JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Rigidity strengthening is a mechanism for protein-ligand binding

Duc D. Nguyen, Tian Xiao, Menglun Wang, and Guo-Wei Wei

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.7b00226 • Publication Date (Web): 30 Jun 2017

Rigidity Strengthening Is a Mechanism for Protein-Ligand Binding

Duc D. Nguyen,[†] Tian Xiao,[†] Menglun Wang,[†] and Guo-Wei Wei[*,†,‡,¶]

[†]*Department of Mathematics*

Michigan State University, MI 48824, USA

[‡]*Department of Biochemistry and Molecular Biology*

Michigan State University, MI 48824, USA

[¶]*Department of Electrical and Computer Engineering*

Michigan State University, MI 48824, USA

E-mail: wei@math.msu.edu

Abstract

Protein-ligand binding is essential to almost all life processes. The understanding of protein-ligand interactions is fundamentally important to rational drug design and protein design. Based on large scale data sets, we show that protein rigidity strengthening or flexibility reduction is a mechanism in protein-ligand binding. Our approach based solely on rigidity is able to unveil a surprisingly apparently long range contribution of apparently four residue layers to protein-ligand binding, which has a ramification for drug and protein design. Additionally, the present work reveals that among various pairwise interactions, the short range ones within the distance of the van der Waals diameter are most important. It is found that the present approach outperforms all the other state-of-the-art scoring functions for protein-ligand binding affinity predictions of two benchmark test sets.

1 Introduction

Protein-ligand binding is fundamental to many biological processes in living organisms. The binding process involves detailed molecular recognition, synergistic protein-ligand corporation, and possible protein conformational change. Agonist binding alternates receptor function and trigger a physiological response, such as transmitter-mediated signal transduction, hormone and growth factor regulated metabolic pathways, and stimulus-initiated gene expression, enzyme production, and cell secretion, to name only a few. The understanding of protein-ligand interactions is essential for drug design and protein design, and has been a central issue in molecular biophysics, structural biology and medicine. A common belief is that protein-ligand binding is driven by binding energy reduction, which is described in intermolecular forces, such as ionic bonds, hydrogen bonds, hydrophobic effects, and van der Waals (vdW) interactions.[1] However, this view has not entirely translated into accurate binding affinity predictions of large scale binding data sets, despite decades of efforts. Other potential mechanisms, such as flexibility reduction or rigidity enhancement, have been neglected in the current modeling and computation.

Current understanding of flexibility with respect to protein-ligand binding is very limited. On the one hand, it is well-known that flexibility or plasticity of proteins as well as ligands facilitates the ligand docking during the binding process.[2,3] On the other hand, protein-ligand binding reduces the system entropy, which favors the disassociation process. Since protein flexibility is intuitively associated with conformational entropy, binding induced flexibility reduction is widely regarded as unfavorable to the protein-ligand binding.[4] The present work offers evidence against this prevalent view.

Thermodynamically, the protein-ligand binding process is described by the binding affinity, i.e., the change in the Gibbs free energy, which can be expressed in terms of enthalpy and entropy changes at a given temperature. The intricate interplay between enthalpy and entropy and over-simplified association between flexibility and entropy have made the role of flexibility in protein-ligand binding elusive. Fortunately, the availability of vast amount

of affinity databases[5,6] makes it possible to directly test our new hypotheses and reexamine existing theories and computational models. Given the importance of protein-ligand binding to a number of biological fields and disciplines, a wide variety of scoring functions have been proposed. One might classifies these scoring functions into four categories, namely physics based, empirically based, knowledge based and machine learning based ones.[7] Physics based models consider vdW and electrostatics interactions between protein and ligand, in addition to hydrogen bonding and solvation effects.[8–10] Physics based models are important for the mechanistic understanding of protein-ligand interactions. Empirical or regression methods regard binding affinity as a superposition of vdW interaction, hydrogen bonding, desolation, and metal chelation, etc.[11–13] Knowledge-based approaches make use of available protein-ligand binding databases to define scoring functions.[14–16] Finally, machine learning strategies take advantages of large scale databases and the progress in regression algorithms to construct scoring functions that outperform other existing binding predictors for large and diverse binding data sets.[17–20] Despite of the success of machine learning strategies over other approaches, their dependence on numerous, sometimes, apparently unphysical descriptors[20] makes the molecular mechanism of protein-ligand binding more elusive than ever before.

The objective of the present work is to elucidate the role of flexibility or rigidity in protein-ligand binding. We postulate that binding induced protein flexibility reduction, or rigidity strengthening, plays a unique role in the protein-ligand binding. This hypothesis guides us to design a purely rigidity based machine learning strategy for the prediction of protein-ligand binding affinities. It is emphasized that at the microscopic scale, due to the lack of experimental measurement data, the concepts of flexibility and rigidity cannot be as rigorously defined as their counterparts at the macroscopic scale. Consequently, the notion of protein flexibility and rigidity is always subjective. In the present work, protein rigidity modeling is carried out using flexibility-rigidity index (FRI).[21] FRI has been introduced as a simple and efficient algorithm for protein thermal fluctuation analysis, which is convention-

ally associated with protein flexibility in the literature.[22–24] It has been shown to be about 20% more accurate and orders of magnitude more efficient than other classic approaches, such as Gaussian network model (GNM),[23] and anisotropic network model (ANM)[24] in the B-factor prediction of proteins[21,25] and protein-nucleic acid complexes.[26] Our postulation is supported by the rigidity index (RI) prediction of 195 protein complexes and cross validation of 4057 complexes. Our rigidity index based binding affinity predictor, RI-Score, is able to outperform all other recent scoring functions, which supports our hypothesis that flexibility reduction or rigidity enhancement is a mechanism in protein-ligand binding.

