Subject Section

KORP-PL: a coarse-grained knowledge-based scoring function for protein-ligand interactions

Maria Kadukova [1,2], Karina dos Santos Machado [1,3], Pablo Chacón [4,*], and Sergei Grudinin [1,*]

[1]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France and
[2]Moscow Institute of Physics and Technology, 141701 Dolgoprudniy, Russia and
[3]Centro de Ciências Computacionais, Universidade Federal do Rio Grande - FURG, Av. Itália, km 08, Rio Grande, RS, Brazil and
[4]Department of Biological Chemical Physics, Rocasolano Institute of Physical Chemistry C.S.I.C, Madrid, Spain.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Despite the progress made in studying protein-ligand interactions and the widespread application of docking and affinity prediction tools, improving their precision and efficiency still remains a challenge. Computational approaches based on the scoring of docking conformations with statistical potentials constitute a popular alternative to more accurate but costly physics-based thermodynamic sampling methods. In this context, a minimalist and fast sidechain-free knowledge-based potential with a high docking and screening power can be very useful when screening a big number of putative docking conformations.

Results: Here we present a novel coarse-grained potential defined by a 3D joint probability distribution function that only depends on the pairwise orientation and position between protein backbone and ligand atoms. Despite its extreme simplicity, our approach yields very competitive results with the state-of-the-art scoring functions, especially in docking and screening tasks. For example, we observed a two-fold improvement in the median 5% enrichment factor on the DUD-E benchmark compared to Autodock Vina results. Moreover, our results prove that a coarse sidechain-free potential is sufficient for a very successful docking pose prediction.

Availability: The standalone version of KORP-PL with the corresponding tests and benchmarks are available at https://team.inria.fr/nano-d/korp-pl/ and https://chaconlab.org/modeling/korp-pl.

Contact: pablo@chaconlab.org, sergei.grudinin@inria.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Binding processes at physiological conditions are driven by thermodynamics laws. Even though their physics is well understood at the theoretical level, practical application of these laws to computational docking problems requires exhaustive thermodynamic sampling. This makes most of the corresponding approaches computationally prohibitive. A popular alternative consists in avoiding exhaustive sampling and approximating binding free energies with knowledge-based and statistical potentials (Verdonk *et al.*, 2003; Velec *et al.*, 2005; Huang and Zou, 2006, 2010; Neudert and Klebe, 2011; Debroise *et al.*, 2017; Kadukova and Grudinin, 2017). These are directly parameterized against available experimental data, rather than derived from the first principles. However, very often these potentials are not physical. For example, it is easy to demonstrate that the binding free energy cannot be decomposed into a sum of pairwise interactions, as the desolvation term is not pairwise-additive (Ben-Naim, 1997). Also, the performance of statistical potentials is much better in docking exercises and rather moderate in screening tests (Li *et al.*, 2018; Su *et al.*, 2018). These observations, as well as a moderate performance of classical statistical potentials on some of the popular docking benchmarks, both protein-protein, and protein-ligand, have triggered further community research in multiple directions. They include the development of coarse-grained and orientation-dependent scoring functions (Zhang and Zhang, 2010; Elhefnawy *et al.*, 2015; Karasikov *et al.*, 2019; Lopez-Blanco and Chacon, 2019; Neudert and Klebe, 2011; Wang *et al.*, 2013).

In addition to the knowledge-based potentials that are most often derived in a statistical and unsupervised manner, a considerable number of scoring functions are based on other principles (Liu and Wang, 2015; Shen *et al.*, 2020). These include physics-based potentials (Brooks *et al.*, 1983; Ewing *et al.*, 2001; Case *et al.*, 2005) that approximate energy

terms and require very careful calibration, as well as a variety of scoring functions obtained using the principles of supervised machine learning. Starting from the classical empirical scoring functions that were trained to fit experimental binding constants with a linear combination of several physics-based descriptors (Böhm, 1994; Wang *et al.*, 2002; Friesner *et al.*, 2006; Trott and Olson, 2010; Quiroga and Villarreal, 2016; Debroise *et al.*, 2017), more and more complex methods based on non-linear models and diverse descriptors have been developed (Li *et al.*, 2013; Wang and Zhang, 2017; Lu *et al.*, 2019; Ashtawy and Mahapatra, 2017; Shen *et al.*, 2020; Wallach *et al.*, 2015; Ragoza *et al.*, 2017; Jiménez *et al.*, 2018; Karlov *et al.*, 2020). Although some of these often demonstrate high performance in affinity prediction and virtual screening, they are also subject to a number of flaws. Indeed, while classical statistical potentials tend to be biased towards the number of contacts between the two molecules, learning on a relatively small number of available high-quality binding constants introduces biases towards experimental affinities. Very complex models, especially those from deep learning, may also introduce overfitting. For example, some recent architectures demonstrate excellent results on the DUD-E virtual screening benchmark if they are trained on a part of it. However, their performance is rather average if they are trained on other data sources (Chen *et al.*, 2019). Surprisingly enough, the classical empirical AutoDock Vina scoring function and its modifications, while being physically interpretable, still achieve stable state-of-the-art results in both pose and affinity predictions.

Protein-ligand methods usually describe molecules using the all-atom representation. Therefore, incorrect positioning of sidechains inside the binding pockets may introduce steric clashes with the ligands and produce false-positive predictions of binding poses. Some of the methods can include optimization of the sidechains in the conformation search (Trott and Olson, 2010; DeLuca *et al.*, 2015; Marze *et al.*, 2018), but this makes the docking process much more computationally expensive. Furthermore, slight inaccuracies in the positioning of the backbone atoms may introduce significant inaccuracies in the positions of the sidechains. A possible way to circumvent this problem is to model a protein molecule without explicit positioning of its sidechains. Indeed, such representations have already been successfully used in various protein structure prediction applications (Liwo *et al.*, 2002; Karasikov *et al.*, 2019; Lopez-Blanco and Chacon, 2019; Senior *et al.*, 2019; Zheng *et al.*, 2019; Kryshtafovych *et al.*, 2019).

