scientific reports

Check for updates

OPEN

New machine learning and physics-based scoring functions for drug discovery

Isabella A. Guedes[1,3], André M. S. Barreto[1], Diogo Marinho[1], Eduardo Krempser[2], Mélaine A. Kuenemann[3], Olivier Sperandio[3,4], Laurent E. Dardenne[1]✉ & Maria A. Miteva[3,5]✉

Scoring functions are essential for modern in silico drug discovery. However, the accurate prediction of binding affinity by scoring functions remains a challenging task. The performance of scoring functions is very heterogeneous across different target classes. Scoring functions based on precise physics-based descriptors better representing protein–ligand recognition process are strongly needed. We developed a set of new empirical scoring functions, named DockTScore, by explicitly accounting for physics-based terms combined with machine learning. Target-specific scoring functions were developed for two important drug targets, proteases and protein–protein interactions, representing an original class of molecules for drug discovery. Multiple linear regression (MLR), support vector machine and random forest algorithms were employed to derive general and target-specific scoring functions involving optimized MMFF94S force-field terms, solvation and lipophilic interactions terms, and an improved term accounting for ligand torsional entropy contribution to ligand binding. DockTScore scoring functions demonstrated to be competitive with the current best-evaluated scoring functions in terms of binding energy prediction and ranking on four DUD-E datasets and will be useful for in silico drug design for diverse proteins as well as for specific targets such as proteases and protein–protein interactions. Currently, the MLR DockTScore is available at www.dockthor.lncc.br.

Structure-based drug design and virtual screening have become common approaches for drug discovery. The predictive performance of scoring functions is essential for such methodologies[1–3]. However, accurate prediction of protein–ligand binding affinity remains a major challenge for current scoring functions. Despite the improvement over the last years of empirical, force-field or knowledge-based scoring functions, most of them still show unsatisfactory correlation with the experimental binding affinity or are based on meaningless description of protein–ligand interactions exhibiting overestimated accuracies in some cases[4–6].

Empirical scoring functions are based on a set of individual contributions or interaction descriptors calibrated by regression or statistical approaches using a training set of experimental affinity data for protein–ligand complexes[7,8]. Improvement of scoring functions can be achieved by developing new terms, training on larger high-quality datasets or using sophisticated machine learning-based algorithms for regression analysis, e.g. XGBoost and LightGBM boosting approaches[9–13]. Next, solvation and entropy contributions are key for ligand binding[14–20]. Although several previous scoring functions have considered such effects[14,15,17,19]common limitations of scoring functions are related to often neglecting them[10,21–23]. New scoring functions based on more precise physics-based descriptors to better represent protein–ligand recognition process are thus needed. Furthermore, a number of studies demonstrated that scoring functions performance is very heterogeneous across different target classes[22–26]. Target-specific scoring functions have shown to achieve better affinity prediction performance than general scoring functions trained over diverse protein families[21–23,27–29].

In this work, we developed a set of new empirical scoring functions, named DockTScore, to estimate protein–ligand binding affinity by explicitly accounting for physics-based interaction terms contributing to the binding free energy. Our models are based on the MMFF94S force field and trained and validated on high-quality large datasets properly curated. DockTScore scoring functions incorporate classical van der Waals and electrostatic energy terms, optimized terms accounting for solvation, lipophilic protein–ligand interactions and an improved estimation of ligand torsional entropy contribution to ligand binding for better describing

[1]Laboratório Nacional de Computação Científica, Petrópolis 25651-075, Brazil. [2]Fundação Oswaldo Cruz, Rio de Janeiro 21040-361, Brazil. [3]Inserm U973, Université Paris Diderot, Paris, France. [4]Structural Bioinformatics Unit, CNRS UMR3528, Institut Pasteur, 75015 Paris, France. [5]Inserm U1268 "Medicinal Chemistry and Translational Research", CiTCoM, UMR 8038, CNRS, Université de Paris, 75006 Paris, France. ✉email: dardenne@lncc.br; maria.mitev@inserm.fr

Protein (short name)	Total[a]	Affinities (kcal mol^{-1})	Training[b]	Test[c]
Bcl2-like/BAX	10	−12.636[d], −5.244[e]	7	3
Bromodomain2/Histone	2	−9.968, −8.561	2	0
Bromodomain4/Histone	11	−9.931, −6.145	9	2
K-Ras/SOS1	1	−4.712	1	0
MDM2-like/P53	20	−12.768, −6.737	14	6
Menin	1	−10.404	0	1
Xiap/Smac	7	−11.278, −5.378	6	1
E1/E2	1	−10.051	1	0
IL2/IL2R	1	−6.910	1	0
LEDGF/Integrase	4	−10.490, −6.676	2	2
ZipA/ftsZ	2	−6.685, −5.544	2	0
Total	60		45	15

Table 1. The iPPIs dataset. [a]Total number of protein–ligand complexes in the dataset. [b]Number of complexes in the training set. [c]Number of complexes in the random test set. [d]Binding affinity of the strongest protein–ligand interactions. [e]Binding affinity of the weakest protein–ligand complex.

of protein–ligand recognition. Firstly, we employed multiple linear regression (MLR)[30,31] to ensure a physical interpretation of the individual term contribution. Then, we developed more sophisticated nonlinear scoring functions using support-vector machine (SVM) for regression (named "SMOReg")[32] and random forest (RF)[33] algorithms using the theory-inspired physics-based terms selected from the initial MLR analysis. The development of scoring functions using physics-based descriptors representing protein–ligand recognition process together with the assessment of the accuracies of different linear and nonlinear models are important to avoid unrealistic overestimations of scoring functions accuracy due to some known biases, especially when training nonlinear models[4,6,34,35].

In addition to general scoring functions appropriate for diverse protein targets, we have developed MLR, SMOReg and RF scoring functions for two specific protein classes: proteases, and protein–protein interactions (PPIs) to be targeted by small-molecule inhibitors (iPPIs). Proteases are key drug targets, for which focused scoring functions have already been developed (e.g. targets such as HIV-1 protease[35]). Interestingly, only one work has been reported thus far aiming at developing a linear scoring function to predict the binding affinity of inhibitors of PPIs[36] using a training set of 27 PPIs complexes. Our MLR DockTScore for iPPIs gave new insights into the determinant factors contributing to inhibiting PPIs by small molecules. Moreover, we report here the first nonlinear scoring functions focusing on iPPIs and developed on 60 PPI complex structures carefully selected and curated. We evaluated the accuracy of affinity prediction and success of virtual screening to discriminate between active and decoys compounds of our scoring functions on four DUD-E datasets.

Methods

Data sets. *Data sets of diverse protein–ligand complexes for general scoring functions.* We trained and tested the general scoring functions appropriate for diverse protein targets based on the PDBbind v2013 refined set (http://www.pdbbind-cn.org/, version 2013), which is composed of 2959 protein–ligand complexes with binding affinity data manually collected from their original source[37–40]. PDBbind is known as the largest dataset of high-quality structures available for the development and validation of docking-scoring methods. The refined set was constructed according to several criteria concerning (i) the quality of the structures, (ii) the binding affinity data and (iii) the nature of the complex. Binding affinities in PDBbind comprise a large interval of values, ranging from 1.2 pM (1.2×10^{-12} M) to 10 mM (1.0×10^{-3} M). We converted the original binding constants to energy unit in kcal mol^{-1}.

The PDBbind core set, a subset of the refined set widely used as benchmarking data for evaluation of docking-scoring methods, was used here to assess the performance of our general scoring functions as an external test set only, not being used during the training step. The core set version 2013 is composed of 195 protein–ligand complexes carefully collected from the refined set for comparative studies of scoring functions[38–40].

Data sets for target-specific scoring functions. We selected a random subset from the PDBbind v2013 refined set according to specific ranges of the EC Number, (Enzyme Commission Number (EC Number) is a system of enzyme nomenclature that numerically classifies enzymes based on the chemical reaction catalyzed.) ranging from 3.4.11.0 to 3.4.25.69, to create a dataset for training and testing the scoring function focused for proteases, resulting in a subset composed of 783 structures (Table S1).

To create the dataset for inhibitors of protein–protein interactions (iPPIs), we took the X-Ray-based iPPIs dataset previously described in Kuenemann and colleagues[41], which was composed of 85 protein–ligand complexes. Here, we collected the binding affinity data from the original sources and manually prepared each complex using the Protein Preparation Wizard from Maestro (Maestro, version 9.7, Schrödinger, LLC, New York, NY, 2014). From the initial 85 iPPIs dataset, 25 complexes were removed due to their low resolution (value higher than 2.5 Å), the presence of covalently bound ligands or absence of affinity data. The remaining 60 structures were suitable for training and testing the specific scoring functions for iPPIs (Table 1).

Training and test sets. All datasets were randomly separated into a training set with 75% of the structures and an independent test set with the remaining 25% structures (Table S1). For the general scoring functions, the core set (N = 195) was extracted from the refined set, initially containing 2959 complexes. Thus, the random selection of complexes for the independent test and training sets was performed exclusively with the remaining 2764 complexes. The random 75% of the 2764 complexes used to train the general scoring functions is called "General::random" training set (N = 2073, Table S1. In addition, we tested the influence of the training data set size on the predictive capacity for the general scoring functions. Thus, we also trained general scoring functions using all the 2764 protein–ligand complexes (called here "General::all", Table S1). In this case, the predictive performance was evaluated only on the v2013 core set (N = 195).

For proteases, the training set was composed of 587 complexes and the test set was composed of 196 distinct complexes, not being used during the training step. Given the smaller size of the iPPI dataset, we characterized the composition of both training and test sets according to the protein families and the range of the binding affinity data (Table 1). Complexes of MDM2-like/P53 interacting with small ligands are the most frequent with 20 available structures, followed by complexes of Bromodomain4/Histone (11 complexes) and Blc2-like/BAX (10 complexes).

Preparation of the structures. Protein–ligand complexes of the v2013 refined set consist of the complete unit taken from Protein Data Bank (PDB)[42] (rcsb.org) and is available as prepared structures following an automatic procedure with some manual inspection performed by Li and colleagues[38]. Originally, the protein–ligand complexes were prepared following a simple protonation scheme considering a neutral pH: (i) all carboxylic acid and phosphate groups were deprotonated, and (ii) all aliphatic amine, guanidine and amidine groups were protonated. As well known, the correct assignment of both protein and ligand protonation/tautomeric states is crucial for correct binding mode and affinity predictions, but is a very time-consuming task for a large number of ligands[43–45]. In this work, we applied an improved protocol for the preparation of the structures of the v2013 refined set using the Protein Preparation Wizard from Maestro (Maestro, version 9.7, Schrödinger, LLC, New York, NY, 2014). Protonation assignment and hydrogen-bond optimization were performed using ProtAssign and PROPKA[46] considering the presence of the bound ligand. Protonation and tautomeric states of the ligand were calculated using Epik[47] (Epik, version 2.7, Schrödinger, LLC, New York, NY, 2014). Metal ions were considered as cofactors, and all waters were removed from the structures. Finally, energy minimization was performed to optimize the hydrogen atoms positions. A special attention was paid for the preparation of the core set due to its importance for the benchmarking studies. The protonation/tautomeric states of the binding-site residues and the bound ligand of the core set were further visually inspected and appropriate corrections were made guided by the original reference corresponding to the respective crystallographic structure and the Protoss program[48]. The curated core set (protein, ligand and cofactors) is freely available in the Supplementary Material. All structures of the iPPIs datasets and the proteases from DUD-E were prepared using the same protocol adopted for the core set.

Physics-based interaction terms. In this work, we implemented and evaluated several physicochemical terms contributing to the binding free energy to obtain pertinent descriptors for the derivation of the empirical scoring functions: protein–ligand electrostatic interactions (E_{coul}), van der Waals interactions (E_{vdW}), lipophilic contact interactions (E_{lipo}), polar (E_{polar_solv}) and nonpolar (E_{np_solv}) solvation contributions, and ligand torsional entropy contribution ($E_{entropy}$).