2 Theory and methods

2.1 Flexibility-rigidity index (FRI)

Consider a biomolecule having N atoms with coordinates given as $\{\mathbf{r_i}|\mathbf{r_i} \in \mathbb{R}^3, i = 1, 2, \ldots, N\}$. We denote $\|\mathbf{r}_i - \mathbf{r}_j\|$ the Euclidean distance between ith and jth atom. We denote r_i the van der Waals radius of ith atom and set $\eta_{ij} = \tau(r_i + r_j)$ as a scale to characterize the distance between the ith and the jth atoms. Where $\tau > 0$ is an adjustable parameter. The atomic rigidity index and flexibility index are expressed as[21,25]

$$\mu_i = \sum_{j=1}^{N} w_j \Phi_\tau \left(\|\mathbf{r}_i - \mathbf{r}_j\| \right) \quad \text{and} \quad f_i = \frac{1}{\mu_i}, \tag{1}$$

where w_j are the particle-type dependent weights, and are set to 1 in the present work. Here Φ is a real-valued monotonically decreasing correlation function satisfying

$$\Phi \left(\|\mathbf{r}_i - \mathbf{r}_i\| \right) = 1, \quad \text{as} \quad \|\mathbf{r}_i - \mathbf{r}_j\| \to 0, \tag{2}$$

$$\Phi \left(\|\mathbf{r}_i - \mathbf{r}_j\| \right) = 0, \quad \text{as} \quad \|\mathbf{r}_i - \mathbf{r}_j\| \to \infty. \tag{3}$$

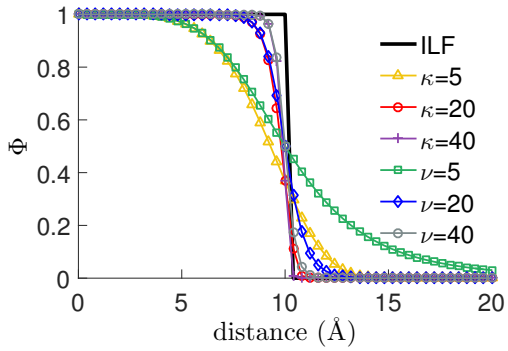


Figure 1: Illustration of FRI correlation functions, which behave like the ideal low filter (ILF) at large κ or ν values.

Commonly used FRI correlation functions include generalized exponential functions

$$\Phi^{\mathrm{E}}_{\kappa,\tau}\left(\|\mathbf{r}_i-\mathbf{r}_j\|\right)=e^{-(\|\mathbf{r}_i-\mathbf{r}_j\|/\eta_{ij})^{\kappa}},\quad \kappa>0 \tag{4}$$

and generalized Lorentz functions

$$\Phi^{\mathrm{L}}_{\nu,\tau}\left(\|\mathbf{r}_i-\mathbf{r}_j\|\right)=\frac{1}{1+(\|\mathbf{r}_i-\mathbf{r}_j\|/\eta_{ij})^{\nu}},\quad \nu>0. \tag{5}$$

As shown in Figure 1, both generalized exponential functions and generalized Lorentz functions approximate the ideal low-pass filter (ILF) as $\kappa\to\infty$ and $\nu\to\infty$, respectively. The atomic rigidity index can be regarded as a contact-based force field.

FRI measures the topological connectivity of the protein-ligand network at every node with appropriate distance-based weights and describes the binding complex with a high level of abstraction. Such an abstraction is ideally suited for the extraction of intrinsic protein-ligand interactions from complex and large-scale binding datasets. One advantage of FRI is that it allows the multiresolution analysis of protein-ligand binding interactions by varying parameter τ, which endows FRI the ability to explore what is the dominant protein-ligand interaction force. Another advantage of FRI is its multiscale analysis via multi-kernel based multiscale FRI (mFRI),[27] which enables FRI to capture different length-scales in various protein-ligand interactions. Finally, by using an ILF representation, the FRI is able to

quantitatively detect the relevant length of interactions.

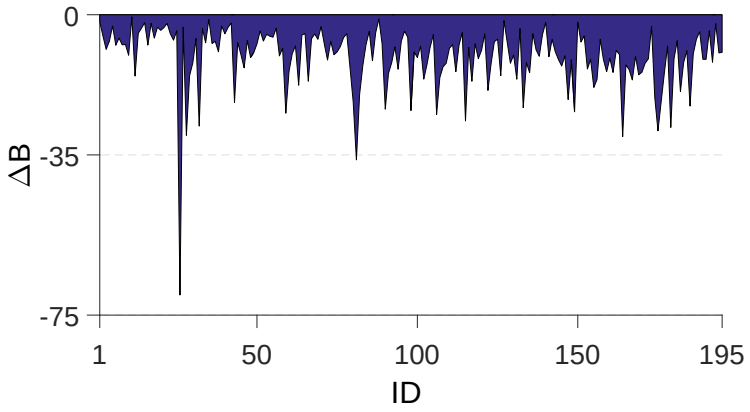


Figure 2: Illustration of relative B-factor change (ΔB) upon ligand binding for the PDBbind v2007 core set ($N = 195$).

2.2 Binding induced flexibility reduction and rigidity strengthening

We first illustrate how the protein flexibility is quantitatively reduced after ligand binding. To this end, we define a protein relative B-factor change ΔB by

$$\Delta B = \sum_j \left[\frac{B_j(\text{Com}) - B_j(\text{Pro})}{B_j(\text{Pro})} \right], \tag{6}$$

where B_j(Com) and B_j(Pro) are respectively the predicted B-factors of the jth heavy atom of the protein-ligand complex (Com) and the original protein (Pro), using the flexibility index.[21] Figure 2 depicts the relative B-factor change (ΔB) against each complex in the PDBBind v2007 core set (N=195).[5] For each complex, all protein heavy atoms within 10 Å from the ligand are considered in this calculation. The average prediction correlation coefficients for the whole data set are 0.62 and 0.63, respectively for proteins and protein-ligand complexes. The decrease in the relative B-factors upon ligand binding indicates that ligand binding reduces protein flexibility and strengthens protein rigidity.

2.3 Rigidity index based scoring functions (RI-Score)

Having shown that ligand binding enhances protein rigidity, it remains to demonstrate that rigidity strengthening is a mechanism for protein-ligand binding. Our hypothesis is that, for blind prediction of protein-ligand binding affinities, if a quantitative model solely based on rigidity analysis is very accurate and more competitive than other state-of-the-art models in the field, it shows that rigidity strengthening is a mechanism for protein-ligand binding. To prove our hypothesis, we exclude electrostatic, van der Waals, hydrogen bond, hydrophobic and hydrophilic interactions used in conventional force field based models,[1,8–10] as well as our method,[20] and consider nothing but protein rigidity change upon ligand binding. According to Eq. (1), we define element-specific protein-ligand rigidity index by collecting cross correlations