Motivated by the excellent results obtained in protein and loop modeling with a sidechain-independent potential KORP (Lopez-Blanco and Chacon, 2019), we propose to adapt its methodology to protein-ligand interactions. The success of KORP is rooted in the consideration of the full six-dimensional (6D) joint probability distribution function that only depends on the relative orientation between protein residues. For the protein-ligand interactions, we reduce the dependence of the pairwise potential to a 3D joint probability of observing an interacting ligand atom at a given relative position and orientation from a protein residue. The proposed method, called KORP-PL, does not require protein sidechain atoms, and only three backbone atoms of the protein residue are needed. As a result, it is relatively fast, as each interaction involves only the computation of two spherical angles and a single distance. Despite its seeming simplicity, our approach yields state-of-the-art results.

2 Methods

2.1 The KORP-PL model

Our starting point was the 6D orientation-dependent knowledge-based potential KORP (Lopez-Blanco and Chacon, 2019), which had been successfully used in protein and loop modeling. The main idea behind using the 6D statistics in KORP is that one can unambiguously define

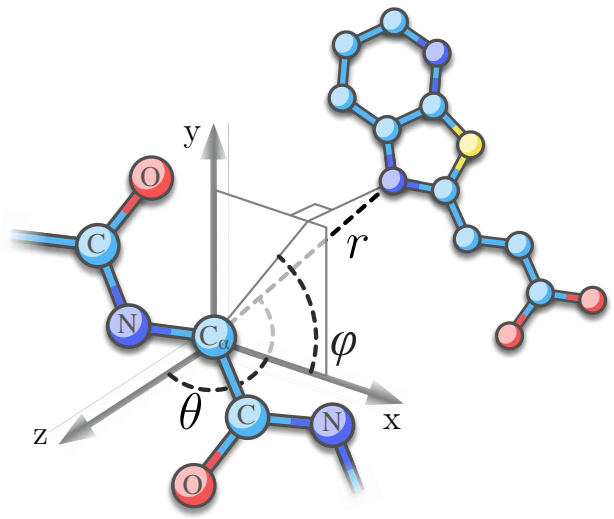

Fig. 1. Schematic view of the relative orientation of a ligand molecule to a protein residue. The residue is represented with a 3D oriented frame built from three backbone atoms. The relative orientation of a ligand atom is described by two spherical angles, θ, the polar angle between the r and z vectors, and φ, the azimuthal angle between x and the projection of r into the xy plane.

local coordinate frames for each of the protein residues. However, this model cannot be applied to small molecules owing to their higher chemical diversity. Therefore, we keep the reduced representation for the protein molecule, and use the all-atom representation for the ligand. As depicted in Fig. 1, the relative position and orientation of a given interaction is specified by the ligand atom coordinates and by a 3D oriented frame (i.e. a local coordinate system) built from three backbone atoms C, C_α, and N of the protein residue. Therefore, only two spherical angles θ and φ and the distance r between the ligand atom and the center of the residue frame located at C_α are required. The interaction $\tilde{E}_{i,j}$ between a residue i and a ligand atom j is then derived from the Boltzmann distribution,

$$\tilde{E}_{i,j}(\theta,\varphi,r) = -RT \ln \frac{P^{\text{obs}}_{aa,lig}(\theta,\varphi,r) + z}{P^{\text{ref}}(\theta,\varphi,r) + z}, \tag{1}$$

where RT is the Boltzmann factor, and $P^{\text{obs}}_{aa,lig}$ is the joint 3D probability of observing a protein amino acid i of a type aa and a ligand atom j of a type lig at a given distance r and orientation (θ,ψ) in a set of crystallographic structures. We should note that there is a full dependence between the three variables in $P^{\text{obs}}_{aa,lig}(\theta,\varphi,r)$. In order to counterbalance nonspecific residue-ligand interactions, we introduce the reference probability P^{ref} regardless of the type of interaction. It corresponds to the *reference state*, defined as an average distribution over the different amino acid and ligand types (Samudrala and Moult, 1998). Also, we add the z constant to both the nominator and denominator of the above expression to prevent numerical instability for low-count statistics. Following the original KORP implementation, we zero-mean normalized individual contributions $\tilde{E}_{ij}(\theta,\varphi,r)$ at every distance to reduce distance-specific biases. More precisely, from each value of the interaction potential at (θ,φ,r) we subtracted an average taken at the same distance r over all θ and φ values,

$$E_{ij}(\theta,\varphi,r) = \tilde{E}_{ij}(\theta,\varphi,r) - \langle \tilde{E}_{ij}(\theta,\varphi,r) \rangle_{\theta,\varphi}. \tag{2}$$

The total protein-ligand interaction potential will be then the sum of all individual contributions E_k within a certain cutoff distance,

$$E = \sum_k E_k(\theta_k,\varphi_k,r_k). \tag{3}$$

The total number of protein residue types is equal to 20 and corresponds to the 20 standard amino acids. The set of 37 ligand atom types comprises 8 carbon types, 12 nitrogen types, 7 oxygen types, 4 sulfur types, 2 phosphorus types, and 4 types describing halogens (see Table S1 of Supporting Information for more details). Each ligand atom type is assigned using the Knodle library (Kadukova and Grudinin, 2016) in the same manner as we did for the Convex-PL scoring function (Kadukova and Grudinin, 2017).

2.2 Training data

We derived KORP-PL using structures of protein-ligand complexes deposited in the PDBBind 2016 general dataset (Wang *et al.*, 2005). We excluded 373 structures intersecting with those from the CASF-2013 and CASF-2016 benchmarks. This resulted in 12,910 selected examples. Also, there were no intersections between PDBBind 2016 and examples from the D3R challenges that we use to compile our benchmark. We did not specifically preprocess the input structures. We did not remove homologous receptor structures since their bound ligands can be very diverse. In fact, previously we did not find any effect of excluding structures in the training set homologous to the ones in the test set on the prediction accuracy (Kadukova and Grudinin, 2017). Nonetheless, we provide additional computational experiments excluding the test set structures from the training set at different levels of similarity.

