Refinement of Pairwise Potentials via Logistic Regression to Score Protein-Protein Interactions

Kiyoto A. Tanemura, Jun Pei, and Kenneth M. Merz, Jr.*

Department of Chemistry, Michigan State University, 578 S. Shaw Lane, East Lansing, Michigan, 48824, United States

Email: merz@chemistry.msu.edu

Acknowledgments

The authors thank the high-performance computing center (HPCC) at Michigan State University for providing their com-putational resources.

Abstract

Protein-protein interactions (PPIs) are ubiquitous and functionally of great importance in biological systems. Hence, the accurate prediction of PPIs by protein-protein docking and scoring tools is highly desirable in order to characterize their structure and biological function. *Ab initio* docking protocols are divided into the sampling of docking poses to produce at least one near-native structure, then to evaluate the vast candidate structures by scoring. Concurrent development in both sampling and scoring is crucial for the deployment of protein-protein docking software. In the present work, we apply a machine learning model on pairwise potentials to refine the task of protein quaternary structure native structure detection among decoys. A decoy set was featurized using the Knowledge and Empirical Combined Scoring Algorithm 2 (KECSA2) pairwise potential. The highly unbalanced decoy set was then balanced using a comparison concept between native and decoy structures. The resultant comparison descriptors were used to train a logistic regression (LR) classifier. The LR model yielded the optimal performance for native detection among decoys compared to conventional scoring functions, while exhibiting lesser performance for the detection of low root mean square deviation (RMSD) decoy structures. Its deployment on an independent benchmark set confirms that the scoring function performs competitively relative to other scoring functions. Scripts used are available at: https://github.com/TanemuraKiyoto/PPI-native-detection-via-LR .

Keywords: protein-protein docking, docking refinement, scoring function, quaternary structure prediction, machine learning

INTRODUCTION

Protein-protein interactions (PPI) are present in the underlying mechanisms of virtually all biochemical processes. In terms of drug discovery, they present an alternative and important drug target to the traditional small binding pockets of enzyme active sites.[1,2] Experimental techniques for protein structural characterization such as X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy have made available high-resolution data for proteins, including those of protein-protein complexes.[3] Protein-protein docking prediction as a computational method complements experimental techniques where experimental approaches do not suffice. Development of protein-protein docking methods advances the understanding of the mechanisms of biologically important functions, as well as enabling the exploitation of the underlying PPI as a target for therapeutic agents.

Ab initio protein-protein docking is generally separated into two phases due to the complexity of the problem. These phases are the sampling of docking poses followed by their evaluation by means of a scoring function.[4] For the sampling phase, the protein subunits may be treated as rigid bodies, as is the case for the docking algorithms ZDOCK, FTDOCK, and GRAMM.[5–7] Backbone flexibility can be modeled during the docking phase through normal mode analysis with modest computational cost, as exemplified by the flexible docking algorithms ATTRACT and SwarmDock.[8,9] Soft surface or pseudoatomic representations are typically employed in both rigid-body or flexible docking protocols to smooth the potential energy surface and allow faster convergence to energy minima. Sidechain flexibility is commonly modelled in the refinement stage after sampling, as is the case in iATTRACT.[10]

Due to the copiousness of the predictions generated during the sampling phase, the scoring function must achieve high computational efficiency and must accurately assign low energy structures to low ranks in the scoring process. Scoring functions belong to broad categories of physics-based, knowledge-based, and machine learning (ML)-based. Physics-based scoring functions are widely used and include ZRANK, ATTRACT, FASTCONTACT, FireDock, GalaxyTongDock, HawkRank, HADDOCK and ClusPro[11–18] Energy terms commonly include van der Waals, electrostatic, and desolvation potentials. Knowledge-based methods instead tend to apply Boltzmann inversion to the

frequency of observed interatomic/interresidue distances to approximate relative energies of PPI docking predictions. This class of scoring function include InterEvScore, SPIDER, and dDFIRE.[19–22]

Compared to classical scoring functions, ML-based methods have the advantage that they do not require prior assumptions between the structural data and protein-protein complex stability. This enables integrated processing of input data, as was exemplified by the PPI scoring function ProQDock and iScore.[23,24] Meanwhile, ML-based models are frequently criticized for being a black-box alternative to well defined scoring functions. This highlights the need to evaluate not only the performance of ML models but also the trends and insights it deduces from the data.

The random forest (RF) refinement methodology for native detection among decoys has been applied by us to protein-folding and protein-ligand decoy detections.[25,26] In short, a dataset consisting of a native conformer and many decoy structures are featurized using conventional pairwise potentials such as the Assisted Model Building with Energy Refinement (AMBER) force field and Knowledge-Based and Empirical Combined Scoring Algorithm (KECSA2) pairwise potentials.[27,28] The extreme skew in representation of decoy and native structures of the dataset is mitigated by comparing the descriptors between the native and decoy structures. The balanced dataset is then suitable to train a RF model for binary classification. This methodology outperformed conventional programs for decoy detection in protein folding and protein-ligand systems. This novel methodology is further explored for scoring PPI prediction, using the logistic regression (LR) classifier which is considered a simpler ML model.