$$\mathrm{RI}^{\alpha}_{\beta,\tau,c}(\mathrm{X}-\mathrm{Y}) = \sum_{k\in \mathrm{X}\in \mathrm{Pro}} \sum_{l\in \mathrm{Y}\in \mathrm{Lig}} \Phi^{\alpha}_{\beta,\tau}\left(\|\mathbf{r}_k - \mathbf{r}_l\|\right),\ \forall \|\mathbf{r}_k - \mathbf{r}_l\| \le c, \tag{7}$$

where $\alpha = \mathrm{E}, \mathrm{L}$ is a kernel index indicating either the exponential kernel (E) or Lorentz kernel (L). Correspondingly, β is kernel order index such that $\beta = \kappa$ when $\alpha = \mathrm{E}$ and $\beta = \nu$ when $\alpha = \mathrm{L}$. To reduce the number of free parameters, the particle-type dependent weights w_j are simply assigned to be 1, thus they are omitted in Eq.(7). We adopt our fast FRI (fFRI) based on the cell lists algorithm[28] with a cutoff distance c to reduce computational complexity.[25] Here, X denotes a type of heavy atoms in the protein (Pro) and Y denotes a type of heavy atoms in the ligand (Lig). Four types of protein heavy atoms, namely $\{\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{S}\}$, and nine types of ligand atoms, i.e., $\{\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{S}, \mathrm{P}, \mathrm{F}, \mathrm{Cl}, \mathrm{Br}, \mathrm{I}\}$, are utilized in this work. Unlike force field based methods which require sophisticated data processing, the current model applies directly to protein structure data.

3 Results and discussion

4 The PDBBind v2007 benchmark

For quantitative prediction of protein-ligand binding affinities, we combine protein-ligand rigidity index in Eq. (7) and machine learning to construct RI-Score. Although machine learning can be a powerful approach for modeling massive datasets, its performance depends crucially on its feature vectors. Therefore, if rigidity strengthening is a mechanism for protein-ligand binding, the proposed RI-Score should be able to do well in the binding affinity prediction of massive experimental data sets. Although a specific machine learning algorithm, Random Forest, is used in this work, other machine learning techniques, such as gradient boosting algorithms, deliver similar results. To validate RI-Score, we consider a benchmark dataset, PDBBind v2007, to validate our RI-Score against a large number of scoring functions.[5,29,30] The core set of PDBBind v2007 consists of 195 protein-ligand complexes involving a total of 65 clusters. Its diversity has found to be a challenge for 16 popular scoring functions implemented in mainstream software packages.[5] This finding has stimulated much interest in the community to develop advanced methods for binding affinity predictions.[20,29,30] In the present work, the PDBBind v2007 core set of 195 protein-ligand complexes is utilized as our test set, while our model is trained on the PDBBind v2007 refined set of 1300 protein-ligand complexes, excluding the PDBBind v2007 core set of 195 complexes[5] (i.e., $N = 1105$).

We first consider exponential kernels which are known for their fast decay and thus can be used to detect the effective length of short-, medium- and long-range interactions. To this end, we investigate the behavior of high-order exponential kernels with large κ values, which are essentially the ideal low pass filter as shown in Figure 1 and thus are able to exclude all interactions beyond the kernel length scale. Figure 3(a) depicts the Pearson correlation coefficients of four high order exponential kernels over a wide range of scales. It is interesting to find surprising oscillations in the Pearson correlation coefficients over different scales. Such

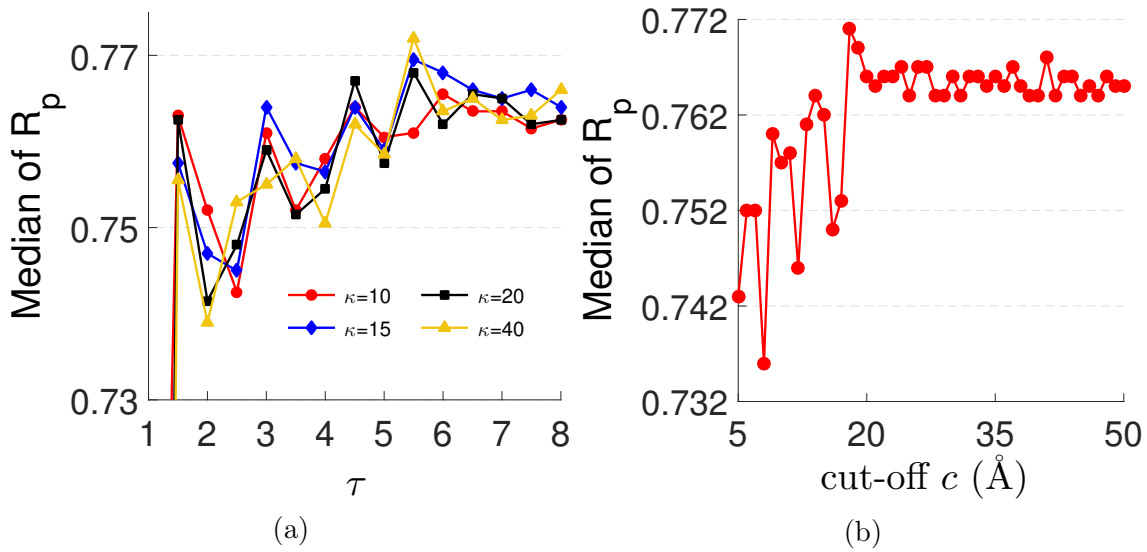


Figure 3: (a) Pearson correlation coefficients (R_p) of $\mathrm{RI}^{\mathrm{E}}_{\kappa,\tau,18}$ are plotted against the choice of τ for for PDBBind v2007 core set over a range of values for κ. Values of R_p under 0.73 are not shown. (b) R_p values of $\mathrm{RI}^{\mathrm{E}}_{40,5.5,c}$ are plotted against different values of cut-off distance c for PDBBind v2007 core set. The obvious oscillations are associated with the numbers of nearest residues included in the rigidity features.

oscillations might indicate the inclusion of different numbers of nearest residue layers. For example, when $\tau = 1.5$, η_{CC} for a $\mathrm{C}-\mathrm{C}$ pair is 5.1Å. Since one of these two carbons belongs to the protein and the other belongs to the ligand, the only hydrophobic interactions that can be accounted are those from the nearest layer of residues. The next peak occurs at $\tau = 3$ (i.e., $\eta_{\mathrm{CC}} = 10.2$Å), which might be due to the inclusion of the interactions of the nearest two layers of residues. It is amazing to note that these oscillations persist for two more peaks to reach the best predictions at $\tau = 5.5$, which suggests that interactions from the apparently nearest four layers of residues effectively contribute to the protein-ligand binding. The present finding has significant ramifications to protein design.