We collected the statistics using interactions within the range of radial distances r of (2 Å, 11 Å). This statistics were divided into 12 bins. The angular statistics were collected into 180 equiareal bins using a uniform angular sampling tessellation described elsewhere (Lopez-Blanco and Chacon, 2019).

2.3 Reweighing the potential for binding affinity predictions

Initial tests demonstrated rather poor performance of KORP-PL in affinity prediction exercises. This is mostly the result of the independence of E_{ij} terms from each other. Motivated by this observation, we devised a reweighing scheme to balance the contributions of each component E_{ij} as

$$E^w = \sum_i^{N_{aa}} \sum_j^{N_{lig}} c_i r_j E_{ij}, \tag{4}$$

where c_i and r_j are the weighting factors for a given amino acid or ligand atom type, respectively, N_{aa} is the number of amino acid types, and N_{lig} is the number of ligand atom types. We computed the weighting factors by fitting experimental binding constants from the PDBBind 2016 general dataset. This was achieved by minimizing the squared error loss using the L-BFGS-B algorithm (Zhu *et al.*, 1997) implemented in scipy (Virtanen *et al.*, 2020). We then iteratively optimized the $\mathbf{c}$ and $\mathbf{r}$ vectors following Algorithm 1 from SI. Details of the optimization are available in SI, including the values of the obtained weights listed in Table S2. Interestingly, higher weights correspond to hydrophobic interactions, which will be mentioned below. Throughout the text, we will refer to the reweighed version of KORP-PL as to KORP-PL^w.

2.4 CASF benchmarks

We assessed KORP-PL on a recent CASF-2016 benchmark (Su *et al.*, 2018), and a smaller but more widely used CASF-2013 benchmark (Li *et al.*, 2018). These benchmarks are the sets of respectively 285 and 195 high-quality crystal structures with the corresponding binding affinities. Four different metrics are used in these benchmarks defined as *docking power*, *scoring power*, *ranking power*, and *screening power*. Docking power corresponds to the ability of a scoring function to predict the native or the best near-native docking pose among a set of computer-generated configurations. Scoring functions are evaluated by the number of the top-ranked predictions (top-1, top-2, and top-3) below a predefined cutoff distance from the crystal structure (1.0, 2.0, and 3.0 Å). Scoring and ranking powers measure the quality of affinity prediction of complexes with known co-crystal structures. Scoring power assesses the correlation of scoring function predictions with the experimental binding affinity data. Ranking power is related to the capability of a scoring function to correctly rank a set of known ligands for a target protein. In CASF-2016, where five known ligands are available for each target protein, it is measured by Spearman's correlation coefficient. However, in CASF-2013 only three ligands per protein are available and ranking power is represented with two numbers characterizing success rates of either correct ranking of all the given ligands, or finding the most affine one. Finally, screening power is related to the ability of a scoring function to identify true binders for a target protein among a set of small molecules. CASF benchmarks suggest two metrics to evaluate this ability. Enrichment Factor (EF) is calculated as a ratio between the total number of true binders observed among a fraction of top-ranked candidates (1%, 5%, and 10%) and the total number of true binders multiplied by this fraction. It represents the ability of a scoring function to correctly find active compounds compared to a random selection. 'Best binder success rate' is a success rate of identifying the highest-affinity binder among the 1%, 5%, or 10% of top-ranked ligands over all the test cases.

2.5 D3R benchmarks

A number of community-wide blind protein-ligand docking challenges were held throughout recent years. For example, the CSAR (Carlson *et al.*, 2016) initiative was carried out in 2010-2014. Later on, it was continued and further developed by the Drug Design Data Resource (D3R) (Gathiaka *et al.*, 2016). The aim of these challenges was the evaluation of docking protocols on previously unpublished structural data. After all participants have submitted their predictions, co-crystal structures become revealed and submissions get evaluated. A considerable effort was made by the D3R community to host data from the previous challenges. In particular, this resource contains all user submissions and answers, i.e. native structures and binding constants, from the recent three blind challenges, namely Grand Challenge 2 (Gaieb *et al.*, 2018), Grand Challenge 3 (Gaieb *et al.*, 2019), and Grand Challenge 4 (Parks *et al.*, 2020). Unfortunately, user submission data from the first D3R challenges is not publicly available.

Thus, we compiled a benchmark from the user submissions and published answers of the three blind challenges. Similar to the CASF benchmarks, it contains pose and affinity prediction exercises. However, this benchmark is different from CASF in several aspects. Unlike the CASF benchmarks, which were created from the data deposited in the Protein Data Bank (Rose *et al.*, 2017), experimental data for each of the D3R challenge targets were provided by a single research group. Co-crystal structures were also visually inspected by the challenge organizers and participants. This allows us to expect higher quality and consistency of this data, especially for the binding constants, which are less trustworthy in the CASF benchmarks, and PDBBind in general. On the contrary, data from the D3R Challenges provides smaller diversity of both proteins and small molecules, since each of the three challenges was focused on one protein target binding with compounds of several chemical series. For example, the affinity prediction test made from the D3R Challenge data is closer to the CASF ranking test than to the scoring one.

For the pose prediction tests, we collected all available user submissions from the pose prediction stages of the three challenges. RMSD values were obtained from the D3R website when possible, otherwise, we computed them using a modified version of symmetry-adapted RDKit's GetBestRMS() function, in which we disabled the ligand alignment, and PyMol's (Schrödinger, LLC, 2011) *align* function to superpose each

protein to its native structure. We excluded several submissions listed in Table S15 of the Supporting Information because of various errors and clustered the rest of submissions with a 0.1 Å threshold without the binding pocket alignment. This clustering was mainly done to remove very similar or equivalent docking poses that were often present in submissions from the same users. Finally, we measured the pose prediction success rates on each test separately with and without the inclusion of the native structures. For the affinity prediction tests, we selected only the native structures and then measured the Spearman's correlation coefficients between predicted and experimental binding constants for each of the Grand Challenges. When the ligand was present in several chains of the co-crystal structure, we scored all of the available complexes and took the average. The number of available submissions and binding constants is summarized in Table S14 of SI.

