of protein knowledge-based potentials. *Proteins: Structure, Function, and Bioinformatics*, 76(1):72–85, 2009.

[22] Kuan Pern Tan, Raghavan Varadarajan, and Mallur S Madhusudhan. Depth: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic acids research*, 39(suppl 2):W242–W248, 2011.

[23] Y. Feng, A. Kloczkowski, and R.L. Jernigan. Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins: Structure, Function, and Bioinformatics*, 68(1):57–66, 2007.

[24] Yaping Feng, Andrzej Kloczkowski, and Robert L Jernigan. Potentials' r'us web-server for protein energy estimations with coarse-grained knowledge-based potentials. *BMC bioinformatics*, 11(1):92, 2010.

[25] Pawel Gniewek, Sumudu P Leelananda, Andrzej Kolinski, Robert L Jernigan, and Andrzej Kloczkowski. Multibody coarse-grained potentials for native structure recognition and quality assessment of protein models. *Proteins: Structure, Function, and Bioinformatics*, 79(6):1923–1929, 2011.

[26] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[27] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

[28] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.

[29] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, and Anna Tramontano. Critical assessment of methods of protein structure prediction (casp) round ix. *Proteins: Structure,*

Function, and Bioinformatics, 79(S10):1–5, 2011.

[30] David Eramian, Narayanan Eswar, Min-Yi Shen, and Andrej Sali. How well can the accuracy of comparative protein structure models be predicted? *Protein Science*, 17(11):1881–1893, 2008.

[31] Ram Samudrala and Michael Levitt. Decoys 'r' us: A database of incorrect conformations to improve protein structure prediction. *Protein Science*, 9(7):1399–1401, 2000.

[32] Eshel Faraggi and Andrzej Kloczkowski. GENN: A GEneral Neural Network for learning tabulated data with examples from protein structure prediction. *Methods in Molecular Biology*, (Artificial Neural Networks: Methods and Applications), 2014.

[33] CASP10. Official group performance ranking, 2012. [Online; http://www.predictioncenter.org/casp10/groups_analysis.cgi (accessed 10-June-2012)].

[34] Michael Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of molecular biology*, 104(1):59–107, 1976.

[35] Burak Erman, Ivet Bahar, Andrzej Kloczkowski, and James E Mark. Lattice model for segmental orientation in deformed polymeric networks. 1. contribution of intermolecular correlations. *Macromolecules*, 23(26):5335–5340, 1990.

[36] DA Hinds and M Levitt. A lattice model for protein structure prediction at low resolution. *Proceedings of the National Academy of Sciences*, 89(7):2536–2540, 1992.

[37] Andrzej Kloczkowski, James E Mark, and Burak Erman. A diffused-constraint theory for the elasticity of amorphous polymer networks. 1. fundamentals and stress-strain isotherms in elongation. *Macromolecules*, 28(14):5089–5096, 1995.

[38] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

[39] Zerrin Bagci, Andrzej Kloczkowski, Robert L Jernigan, and Ivet Bahar. The origin and extent of coarse-grained regularities in protein internal packing. *Proteins: Structure, Function, and Bioinformatics*, 53(1):56–67, 2003.

[40] Andrzej Kolinski et al. Protein modeling and structure prediction with a reduced representation. *ACTA BIOCHIMICA POLONICA-ENGLISH EDITION-*, 51:349–372, 2004.

[41] Taner Z Sen, Yaping Feng, John V Garcia, Andrzej Kloczkowski, and Robert L Jernigan. The extent of cooperativity of protein motions observed with elastic network models is similar for

atomic and coarser-grained models. *Journal of chemical theory and computation*, 2(3):696–704, 2006.

[42] Robert L Jernigan and Andrzej Kloczkowski. Packing regularities in biological structures relate to their dynamics. In *Protein Folding Protocols*, pages 251–276. Springer, 2006.

[43] Peter Minary and Michael Levitt. Probing protein fold space with a simplified model. *Journal of molecular biology*, 375(4):920, 2008.

[44] Tanja Kortemme, Alexandre V Morozov, and David Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. *Journal of molecular biology*, 326(4):1239–1259, 2003.

[45] Alexandre V Morozov, Tanja Kortemme, Kiril Tsemekhman, and David Baker. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proceedings of the National Academy of Sciences of the United States of America*, 101(18):6946–6951, 2004.

[46] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[47] Joanna I Sułkowska, Faruck Morcos, Martin Weigt, Terence Hwa, and José N Onuchic. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, 109(26):10340–10345, 2012.

[48] Sanzo Miyazawa. Prediction of contact residue pairs based on co-substitution between sites in protein structures. *PloS one*, 8(1):e54252, 2013.