We note that when $\tau = 1.5$, Pearson correlation coefficients obtained by all exponential kernels are already better than those attained by most other methods in the field.[5,29,30] This good performance is due to the fact that at $1.5(r_i + r_j)$, all the hydrogen bonds, most van der Waals interactions, and good portion of electrostatic interactions are included in our RI-Score. It is well known that hydrogen bond interactions can be effective from 2.2 to 4 Å,

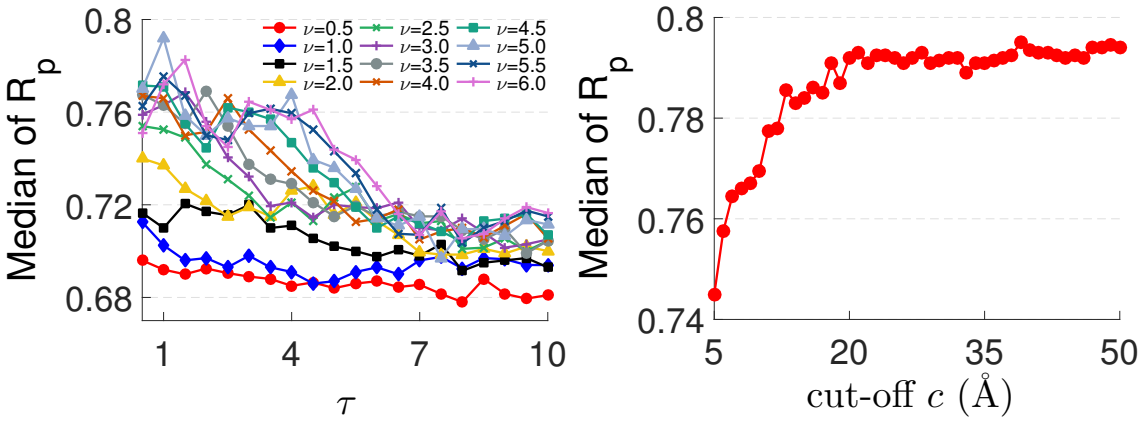


Figure 4: (a) Pearson correlation coefficients (R_p) of $\mathrm{RI}^{\mathrm{L}}_{\nu,\tau,40}$ are plotted against the choice of τ for for PDBBind v2007 core set over a range of ν values. (b) R_p values of $\mathrm{RI}^{\mathrm{L}}_{5,1,c}$ for PDBBind v2007 core set are plotted against different values of cut-off distance c.

with donor-acceptor distances of 2.2-2.5Å as strong, mostly covalent, 2.5-3.2Å as moderate, mostly electrostatic, and 3.2-4.0Å as weak, electrostatic.[31] Their corresponding energies are about 40-14, 15-4, and less than 4 kcal/mol, respectively.[31]

The impact of cutoff distance as shown in Figure 3(b) also show an oscillatory pattern. In fact, this oscillation is consistent with that shown in Figure 3(a).

Unlike exponential kernels, Lorentz kernels are well known for their slow decay, which is able to capture interactions over a wide range of distances. Figure 4 shows the influence of the power of the Lorentz kernel, the scale, and cutoff distance to the blind prediction accuracy in terms of Pearson correlation coefficients with respect to the experimental binding affinity data. With a sufficiently large cutoff values $c = 40$Å, best prediction is obtained at $\nu = 5$ and $\tau = 1$. This result reveals that most important protein-ligand interactions occur approximately at van der Waals distances, which strongly indicates that *short range hydrogen bond interactions are relatively more important than long range interactions*. When $\nu = 5$ and $\tau = 1$, a cutoff value of 20Å, which contains four layers of residues, is found to be large enough to include all the essential protein-ligand interactions.

Protein-ligand interactions intrinsically involve multiple characteristic length scales, such as those for three types of hydrogen bonds, van der Waals bonds, and hydrophobic inter-

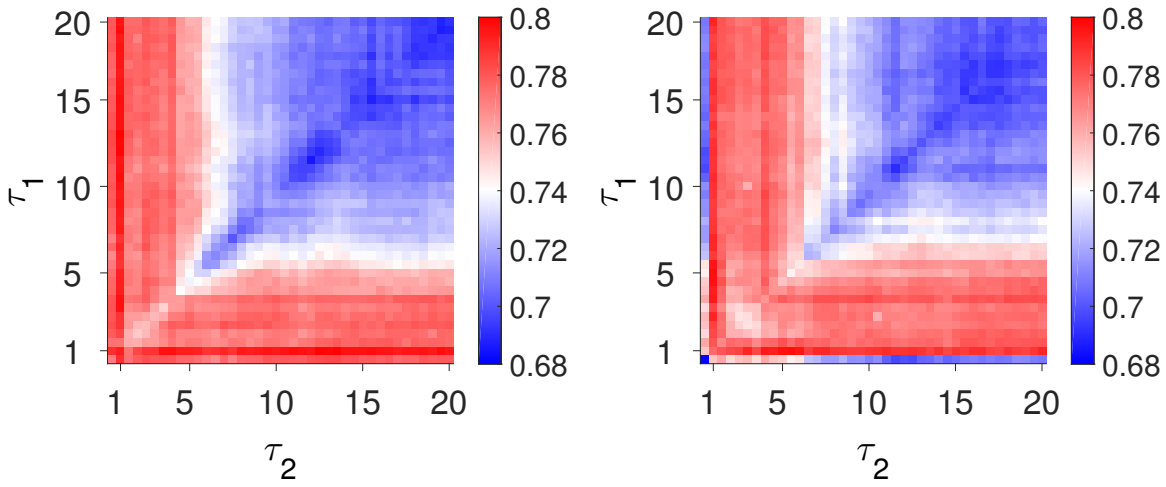


Figure 5: An illustration of multiscale behavior in protein-ligand binding prediction. Median values of Pearson correlation coefficients of 2-scale models on PDBBind v2007 core set are plotted against different scale values τ_1 and τ_2: (a) $\mathrm{RI}^{\mathrm{LL}}_{5,\tau_1,\max(39,3.7\tau_1);5,\tau_2,\max(39,3.7\tau_2)}$; (b) $\mathrm{RI}^{\mathrm{EE}}_{2.5,\tau_1,\max(12,3.7\tau_1);2.5,\tau_2,\max(12,3.7\tau_2)}$. Scale parameters τ_1 and τ_2 vary from 0.5 to 20 with increments of 0.5. Obvious, the combination of a relatively small-scale kernel and a relatively large-scale kernel delivers best prediction, which indicates the importance of incorporating multiscale in protein-ligand binding prediction.