2.6 DUD-E benchmark

The DUD-E benchmark (Mysinger *et al.*, 2012), the successor of the DUD benchmark (Huang *et al.*, 2006), is a very popular approach for assessing virtual screening abilities of various scoring functions and docking protocols. It consists of 102 targets, a set of active compounds per target known to bind it, and 50 inactive compounds, or decoys, per each active one. The total number of active compounds for all 102 targets equals to 22,886. For each target, one protein-ligand complex is provided and can be used for the identification of the binding pocket and molecular docking. The benchmark also contains 3D conformers of all the active and inactive compounds. Unlike the CASF benchmarks and the D3R-based benchmark that we have derived specifically for structure-based scoring functions assessment, evaluation on DUD-E requires a pose sampling stage. Therefore, we firstly performed molecular docking using AutoDock Vina with default settings except for the *exhaustiveness* that was set to 10, and then re-scored the obtained poses with KORP-PL and KORP-PL^w.

We should note that the DUD-E benchmark contains several targets with co-factors that seem to be crucial for binding. We have excluded from the evaluation 12 complexes listed in Table S22 that contain HEM, NAD, NAP, FAD, ADP, and FMN, since KORP-PL is not parametrized to predict interactions with co-factors.

3 Results and discussion

3.1 CASF benchmarks

Figure 2 shows the results obtained on the docking, scoring, ranking, and virtual screening tests from the CASF-2016 benchmark. The results obtained on the exercises from the CASF-2013 can be found in Figure S1 and Tables S3-S5 of SI. We can see that KORP-PL performs exceptionally well in the pose prediction exercise, despite being a coarse-grained scoring function. Indeed, for the CASF-2016 benchmark, its success rate in finding a near-native pose within 2 Å RMSD as the best prediction is 85.6%. This is better than the success rates of all other tested scoring functions.

Figures 2 (d-e) demonstrate the top-ranked performance of KORP-PL in both screening tests. These results are especially notable if considering the enrichment factor metric, where all other tested scoring functions perform rather poorly. For example, CASF-2016 Top1% EF for KORP-PL is 22.23, while the third-best Top1% EF is 11.91 for ChemPLP@GOLD. Figure 2 (b) compares KORP-PL binding affinity predictions. They turned out to be worse than average. As a consequence, ranking power results (Fig. 2 (c)) are also worse or close to average when compared with the other scoring functions. To investigate the reasons leading to such rather poor performance, we plotted binding affinities predicted by KORP-PL versus the experimental binding constants. Figure 3 shows them colored according to the hydrophobic scale of the protein binding pockets suggested by Su *et al.* (2018). We can see that KORP-PL often underestimates affinity values for complexes with hydrophobic pockets. We suppose that it happens due to the way we compute the *reference state* inherited from the original KORP 6D potential. Indeed, the 6D residue-residue interactions have a strong angular dependence, which is not the case for the protein-ligand setting. For example, the subtraction of the angular average in Eq. 2 will result in a near-zero potential for non-directional contributions. This is precisely the case for some of the hydrophobic interactions. It motivated us to introduce the reweighing scheme (see Eq. 4), which allowed us to partially compensate for this effect. Indeed, the KORP-PL^w potential performed considerably better than KORP-PL on the scoring tests. However, its performance is still far from perfect and this is a subject for further investigation. We should also note that moderate performance of various scoring functions in affinity prediction tasks can be partially explained by the fact that experimental uncertainties of binding affinity data in current databases are often larger than one order of magnitude (Wätzig *et al.*, 2015). Such significant scatter is the result of different methodologies and accuracy of binding assays used in different research groups. SI Table S13 contains further analysis of the correlation between the KORP-PL scores and a number of ligand properties computed for the CASF-2016 complexes.

CASF benchmarks are derived from the PDBBind database and thus contain complexes similar to our training set. Thus, it is interesting and important to learn how much our results can *overfit* the input data. Therefore, we ran additional experiments and modified the training set by augmenting it with the intersection with the test set, and also removing a number of complexes based on the protein (Zhang and Skolnick, 2004; Ritchie *et al.*, 2012) and ligand (Landrum, 2006) shape similarity. After, we recomputed the CASF docking and screening tests to investigate the possible overfitting. These results are listed in Tables S6-S12 and discussed in Supporting Information. Overall, removing the closest complexes (pocket TM-score > 0.8 and ligand shape Jaccard distance < 0.2) affects the metrics only marginally. Further elimination of about a thousand of more distant complexes (pocket TM-score > 0.5 and ligand shape Jaccard distance < 0.4) worsens the overall performance. Notably, high-quality docking predictions (Q1) are affected more than the low-quality ones (Q2-3). This indicates that for a successful high-resolution pose prediction, the training set must contain complexes with interactions that somewhat resemble those in the test set. Indeed, any statistical (Boltzmann in our case) approximation is limited if some features are not present or their distribution is unbalanced in the training set.

3.2 D3R benchmarks

Figure 4 demonstrates very good performance of KORP-PL in all pose prediction exercises derived from the D3R Challenges. KORP-PL also showed good results in the Grand Challenge 2 and Grand Challenge 4 affinity ranking tasks. However, we obtained near-zero correlations in affinity prediction of the cathepsin S complexes from the Grand Challenge 3.