MATERIALS AND METHODS

Decoy Set

The Critical Assessment of Scoring Function (CASF) - PPI decoy set was employed in this study, which consists of 273 systems with 2000 decoys each.[29] The decoys were generated previously by rigid-body docking using the FTDock software.[6] The CASF-PPI decoy set was more suitable than other decoy sets to train the ML-based scoring function because it removed artifacts and complications due to the docking protocol by using subunits of the bound PPI crystal structure as inputs to rigid-body docking.

Therefore, each PPI system consisted of one high quality native structure and decoy sets consistent in all respect but for the binding mode.

Featurization of the PPI Complexes

Let any PPI complex be described as an n-body system and let all independent pairwise probabilities be known. The overall probability of the PPI complex is described as,

$$p_n = \prod_{i,j;\, i \neq j}^{n} c_{ij} p_{ij} \tag{1}$$

in which p_{ij} is the independent probability of particle pair i and j, and c_{ij} is its empirical scaling constant. As a PPI complex, the overall probability can be further be decomposed to bond, angle, torsion, and nonbonding interactions as follows,

$$p_{complex} = \left(\prod_{bond} c_{ij} p_{ij} \right) \left(\prod_{angle} c_{kl} p_{kl} \right) \left(\prod_{torsion} c_{mn} p_{mn} \right) \left(\prod_{nonbond} c_{pq} p_{pq} \right) \tag{2}$$

in which $c_{\alpha\beta}$ corresponds to the scaling constant and $p_{\alpha\beta}$ corresponds to the pairwise probability of bond (ij), angle (kl), torsion (mn), and nonbonding interaction (pq). The present work deals with decoy structures generated using rigid-body docking, thus bond, angle, and torsional probabilities are constant between a native structure and its decoys. The equation can then be rewritten as,

$$p_{complex} = C \times \left(\prod_{nonbond} c_{pq} p_{pq} \right) \tag{3}$$

for some constant C. Taking the natural logarithm yields,

$$\ln p_{complex} = \ln C + \sum_{nonbond} \ln c_{pq} p_{pq} \tag{4}$$

Pairwise nonbonding interaction probabilities were obtained using KECSA2 pairwise potentials.[28] The potential of nonbonding interactions between atoms A and B with distance r_i were described using the following Lennard-Jones type equation:

$$E_{AB}(r_i) = \varepsilon_1 \left(\frac{\sigma}{r_i}\right)^{\alpha} - \varepsilon_2 \left(\frac{\sigma}{r_i}\right)^{\beta} \quad (5)$$

in which parameters ε_1, ε_2, σ, α, and β were obtained from the KECSA2 database. Then the potentials were translated to relative probabilities by Boltzmann distribution. Constant C and scaling factors c_{pq} are cancelled upon generation of the comparison descriptors.

Generation of Comparison Descriptors

The skewed representation of native and decoy PPI complexes are balanced using a comparison method, previously applied to protein tertiary structure and protein-ligand decoy sets.[25,26] The native structure was assumed to be more stable than the decoy, thus subtracting the logarithmic probability of the decoy from the native structure would result in a more positive vector and *vice versa*. By performing this subtraction, a balanced comparison dataset is generated, consisting of 4000 descriptors per system. A target label of '0' was appended for descriptors generated by decoy minus native, and '1' was appended for comparison descriptors for the other direction. The comparison descriptors were used as a balanced dataset to train the LR model.

Training and Evaluation of the LR Classifier

The LR classifier model (sklearn.linear_model.LogisticRegression) was trained on various fractions of the shuffled dataset.[30] Default values were used for hyperparameters. Training and validation sets were standardized by zero mean and unit variance. Each model was subjected to five-fold cross validation to obtain training and validation accuracies. Several replicates of the operation were performed to ensure sufficient coverage of the decoy set that was partitioned into the training set. Test accuracies were computed from the models refit on the training/validation sets. Accuracy is defined as,

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

for the count of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions. Accuracy was amenable for evaluating the performance of the LR classifier because the representation of the target label was balanced.

Scoring and Ranking of Decoy Sets via the LR Classifier

The PPI complexes for each system were ranked as follows. Let $\boldsymbol{x}_i$ be a descriptor of structure i. Then comparison descriptors were generated by $\boldsymbol{x}_i - \boldsymbol{x}_j$ for all $j \neq i$, thereby generating 2000 comparison descriptors. These comparison descriptors were classified by the LR classifier as either '0' or '1'. Finally, these labels were summed to provide a score for $\boldsymbol{x}_i$, in which the greater value was predicted to be the more native-like structure. Once all structures were scored, structures were ranked by their score in descending order.