Electrostatic and van der Waals protein–ligand interactions. The protein–ligand electrostatic and van der Waals interactions are calculated using the MMFF94S force field[49,50]. The MMFF94S force field was parameterized using high-quality ab initio quantum–mechanical data and demonstrated to accurately reproduce protein–ligand binding geometry in docking studies[51,52]. The electrostatic interaction E_{coul} was calculated using:

$$E_{coul} = \frac{332.0716 q_i q_j}{\varepsilon\left(R_{ij} + \delta_{elec}\right)}$$

where q_i and q_j are the partial charges of atoms i and j, ε is the dielectric constant, R_{ij} is the distance between the centers of the atoms i and j, and $\delta_{elec} = 0.05$ is the electrostatic buffering constant. The partial charges q_i and q_j are calculated through a bond-charge-increment method starting from an initial formal charge of the atom i (q_i^0) and adding the bond-charge-increment contributions (ω_{ki}), which reflect the polarity of the covalent bonds of the atoms i and k:

$$q_i = q_i^0 + \sum \omega_{ki}$$

In this work, we evaluated two sigmoidal distance-dependent dielectric functions to consider the electrostatic screening due to the dielectric medium of protein–ligand complexes. The first one developed by Hingerty and colleagues[53] is currently implemented in the MMFF94S functional form used by the DockThor program for protein–ligand docking[51,52] (available as a web server at https://www.dockthor.lncc.br):

$$\varepsilon(r) = 78 - 77\left(\frac{r}{2.5}\right)^2 \frac{e^{r/2.5}}{\left(e^{r/2.5} - 1\right)^2}$$

where r is the internuclear separation between the atoms i and j.

The second dielectric function was formulated by Ramstein and Lavery, allowing to change both the maximal value of the dielectric constant (D) and the limiting value of the dielectric (D_i) when the interatomic distance approaches 0 ($\varepsilon(r) \rightarrow D_i$ when $r \rightarrow 0$)[54]. Here, we tested D_i values of 1 and 4 to simulate the relatively low dielectric at the interior of protein binding sites[55].

$$\varepsilon(r) = D - \left(\frac{D - D_i}{2}\right)\left[(rs)^2 + 2rs + 2\right]e^{-rs}$$

r is the internuclear separation between the atoms i and j, $s = 0.16$ is the slope of the sigmoidal segment and $D = 78$.

The van der Waals potential (E_{vdW}) as implemented in the MMFF94S force field representing a "Buffered 14–7" form[50] includes specific buffering constants δ_{vdW} and $\gamma = 0.12$:

$$E_{vdW} = \varepsilon_{ij}\left(\frac{(1+\delta_{vdW})R_{ij}^{*}}{R_{ij} + \delta_{vdW}R_{ij}^{*}}\right)^{7}\left(\frac{(1+\gamma)R_{ij}^{*7}}{R_{ij}^{7} + \gamma R_{ij}^{*7}} - 2\right)$$

where R_{ij} is the interatomic distance (Å), ε_{ij} is the well depth (kcal mol^{-1}) and R_{ij}^{*} is the minimum-energy separation (Å), which depends on the MMFF94S types of the atoms i and j. The original buffering constant $\delta_{vdW} = 0.07$ was replaced in this work by $\delta_{vdW} = 0.67$, which was empirically obtained to produce a more softened version of the van der Waals potential noted as E_{vdWS}.

Lipophilic protein–ligand interactions. We developed two descriptors E_{lipo} to calculate the lipophilic contact interactions effect E_{lipo} by summing all hydrophobic atom pairs between the ligand and the protein following the previously proposed functional forms in ChemScore[56] and X-Score[57] scoring functions. For each of them, the atoms considered for lipophilic contacts were: (i) all carbon atoms, or (ii) any non-hydrogen atom with MMFF94S partial charge q in the interval $-0.4 < q < +0.4$. We empirically estimated this range of partial charges through analysis of several protein–ligand complexes parameterized with the MMFF94S force field. The E_{lipo} descriptor for each lipophilic contact following e.g. the ChemScore is calculated by:

$$E_{lipo} = \begin{cases} 1, & d \leq d_{vdW} + 0.5\text{Å} \\ 1 - \frac{d - d_{vdW} + 0.5}{3}, & d_{vdW} + 0.5\text{Å} < d \leq d_{vdW} + 3.5\text{Å} \\ 0, & d > d_{vdW} + 3.5\text{Å} \end{cases}$$

where d is the distance between the pairs of atoms and d_{vdW} is the sum of their van der Waals radii.

Polar and nonpolar solvation contributions. In this work, the solvation contribution was calculated using a polar solvation term, which accounts for the loss of polar interactions of the charged groups of both protein and ligand with the solvent, and a nonpolar solvation term, which reflects the desolvation of the hydrophobic protein and ligand groups due to binding. The polar solvation term E_{polar_solv} was calculated by summing up the number of charged atoms becoming buried after the complex formation and not interacting with a charged atom in the protein–ligand complex. In this term, two charged atoms were considered as interacting if the distance between them (d) was equal to or lower than $d_{vdW} + 1.0$Å, where d_{vdW} is the sum of their van der Waals radii. A charged atom was defined as a non-hydrogen and a non-carbon atom with a partial charge $|q| > 0.8$.

The nonpolar solvation E_{np_solv} was calculated based on the total loss of the solvent-accessible surface area (SAS) of the protein and the ligand due to the binding converted into energy (E_{np_solv} in kcal mol^{-1}) following Kuhn and Kollman[58]. The SAS of atoms in the free and complexed states was calculated with the program MSMS[59].

$$E_{np\_solv} = G_{np}complex - \left(G_{np}protein + G_{np}ligand\right)_{free}$$

where G_{np} is calculated by:

$$G_{np} = 0.0092 * SAS + 0.82$$

Ligand torsional entropy contribution. We revisited here the ligand torsional entropy term based on the conformational component of the ligand entropy and arising from the loss of the torsional degrees of freedom for a flexible ligand upon binding. Instead of a crude approximation based on the total number of all rotatable bonds[14–17,19], we propose an improved estimation of the lost torsional freedom of the ligand by considering only the rotatable bonds, which become "frozen" due to binding. Similar approaches were previously adopted to approximate protein side-chain entropic contributions[15,60].

The bonds are considered as "frozen" based on the change of the solvent-accessible surface areas of the ligand atoms directly involved in each rotatable bond, aiming to penalize only dihedrals that are unable to rotate after the complex formation.

Firstly, each rotatable bond of the ligand (Fig. 1A) is divided into two sides for the two atoms i and j (Fig. 1B,C). Each side is composed of (i) the atom i, which is directly involved in the bond (symbol *), and (ii) the first neighbors of the atom i (symbol +). The same procedure is applied to the other side (atom j). The change of the SAS (ΔSAS) for each side upon the binding is computed taking into account all atoms of the side. If SAS

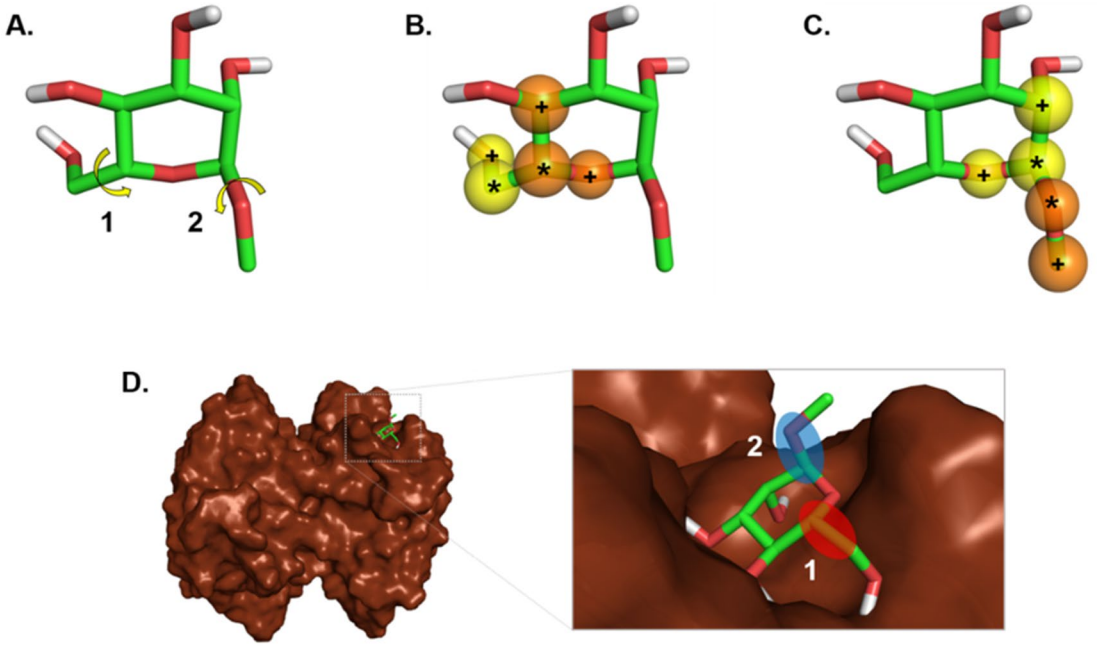

Figure 1. Illustration of the algorithm for computing the ligand torsional entropy term. (**A**) Selection of the rotatable bonds in the ligand. (**B** and **C**). Each rotatable bond is divided into two sides (i in yellow and j in orange) and the root (*) and the neighboring (+) atoms are detected. (**D**) A rotatable bond is considered as frozen if both sides become buried with more than 50% due to the binding (case 1). If at least one side does not become buried with more than 50% due to the binding the rotatable bond is not taken into consideration (case 2).

decreases ≥ 50% for the two sides, the rotatable bond is considered as frozen due to the binding. We consider that a hiding of a rotatable bond by more than 50% is significant for the ligand flexibility, and thus critical to the change of the ligand entropy due to the binding. In fact, the protein receptor is kept rigid during the docking and slight protein movements could compensate for a small change of the SAS of a ligand rotatable bond. We thus take into consideration only those bonds becoming frozen due to the binding for the ligand torsional entropy contribution estimation (Fig. 1D).

Derivation of linear scoring functions. We performed the selection of the descriptors based on the assumption that the major contributions to the free energy of binding are the intermolecular interactions, represented by the van der Waals and electrostatic interactions between the protein and the ligand, and the solvation and entropy changes due to the binding. We developed thus independent descriptors accounting for van der Waals and electrostatic interactions, protein–ligand lipophilic contacts, the change of the conformational entropy of the ligand, and polar/nonpolar solvation contribution to the binding (see their definition in "Physics-Based Interaction Terms"). Then, we selected the best descriptors (see below), assuring that all above mentioned classes of interactions have been present in the final scoring functions, instead of using a combinatorial or sequential descriptors selection.

We applied multiple linear regression (MLR) ensuring a physical interpretation of the individual terms' contributions. A tenfold cross-validation was used to select the best performing physics-based descriptors. This initial descriptor selection was applied only for the derivation of the general scoring function since it was trained with the largest training set containing diverse protein–ligand complexes. We started with the basic function F_{MMFF} containing the electrostatic term with the Ramstein dielectric function tending to 4, $Di = 4$, (E_{coul4}) and the soft van der Waals term (E_{vdWS}) based on the original MMFF94S force field. These two terms were selected since they achieved the best correlation among four combinations tested for the electrostatic and vdW terms (see Table S2).