D3R Grand Challenge 3 pose prediction test turned out to be an interesting case. In this exercise, the binding site is exposed to solvent and is surrounded by water molecules in the co-crystal structure as well as in some of the user submissions. We should specifically mention that we do not take explicit water molecules into account. KORP-PL showed excellent results in the pose prediction exercise compared to AutoDock Vina and Convex-PL scoring functions. Although we cannot directly compare the pose prediction results with the full protocols evaluated in the challenge, only a few of them were successful, especially if no visual inspection and ligand-based methods were used (Gaieb *et al.*, 2019). This means that the selection of correct binding poses for the cathepsin S inhibitors could be a challenge for many scoring functions. For example, as can

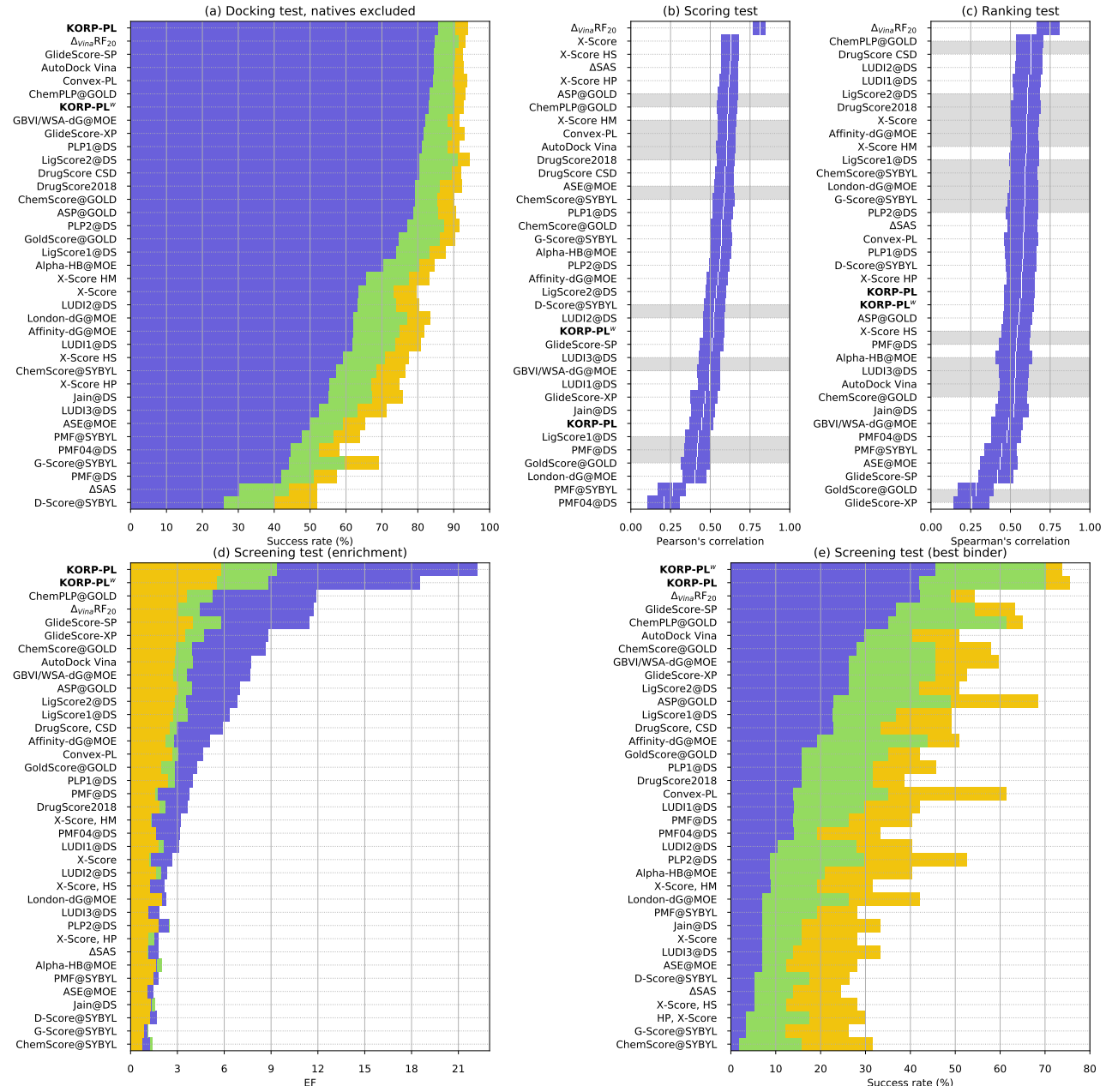

Fig. 2. CASF-2016 benchmark results. (a) The success rate of finding a near-native pose within 2 Å RMSD in Top 1 (blue), Top 2 (green), and Top 3 (yellow) predictions. Native poses are excluded. (b) Pearson's correlation with confidential values between predicted scores and experimental $\log K_a$. Scoring functions sharing the same gray bar are not distinguishable at $\alpha = 0.1$ in the post-hoc Friedman test (Su et al., 2018). (c) Spearman's rank correlation with confidential values among the 57 clusters. (d) Enrichment factors computed considering 1% (blue), 5% (green), and 10% (yellow) of the top-ranked compounds. (e) The success rate of identifying the highest-affinity binder among the 1% (blue), 5% (green), or 10% (yellow) top-ranked ligands. All results except KORP-PL and Convex-PL were taken from the supplementary information of the CASF-2016 benchmark paper (Su et al., 2018). The results of KORP-PL, KORP-PL^{w}, and Convex-PL are available in Tables S3-S5 of SI.

be seen in Figure 4, Convex-PL failed in many cases to detect the correct binding mode, while AutoDock Vina and the simplistic ΔSAS were almost completely incapable of doing it. This could be caused by a combination of the following reasons. Firstly, we believe that by its design, KORP-PL is able to better catch directed interactions from target complexes, such as hydrogen and halogen bonding, and π-stacking (Salentin *et al.*, 2015). Secondly, all the incorrect poses are located deeper in the binding pocket, forming more contacts than the native conformation. Most of the scoring functions tend to be biased towards the total number of protein-ligand contacts, which could lead to incorrect predictions for Convex-PL, Vina, and ΔSAS. As we have already discussed, KORP-PL underestimates some of the non-orientational hydrophobic interactions. However, in this particular case of D3R Grand Challenge 3, it helps to predict ligand positions that are not very buried in protein pockets.