Evaluation of Ranks Assigned by the Scoring Function

Metrics used to evaluate the scoring functions include success rate (SR), modified success rate (Y), native ranking, first root mean square deviation (RMSD), first decoy RMSD, and Spearman correlation coefficient of ligand RMSD and rank. CAPRI criteria were used to define near-native structures.[31]

Performances of the LR scoring function were compared to the scoring functions: ATTRACT, dDFIRE, FASTCONTACT, and ZRANK.[11–13,21] The scores for each scoring function were previously calculated and included with the CASF-PPI decoy set.[29]

Success Rate

SR is the probability of finding a near native structure in the top N predictions. Let $X = \{X_1, X_2, \cdots, X_n\}$ be the set of all PPI systems in the dataset. Let $h(X_i, N)$ be the number of near native structures in the top N predictions of system X_i. Then we write the success rate as,

$$SR(N) = \frac{\sum_{i=1}^{n}(h(X_i, N) > 0)}{n} \tag{7}$$

Modified Success Rate

Y takes into account the fraction of near-native structures identified in the top N predictions.[32] Additionally, it assigns a higher score to the lower ranking of near-native structure. It is defined as,

$$Y(N) = \frac{\sum_{i=1}^{n} F(X_i, N)}{n}; F(X_i, N) = \frac{\sum_{j=1}^{N}(1 + top_j)}{h(X_i, |X_i|)};$$

$$top_j = rank_j^{-1} \tag{8}$$

Spearman Correlation Coefficient of ligand RMSD and rank

The ideal funnel shape between ligand RMSD and rank of the scoring function are quantified using the Spearman correlation coefficient (ρ). A more positive ρ is ideal, as it suggests the lower rank is associated with lower ligand RMSD and *vice versa*.

Fold Enrichment of Near-Native Predictions in First Ranked Structures

We measure the representation of various qualities of predictions in the first ranked structures using the mean of the quotient of observed near-native prediction divided by its expectation E, and refer to it as fold enrichment FE. The expectation is the probability of arbitrarily choosing a near-native structure by sampling exactly one structure from all predictions of a given PPI system. We define the fold enrichment as,

$$FE(X) = \frac{\sum_{i=1}^{n} \frac{h(X_i, 1)}{E(X_i)}}{n}; E(X_i) = \frac{h(X_i, |X_i|)}{|X_i|} \tag{9}$$

Analysis and Selection of Feature Coefficients

Median values of coefficients assigned to each feature were obtained from LR classifiers trained on 0.99 of the data. The features consist of pairwise nonbonding interactions of heavy atoms of residues. The residues were assigned to the following subjective categories: anionic, cationic, polar, nonpolar, aromatic, flexible, and small (Table S1). The interactions of the residues were assigned by their categories (e.g. cationic-anionic). Distributions of the types of interactions were assessed.

Features with median coefficients of the greater magnitude were considered more salient. Subsets of features were taken as fractions of top salient features, which were used to train LR

classifiers. The performance of the LR classifiers with reduced dimensions were assessed via the aforementioned metrics.

Flexible Docking of Weng Benchmark 5.0 by ATTRACT

An independent decoy set was generated as a benchmark to assess the performance of various scoring functions. Unbound subunits of the Weng Benchmark 5.0 (BM5) were subjected to ATTRACT flexible docking with the iATTRACT interface refinement.[8,10,33] Bash script for performing docking was obtained from the ATTRACT web interface.[34] 1000 structures were generated using five normal modes. The complexes with the following PDB IDs were not further considered because they required repair of missing atoms in the input files to perform docking: 1F51, 1F6M, 1FC2, 1FCC, 1NSN, 1QFW, 1RLB, 1SYX, 2CFH, 3R9A, 4FZA, 4GAM. The complex, 1N2C, was omitted due to complications with the size of the input files. Because the weights of terms in the ZRANK scoring function were fitted on structures in the Weng Benchmark 1.0, structures from Benchmark 1.0 were removed for fair comparisons between scoring functions.[11,35] Out of 186 systems subjected to protein-protein docking, 135 structures contained at least one acceptable structure.

RESULTS AND DISCUSSION

LR Classifier Achieved High Accuracy on CASF-PPI Test Set

A learning curve was plotted for the LR classifier trained on various fractions of the decoy set (Figure 1). As the fraction of data used as the training set was increased to 0.99, the validation and test score approached an accuracy of 0.99. The small differences between training accuracy and validation or test accuracy illustrate the proficient model performance generalized to the remainder of the CASF-PPI decoy set. The narrow range in each accuracy indicated the model performance was stable. Near optimal performance of the LR classifier on the decoy set was observed with a training set fraction of 0.7 or greater.