Then, each of the remaining physics-based descriptors (lipophilic contacts, entropy, polar solvation and nonpolar solvation) was individually added to the basic function F_{MMFF} one at a time, to find the best variation for each of them leading to the best correlation on cross-validation experiments. Thus, the combinations evaluated herein were: F_{MMFF} + lipophilic contacts (4 variants), F_{MMFF} + entropy, F_{MMFF} + polar solvation, and F_{MMFF} + nonpolar solvation. The correlations obtained for all combinations are present in the Supplementary Material (Tables S3 and S4). We considered the best variation of each specific term to finally combine them into the general scoring function ($F_{final} = F_{MMFF}$ + lipophilic contacts + ligand conformational entropy + polar solvation + nonpolar solvation). Next, the best combination of terms of the general scoring function was applied to the class-specific scoring functions, and was also used for the descriptors in the development of nonlinear scoring functions with machine learning methods.

Derivation of nonlinear scoring functions. In this work, we also developed nonlinear scoring functions using the Support Vector Machine for Regression (SMOReg) and Random Forest (RF) algorithms. Such scoring functions were trained using the same physics-based descriptors selected for the final linear scoring functions.

Support Vector Machine (SVM) aims to find the hyperplane that maximizes the margin of separation between data classes. In particular, in the kernel application the original nonlinear separable data can be transformed to a linear hyperplane separable problem on a higher dimension space[61]. SMOReg uses the sequential minimal optimization (SMO) for training support-vector machines (SVM) models in regression problems. In regression problems, all prediction errors less than a value of ε are ignored (*insensitive-loss function*)[30,62]. This strategy reduces the risk of overfitting on the training set and is controlled by the complexity parameter *C*, which is user-defined together with ε.

Random Forests (RF) were introduced by Breiman in 2001 as a powerful strategy for ensemble learning[33]. The RF combines several random trees (*numTrees*) in a bagging ensemble model, often leading to excellent results in diverse classification problems[33,62]. The output variable of a RF model is usually an average value of the predictions of the regression trees (as used in this work), where the node splitting is performed using a finite subset of features randomly chosen (*numFeatures*).

All the machine-learning procedures were carried out using the Weka v3.8.3 package[30]. We explored diverse configurations of SMOReg and RF on a tenfold cross-validation procedure. For SMOReg, we varied the complexity parameter *C*, tolerance in loss function epsilon (ε), kernel (*puk* or *rbf*), gamma (γ) of the *rbf* kernel, and sigma (σ) and omega (ω) of the *puk* kernel. In the RF training, we explored the number of trees (*numTrees*) and the number of features that are randomly chosen for splitting the parent node (*numFeatures*).

The tested learning parameters and their optimal values found are present in Tables S5 and S6, respectively (see Supporting Information).

Validation of the scoring functions. *Binding affinity accuracy.* The best model of each machine-learning algorithm was selected according to the Pearson's Correlation Coefficient (*R*) using the tenfold cross-validation strategy. Then, we applied the scoring functions to the respective test sets to validate their affinity predictability according to *R* and root mean squared error (*RMSE*). Both *R* and *RMSE* were calculated using the experimental and predicted free energy of binding (ΔG_{bind}):

$$R = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(t_i - \bar{t})}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{N} (t_i - \bar{t})^2}}$$

where y_i and t_i are respectively the predicted and the experimental binding affinities for the i-th complex, $\bar{y}$ and $\bar{t}$ are the arithmetic average values for y and t and N is the number of points in the data set.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - t_i)^2}$$

where N is the number of points in the dataset, y_i is the predicted binding affinity and t_i is the experimental binding affinity.

Virtual screening experiments. In order to evaluate the success of our scoring functions to discriminate active and decoys compounds, we performed docking experiments using the protein–ligand docking program DockThor[51,52] and re-scoring with DockTScore on core set and the DUD-E datasets[63] for the proteases FA7 (coagulation factor VII, PDB code 1W7X), RENI (renin, PDB code 3G6Z), TRYB1 (tryptase β1, PDB code 2ZEC), and UROK (urokinase-type plasminogen activator, PDB code 1SQT), and the kinases AKT2 (serine/threonine-protein kinase AKT2, PDB code 3D0E), KIT (stem cell growth factor receptor, PDB code 3G0E) and MK01 (MAP kinase ERK2, PDB code 2OJG). Proteases were selected to evaluate the screening success of the DockTScore general and target-specific scoring functions trained on the PDBbind refined set due to the large size of the training set used to calibrate the focused scoring functions for proteases. The protease and kinase datasets from DUD-E were chosen according to the following criteria: (i) no metal ions interacting with the ligand, and (ii) co-crystallized ligand successfully redocked with the top-energy solution with RMSD ≤ 2.0 Å. For PPIs, we constructed screening datasets for Bcl2-like/BAX and MDM2/p53 systems composed of actives taken from the iPPI-DB[64] database (https://ippidb.pasteur.fr/) and inactive compounds taken from the BDM chemical library available at ChemREST (https://chem-rest.pasteur.fr/#?&versioned_sources=8&used_filters=). The iPPI-DB is a database that contains the structure, some physicochemical characteristics, the pharmacological data and the profile of about 2000 modulators of protein–protein interactions. It contains exclusively small molecules and therefore no peptides. BDM compounds have been previously shown to be negative on MDM2 and Bcl2 interactions via fluorescence polarization assays[65]. For the PPIs screening datasets, we selected only the compounds without chiral centers and having only one protonation/tautomer state as predicted by Epik. Following the DUD-E sets construction, we selected randomly 50 inactives for each active compound to keep an adequate balance between actives and inactives to evaluate the scoring functions performance on virtual screening experiments. The PDB codes 3QKD and 4IPF were used for the receptor structures of the Bcl-2-like protein 1 and MDM2, respectively.

The docking poses were generated with the program DockThor for protein–ligand docking freely available as a web server at https://dockthor.lncc.br). The DockThor program uses a grid box to define the search space, the DMRTS genetic algorithm as the search algorithm, and an MMFF94S-based scoring function for pose

Scoring functions	E_{coul4}	E_{vdWS}	E_{lipo}	$E_{entropy}$	E_{polar_solv}	E_{np_solv}	c_0
General::random[a]	0.0039	0.0386	−0.0111	0.0560	0.1025	0.0169	−5.5197
General::all[b]	0.0045	0.0343	−0.0104	0.0605	0.0987	0.1180	−5.5178

Table 2. Coefficients of the terms obtained for the general scoring functions trained with MLR. [a]Scoring function trained with the random training set (N = 2073). [b]Scoring function trained with the refined set minus core set (N = 2764).

prediction[51,52]. Configuration of the search space of each protein target was automatically determined according to the reference ligand: (i) the center of coordinates was defined as the center of coordinates of the ligand, (ii) the grid size was defined as the largest axis value of the ligand plus a tolerance of 6 Å on each dimension, (iii) the discretization (*i.e.* spacing between two points of the grid) was set to the default value of 0.25 Å except for the cases where the grid size was greater than 26 Å. The parameters of the search algorithm were set as follows for redocking experiments: (i) 24 docking runs, (ii) 1,000,000 evaluations on each docking run, (iii) initial population of 1,000 individuals. The MMFF94S-based scoring function for ranking the docking poses (E_{total}) consists of (i) the torsional, electrostatic and Buf-14–7 van der Waals potential terms for the internal energy, and (ii) the electrostatic and Buf-14–7 van der Waals potential terms for the intermolecular interactions. The docking poses are clustered using our in-house tool *dtstatistic* using a criterion of diversity equals to 2.0 Å.

The screening experiments were performed using the computational facilities provided by the Brazilian SINAPAD (*Sistema Nacional de Alto Desempenho*, https://www.lncc.br/sinapad/) high-performance platform and the Supercomputer SDumont. We used a set of GA parameters named "virtual screening" for the screening experiments used to reduce the computational cost, consisting of 12 docking runs, 500,000 GA evaluations and initial population of 750 individuals. The top-energy docking pose ranked by the total energy E_{total} were selected for the virtual screening experiments and binding affinity predictions.

The screening success was evaluated according to the area under the curve for the receiver operation characteristics (ROC AUC), the enrichment factor at 1% of the screened libraries (*i.e.,* $EF_{1\%}$), and the Boltzmann-enhanced discrimination of ROC values ($\alpha = 20$ and $\alpha = 100$, respectively BEDROC20 and BEDROC100)[66] using the open-source tool for virtual screening analysis Rocker[67].

Results

Performance of physics-based terms for the scoring functions. The best correlation between the predicted and experimental affinities (R = 0.493) using tenfold cross-validation on the General::random training set (N = 2073) with MLR for a scoring function accounting only for E_{vdW} and E_{coul} was obtained with our softened version of the Buf-14-7 van der Waals potential (E_{vdWS}, with $\delta_{vdW} = 0.67$) and the electrostatic term using the sigmoidal dielectric function of Ramstein and Lavery[58] with $D_i = 4$ (Table S2), noted here as E_{coul4}. The scoring function composed of only E_{vdWS} and E_{coul4} terms is noted in this work as the "basic scoring function" F_{MMFF}. No correlation was obtained in cross-validation experiments (R = 0.053) using only the two original MMFF94S force field terms E_{vdW} Buf-14–7 (with $\delta_{vdW} = 0.07$) and E_{coul} ($D_i = 1$). It is interesting to note that the best correlation was obtained with the softened version E_{vdWS} , which is expected because no energy minimization of the complex structures was performed. Soft vdW potentials are more permissive for small clashes that can be present, in particular in structures generated by molecular docking without subsequent energy minimization. For X-ray derived structures shorter non-bonded atom–atom distances may be present when compared to energy minimized structures through classical force fields optimizations. Indeed, when dealing with non-optimized structures such as those used in X-ray models, it is indicated to softening the Buf-14-7 potential increasing the δ_{vdW} buffering constant[50]. The E_{lipo} lipophilic contact term provided better results when nonpolar atoms were defined based on the MMFF94S partial charges instead of considering only carbon atoms, achieving here a Pearson correlation of R = 0.538 when added to the F_{MMFF} basic scoring function (Table S3). This result indicated that our description of the atom types according to their partial atomic charges, specific for the MMFF94S force field is relevant. Adding our original and simple term for the polar solvation also improved the accuracy of the basic scoring function F_{MMFF} (R = 0.514). Similarly, adding the nonpolar solvation term to F_{MMFF} improved the correlation in tenfold cross-validation experiments (R = 0.503). In the same line, our proposed improved term for ligand torsional entropy contribution demonstrated to be important for the affinity prediction when associated with the basic scoring function, improving its correlation on cross-validation experiments (R = 0.507). The observed improvement due to our individual physics–based terms permitted their validation for further training of the general and target-specific empirical scoring functions.

General scoring functions. The MLR coefficients obtained for the general scoring functions considering all validated six terms are shown in (Table 2). As expected, the coefficients are in accordance with the physical meaning of the corresponding terms (*i.e.,* favorable or unfavorable contribution). Energy terms such as van der Waals, electrostatic and nonpolar solvation increase the binding affinity when the associated coefficients have positive values and the corresponding interactions for E_{coul} and E_{np_solv} are favorable for the binding. The empirical term related to the counting of the lipophilic atom pairs has a favorable contribution as the associated coefficient has a negative value. The polar solvation and the entropy terms are unfavorable as the coefficients are positive.