actions, which can still be very important for residues far away as indicated above. In our earlier study, we found that multiscale FRI (mFRI) that utilizes two or three FRI kernels parametrized with different length scales can significantly improve the B-factors prediction. In this work, we are interested in examining the impact of multiscale rigidity to binding free energy prediction. Machine learning approach makes it particularly convenient to incorporate multiscale effect in predictive models. One just needs to construct one additional set of features at a desirable scale. To this end, we construct two sets of RI feature vectors, with one of the set parametrized at τ_1 and the other at τ_2. In general, we denote these feature vectors by $\mathrm{RI}^{\alpha_1\alpha_2}_{\beta_1,\tau_1,c_1;\beta_2,\tau_2,c_2}$ as a straightforward extension of our notation. The two-scale models consist of 72 descriptors. In which, each kernel α_i $(i = 1, 2)$ is used to generate 36 descriptors, presenting cross correlations between heavy atoms in protein and ligands as described in Eq. 7. To reduce the number of degrees of freedom, values of ν(or κ) and c are adopted from the corresponding single-kernel model. As a result, we only need

to search for two scale parameters τ_1 and τ_2. Figure 5 illustrates the impact of two-scale RI feature based RI-Score predictions of the PDBBind v2007 core set of 195 complexes. The median of R_p of RI^{LL} and $\mathrm{RI}^{\mathrm{EE}}$ are plotted against different pairs of scale parameters (τ_1, τ_2) varying from (0.5,0.5) to (20,20). First, it is easy to see that two-scale predictions are typically better than the single scale one. The best predictions are typically achieved at the combination of a relatively small-scale kernel and a relatively large-scale kernel. A scale optimized model, $\mathrm{RI}^{\mathrm{LL}}_{5,1,39;5,12.5,47}$, delivers the median and the best Pearson correlation coefficient $(R_p^m, R_p^b) = (0.800, 0.811)$. Similarly a scale optimized model, $\mathrm{RI}^{\mathrm{EE}}_{2.5,1,12;2.5,5.5,21}$, achieves $(R_p^m, R_p^b) = (0.797, 0.809)$. More details are reported in Table SI in Supporting Information. These findings again confirm the importance of incorporating multiscale in the protein-ligand predictions.

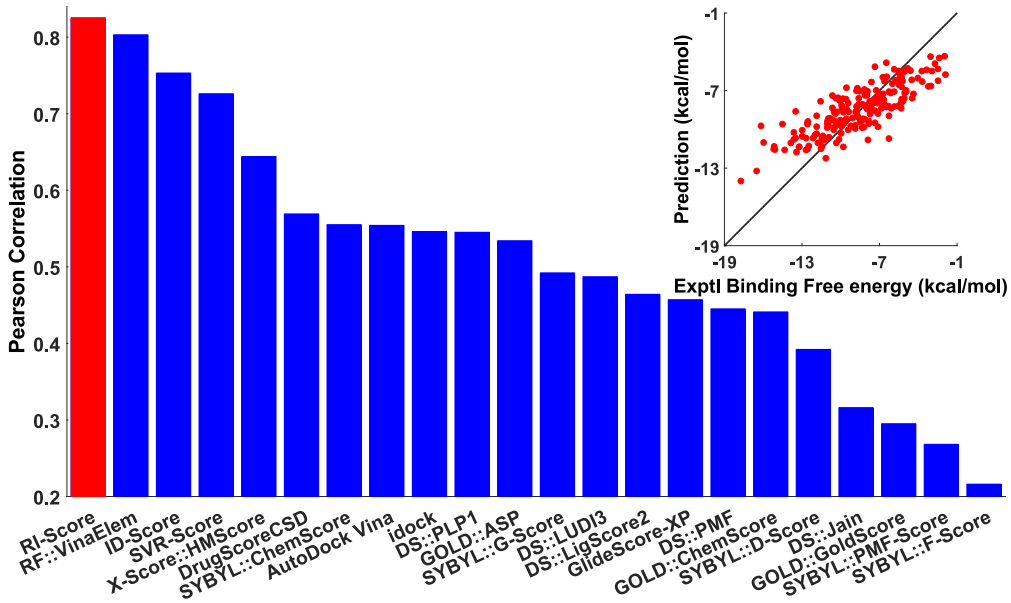


Figure 6: Performance comparison between different scoring functions on the PDBBind v2007 core set. The performances of other scoring functions are adopted from Refs.[5,29,30,32,33] The RMSE and Pearson correlation coefficient of our prediction, $\mathrm{RI}^{\mathrm{LLE}}_{5,1,39;5,13,49;2.5,1,12}$, are 1.99 kcal/mol and 0.825, respectively.

Encouraged by the above two-scale results, we have also explored the utility of three-scale RI-Score models for the prediction of the PDBBind v2007 core set. In each three-scale model, a total of 108 features is generated by three corresponding kernels. Certainly, a

complete search of the parameter space is prohibitively expensive. We therefore focus on exploring the scale parameters τ_1 and τ_2 for certain sets of other parameters. It is found that $\mathrm{RI}^{\mathrm{LLE}}_{5,1,39;5,13,49;2.5,1,12}$ is the best model with $(R_p^m, R_p^b) = (0.803, 0.825)$. The corresponding RMSE of our best prediction is 1.99 kcal/mol. One can realize that the models involving Lorentz kernels usually perform better than ones consist of only exponential kernels. This finding is consistent with the B-factor predictive performance reported in our previous work.[21,25] More details of predicted energies by RI-Score are provided in the Supporting Information.

Predictions from different scoring functions on PDBBind v2007 core set have been presented in the literature.[5,29,30] Figure 6 plots the performance of these scoring functions adopted from Ref.[5,29,30] along with our prediction, i.e., $\mathrm{RI}^{\mathrm{LLE}}_{5,1,39;5,13,49;2.5,1,12}$ (highlighted with red color). Clearly, RI-Score outperforms all the other scoring functions. One might argue that this benchmark is unfair to 16 popular physical and empirical models since their parameters were trained on different data sets. However, these methods were designed to be training-set independent. A user cannot change their parametrization. Therefore, they should be allowed to be used and compared on any data set. In fact, a rich diversity of protein clusters in the PDBBind data sets (v2007 core set has 65 clusters) gives rise to a difficult task for these approaches to fit their free parameters to accommodate such diverse sets. Whereas, machine learning scoring functions, such as RF-Score[32] and our model in this work, takes the advantages of advanced regression algorithms and the liberty of adaptive training practice to identify the relationship between unseen and trained complexes in large and diverse data sets. However, the performance of machine learning methods highly depends on the statistical similarity between the training set and the test set.