Analysis and Selection of Salient Features

The input data was standardized to zero mean and unit variance. Thus, the magnitude and sign of the coefficients provide a measure for salient features in detecting native structures. The coefficients plotted in decreasing order displayed a **logit** shape, with the magnitudes rapidly decreasing toward the center of the distribution (Figure 2). This suggests there was a relatively small subset of highly salient features, while the majority of features contributed moderately to the classification task. The range of coefficients for each feature was narrow, suggesting the similarity in coefficients between models.

To generalize the types of interactions contributing to the classification, the interacting residues were categorized to broad classes and the type of interacting residues were recorded. The density plot displayed the representation of a given type of interaction over the features ordered by their coefficient values (Figure 2). Qualitatively, interactions between charged residues displayed the highest representations at the ends of the distribution, representing coefficients with greater magnitude and corresponding with features with greater effect on the classification. This is consistent with ionic interactions as the strong, specific, and dynamic nonbonding interaction characteristic to PPIs.[36] As expected, opposing charges were favored by having a greater density on the side of positive coefficients, while like charges had high representation for negative coefficients.

Aromatic-small residue interactions were also notable on the positive end of coefficients. The shape complementarity achieved by hydrophobic residues of opposing sizes has been attributed as a key factor affecting PPIs.[37] The relatively high representation may signify the importance of shape complementarity of hydrophobic residues between interacting proteins. The distribution is contrasted from the fairly even distribution of nonpolar-nonpolar interaction, further emphasizing the importance of shape complementarity.

Furthermore, the present LR classifiers represented polar-polar residue interaction with negative coefficients. A greater change in potential energy upon the exclusion of water molecules from PPI interface upon binding would be expected if favorable interaction between solvent and polar residue were absent. The higher density of polar-polar residue with negative coefficients suggests polar residues

are underrepresented in native PPI interfaces. Thus, the coefficients may implicitly suggest that the desolvation energy to be a contributor to the detection of native PPI complexes.

In summary, the LR classifier appeared to prioritize charge and geometric complementarity while disfavoring polar-polar interaction. A more complete plot is available in the SI (Figure S1) Dimensionality reduction was pursued by training logistic regression classifiers with features associated with coefficients of greater magnitude. Specifically, the features were ordered by the magnitude of the median coefficient, then various fractions of the salient features were used as input data. Performance of models trained on 0.9 of the data set are reported (Table I). Test accuracy and native ranks appeared to peak when around 0.1 of features were selected. Meanwhile, first decoy RMSD was constant between the fractions. The improvement in performance may be due to reducing noise arising from superfluous features. Refer to SI (Figures S2 - S4) for the full performance metrics.

The performance of the LR scoring function trained on the 0.1 top features was selected for further investigation. By use of a simple ML model coupled with the use of only 0.1 of the most salient features, we arrive to a scoring function which ranks thousands of structures in the order of minutes. The performance of the new scoring function was benchmarked.

The LR Scoring Function Was Sensitive towards Native Structures while less responsive to Near-Native Structures

The performance of the LR scoring function trained on the 0.1 top features was compared to those of conventional scoring functions (Figure 3). The SR of the LR scoring function was higher than other scoring functions. There were little improvements in SR as the threshold quality was relaxed down to acceptable predictions. Unlike other scoring functions, the LR scoring function displayed an early saturation of SR at about $N = 10$. A plateau in Y accompanies this trend, in which the LR scoring function displayed little improvement above $N = 10$ for thresholds *native* and *high*.

The sensitivity of the LR classifier to native structures was further exemplified in the distribution of native ranks between the various scoring functions (Table II). The LR scoring function yielded the

lowest mean native ranks compared to the other scoring functions. The RMSD of low-ranking structures illustrated a different trend. The mean of the first RMSD for the LR scoring function was low due to the superior native ranking compared to other scoring functions. Yet the insensitivity of the LR scoring function to near native decoy structures was apparent in the distribution of the first decoy RMSD, in which the LR scoring function displayed the greatest mean compared to other scoring functions. ZRANK notably exhibited the greatest performance for assigning near native structures to low ranks. Plots of the distributions are available in the SI (Figure S5).