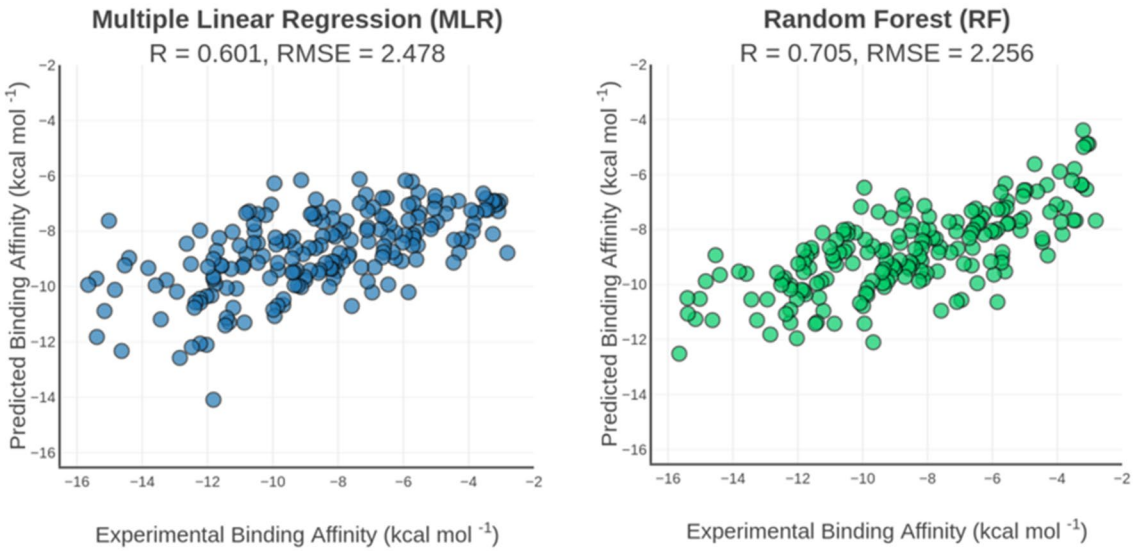

Figure 2. Correlation plot of the experimental and predicted binding affinities by the MLR (left) and RF (right) general scoring functions. Models trained on the PDBbind v2013 refined set (N = 2764) and evaluated on curated v2013 core set (N = 195). R is the Pearson's correlation coefficient and RMSE is the root mean squared error given in kcal mol^{-1}.

MLR general scoring function trained with the random training set (N = 2073) exhibited a good performance on tenfold cross-validation experiments (R = 0.548) and on the curated core set (R = 0.602), and a lower performance on the random test set (R = 0.494) (Table S7). Our MLR general scoring function has predictive capacity comparable to the best evaluated linear scoring functions, with performance close to X-Score:HMScore (R = 0.614) and X = Score::SAS (R = 0.606) reported in the v2013 core set benchmark paper[39].

According to the tenfold cross-validation in the random general training set (N = 2073), it is seen that the SMOReg and RF models outperformed the MLR model, providing significantly better performances with R = 0.653 and R = 0.655, respectively (Table S7). These results confirm previous findings that nonlinear regression may better predict binding affinities than MLR and that the additive assumption adopted in the linear scoring functions could be too restrictive[68]. Using two different size training sets, the *General::all* one (N = 2764) and the *General::random* one (N = 2073) did not change the predictive performance of MLR model (R = 0.601 vs R = 0.602) while the larger training set improved the predictive performance of the SMOReg and RF models on the core set (Fig. 2 and Table S7), respectively $R_{SMOReg} = 0.668$ vs $R_{SMOReg} = 0.687$ and $R_{RF} = 0.678$ vs $R_{RF} = 0.705$. These results are consistent with other studies evaluating the influence of the training size, indicating that nonlinear scoring functions increase performance when more data is included in the training set while linear models seem to be less sensitive to the training set size[69,70].

Target-specific scoring functions. *Proteases.* The linear scoring function for proteases exhibited good performance on the cross-validation experiments (R = 0.614) and on the independent test set (R = 0.653) (Fig. 3). All coefficients were very similar to those obtained for the general scoring function and their signals were in accordance with the physical meaning of the corresponding terms (Table 3). Likewise to the results observed for general scoring function, the nonlinear models for proteases exhibited significant improvements in the prediction capacity for both tenfold cross-validation experiment ($R_{SMOReg} = 0.749$ and $R_{RF} = 0.735$) and the independent test set ($R_{SMOReg} = 0.730$ and $R_{RF} = 0.723$).

Protein–protein interactions (PPI). For the iPPI linear scoring function, the representation of solvation as two independent terms leads to an unexpected favorable contribution of polar solvation instead of penalizing the buried charged atoms not involved in charge-charge interactions (Table 4). Thus, we decided to consider a single term for both polar and nonpolar solvation (called "oneSolv"), which has the same functional form of the nonpolar term but taking into account all heavy atoms, i.e., both polar and nonpolar ones. The solvation term "oneSolv" performed slightly better for the PPI-specific scoring function on cross-validation than using two solvation terms (R = 0.552 versus R = 0.545). Comparing the magnitude of the coefficients in the "oneSolv" model, the entropic and electrostatic terms exhibited a significantly higher contribution for iPPIs (Table 4). It has been widely demonstrated that iPPIs have higher hydrophobicity, aromaticity and molecular weight compared to enzyme inhibitors, as usually interacting within flatter, larger and more hydrophobic binding sites than the enzyme catalytic sites[41,71,72]. Given this, it is expected that the hydrophobic effect due to the binding represented here by the lipophilic contact and "oneSolv" solvation terms exhibit a strongly favorable contribution for this class of complexes. The unfavorable contribution of the E_{vdWS} term might be due to some overlapping with the lipophilic contact and the "oneSolv" solvation terms. Further, a larger dataset set would allow to better evaluate the solvation contribution for inhibiting PPI.

Figure 3. Correlation plot of experimental and predicted binding affinities by MLR (left) and SMOReg (right) specific scoring functions for proteases. The scoring functions were evaluated on the independent test set for proteases (N = 196). R is the Pearson's correlation coefficient and RMSE is the root mean squared error given in kcal mol^{-1}.

Scoring functions	E_{coul4}	E_{vdWS}	E_{lipo}	$E_{entropy}$	E_{polar_solv}	E_{np_solv}	c0
Proteases	0.0089	0.0399	−0.1120	0.0153	0.0515	0.0809	−4.8954

Table 3. Coefficients of the terms obtained for the protease-specific scoring functions trained with MLR.

Scoring functions	E_{coul4}	E_{vdWS}	E_{lipo}	$E_{entropy}$	E_{polar_solv}	E_{np_solv}	c_0
iPPIs	0.0505	0.0024	−0.0130	0.1967	−0.1698	1.0569	−0.7898
iPPIs-oneSolv	0.0335	−0.0207	−0.0153	0.2038	1.1227		−1.1397

Table 4. Coefficients of the terms obtained for the iPPI-specific scoring functions trained with MLR.

Regarding the ligand entropy, it is clearly unfavorable for the binding. We expect that our improved entropic term penalizing only frozen rotatable bonds instead of all rotatable bonds is particularly important for the PPI class taken into account the large size of iPPIs and thus a possibly larger number of rotatable bonds. To confirm this hypothesis, we evaluated the linear scoring function for iPPIs on tenfold cross-validation experiments using the commonly used total number of rotatable bonds instead of the number of frozen torsions, and we obtained a slightly reduced correlation (R = 0.515). In this context, our entropic term demonstrated to be more appropriate for iPPIs than the total number of rotatable bonds.

As expected, the nonlinear scoring functions specific for iPPIs, mainly the SMOReg model, improved the predictive performance when compared with the MLR model (Fig. 4), obtaining correlations of $R_{SMOReg} = 0.600$ and $R_{RF} = 0.666$ on the tenfold cross-validation, and $R_{SMOReg} = 0.613$ and $R_{RF} = 0.478$ on the test set. Curiously, despite the RF performing better on the tenfold cross-validation, the SMOReg model achieved a real improvement on the test set.

Virtual screening. In general, the DockTScore functions performed well in virtual screening experiments for the proteases (Table 5 and Fig. 5). According to the results, the best models achieved AUC ROC values better than 0.70 in most of the cases, while the early recognition of active compounds according to the $EF_{1\%}$ and the BEDROC values was variable between the different proteases studied, keeping in mind that BEDROC100 is very exigent for the early recognition of actives. Following the same trend observed for the binding affinity prediction, the nonlinear models generally performed better than the MLR models in terms of the screening success. Best results were obtained when using the specific scoring functions for proteases with the SMOreg model being the best-performing scoring function to distinguish actives from decoys. As an exception, the general and target-specific scoring functions exhibited low predictive performance for the TRYB1 target, with AUC ROC values lower than 0.651, a maximum EF1% only of 8, BEDROC20 of 0.203, BEDROC100 of 0.167. In this case, the accuracy is very low, taking into consideration that depending on the library size, often one can screen experimentally about 1% of the in silico screened compounds. The TRYB1 is a particular case, its binding

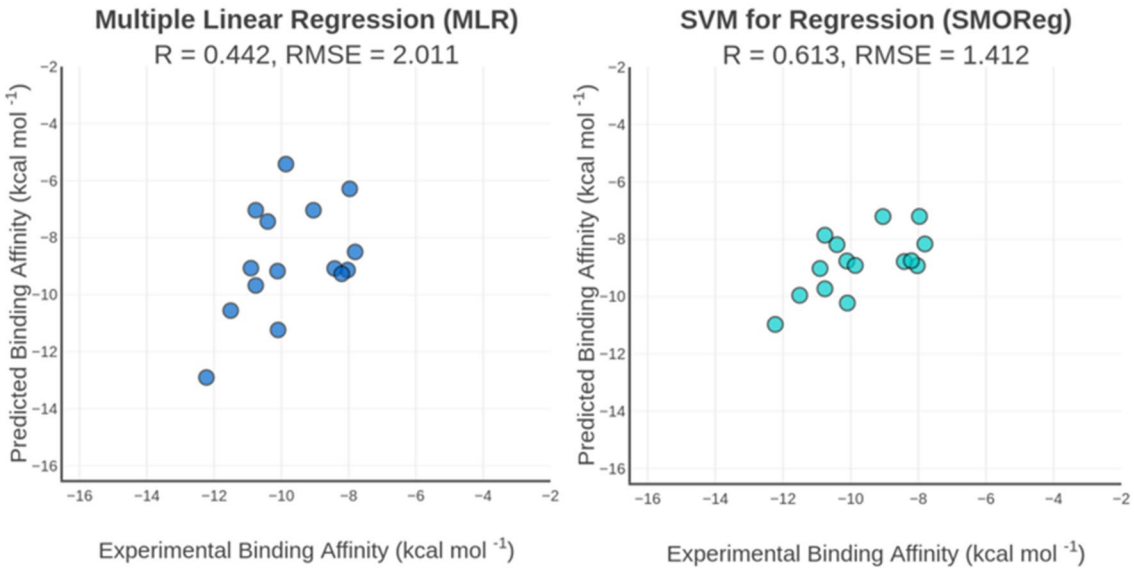

Figure 4. Correlation plot of predicted and predicted binding affinity by MLR (left) and SMOReg (right) specific scoring functions for iPPIs using one solvation term evaluated on the independent test set for iPPIs (N = 15). R is the Pearson's correlation coefficient and RMSE is the root mean squared error given in kcal mol^{-1}.