4.1 The PDBBind v2013 benchmark

It remains to show that the outstanding performance of proposed RI-Score is not limited a specific data set. To this end, we consider PDBBind v2013 core set of 195 protein-ligand

complexes as our test set.[6] Involving 65 protein clusters, this core set is another popular benchmark test for scoring functions. A comparative assessment of 20 mainstream scoring functions was carried out by Li *et al* in the literate to examine their scoring power, ranking power and docking power.[6] The relative low scoring power of current scoring functions revealed by this study has reminded the community of the importance to develop alternative scoring functions. In the present work, the PDBBind v2015 refined set of 3706 protein-ligand complexes, excluding the PDBBind v2013 core set, is employed as our training set. Our best model for the PDBBind v2013 core set which is the same as PDBBind v2015 core set, is given by $RI^{LLE}_{6,4,15;6,10,37;3.5,1.5,8}$ with a Pearson correlation coefficient and RMSE of 0.782 and 2.051 kcal/mol, respectively. More details of RI-Score predictions are provided in Table S3 of the Supporting Information.

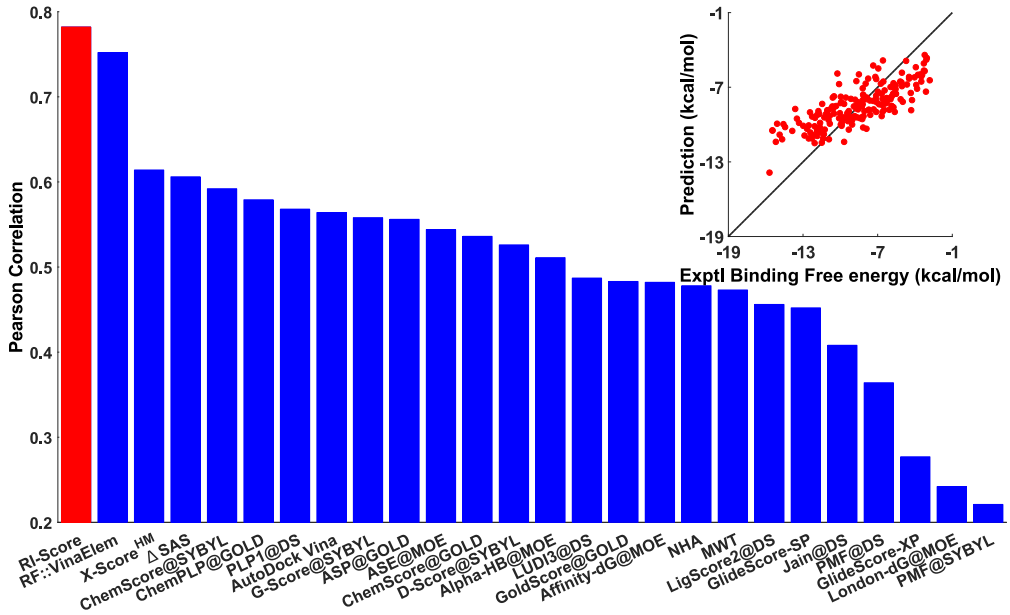


Figure 7: Performance comparison between different scoring functions on the PBDBind v2013 core set. The performances of RF::VinaElem is adopted from.[34] Results of 20 other scoring functions were reported in Ref.[6] The RMSE and Pearson correlation coefficient of our prediction, $RI^{LLE}_{6,4,15;6,10,37;3.5,1.5,8}$, are 2.051 kcal/mol and 0.782, respectively.

.

For comparison, Figure 7 plots the performance of our method and 25 other scoring functions on this benchmark test with data directly taken from Refs.[6,34] Again, RI-Score is found to be significantly more accurate than other methods. As discussed before, unlike

machine learning based approaches, 20 empirical methods were not trained on the PDBBind database, although they are designed to be training-set independent.

4.2 The PDBBind v2016 data sets

The PDBBind v2016 is the latest version of PDBBind database and is available on the public website `http://www.pdbbind.org.cn`. More specifically, the PDBBind v2016 is an extended version of the PDBBind v2015 with the number of complexes in the refined set increasing from 3706 complexes in v2015 up to 4057 complexes in v2016. For simplicity, we use the best RI parameters found in our PDBBind v2013 benchmarking, i.e., $\mathrm{RI}^{\mathrm{LLE}}_{6,4,15;6,10,37;3.5,1.5,8}$, for all predictions related to PDBBind v2016 sets. To validate our model, we carry out five-fold and ten-fold cross validations on 4057 complexes in PDBBind v2016 refined set. In the five-fold cross validation, the average of Pearson correlations R_p and RMSEs are 0.747 and 1.827 kcal/mol, respectively. In the ten-fold cross validation, the performance of RI-Score marginally improves with $R_p = 0.754$ and RMSE = 1.800 kcal/mol.

The PDBBind v2016 core set consists of 290 protein-ligand complexes in 58 clusters and is used to illustrate the predictive power of the RI-Score. We perform two predictions on the PDBBind v2016 core set by respectively employing PDBBind v2015 and PDBBind v2016 refined sets as training data. Similarly to the previous numerical experiments, all complexes in the core set are excluded from the training data. The Pearson correlation R_p and RMSE of RI-Score on the PDBBind v2016 core set are 0.811 and 1.854 kcal/mol when using the PDBBind v2015 refined set ($N = 3436$) as a training data. When the PDBBind v2016 refined set ($N = 3767$) is utilized, the performance of our method is slightly improving with values of R_p and RMSE found to be 0.815 and 1.854 kcal/mol, respectively. By comparing the performances of RI-Score on the core set to those on five-fold and ten-fold cross validations, one can see that while Pearson correlations are getting better, the RMSEs remain similar. This observation ensures there is no over-fitting concern in the proposed model. Since currently no other scoring function result is available on these data sets in the

literature, Table 1 only presents our RI-Score predictions.