A similar comparison was present in the Spearman correlation coefficient between ligand RMSD and rank (Figure 4). If only structures with ligand RMSD up to 5 Å were considered, all scoring functions displayed a ρ near 0.5. While other scoring functions preserved a distribution centered at a positive value when structures up to 10 Å were considered, ρ for ATTRACT and the LR scoring functions returned to a distribution centered at 0.0. All distributions were centered at 0.0 if structures up to 20 Å were considered. The relatively early erosion of correlation for ATTRACT and LR scoring functions emphasize they are less responsive to near-native decoy structures than other scoring functions. In the context of docking protocols which may generate near native predictions but not necessarily close matches to the native structure, the sensitivity to near native structures exhibited particularly by ZRANK may be more desirable as the accompanying scoring function for rigid-body docking. The superior performance of the LR scoring function on detecting native structure suggested its utility will lie in flexible docking predictions, in which the structures are predisposed to be of higher quality than its early stage rigid body counterparts. This led to the independent assessment of the LR scoring function on decoys generated by the ATTRACT flexible docking protocol with the iATTRACT interface optimization.[8,10]

The LR Scoring Function Performed Competitively on ATTRACT PPI Docking Predictions

We confirmed the performance of the LR scoring function generalizes on the CASF-PPI dataset due to its performance on the validation and test sets partitioned from the CASF-PPI dataset. The

performance was further assessed on ATTRACT flexible docking predictions of PPI complexes in the Weng Benchmark 5.0.[33] The independent dataset served as assessment for whether the performance of the LR model was inflated by correlation in docking protocol between the training and test sets, or biases arising by possible artifacts in docking which were not present in the native structure.

The success rate and modified success rate were determined for predictions ranked by the LR scoring function, ATTRACT, and ZRANK (Figure 5). ATTRACT and ZRANK were selected for comparison because they performed competitively on the CASF-PPI dataset. While the performance for acceptable and medium quality predictions are consistent between the scoring functions assessed, the LR scoring function displayed a greater success rate and modified success rate for high quality predictions. Due to the scarcity of high-quality predictions, the evidence is anecdotal and qualitative. Nevertheless, the LR scoring function appears competitive to other scoring functions on realistic docking predictions.

The LR scoring function is largely rewarded by ranking a greater fraction of high-quality structures in $N = 1$. We sought to compare the quality of predictions ranked first by each scoring functions. We use the observed quality divided by the probability of arbitrarily choosing a prediction with the quality. The mean of these quotients provides a measure of the representation of each quality of prediction in $N = 1$ compared to expectation of a random ordering, and refer to it as fold enrichment.

The fold enrichment is summarized (Table III). Generally, there is an increase in representation as the quality is improved, suggesting the sensitivity of the ZRANK and LR scoring functions to higher quality predictions. The LR scoring function showed a high representation of high & medium-quality predictions in $N = 1$, while ZRANK was advantageous for representing acceptable quality predictions as top structures. The greater fold enrichment of high & medium-quality predictions by the LR scoring function illustrates its greater sensitivity towards high quality predictions when compared to other scoring functions.

CONCLUSIONS

The LR scoring function or PPI prediction illustrates the RF refinement of pairwise potentials extends to protein quaternary structure prediction and performs competitively to conventional scoring functions for the task of native detection among PPI decoys. The salient features were consistent with terms present in various physics-based scoring functions. The utility of the scoring function was highlighted on ATTRACT flexible docking predictions, in which the representation of high-quality structures was greater at the top ranked predictions compared to other scoring functions.

References

1. Laraia L, McKenzie G, Spring DR, Venkitaraman AR, Huggins DJ. Overcoming Chemical, Biological, and Computational Challenges in the Development of Inhibitors Targeting Protein-Protein Interactions. *Chem Biol*. 2015;22(6):689-703. doi:10.1016/j.chembiol.2015.04.019
2. Bakail M, Ochsenbein F. Targeting protein-protein interactions, a wide open field for drug design. *Comptes Rendus Chim*. 2016;19(1-2):19-27. doi:10.1016/j.crci.2015.12.004
3. Dobson CM. Biophysical Techniques in Structural Biology. *Annu Rev Biochem*. 2019;88(1):25-33. doi:10.1146/annurev-biochem-013118-111947
4. Vajda S, Hall DR, Kozakov D. Sampling and scoring: A marriage made in heaven. *Proteins Struct Funct Bioinforma*. 2013;81(11):1874-1884. doi:10.1002/prot.24343
5. Pierce BG, Hourai Y, Weng Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS One*. 2011;6(9):0-5. doi:10.1371/journal.pone.0024657
6. Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol*. 1997;272(1):106-120. doi:10.1006/jmbi.1997.1203
7. Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res*. 2006;34(WEB. SERV. ISS.):310-314. doi:10.1093/nar/gkl206
8. De Vries S, Zacharias M. Flexible docking and refinement with a coarse-grained protein model using ATTRACT. *Proteins Struct Funct Bioinforma*. 2013;81(12):2167-2174. doi:10.1002/prot.24400
9. Moal IH, Bates PA. SwarmDock and the use of normal modes in protein-protein Docking. *Int J Mol Sci*. 2010;11(10):3623-3648. doi:10.3390/ijms11103623
10. Schindler CEM, de Vries SJ, Zacharias M. iATTRACT: Simultaneous global and local interface optimization for protein-protein docking refinement. *Proteins Struct Funct Bioinforma*. 2015;83(2):248-258. doi:10.1002/prot.24728