Target	Metrics	General SFs			Protease-specific SFs		
		MLR	SMOreg	RF	MLR	SMOreg	RF
FA7	AUC	0.789	0.860	0.875	0.818	0.893	0.869
ac = 112	EF1% (max = 52.973)	8.979	9.876	8.979	12.570	17.059	17.059
dec = 5,821	BEDROC20	0.299	0.346	0.328	0.350	0.478	0.397
tot = 5,933	BEDROC100	0.181	0.181	0.165	0.230	0.333	0.310
RENI	AUC	0.786	0.769	0.763	0.807	0.771	0.782
ac = 73	EF1% (max = 86.425)	16.462	20.577	10.975	17.834	16.462	8.231
dec = 6,236	BEDROC20	0.300	0.334	0.271	0.349	0.346	0.268
tot = 6,309	BEDROC100	0.253	0.281	0.155	0.283	0.207	0.119
TRYB1	AUC	0.619	0.649	0.614	0.651	0.651	0.633
ac = 147	EF1% (max = 51.633)	1.359	1.359	2.038	4.076	7.473	8.153
dec = 7,443	BEDROC20	0.099	0.103	0.080	0.141	0.203	0.169
tot = 7,590	BEDROC100	0.037	0.040	0.046	0.080	0.167	0.167
UROK	AUC	0.740	0.774	0.775	0.762	0.814	0.788
ac = 129	EF1% (max = 69.837)	7.760	8.536	6.208	11.640	14.743	10.088
dec = 8,880	BEDROC20	0.262	0.306	0.295	0.295	0.352	0.283
tot = 9,009	BEDROC100	0.123	0.147	0.118	0.179	0.232	0.182

Table 5. Screening success of the general and target-specific scoring functions trained with MLR, SMOreg and RF for the FA7, RENI, TRYB1 and UROK datasets from DUD-E. *ac*, *dec* and *tot* are the number of active, decoy compounds and the total number of molecules in the final dataset (i.e., compounds that were docked and rescored with DockThor and DockTScore, respectively). Only the top-scored protonation state of a compound according to each scoring function (SF) was kept.

site is remarkably exposed to the solvent. It is located in the interface of the two TRYB1 monomers belonging to the active tetramer[8] sharing thus PPI-like properties. The co-crystallized ligand is bound with only one "frozen" rotatable bond in the dimer out of four rotatable bonds (Fig. 6). Therefore, we also evaluated the performance of the DockTScore PPI-specific scoring functions on the TRYB1 target (Fig. 7). Interestingly, the PPI-specific MLR scoring function outperformed the other scoring functions evaluated (*i.e.,* general and protease-specific, linear and nonlinear), achieving an AUC ROC curve of 0.762 ($SMOreg_{protease}$ was 0.651), $EF_{1\%}$ = 15.626 ($SMOreg_{protease}$ was 7.473), BEDROC20 = 0.291 ($SMOreg_{protease}$ was 0.203) and BEDROC100 = 0.272 ($SMOreg_{protease}$ was 0.167).

The screening of actives and inactives on the two PPIs datasets resulted in AUC values better than 0.70 for the two targets for almost all scoring functions (Table 6 and Fig. 8), while the early recognition problem was successfully addressed only for the Bcl2-like system, reaching high BEDROC values of 0.474 (α = 20) for SMOreg and 0.539 (α = 100) for MLR. For the Bcl2-like protein/BAX system, the SMOreg scoring functions generally

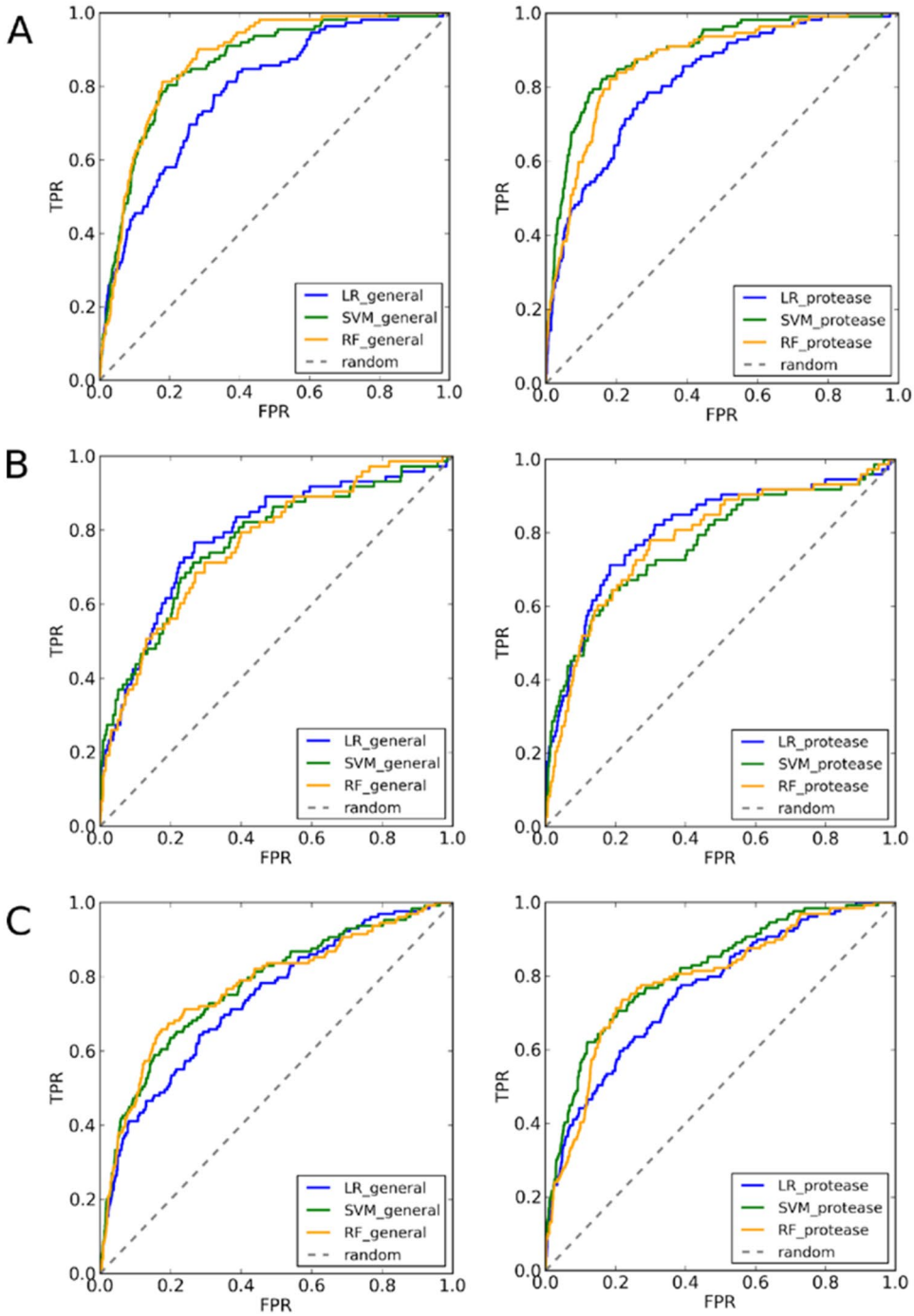

Figure 5. AUC ROC curves of the general (left) and protease-specific scoring functions (right) trained with MLR, SMOreg and RF for the FA7 (**A**), RENI (**B**), and UROK (**C**) datasets from DUD-E.

outperformed the other machine learning methods, whereas the PPI-specific scoring functions improved the $EF_{1\%}$ and BEDROC for all algorithms. Interestingly, the linear PPI-specific scoring function, with a satisfactory AUC ROC value of 0.709, obtained the best $EF_{1\%}$ value and the highest BEDROC100 value of 0.539. In the case of MDM2 target, the nonlinear general scoring functions outperformed the specific models in terms of AUC ROC, whereas the RF-based achieved the best overall screening performance. However, for this target all methods exhibited insufficient early recognition capacity according to the $EF_{1\%}$ and the BEDROC values.

In addition to the proteases and PPIs targets, we also evaluated the performance of our general scoring functions trained with MLR, SMOreg and RF on three protein kinases datasets taken from DUD-E. Kinases are considered as challenging targets mainly due to binding site flexibility, which frequently leads to induced-fit effects due to ligand binding. Although DockTScore is not developed to deal with the receptor flexibility, our scoring functions exhibited satisfactory performances for two out of three kinases in virtual screening experiments, with

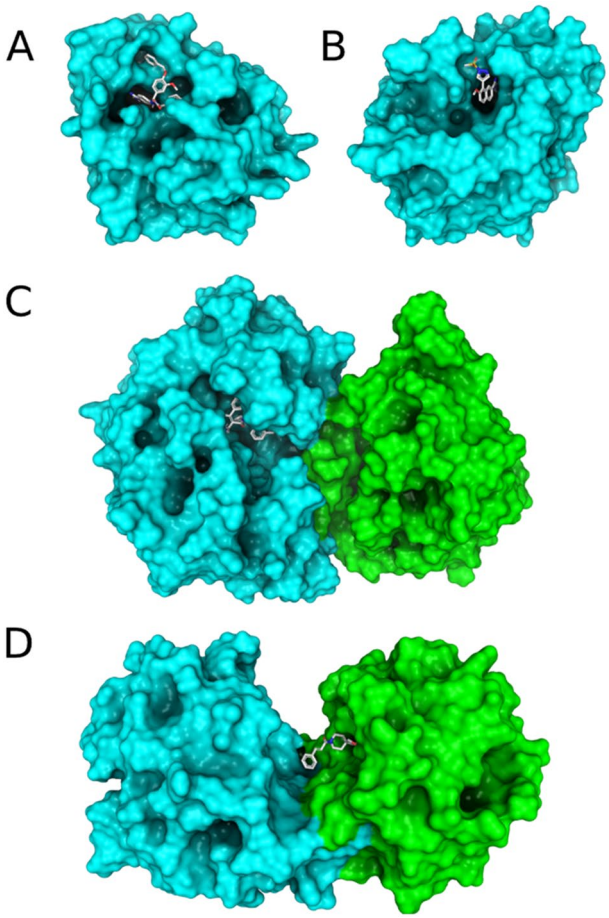

Figure 6. Surface representation of the binding sites of the proteases (**A**) FA7, (**B**) UROK, (**C**) RENI, and (**D**) TRYB1 colored by chain. The co-crystallized ligand is represented as sticks.

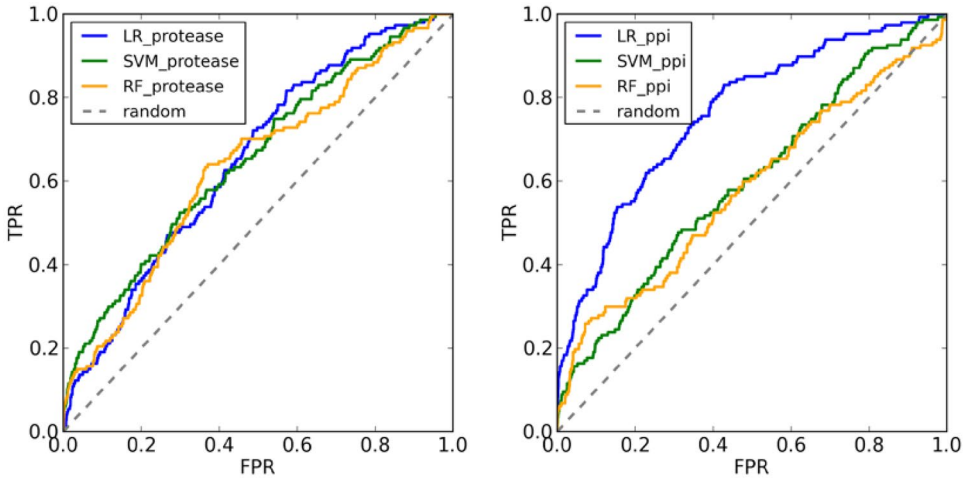

Figure 7. AUC ROC curves of the protease-specific (left) and PPI-specific (right) scoring functions trained with MLR, SMOreg and RF for the TRYB1 datasets from DUD-E.