Table 1: Performances of $RI^{LLE}_{6,4,15;6,10,37;3.5,1.5,8}$ on the PDBBind v2016 data sets. [a] The average results of 5-fold cross validations. [b] The average results of 10-fold cross validations. [c] Predictions obtained by using the non-overlap part of PDBBind v2015 refined set as training data. [d] Predictions obtained by using the non-overlap part of PDBBind v2016 refined set as training data.

Data set	R_p	RMSE (kcal/mol)
PDBBind v2016 refined set	0.747^a	1.827^a
	0.754^b	1.800^b
PDBBind v2016 core set	0.811^c	1.854^c
	0.815^d	1.854^d

5 Conclusion

Protein-ligand binding is of paramount importance to biomolecular functions and biomedicine. The molecular mechanism of protein-ligand binding remains an active research topic, despite of enormous effort in the past few decades. Using an intuitively defined rigidity function, the present work qualitatively demonstrates that protein-ligand binding gives rise to protein rigidity enhancement. Further study shows that rigidity index alone is able to quantitatively describe protein-ligand binding affinities. The fact that rigidity index based scoring function, RI-Score, offers the best binding affinity prediction over two benchmark data sets, i.e., PDBBind v2007 core set and PDBBind v2013 core set, over a variety of other popular scoring functions indicates that rigidity strengthening is a mechanism in protein-ligand binding. Therefore, protein flexibility reduction may drive protein-ligand binding. It is well known that protein flexibility is associated with many protein functions. The present finding may suggest that agonist protein receptor binding inhibits protein functions through protein flexibility reduction. Additionally, a correlation of the nearest four layers of residues to protein-ligand binding affinities is discovered, which has a nontrivial ramification to drug and protein design.

The present finding on rigidity strengthening can be explained in terms of free energy (or enthalpy) driven protein-ligand binding indicated by short range interactions. However, the present model does not explicitly include the interactions of water molecules with protein and ligand in the description. The absence of such interactions implies the missing of more detailed entropy counts. The interplay between entropy and enthalpy can be more complex. The present result may not always be true for all protein-ligand binding cases. Entropy driven protein-ligand binding manifested by hydrophobic effects is possible in certain circumstances. This mechanism becomes common in protein-protein binding.

Supporting Information Available

See `SI_RI-score-5.pdf` for discussion of dataset preparation, and some additional results for PDBBind v2007 core set , PDBBind v2013 core set and PDBBind v2016 core set.
RI-Score online server is publicly available at `http://weilab.math.msu.edu/RI-Score`.
This material is available free of charge via the Internet at `http://pubs.acs.org/`.

Acknowledgement

This work was supported in part by NSF grants IIS-1302285 and DMS-1721024 and MSU Center for Mathematical Molecular Biosciences Initiative.

References

(1) Gilson, M. K.; Zhou, H. X. Calculation of Protein-ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(2) Zavodszky, M. I.; Kuhn, L. A. Side-chain Flexibility in Protein–ligand Binding: The Minimal Rotation Hypothesis. *Protein Sci.* **2005**, *14*, 1104–1114.

(3) Tuffery, P.; Derreumaux, P. Flexibility and Binding Affinity in Protein–Ligand, Protein–Protein and Multi-component Protein Interactions: Limitations of Current Computational Approaches. *J. R. Soc., Interface* **2012**, *9*, 20–33.

(4) Grünberg, R.; Nilges, M.; Leckner, J. Flexibility and Conformational Entropy in Protein–Protein Binding. *Structure* **2006**, *14*, 683–693.

(5) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions On a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49*, 1079–1093.

(6) Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54*, 1717–1736.

(7) Liu, J.; Wang, R. Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, *55*, 475–482.

(8) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33*, 889–897.

(9) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **1995**, *38*, 2681–2691.

(10) Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V. Medusascore: An Acurate Force Field-based Scoring Function for Virtual Drug Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1656–1662.

(11) Zheng, Z.; Merz Jr, K. M. Ligand Identification Scoring Algorithm (LISA). *J. Chem. Inf. Model.* **2011**, *51*, 1296–1306.

(12) Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. Empirical Free Energy Calculations of Ligand–Protein Crystallographic Complexes. I. Knowledge Based Ligand–Protein Interaction Potentials Applied to the Prediction of Human Immunodeficiency Virus Protease Binding Affinity. *Protein Eng.* **1995**, *8*, 677–691.

(13) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput. Aided. Mol. Des.* **1997**, *11*, 425–445.

(14) Muegge, I.; Martin, Y. A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791–804.

(15) Velec, H.; Gohlke, H.; Klebe, G. DrugScore (CSD)-Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-native Ligand Poses and Better Affinity Prediction. *J. Med. Chem.* **2005**, *48*, 6296–303.

(16) Huang, S. Y.; Zou, X. An Iterative Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions: I. Derivation of Interaction Potentials. *J. Comput. Chem.* **2006**, *27*, 1865–1875.

(17) Kinnings, S. L.; Liu, N.; Tonge, P. J.; Jackson, R. M.; Xie, L.; Bourne, P. E. A Machine Learning Based Method to Improve Docking Scoring Functions and Its Application to Drug Repurposing. *J. Chem. Inf. Model.* **2011**, *51*, 408–419.

(18) Ashtawy, H. M.; Mahapatra, N. R. A Comparative Assessment of Ranking Accuracies of Conventional and Machine-Learning-Based Scoring Functions for Protein-Ligand Binding Affinity Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2012**, *9*, 1301–1313.

(19) Brenner, M. P.; Colwell, L. J. Predicting Protein–ligand Affinity with a Random Matrix Framework. *Proc. Natl. Acad. Sci.* **2016**, 13564–13569.

(20) Wang, B.; Zhao, Z.; Nguyen, D. D.; Wei, G. W. Feature Functional Theory - Binding Predictor (FFT-BP) for the Blind Prediction of Binding Free Energy. *Theor. Chem. Acc.* **2017**, *136*, 55.

(21) Xia, K. L.; Opron, K.; Wei, G. W. Multiscale Multiphysics and Multidomain Models — Flexibility and Rigidity. *J. Chem. Phys.* **2013**, *139*, 194109.

(22) Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. Protein Flexibility Predictions Using Graph Theory. *Proteins: Struct., Funct., Bioinf.* **2001**, *44*, 150–165.

(23) Bahar, I.; Atilgan, A. R.; Erman, B. Direct Evaluation of Thermal Fluctuations in Proteins Using a Single-Parameter Harmonic Potential. *Folding Des.* **1997**, *2*, 173 – 181.