11. Pierce B, Weng Z. ZRANK: Reranking Protein Docking Predictions With an Optimized Energy Function. *Proteins Struct Funct Bioinforma*. 2007;67:1078-1086. doi:10.1002/prot

12. Fiorucci S, Zacharias M. Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. *Proteins Struct Funct Bioinforma*. 2010;78(15):3131-3139. doi:10.1002/prot.22808

13. Camacho CJ, Zhang C. FastContact: Rapid estimate of contact and binding free energies. *Bioinformatics*. 2005;21(10):2534-2536. doi:10.1093/bioinformatics/bti322

14. Andrusier N, Nussinov R, Haim J. Wolfson. FireDock: Fast interaction refinement in molecular docking. *Proteins Struct Funct Bioinforma*. 2007;69:139-159. doi:10.1002/prot.21495

15. Park T, Baek M, Lee H, Seok C. GalaxyTongDock: Symmetric and asymmetric ab initio protein–protein docking web server with improved energy parameters. *J Comput Chem*. 2019;40(27):2413-2417. doi:10.1002/jcc.25874

16. Feng T, Chen F, Kang Y, et al. HawkRank: A new scoring function for protein-protein docking based on weighted energy terms. *J Cheminform*. 2017;9(1):1-15. doi:10.1186/s13321-017-0254-7

17. Vries SJ de, Dijk ADJ van, Krzeminski M, et al. HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets Sjoerd. *Proteins Struct Funct Bioinforma*. 2007;69:726-733. doi:10.1002/prot

18. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: An automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*. 2004;20(1):45-50. doi:10.1093/bioinformatics/btg371

19. Andreani J, Faure G, Guerois R. InterEvScore: A novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics*. 2013;29(14):1742-1749. doi:10.1093/bioinformatics/btt260

20. Khashan R, Zheng W, Tropsha A. Scoring protein interaction decoys using exposed residues (SPIDER): A novel multibody interaction scoring function based on frequent geometric patterns

of interfacial residues. *Proteins Struct Funct Bioinforma*. 2012;80(9):2207-2217. doi:10.1002/prot.24110

21. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*. 2009;11(11):2714-2726. doi:10.1110/ps.0217002
22. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins Struct Funct Genet*. 2008;72(2):793-803. doi:10.1002/prot.21968
23. Basu S, Wallner B. Finding correct protein-protein docking models using ProQDock. *Bioinformatics*. 2016;32(12):i262-i270. doi:10.1093/bioinformatics/btw257
24. Geng C, Jung Y, Renaud N, Honavar V, Bonvin AMJJ, Xue LC. iScore: A novel graph kernel-based function for scoring protein-protein docking models. *Bioinformatics*. 2019;36(June 2019):112-121. doi:10.1093/bioinformatics/btz496
25. Pei J, Zheng Z, Merz KM. Random Forest Refinement of the KECSA2 Knowledge-Based Scoring Function for Protein Decoy Detection. *J Chem Inf Model*. 2019;59(5):1919-1929. doi:10.1021/acs.jcim.8b00734
26. Pei J, Zheng Z, Kim H, et al. Random Forest Refinement of Pairwise Potentials for Protein-Ligand Decoy Detection. *J Chem Inf Model*. 2019;59(7):3305-3315. doi:10.1021/acs.jcim.9b00356
27. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. 20035_Ftp. *J Comput Chem*. 2004;56531(9):1157-1174.
28. Zheng Z, Merz KM. Development of the knowledge-based and empirical combined scoring algorithm (KECSA) to score protein-ligand interactions. *J Chem Inf Model*. 2013;53(5):1073-1083. doi:10.1021/ci300619x
29. Han L, Yang Q, Liu Z, Li Y, Wang R. Development of a new benchmark for assessing the scoring functions applicable to protein-protein interactions. *Future Med Chem*. 2018;10(13):1555-1574. doi:10.4155/fmc-2017-0261

30. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn. *GetMobile Mob Comput Commun*. 2015;19(1):29-33. doi:10.1145/2786984.2786995

31. Méndez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins Struct Funct Genet*. 2003;52(1):51-67. doi:10.1002/prot.10393

32. Zhang Q, Feng T, Xu L, et al. Recent Advances in Protein-Protein Docking. *Curr Drug Targets*. 2016;17(14):1586-1594.