AUC ROC values higher than 0.745 (Table 7 and Fig. 9). It is interesting to note that for AKT2 and MK01 targets, the MLR function showed better values for early the recognition metrics (e.g., EF, BEDROC20 and BEDROC100) than the SMOreg (AKT2and MK01) and RF (only for MK01) nonlinear functions. However, for the KIT target all the functions achieved insufficient performance for all evaluated metrics. It is important to note that in the screening experiments, we used a softened version of the MMFF94S Buf-14-7 force field to implicitly account

Target	Metrics	General SFs			PPI-specific SFs		
		MLR	SMOreg	RF	MLR	SMOreg	RF
Bcl2-like protein/BAX	AUC	0.755	0.838	0.740	0.709	0.801	0.716
ac = 98	EF1% (max = 51.510)	22.664	20.604	20.604	29.876	23.695	22.664
inac = 4,950	BEDROC20	0.370	0.375	0.330	0.471	0.474	0.418
tot = 5,048	BEDROC100	0.386	0.368	0.378	0.539	0.445	0.430
MDM2/p53	AUC	0.741	0.791	0.794	0.736	0.654	0.553
ac = 114	EF1% (max = 50.991)	4.400	4.400	6.154	2.637	1.758	5.275
inac = 5,699	BEDROC20	0.204	0.251	0.262	0.163	0.114	0.117
tot = 5,813	BEDROC100	0.010	0.112	0.124	0.068	0.042	0.090

Table 6. Screening success of the general and PPI-specific scoring functions trained with MLR, SMOreg and RF evaluated on the Bcl2-like protein/BAX and MDM2/p53 datasets. *ac*, *inac* and *tot* are the number of active, inactive compounds and the total number of molecules in the final dataset (*i.e.,* compounds that were docked and rescored with DockThor and DockTScore, respectively). Only the top-scored protonation state of each compound according to each scoring function (SF) was kept.

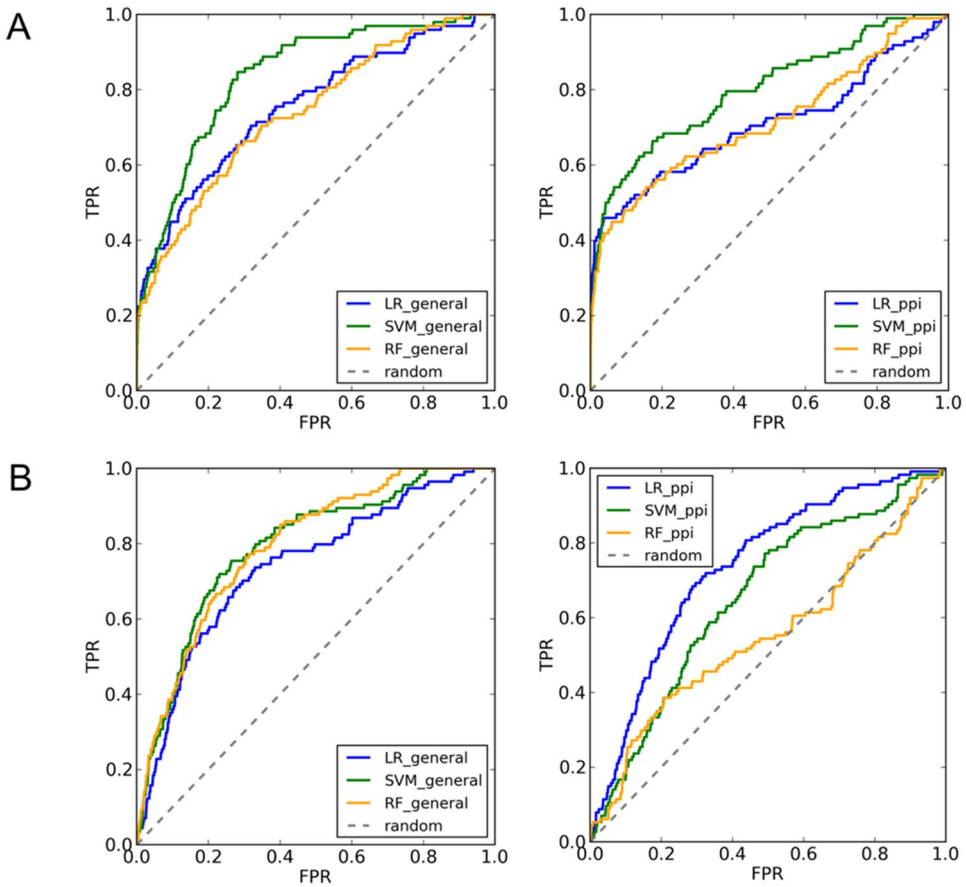

Figure 8. AUC ROC curves of the general (left) and PPI-specific (right) scoring functions trained with MLR, SMOreg and RF for the Bcl2-like/BAX (**A**) and MDM2/p53 (**B**) datasets.

for the protein flexibility to some extend explicitly permitting small clashes by reducing the repulsive energy between the protein–ligand atoms. However, the use of strategies that account for large movements of the binding site, such as ensemble docking with more than one representative structure of the protein, might be necessary to achieve better screening results on highly flexible systems such as kinases.

Target	Metrics	General SFs		
		MLR	SMOreg	RF
AKT2	AUC	0.769	0.800	0.814
ac = 116	EF1% (max = 60.414)	24.166	15.535	13.809
dec = 6,892	BEDROC20	0.421	0.378	0.379
tot = 7,008	BEDROC100	0.394	0.288	0.269
KIT	AUC	0.640	0.635	0.657
ac = 166	EF1% (max = 63.934)	3.016	2.413	5.428
dec = 10,447	BEDROC20	0.148	0.146	0.176
tot = 10,613	BEDROC100	0.063	0.043	0.090
MK01	AUC	0.786	0.766	0.745
ac = 78	EF1% (max = 59.308)	10.314	12.893	7.736
dec = 4,548	BEDROC20	0.352	0.364	0.340
tot = 4.626	BEDROC100	0.153	0.220	0.193

Table 7. Screening success of the general scoring functions trained with MLR, SMOreg and RF evaluated on the AKT2, KIT, and MK01 datasets from DUD-E. *ac*, *dec* and *tot* are the number of active, decoy compounds and the total number of molecules in the final dataset (*i.e.,* compounds that were docked and rescored with DockThor and DockTScore, respectively). Only the top-scored protonation state of each compound according to each scoring function (SF) was kept.

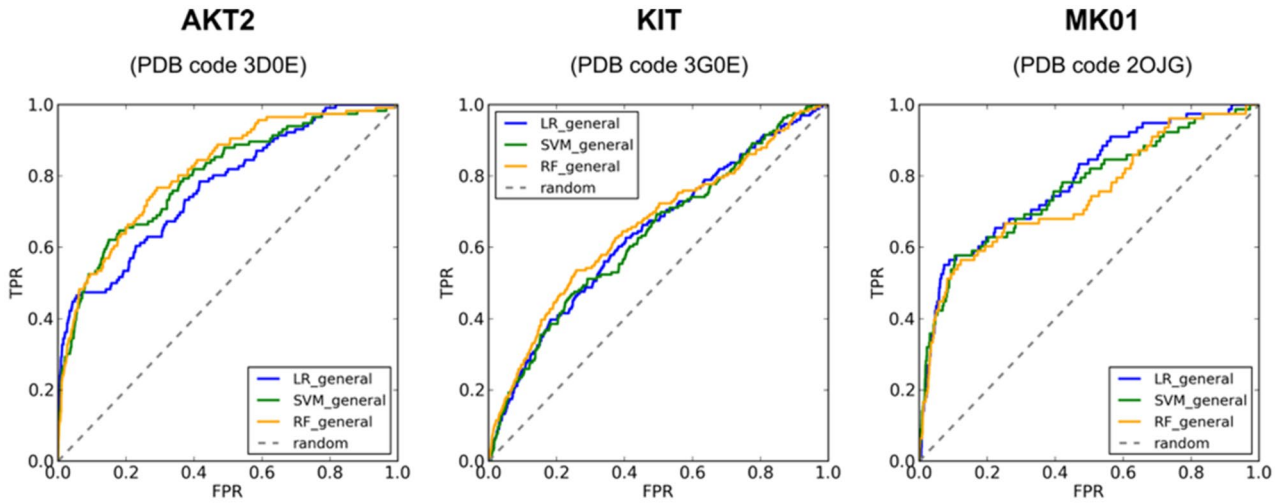

Figure 9. AUC ROC curves for the general scoring functions trained with MLR, SMOreg and RF evaluated on the AKT2, KIT and MK01 kinase datasets from DUD-E.

Discussion

We validated our physics-based terms for the general scoring functions using MLR. Despite its simplest form, MLR has the advantage to provide practical insights into relationships between the predicted binding affinity and the individual contribution of each specific term to the scoring function. In this work, all decisions regarding the selection of terms and machine-learning algorithms were made based on cross-validation experiments on the training set. The strategy of selecting random and independent test sets as employed here is particularly important in order to avoid performance overestimation. The performance of binding affinity prediction of the DockTScore general scoring functions are comparable with other well-known empirical scoring functions also tested on the v2013 PDBBind core set, *e.g.*, X-score::HMScore (R = 0.644)[57], Surflex-Dock (R = 0.388)[73], $VinaRF_{20}$ (R = 0.686)[74], and RF::VinaElem (R = 0.752) (Fig. 10). We obtained better performance on the carefully prepared core set compared to the random test set. One reason is that the selection of the complexes to form the *core set* ensured that all protein families in this benchmarking set were also present in the training set. Also, we believe that a correct preparation of the system, like the protonation state assignment as done for the *core set*, is important for proper binding energy prediction and a reliable assessment of scoring functions based on a more sophisticated protein–ligand interactions description.

Interestingly, the RF-Score::VinaElem (R = 0.752)[78], a nonlinear scoring function based on 36 element-element distance counts, the five Vina scoring function energy terms and the number of rotatable bonds in the ligand, showed highest performance in comparison with other well-established scoring functions validated on the same

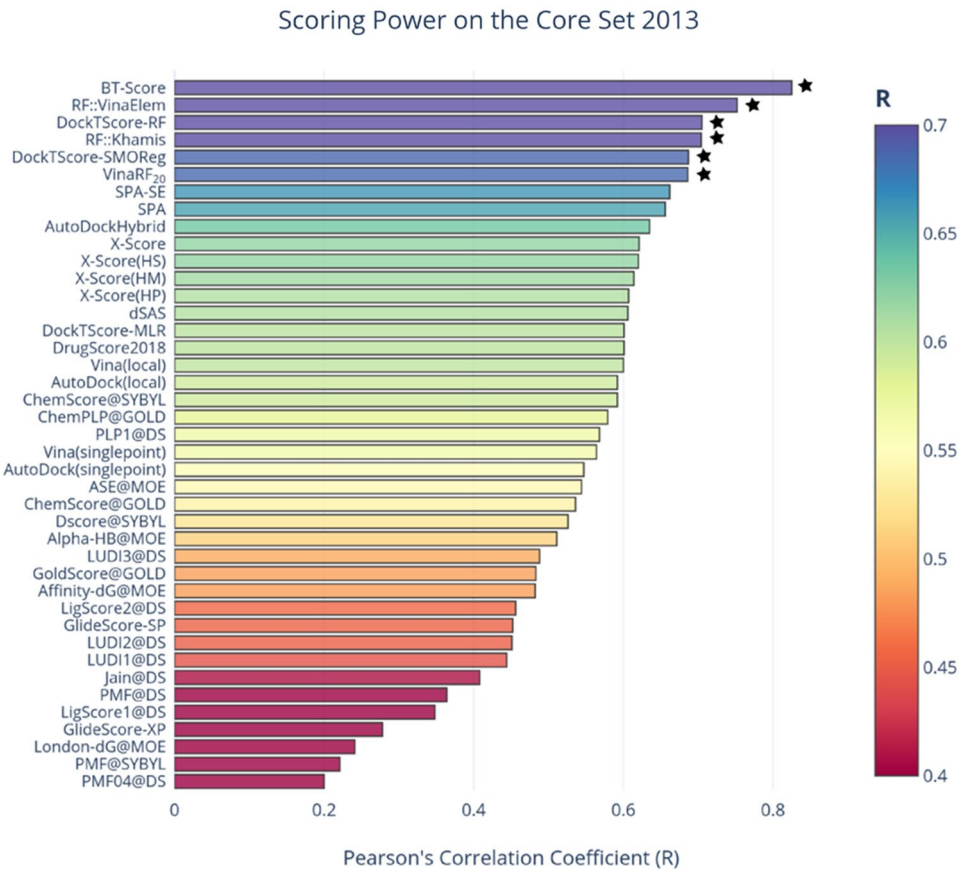

Figure 10. Scoring power of DockTScore linear and nonlinear models compared to the scoring functions evaluated on the core set 2013. Performances collected from the literature: BT-Score[75], CompSPA[76], AutoDockHybrid[23], and the remaining were recalculated from raw data in the recent work of T. Gaillard[77]. Nonlinear models are highlighted with a star. Scoring functions with Pearson's correlation coefficients higher or equal than 0.7 are colored purple and those lower than or equal to 0.4 are colored red.

v2013 core set[39]. On the other hand, it has been recently suggested that linear scoring functions, which can be less-accurate for binding affinity prediction but are composed of meaningful protein–ligand interaction terms, can be more robust than nonlinear scoring functions based only on element-element distance counts[4]. Definitely, element-element pair approaches are less sensitive to the proper dataset preparation, discarding the necessity of the time-consuming task of a careful assignment of the protonation states and atom types. However, scoring functions based on the calculation of physics-based binding energy terms might capture free energy changes arising from subtle protein–ligand interaction changes, useful particularly for hit-to-lead optimization.