3.3 DUD-E benchmark

To evaluate the performance of KORP-PL in large-scale virtual screening tasks, we have assessed it on 90 targets from the DUD-E benchmark. As it can be seen in Table 1 and Figure 5, KORP-PL and KORP-PL^{w} outperform

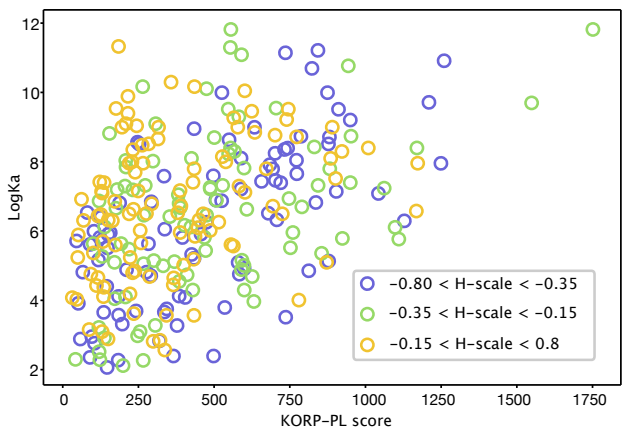

Fig. 3. Scatter plot of KORP-PL scores versus the $\log K_a$ constants from CASF-2016 benchmark. Each point is colored according to the hydrophobicity of the protein binding pocket (H-scale, in logD units) as defined in (Su et al., 2018). The Pearson correlation coefficients between KORP-PL scores and $\log K_a$ constants, computed for three different H-scale groups, are: 0.63 for H-scales between -0.80 and -0.35, 0.45 for H-scales between -0.35 and -0.15, and 0.31 for H-scales between -0.15 and 0.8.

AutoDock Vina in all the metrics, being almost twice better if considering the enrichment factors. This makes KORP-PL comparable to some recent structure-based deep-learning models that demonstrate excellent virtual screening performance (Ragoza *et al.*, 2017). However, according to Chen *et al.* (2019) that we have mentioned in the introduction, such scoring functions tend to achieve high performance on the DUD-E benchmark only if they have been originally trained on it, and thus probably learn hidden biases such as the decoy selection criteria that were used upon the benchmark construction.

Scoring function	ROC AUC		EF5%		BEDROC $_{\alpha=20}$	
	median	average	median	average	median	average
AutoDock Vina	0.731	0.714	3.691	4.528	0.234	0.264
KORP-PL	0.816	0.785	9.083	8.637	0.502	0.472
KORP-PL^w	0.818	0.786	8.839	8.423	0.458	0.465

Table 1. ROC AUC scores, 5% enrichment factors, and BEDROC (Truchon and Bayly, 2007) values computed for the 90 targets from the DUD-E dataset. Twelve targets with co-factors in the binding pocket were excluded from the 102 original targets. It is important to note here that our results for AutoDock Vina are slightly lower than those reported in Ragoza et al. (2017), where the median and average ROC AUC, and median and average EF5% are equal to 0.740, 0.717, 4.228, and 4.485, respectively. This could be caused by the differences in the binding pocket detection or other docking protocol settings. Per-target evaluation results can be found in Table S23 of SI.

3.4 Computational details

KORP-PL is implemented in C++ and available as a binary for macOS and Linux operating systems. It takes about 25 milliseconds on a single core of Linux Intel(R) Xeon(R) CPU E5-2609 @ 2.40GHz to score a protein-ligand complex from the CASF-2013 core set containing a single ligand pose of 25 heavy atoms on average. However, energy computation itself takes only 2 milliseconds and the rest of the runtime is spent on the complex file parsing. As the method does not require positions of the sidechain atoms, it can be readily applied to scoring protein models represented only by their backbones.

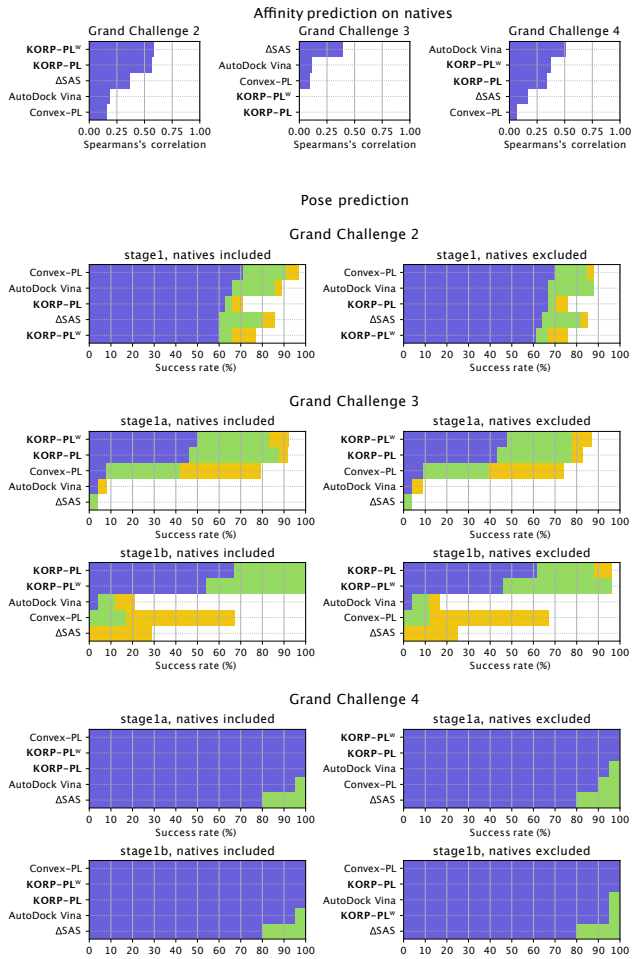

Fig. 4. D3R pose prediction and scoring results. Success rates of finding a pose within 2 Å RMSD from the native conformation among the 1%, 5%, and 10% of top-ranked poses are shown in blue, green, and yellow, respectively. Scoring power is represented by the Spearman's correlation coefficient between the predicted and experimental binding constants. These success rates are computed with respect to the actual number of ligands, for which the poses with the desired RMSD values were present in the user submissions. Due to this fact, for example, the KORP-PL success rate in Grand Challenge 2 is higher when the native poses are excluded. The results of KORP-PL, KORP-PL^w, Convex-PL, AutoDock Vina, and ΔSAS evaluation are listed in Tables S16-S21 of SI. The pose prediction stage of all the three challenges was called 'Stage 1', the affinity prediction stage was called 'Stage 2'. However, receptor flexibility turned out to be a considerable issue for many approaches (Kadukova and Grudinin, 2018), and in both Grand Challenge 3 and Grand Challenge 4, Stage 1 was subdivided into Stage 1a, where neither ligand, nor receptor 3D structure was known, and Stage 1b, where the receptor 3D structure was revealed. We evaluated these stages separately.