33. Vreven T, Moal IH, Vangone A, et al. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol*. 2015;427(19):3031-3041. doi:10.1016/j.jmb.2015.07.016

34. De Vries SJ, Schindler CEM, Chauvot De Beauchêne I, Zacharias M. A web interface for easy flexible protein-protein docking with ATTRACT. *Biophys J*. 2015;108(3):462-465. doi:10.1016/j.bpj.2014.12.015

35. Chen R, Mintseris J, Weng Z. A Protein – Protein Docking Benchmark. 2003;91(November 2002):88-91.

36. Kumar S, Nussinov R. Close-Range Electrostatic Interactions in Proteins. *ChemBioChem*. 2002;3:604-617.

37. Li Y, Zhang X, Cao D. The role of shape complementarity in the protein-protein interactions. *Sci Rep*. 2013;3:3-9. doi:10.1038/srep03271

Figure/Table Legends

Table I. Performance metrics of the LR classifier trained on various fractions of top salient features. For each quantity, the mean, 25th percentile (lower) and 75th percentile (upper) are reported. Four digits are reported all values except for lower/upper quantiles of native rank, which are natural numbers.

Table II. Performance metrics of the LR classifier trained on 0.1 top salient features in comparison to other scoring functions (SF). For each quantity, the mean, 25th percentile (lower) and 75th percentile (upper) are reported. Four digits are reported all values except for lower/upper quantiles of native rank, which are natural numbers.

Table III. Fold enrichment in top scoring structures are reported for various scoring functions (SF) and quality of prediction. All values were rounded to three significant figures. Sample sizes by quality: $n_{high} = 6, n_{medium} = 63, n_{acceptable} = 134, n_{incorrect} = 186$.

Figure 1. The distributions of training (green), validation (blue), and test (red) accuracies of LR classifiers trained on various fractions of the decoy set. The baseline is the accuracy achieved by the linear, unweighted sum of KECSA2 potentials with no LR refinement. Training and validation accuracies were obtained as the mean of the five-fold cross validation results. Test accuracy was calculated from LR classifiers refit on the training and validation set.

Figure 2. (top) Median coefficients for each feature in descending order. The curve displayed a logit shape, with the magnitude of the coefficients rapidly decreasing toward the center of the distribution. Maximum and minimum for each coefficient were plotted as a gray ribbon. (bottom) Density of various types of interactions applied to the coefficients ordered in decreasing order. The coordinates for coefficient values 0.1, 0.0, and -0.1 were indicated.

Figure 3. Comparison of scoring functions by success rate (top) and modified success rate (bottom) for various threshold of near-native structures.

Figure 4. Comparison of scoring functions via Spearman correlation coefficient between ligand RMSD and rank assigned by scoring function under various maximum RMSD values.

Figure 5. Comparison of scoring functions via success rate (top) and modified success rate (bottom) using various thresholds for near-native structures.

Tables

Table I

fraction	test accuracy			native rank			First decoy ligand RMSD (Å)		
	mean	lower	upper	mean	lower	upper	mean	lower	upper
1.00	0.9969	0.9968	0.9984	7.216	1	3	22.33	12.63	30.88
0.50	0.9993	0.9990	0.9997	2.344	1	1	23.68	13.03	31.92
0.40	0.9990	0.9992	0.9999	3.020	1	1	23.96	13.11	32.46
0.30	0.9994	0.9995	0.9999	2.285	1	1	23.96	13.80	31.92
0.20	0.9995	0.9997	0.9999	2.044	1	1	22.63	13.10	31.08
0.10	0.9997	0.9996	0.9998	1.665	1	1	22.34	12.91	31.10
0.05	0.9992	0.9990	0.9997	2.585	1	2	23.03	13.23	32.99

Table II

SF	native rank			First ligand RMSD (Å)			First decoy ligand RMSD (Å)		
	mean	lower	upper	mean	lower	upper	mean	lower	upper
LR	1.665	1	1	5.343	0.000	0.000	22.34	12.91	31.10
ATTRACT	16.70	1	3	7.256	0.000	12.98	19.90	9.952	28.89
dDFIRE	68.06	1	109	12.11	0.000	24.07	15.58	1.496	27.85
FASTCONTACT	129.2	5	116	15.34	2.064	26.43	15.90	2.744	26.58
ZRANK	13.52	1	3	6.047	0.000	3.336	9.667	0.982	15.75

Table III

SF	quality of prediction			
	high	medium	acceptable	incorrect
LR	292	73.0	14.2	0.929
ATTRACT	20.8	21.3	9.39	0.951
ZRANK	125	41.3	22.3	0.902