It is widely recognized that target-specific scoring functions increase the efficiency of virtual screening exercises[21,24,27]. Different targeted docking-scoring strategies have been employed during the last decade. Some recent studies focused on combining scoring and pharmacophore/fingerprint filtering showed to improve target-specific pose/ligand selection[22,79,80]. We decided to develop new target-specific scoring functions for two protein classes to directly improve the prediction of the binding affinity by considering physics-based protein–ligand interaction terms. We obtained a remarkable improvement for the best nonlinear scoring function specific for PPIs (*i.e.,* the SMOReg model) compared to the general scoring function, achieving a significantly higher performance $R = 0.613$ against $R = 0.431$ obtained by the SMOReg general scoring function. For protease, such direct comparison is not reliable since most of the protease complexes present in the respective test set were also present in the training set used to derive the general scoring functions. Specific scoring functions have already been developed for well-established key protease targets as HIV-1 protease[28] and their performances are comparable with our SMOreg models. The advantage of our targeted scoring functions for proteases compared to the above-cited studies is the physical interpretability of the terms describing the protein–ligand interactions and good performances on virtual screening experiments evaluated with AUC ROC, $EF_{1\%}$ and BEDROC metrics for the screening assessment.

Despite the insufficient accuracy exhibited by our linear scoring function specific for iPPIs on the independent test set, it served as a basis for the development of nonlinear models using SMOReg and RF techniques. As expected, the nonlinear scoring function specific for iPPIs, in particular SMOReg, showed a significant improvement of the predictive performance when compared with the MLR model in terms of binding affinity prediction. However, analyzing the virtual screening metrics for the Bcl2 target, we observe distinct results.

The AUC values obtained using the SMOReg specific functions are better than the values obtained using MLR specific ones, yet the MLR specific function outperformed following the early recognition metrics (principally for EF1% and BEDROC 100). Thus, both the affinity prediction and ranking of compounds are important to properly evaluate the scoring functions performance. Our PPI-specific scoring functions were trained with 45 different PPI complexes covering thus a larger PPI interaction space than the previously used one for the only one reported linear scoring function specific for iPPIs HADDOCK2P2I[36]. The two PPI-specific scoring functions SMOReg and HADDOCK2P2I seem to perform similarly in terms of binding affinity prediction, yet the studies have been done on different PPI targets. To the best of our knowledge, the present SMOReg DockTScore is the first reported nonlinear scoring function tailored for the iPPI class that facilitates further optimization of the terms and the machine-learning algorithm used for training. In addition, the screening results obtained for the two PPI systems indicate that our PPI-specific scoring function trained with MLR is sufficiently robust to be used in virtual screening experiments, despite being trained with a small training set. Taking into consideration the very few scoring functions dedicated to score properly inhibitors of PPI both HADDOCK2P2I and DockTScore scoring functions can be very helpful e.g., for consensus scoring strategies. Furthermore, the growth of the number of experimentally derived iPPI structures available with associated affinity data enables the further development of more robust scoring functions specific for PPIs.

The variable performances achieved by the DockTScore models on the screening validation for the three different classes of proteins (*e.g.,* proteases, PPIs and kinases) are in agreement with other works published in the literature showing that the accuracy of scoring functions is strongly target-dependent. Further, although our scoring functions consider most of the interactions key for ligand binding, yet we do not take into account some contributions like the vibrational entropy[16] or particular cases as water molecules present in the binding pocket. The vibrational entropy is strongly related to the protein flexibility and to solvent entropy, and their precise estimation is not evident to be included in classical scoring functions. Other approaches as molecular dynamics or normal mode analysis can help to resolve such problems, however they are unpractical for a huge number of ligands and thus they are out of the scope of this work. Kinases are known to be very flexible proteins, and in our study KIT is the kinase protein for which our models exhibited the lowest performances on both AUC ROC and early recognition capacity evaluated through EF1% and BEDROC. The protein conformation of KIT provided by the DUD-E database and used here as the reference structure is complexed with the kinase inhibitor sunitinib. That KIT state corresponds to a more closed conformation of the ATP-binding site. The superposition of the autoinhibited KIT complexed with sunitinib (PDB code 3G0E) and the KIT-ponatinib complex (PDB code 4U0I), ponatinib being larger than sunitinib, shows an induced inactive DFG-out conformation of the enzyme, illustrating thus two possible distinct conformations adopted by the enzyme due to different ligands (Figure S1). Such results reinforce the importance of a careful selection of the receptor conformation to be used for virtual screening campaigns and the consideration of the protein flexibility to some extent[81].

Next, many inhibitors of proteases such as TRYB1 and UROK are known to displace water molecules interacting with catalytic residues of the binding site, however, in some cases such molecules can serve as a bridge between the receptor and the ligand. The analysis of the nine experimental complexes used in the virtual screening experiments showed that some of them contain ligands able to displace water molecules (*e.g.,* the proteases) and/or contain bridging waters in the experimental structure of the protein used in the virtual screening experiments (*e.g.,* FA7, TRYB1, and MDM2). In the case of MDM2-like protein, there is a complex network of water molecules mediating hydrogen bonds with the receptor important for the ligand binding. Such data support the importance of the enthalpic and entropic contributions of the water molecules in the binding pocket for the binding energy. The consideration of the contribution arising from bridged water molecules is a complex problem usually treated with more sophisticated methods that take into account the flexibility of the entire system and explicit water molecules. We have previously developed the AMMOS2 web server[82], which permits to take into consideration the presence of explicit water molecules in the binding pocket in order to optimize the predicted protein–ligand interactions.

The better performance of our MLR scoring function specific for PPIs on the protease TRYB1 dataset indicates that it could be applied on targets with similar profiles with those observed for PPI interfaces, such as those with highly solvent-exposed binding sites. We have recently reported similar observations when analyzing solvent-exposed co-crystallized ligands to support the design of novel protein–protein interaction inhibitors[83]. Our scoring function specific for PPIs also reinforces the fact that nonlinear scoring functions are more dependent on larger training sets, while robust linear models can be developed even when scarce data for training is available. Future growth of data for new PPI interfaces including dimer interfaces will allow to develop more robust nonlinear scoring functions specific for protein targets with binding site profiles similar to those found in PPIs.

Conclusion

In this work, we developed general and target-specific scoring functions using physics-based features for predicting binding affinities of protein–ligand complexes. Target-specific scoring functions were derived to account for binding characteristics specific for a target class of interest, focusing here on proteases and protein–protein interactions (PPIs). With regard to the increasing interest toward targeting PPIs by small-molecule inhibitors, here we reported the first and well-performing SVM-based scoring function specific for PPI binding sites that can serve as a valuable tool for discovering new iPPIs. Improved solvation and ligand torsional entropy terms were implemented in DockTScore for a reliable representation of ligand binding. DockTScore scoring functions demonstrated to be competitive with state-of-the-art scoring functions in reported benchmarking studies. As expected, the nonlinear scoring functions generally performed better than the respective MLR models. Finally, we demonstrated that the scoring functions developed in this work also exhibited good performances on virtual screening experiments to distinguish actives from inactive/decoy compounds for various protein targets.

DockTScore functions are independent of docking software and can be used for affinity prediction or consensus scoring to improve the performance of docking-scoring approaches on virtual screening experiments. Currently, the MLR DockTScore predictions are provided for the DockThor docking at the DockThor-VS web server (available at www.dockthor.lncc.br). All the developed scoring functions in this work are under implementation in a dedicated web server.

Data availability

The curated PDBbind core set v2013, manually prepared to insure the correct protonation states of the protein–ligand complexes, is freely available at www.dockthor.lncc.br.

Received: 2 November 2020; Accepted: 20 January 2021
Published online: 04 February 2021