4 Conclusion

This paper presents KORP-PL – a novel knowledge-based scoring function for protein-ligand interactions based on the backbone-only receptor and full-atom ligand representations. The receptor representation is adopted from the KORP scoring function, which was designed to model interactions in a protein molecule with a set of oriented coordinate frames built on each protein residue. KORP-PL interaction potential is then derived using statistics of relative orientations and positions of ligand atoms in the local coordinate systems of protein residues. We have demonstrated for the first time that a coarse-grained sidechain-free protein representation can be successfully used for very accurate predictions of ligand binding poses. Indeed, KORP-PL shows excellent

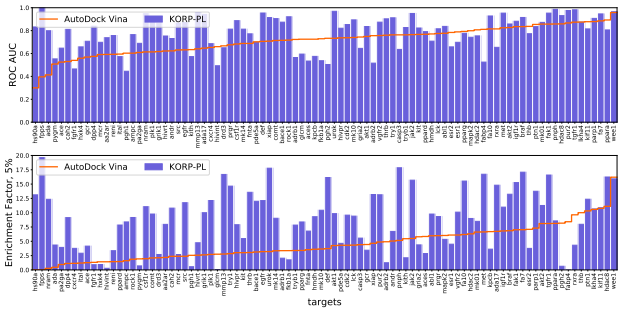

Fig. 5. ROC AUC scores and 5% enrichment factors computed for the 90 targets from the DUD-E dataset with KORP-PL and AutoDock Vina.

pose prediction and screening results in CASF-2013 and CASF-2016 benchmarks, and even in pose prediction benchmarks compiled from the D3R Grand Challenges. KORP-PL also demonstrates outstanding results in the DUD-E virtual screening benchmark, where it considerably outperforms AutoDock Vina. Our affinity prediction performance is, however, lower than average, and much more work is required to advance developments in this direction. Overall, this work proposes a very efficient solution to circumvent the long-standing problem of sampling protein sidechain conformations in molecular docking. This paves the way for the development of a new generation of flexible docking approaches.

Funding

This work was partially supported by the Russian Foundation for Basic Research according to the research project #18-54-00030 Bel_a, Spanish grants BFU2016-76220-P and PID2019-109041GB-C21 (AEI/FEDER, UE), and Inria associate team Flexmol. KSM was supported by CAPES grant PE 88881.207869/2018-01.

References

Ashtawy, H. M. and Mahapatra, N. R. (2017). Task-specific scoring functions for predicting ligand binding poses and affinity and for screening enrichment. *J. Chem. Inf. Model.*, **58**(1), 119–133.

Ben-Naim, A. (1997). Statistical Potentials Extracted from Protein Structures: Are These Meaningful Potentials? *J. Chem. Phys.*, **107**(9), 3698–3706.

Böhm, H.-J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.*, **8**(3), 243–256.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). Charmm: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.*, **4**(2), 187–217.

Carlson, H. A., Smith, R. D., Damm-Ganamet, K. L., Stuckey, J. A., Ahmed, A., Convery, M. A., Somers, D. O., Kranz, M., Elkins, P. A., Cui, G., Peishoff, C. E., Lambert, M. H., and Dunbar, Jr, J. B. (2016). CSAR 2014: A benchmark exercise using unpublished data from pharma. *J. Chem. Inf. Model.*, **56**(6), 1063–1077.

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005). The Amber Biomolecular Simulation Programs. *J. Comput. Chem.*, **26**(16), 1668–1688.

Chen, L., Cruz, A., Ramsey, S., Dickson, C. J., Duca, J. S., Hornak, V., Koes, D. R., and Kurtzman, T. (2019). Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening. *PloS one*, **14**(8), e0220113.

Debroise, T., Shakhnovich, E. I., and Chéron, N. (2017). A hybrid knowledge-based and empirical scoring function for protein–ligand interaction: SMoG2016. *J. Chem. Inf. Model.*, **57**(3), 584–593.

DeLuca, S., Khar, K., and Meiler, J. (2015). Fully flexible docking of medium sized ligand libraries with RosettaLigand. *PloS one*, **10**(7).

Elhefnawy, W., Chen, L., Han, Y., and Li, Y. (2015). ICOSA: A distance-dependent, orientation-specific coarse-grained contact potential for protein structure modeling. *J. Mol. Biol.*, **427**(15), 2562–2576.

Ewing, T. J., Makino, S., Skillman, A. G., and Kuntz, I. D. (2001). Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.*, **15**(5), 411–428.

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C., and Mainz, D. T. (2006). Extra Precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem*, **49**(21), 6177–6196.

Gaieb, Z., Liu, S., Gathiaka, S., *et al.* (2018). D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput.-Aided Mol. Des.*, **32**(1), 1–20.

Gaieb, Z., Parks, C. D., Chiu, M., *et al.* (2019). D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *J. Comput.-Aided Mol. Des.*, **33**(1), 1–18.

Gathiaka, S., Liu, S., Chiu, M., *et al.* (2016). D3R Grand Challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J. Comput.-Aided Mol. Des.*, **30**(9), 651–668.

Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). Benchmarking sets for molecular docking. *J. Med. Chem*, **49**(23), 6789–6801.

Huang, S.-Y. and Zou, X. (2006). An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.*, **27**(15), 1866–1875.

Huang, S.-Y. and Zou, X. (2010). Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J. Chem. Inf. Model.*, **50**(2), 262–273.

Jiménez, J., Skalic, M., Martinez-Rosell, G., and De Fabritiis, G. (2018). K_{DEEP}: Protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J. Chem. Inf. Model.*, **58**(2), 287–296.