[49] Bin Qian, Angel R Ortiz, and David Baker. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15346–15351, 2004.

TABLE I: Types of protein features used

Name	Description	Number of features
res2res	Residue type inverse distance between pairs partial sums	3*21*21
typ233w	Atom type residue delineated inverse distance to solvent partial sums	4*234
4bod	Four body potential [23–25]	5
dfire2	DFire2.0 [10, 11]	3
rwplus	RWPlus [12]	3

TABLE II: Test 1: CASP8 and CASP9 as blind test set

	Test 1				
Parameter	DFire2	RWPlus	4Body	Seder^{w}	Seder2^{wo}
R1n^{a}	76(78)	57(73)	33(47)	29(51)	14(33)
R1no^{b}	60(58)	60(57)	66(55)	63(54)	56(57)
Correlation^{c}	0.54(0.22)	0.56(0.22)	0.44(0.22)	0.51(0.25)	0.83(0.16)
ATM1n^{d}	0.67(0.27)	0.69(0.28)	0.62(0.29)	0.74(0.28)	0.90(0.18)
ATM5n^{e}	0.63(0.26)	0.64(0.26)	0.61(0.27)	0.66(0.26)	0.73(0.23)
ATM1no^{f}	0.60(0.25)	0.60(0.25)	0.57(0.26)	0.60(0.25)	0.67(0.23)
ATM5no^{g}	0.61(0.25)	0.61(0.25)	0.59(0.26)	0.62(0.25)	0.67(0.22)

Training to CASP T0386 testing from CASP T0387. Standard deviation over the tested proteins is given in parenthesis. ^{a}Rank of top 1 with native in pool of candidates ^{b}Rank of top 1 without native in pool of candidates ^{c}Pearson correlation between fitness score and predicted values ^{d}Average TM-score for top 1 model with native in pool of candidates ^{e}Average TM-score for top 5 model with native in pool of candidates ^{f}Average TM-score for top 1 model without native in pool of candidates ^{g}Average TM-score for top 5 model without native in pool of candidates ^{w}Seder with natives in training set ^{wo}Seder without natives in training set

TABLE III: Test 2: only CASP9 as blind test set

	Test 2				
Parameter	DFire2	RWPlus	4Body	Seder[w]	Seder2[wo]
R1n[a]	91(79)	56(74)	28(37)	25(51)	22(48)
R1no[b]	57(57)	59(58)	60(51)	62(63)	51(51)
Correlation[c]	0.52(0.21)	0.52(0.21)	0.39(0.23)	0.57(0.25)	0.83(0.17)
ATM1n[d]	0.64(0.25)	0.68(0.28)	0.59(0.29)	0.73(0.29)	0.88(0.18)
ATM5n[e]	0.62(0.24)	0.63(0.25)	0.58(0.27)	0.64(0.26)	0.70(0.25)
ATM1no[f]	0.61(0.24)	0.61(0.25)	0.54(0.26)	0.58(0.25)	0.65(0.25)
ATM5no[g]	0.61(0.23)	0.61(0.24)	0.55(0.26)	0.60(0.25)	0.64(0.24)

Training to CASP T0514 testing from CASP T0515. Standard deviation over the tested proteins is given in parenthesis. [a]Rank of top 1 with native in pool of candidates [b]Rank of top 1 without native in pool of candidates [c]Pearson correlation between fitness score and predicted values [d]Average TM-score for top 1 model with native in pool of candidates [e]Average TM-score for top 5 model with native in pool of candidates [f]Average TM-score for top 1 model without native in pool of candidates [g]Average TM-score for top 5 model without native in pool of candidates [w]Seder with natives in training set [wo]Seder without natives in training set

TABLE IV: Prediction accuracy contribution from different inputs

	Seder1^{w}		Seder2^{wo}	
Feature Removed	Correlation^{a}	MAE^{b}	Correlation^{a}	MAE^{b}
typ233w	0.305	29.6	0.313	28.1
res2res	0.483	15.0	0.466	15.0
4bod	0.527	18.7	0.542	18.3
dfire2	0.524	18.1	0.540	17.7
rwplus	0.532	17.9	0.547	17.5
All Features	0.546	17.2	0.559	16.8

Effect of eliminating each of the input features on the accuracy of predicting the fitness score. Note: the worse the results without the feature the more important the feature is. ^{a}Pearson correlation between predicted and native fitness score ^{b}MAE between predicted and real fitness score. Reported as percent of the values range. ^{w}Seder with natives in training set ^{wo}Seder without natives in training set