(24) Atilgan, A. R.; Durrell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80*, 505 – 515.

(25) Opron, K.; Xia, K. L.; Wei, G. W. Fast and Anisotropic Flexibility-Rigidity Index for Protein Flexibility and Fluctuation Analysis. *J. Chem. Phys.* **2014**, *140*, 234105.

(26) Opron, K.; Xia, K. L.; Burton, Z.; Wei, G. W. Flexibility-rigidity Index for Protein-nucleic Acid Flexibility and Fluctuation Analysis. *J. Comput. Chem.* **2016**, *37*, 1283–1295.

(27) Opron, K.; Xia, K. L.; Wei, G. W. Communication: Capturing Protein Multiscale Thermal Fluctuations. *J. Chem. Phys.* **2015**, *142*.

(28) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford: Clarendon Press, 1987.

(29) Li, G.-B.; Yang, L.-L.; Wang, W.-J.; Li, L.-L.; Yang, S.-Y. id-score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein-Ligand Interactions. *J. Chem. Inf. Model.* **2013**, *53*, 592–600.

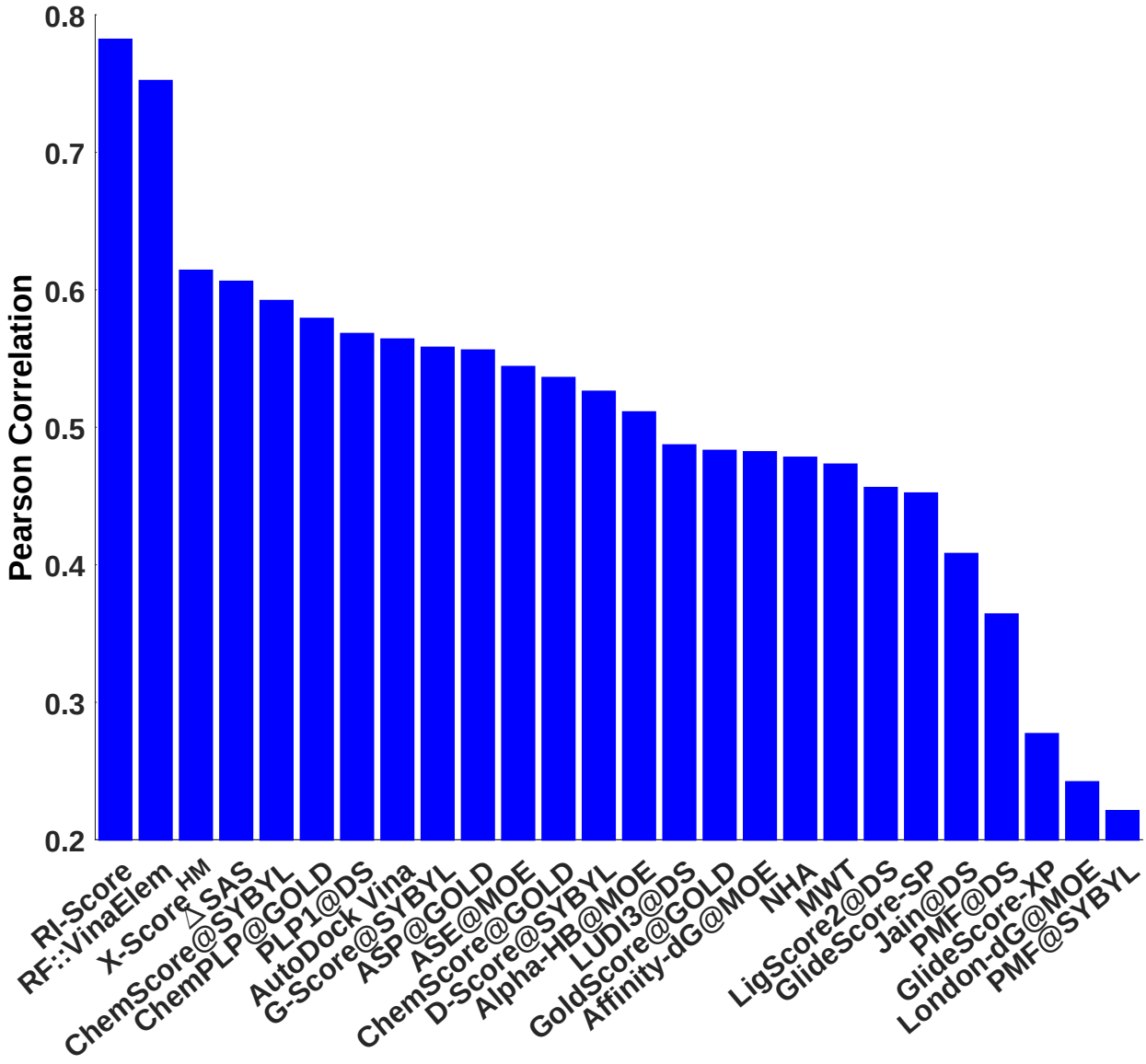
(30) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Improving Autodock Vina Using Random Forest: the Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol. Inf.* **2015**, *34*, 115–126.

(31) Jeffrey, G. A.; Jeffrey, G. A. *An Introduction to Hydrogen Bonding*; Oxford university press New York, 1997; Vol. 12.

(32) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein -ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26*, 1169–1175.

(33) Li, H.; Leung, K.; Ballester, P.; Wong, M. H. istar: A Web Platform for Large-Scale Protein-Ligand Docking. *Plos One* **2014**, *9*.

(34) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Low-Quality Structural and Interaction Data Improves Binding Affinity Prediction via Random Forest. *Molecules* **2015**, *20*, 10947–10962.


Pearson Correlation
0.8
0.7
0.6
0.5
0.4
0.3
0.2
RI-Score
RF::VinaElem
X-Score HM
SAS
ChemScore@SYBYL
ChemPLP@GOLD
PLP1@DS
AutoDock Vina
G-Score@SYBYL
ASP@GOLD
ASE@MOE
ChemScore@GOLD
D-Score@SYBYL
Alpha-HB@MOE
LUDI3@DS
GoldScore@GOLD
Affinity-dG@MOE
NHA
MWT
LigScore2@DS
GlideScore-SP
Jain@DS
PMF@DS
GlideScore-XP
London-dG@MOE
PMF@SYBYL