Figures

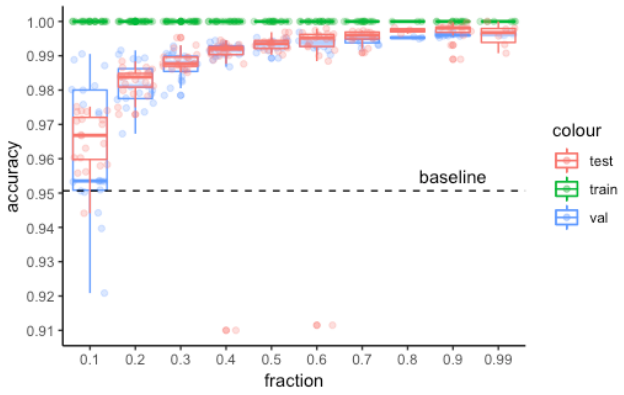

Figure 1

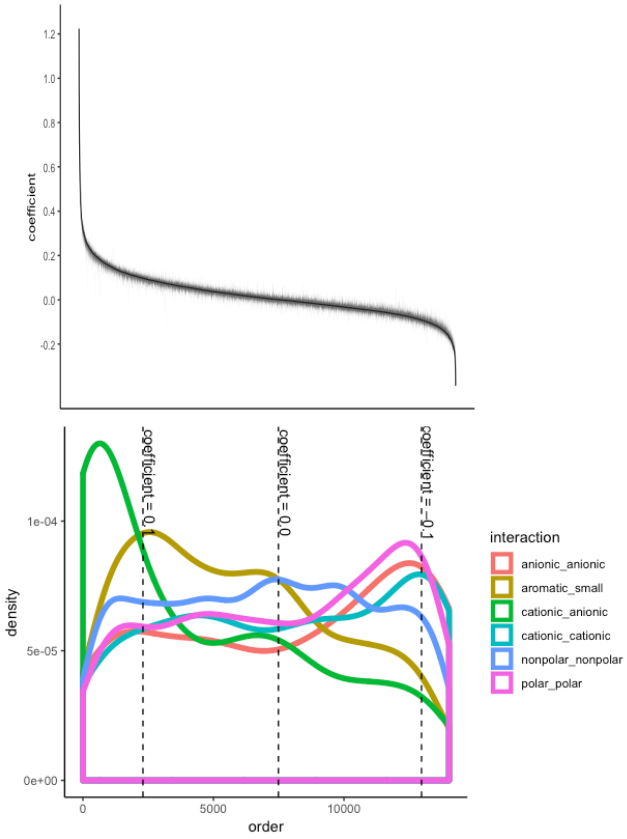

Figure 2

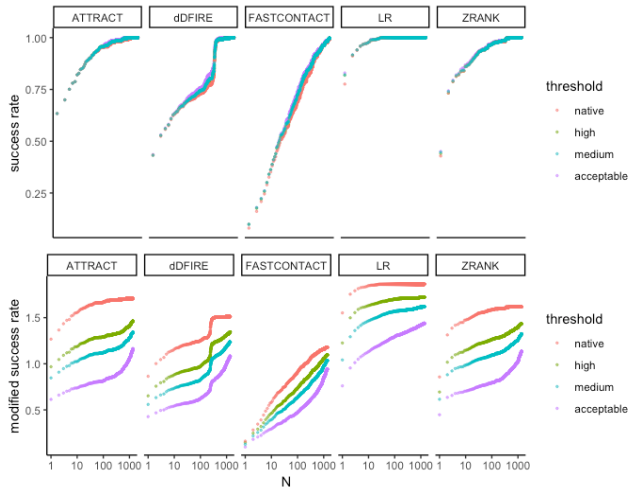

Figure 3

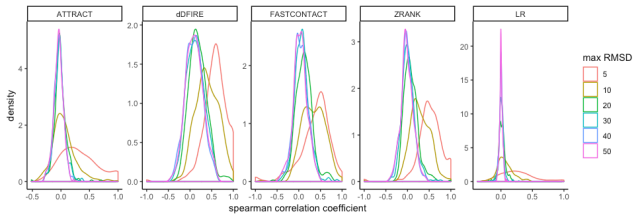

Figure 4

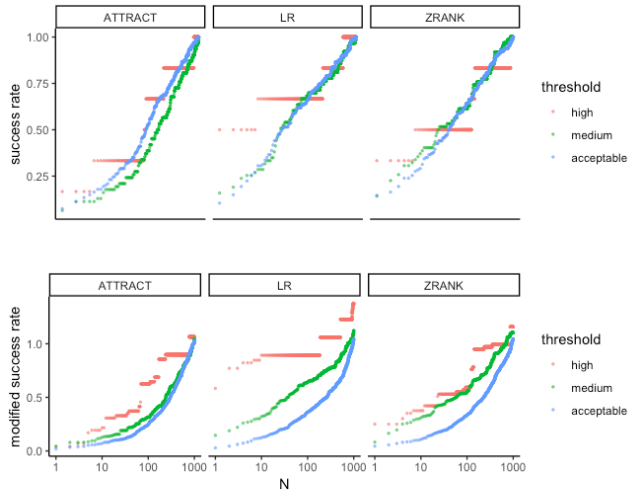

Figure 5