Kadukova, M. and Grudinin, S. (2016). Knodle: A Support Vector Machines-based automatic perception of organic molecules from 3D coordinates. *J. Chem. Inf. Model.*, **56**(8), 1410–1419.

Kadukova, M. and Grudinin, S. (2017). Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization. *J. Comput.-Aided Mol. Des.*, **31**(10), 943–958.

Kadukova, M. and Grudinin, S. (2018). Docking of small molecules to farnesoid X receptors using AutoDock vina with the Convex-PL potential: lessons learned from D3R Grand Challenge 2. *J. Comput.-Aided Mol. Des.*, **32**(1), 151–162.

Karasikov, M., Pagès, G., and Grudinin, S. (2019). Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. *Bioinformatics*, **35**(16), 2801–2808.

Karlov, D. S., Sosnin, S., Fedorov, M. V., and Popov, P. (2020). graphDelta: MPNN scoring function for the affinity prediction of protein–ligand complexes. *ACS omega*, **5**(10), 5150–5159.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)– Round XIII. *Proteins: Struct. Func. Bioinfo.*, **87**(12), 1011 –1020.

Landrum, G. (2006). RDKit: Open-source cheminformatics. http://www.rdkit.org.

Li, G.-B., Yang, L.-L., Wang, W.-J., Li, L.-L., and Yang, S.-Y. (2013). ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein–Ligand Interactions. *J. Chem. Inf. Model.*, **53**(3), 592–600. PMID: 23394072.

Li, Y., Su, M., Liu, Z., Li, J., Liu, J., Han, L., and Wang, R. (2018). Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark. *Nat. Protoc.*, **13**(4), 666.

Liu, J. and Wang, R. (2015). Classification of current scoring functions. *J. Chem. Inf. Model.*, **55**(3), 475–482. PMID: 25647463.

Liwo, A., Arłukowicz, P., Czaplewski, C., Ołdziej, S., Pillardy, J., and Scheraga, H. A. (2002). A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: application to the UNRES force field. *Proc. Natl. Acad. Sci. U.S.A.*, **99**(4), 1937–1942.

Lopez-Blanco, J. R. and Chacon, P. (2019). KORP: knowledge-based 6D potential for fast protein and loop modeling. *Bioinformatics*, **35**(17), 3013–3019.

Lu, J., Hou, X., Wang, C., and Zhang, Y. (2019). Incorporating explicit water molecules and ligand conformation stability in machine-learning scoring functions. *J. Chem. Inf. Model.*, **59**(11), 4540–4549.

Marze, N. A., Roy Burman, S. S., Sheffler, W., and Gray, J. J. (2018). Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics*, **34**(20), 3461–3469.

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem*, **55**(14), 6582–6594.

Neudert, G. and Klebe, G. (2011). DSX: a knowledge-based scoring function for the assessment of protein–ligand complexes. *J. Chem. Inf. Model.*, **51**(10), 2731–2745.

Parks, C. D., Gaieb, Z., Chiu, M., *et al.* (2020). D3R Grand Challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput.-Aided Mol. Des.*, **34**(2), 99–119.

Quiroga, R. and Villarreal, M. A. (2016). Vinardo: A scoring function based on AutoDock Vina improves scoring, docking, and virtual screening. *PloS one*, **11**(5).

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.*, **57**(4), 942–957.

Ritchie, D. W., Ghoorah, A. W., Mavridis, L., and Venkatraman, V. (2012). Fast protein structure alignment using gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, **28**(24), 3274–3281.

Rose, P. W., Prlić, A., Altunkaya, A., *et al.* (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, **45**(D1), D271–D281.

Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., and Schroeder, M. (2015). PLIP: fully automated protein–ligand interaction profiler. *Nucleic acids research*, **43**(W1), W443–W447.

Samudrala, R. and Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, **275**(5), 895–916.

Schrödinger, LLC (2011). The PyMOL molecular graphics system, version 1.3.

Senior, A. W., Evans, R., Jumper, J., *et al.* (2019). Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (CASP13). *Proteins: Struct. Func. Bioinfo.*, **87**(12), 1141–1148.

Shen, C., Ding, J., Wang, Z., Cao, D., Ding, X., and Hou, T. (2020). From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **10**(1), e1429.

Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. (2018). Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.*, **59**(2), 895–913.

Trott, O. and Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **31**(2), 455–461.

Truchon, J.-F. and Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.*, **47**(2), 488–508.

Velec, H. F. G., Gohlke, H., and Klebe, G. (2005). DrugScoreCSD: 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*, **48**(20), 6296–6303.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved protein–ligand docking using GOLD. *Proteins: Struct. Func. Bioinfo.*, **52**(4), 609–623.

Virtanen, P., Gommers, R., *et al.* (2020). SciPy 1.0–fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**(3), 261–272.

Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*.

Wang, C. and Zhang, Y. (2017). Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *J. Comput. Chem.*, **38**(3), 169–177.

Wang, R., Lai, L., and Wang, S. (2002). Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.*, **16**(1), 11–26.

Wang, R., Fang, X., Lu, Y., Yang, C.-Y., and Wang, S. (2005). The PDBbind database: methodologies and updates. *J. Med. Chem*, **48**(12), 4111–4119.

Wang, S.-H., Wu, Y.-T., Kuo, S.-C., and Yu, J. (2013). Hotlig: A molecular surface-directed approach to scoring protein–ligand interactions. *J. Chem. Inf. Model.*, **53**(8), 181–2195.

Wätzig, H., Oltmann-Norden, I., *et al.* (2015). Data quality in drug discovery: the role of analytical performance in ligand binding assays. *J. Comput.-Aided Mol. Des.*, **29**(9), 847–865.

Zhang, J. and Zhang, Y. (2010). 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.

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Struct. Func. Bioinfo.*, **57**(4), 702–710.

Zheng, W., Li, Y., Zhang, C., Pearce, R., Mortuza, S., and Zhang, Y. (2019). Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins: Struct. Func. Bioinfo.*, **87**(12), 1149–1164.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, **23**(4), 550–560.