References

1. Li, J., Fu, A. & Zhang, L. An overview of scoring functions used for protein-ligand interactions in molecular docking. *Interdiscip. Sci. Comput. Life Sci.* **11**, 320–328 (2019).
2. Adeshina, Y. O., Deeds, E. J. & Karanicolas, J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proc. Natl. Acad. Sci.* **117**, 18477–18488 (2020).
3. Guedes, I. A., de Magalhães, C. S. & Dardenne, L. E. Receptor–ligand molecular docking. *Biophys. Rev.* **6**, 75–87 (2014).
4. Gabel, J., Desaphy, J. & Rognan, D. Beware of machine learning-based scoring functions—on the danger of developing black boxes. *J. Chem. Inf. Model.* **54**, 2807–2815 (2014).
5. Wang, Z. *et al.* Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys Chem Chem Phys* **18**, 12964–12975 (2016).
6. Sieg, J., Flachsenberg, F. & Rarey, M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* **59**, 947–961 (2019).
7. Guedes, I. A., Pereira, F. S. S. & Dardenne, L. E. Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front. Pharmacol.* **9**, 1–18 (2018).
8. Pason, L. P. & Sotriffer, C. A. Empirical scoring functions for affinity prediction of protein-ligand complexes. *Mol. Inform.* **35**, 541–548 (2016).
9. Wójcikowski, M., Ballester, P. J. & Siedlecki, P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.* **7**, 46710 (2017).
10. Yan, Y., Wang, W., Sun, Z., Zhang, J. Z. H. & Ji, C. Protein-ligand empirical interaction components for virtual screening. *J. Chem. Inf. Model.* **57**, 1793–1806 (2017).
11. Jiménez Luna, J., Skalic, M., Martinez-Rosell, G. & De Fabritiis, G. KDEEP: Protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J. Chem. Inf. Model.* https://doi.org/10.1021/acs.jcim.7b00650 (2018).
12. Li, H. *et al.* Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. *Bioinforma. Oxf. Engl.* **35**, 3989–3995 (2019).
13. Zhao, Q., Ye, Z., Su, Y. & Ouyang, D. Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques. *Acta Pharm. Sin. B* **9**, 1241–1252 (2019).
14. Böhm, H. J. 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**, 243–256 (1994).
15. Schapira, M., Totrov, M. & Abagyan, R. Prediction of the binding energy for small molecules, peptides and proteins. *J. Mol. Recognit. JMR* **12**, 177–190 (1999).
16. Chang, C. A., Chen, W. & Gilson, M. K. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci.* **104**, 1534–1539 (2007).
17. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **28**, 1145–1152 (2007).
18. Chen, J., Brooks, C. L. & Khandogin, J. Recent advances in implicit solvent based methods for biomolecular simulations. *Curr. Opin. Struct. Biol.* **18**, 140–148 (2008).
19. Huang, S.-Y. & Zou, X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J. Chem. Inf. Model.* **50**, 262–273 (2010).
20. Kar, P., Lipowsky, R. & Knecht, V. Importance of polar solvation and configurational entropy for design of antiretroviral drugs targeting HIV-1 protease. *J. Phys. Chem. B* **117**, 5793–5805 (2013).
21. Seifert, M. H. J. Robust optimization of scoring functions for a target class. *J. Comput. Aided Mol. Des.* **23**, 633–644 (2009).
22. Politi, R., Convertino, M., Popov, K., Dokholyan, N. V. & Tropsha, A. Docking and scoring with target-specific pose classifier succeeds in native-like pose identification but not binding affinity prediction in the CSAR 2014 benchmark exercise. *J. Chem. Inf. Model.* **56**, 1032–1041 (2016).
23. Ericksen, S. S. *et al.* Machine learning consensus scoring improves performance across targets in structure-based virtual screening. *J. Chem. Inf. Model.* **57**, 1579–1590 (2017).
24. Seifert, M. H. J. Targeted scoring functions for virtual screening. *Drug Discov. Today* **14**, 562–569 (2009).
25. Palacio-Rodríguez, K., Lans, I., Cavasotto, C. N. & Cossio, P. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Sci. Rep.* **9**, 5142 (2019).
26. Su, M., Feng, G., Liu, Z., Li, Y. & Wang, R. Tapping on the black box: how is the scoring power of a machine-learning scoring function dependent on the training set?. *J. Chem. Inf. Model.* https://doi.org/10.1021/acs.jcim.9b00714 (2020).
27. Wang, D. *et al.* Improving the virtual screening ability of target-specific scoring functions using deep learning methods. *Front. Pharmacol.* **10**, (2019).
28. Wang, W.-J., Huang, Q., Zou, J., Li, L.-L. & Yang, S.-Y. TS-chemscore, a target-specific scoring function, significantly improves the performance of scoring in virtual screening. *Chem. Biol. Drug Des.* **86**, 1–8 (2015).
29. Logean, A., Sette, A. & Rognan, D. Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. *Bioorg. Med. Chem. Lett.* **11**, 675–679 (2001).
30. Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. *Data mining: practical machine learning tools and techniques*. (2017).
31. Lai, T. L., Robbins, H. & Wei, C. Z. Strong consistency of least squares estimates in multiple regression. *Proc. Natl. Acad. Sci. USA* **75**, 3034–3036 (1978).
32. Shevade, S. K., Keerthi, S. S., Bhattacharyya, C. & Murthy, K. K. Improvements to the SMO algorithm for SVM regression. *IEEE Trans. Neural Netw. Publ. IEEE Neural Netw. Counc.* **11**, 1188–1193 (2000).
33. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
34. Réau, M., Langenfeld, F., Zagury, J.-F., Lagarde, N. & Montes, M. Decoys selection in benchmarking datasets: overview and perspectives. *Front. Pharmacol.* **9**, 11 (2018).

35. Pintro, V. O. & de Azevedo, W. F. Optimized virtual screening workflow: towards target-based polynomial scoring functions for HIV-1 protease. *Comb. Chem. High Throughput Screen.* **20**, 820–827 (2017).
36. Kastritis, P. L., Rodrigues, J. P. G. L. M. & Bonvin, A. M. J. J. HADDOCK $_{2P2I}$: A biophysical model for predicting the binding affinity of protein–protein interaction inhibitors. *J. Chem. Inf. Model.* **54**, 826–836 (2014).
37. Liu, Z. *et al.* PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).
38. Li, Y. *et al.* Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J. Chem. Inf. Model.* **54**, 1700–1716 (2014).
39. 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.* **54**, 1717–1736 (2014).
40. Li, Y. *et al.* Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark. *Nat. Protoc.* **13**, 666–680 (2018).
41. Kuenemann, M. A., Bourbon, L. M. L., Labbé, C. M., Villoutreix, B. O. & Sperandio, O. Which three-dimensional characteristics make efficient inhibitors of protein-protein interactions?. *J. Chem. Inf. Model.* **54**, 3067–3079 (2014).
42. Burley, S. K. *et al.* RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464–D474 (2019).
43. Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **27**, 221–234 (2013).
44. Liu, Z. *et al.* Forging the basis for developing protein-ligand interaction scoring functions. *Acc. Chem. Res.* **50**, 302–309 (2017).
45. Su, M. *et al.* Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.* **59**, 895–913 (2019).
46. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: consistent treatment of internal and surface residues in empirical pK_a predictions. *J. Chem. Theory Comput.* **7**, 525–537 (2011).
47. Shelley, J. C. *et al.* Epik: a software program for pKa prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.* **21**, 681–691 (2007).
48. Bietz, S., Urbaczek, S., Schulz, B. & Rarey, M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J. Cheminformatics* **6**, 12 (2014).
49. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).
50. Halgren, T. A. The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *J. Am. Chem. Soc.* **114**, 7827–7843 (1992).
51. dos Santos, K. B., Guedes, I. A., Karl, A. L. M. & Dardenne, L. Highly Flexible Ligand docking: benchmarking of the DockThor program on the LEADS-PEP protein-peptide dataset. *J. Chem. Inf. Model.* acs.jcim.9b00905 (2020) doi:https://doi.org/10.1021/acs.jcim.9b00905.
52. de Magalhães, C. S., Almeida, D. M., Barbosa, H. J. C. & Dardenne, L. E. A dynamic niching genetic algorithm strategy for docking highly flexible ligands. *Inf. Sci.* **289**, 206–224 (2014).
53. Hingerty, B. E., Ritchie, R. H., Ferrell, T. L. & Turner, J. E. Dielectric effects in biopolymers: the theory of ionic saturation revisited. *Biopolymers* **24**, 427–439 (1985).
54. Ramstein, J. & Lavery, R. Energetic coupling between DNA bending and base pair opening. *Proc. Natl. Acad. Sci. USA* **85**, 7231–7235 (1988).
55. Gilson, M. K. & Honig, B. H. The dielectric constant of a folded protein. *Biopolymers* **25**, 2097–2119 (1986).
56. 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.* **11**, 425–445 (1997).
57. Wang, R., Lai, L. & Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.* **16**, 11–26 (2002).
58. Kuhn, B. & Kollman, P. A. Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *J. Med. Chem.* **43**, 3786–3791 (2000).
59. Sanner, M. F., Olson, A. J. & Spehner, J.-C. Fast and robust computation of molecular surfaces. in 406–407 (ACM Press, 1995). doi:https://doi.org/10.1145/220279.220324.
60. Abagyan, R. & Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983–1002 (1994).
61. Bennett, K. P. & Campbell, C. Support vector machines: hype or hallelujah?. *ACM SIGKDD Explor. Newsl.* **2**, 1–13 (2000).
62. Witten, I. H. & Frank, E. *Data mining: practical machine learning tools and techniques*. (Morgan Kaufman, 2005).
63. Mysinger, M. M., Carchia, M., Irwin, John. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
64. Labbé, C. M. *et al.* iPPI-DB: an online database of modulators of protein–protein interactions. *Nucleic Acids Res.* **44**, D542–D547 (2016).
65. Reynès, C. *et al.* Designing focused chemical libraries enriched in protein-protein interaction inhibitors using machine-learning methods. *PLOS Comput. Biol.* **6**, e1000695 (2010).
66. Truchon, J.-F. & Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **47**, 488–508 (2007).
67. Lätti, S., Niinivehmas, S. & Pentikäinen, O. T. Rocker: open source, easy-to-use tool for AUC and enrichment calculations and ROC visualization. *J. Cheminformatics* **8**, 45 (2016).
68. Williams, D. H. & Bardsley, B. Estimating binding constants: the hydrophobic effect and cooperativity. *Perspect. Drug Discov. Des.* **17**, 43–59 (1999).
69. Ain, Q. U., Aleksandrova, A., Roessler, F. D. & Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening: Machine-learning SFs to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* n/a-n/a (2015) doi:https://doi.org/10.1002/wcms.1225.
70. Fresnais, L. & Ballester, P. J. The impact of compound library size on the performance of scoring functions for structure-based virtual screening. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbaa095 (2020).
71. Lagorce, D., Douguet, D., Miteva, M. A. & Villoutreix, B. O. Computational analysis of calculated physicochemical and ADMET properties of protein–protein interaction inhibitors. *Sci. Rep.* **7**, (2017).
72. Morelli, X., Bourgeas, R. & Roche, P. Chemical and structural lessons from recent successes in protein–protein interaction inhibition (2P2I). *Curr. Opin. Chem. Biol.* **15**, 475–481 (2011).
73. Cheng, T., Li, X., Li, Y., Liu, Z. & Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **49**, 1079–1093 (2009).
74. Wang, C. & Zhang, Y. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *J. Comput. Chem.* **38**, 169–177 (2017).
75. Ashtawy, H. M. & Mahapatra, N. R. Task-specific scoring functions for predicting ligand binding poses and affinity and for screening enrichment. *J. Chem. Inf. Model.* **58**, 119–133 (2018).
76. Yan, Z. & Wang, J. Optimizing the affinity and specificity of ligand binding with the inclusion of solvation effect. *Proteins Struct. Funct. Bioinforma.* **83**, 1632–1642 (2015).

77. Gaillard, T. Evaluation of AutoDock and AutoDock Vina on the CASF-2013 Benchmark. *J. Chem. Inf. Model.* **58**, 1697–1706 (2018).
78. Li, H., Leung, K.-S., Wong, M.-H. & Ballester, P. Low-quality structural and interaction data improves binding affinity prediction via random forest. *Molecules* **20**, 10947–10962 (2015).
79. Kooistra, A. J. *et al.* Function-specific virtual screening for GPCR ligands using a combined scoring method. *Sci. Rep.* **6**, (2016).
80. Martin, E. J. & Sullivan, D. C. Surrogate AutoShim: predocking into a universal ensemble kinase receptor for three dimensional activity prediction, very quickly, without a crystal structure. *J. Chem. Inf. Model.* **48**, 873–881 (2008).
81. Cleves, A. E. & Jain, A. N. Structure- and ligand-based virtual screening on DUD-E+: performance dependence on approximations to the binding pocket. *J. Chem. Inf. Model.* **60**, 4296–4310 (2020).
82. Labbé, C. M. *et al.* AMMOS2: a web server for protein–ligand–water complexes refinement via molecular mechanics. *Nucleic Acids Res* **45**, W350–W355 (2017).
83. Trisciuzzi, D. *et al.* Analysis of solvent-exposed and buried co-crystallized ligands: a case study to support the design of novel protein-protein interaction inhibitors. *Drug Discov Today.* **24**, 551–559 (2019).

Acknowledgements

The authors thank CNPq (Grant 308202/2016-3), Faperj (Grant E-26/010.001229/2015), PCI-LNCC (Grants 300463/2019-7 and 312604/2016-5), the French ANR agency (Grant ToxME), INSERM institute and University of Paris for financial support. We gratefully acknowledge the support of the Brazilian Sistema Nacional de Processamento de Alto Desempenho (SINAPAD) and the availability of the computational resources provided by the Supercomputer SDumont (LNCC/MCTIC).

Author contributions

Conceptualization: M.A.M, I.A.G. and L.E.D; methodology, M.A.M, I.A.G., A.M.S.B. and L.E.D.; software, I.A.G., E.K., D. M. and L.E.D.; validation and analysis, I.A.G. and L.E.D.; investigation, I.A.G., M.A.M and L.E.D.; resources, M.A.K., O.S., L.E.D.; data curation, I.A.G., M.A.K., O.S., writing—original draft preparation, I.A.G., L.E.D and M.A.M..; writing-review and editing, I.A.G., O.S., L.E.D. and M.A.M. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by CNPq, grant numbers 307634/2019-1 and 306894/2019-0; by FAPERJ, grant numbers E-26/010.001229/2015 and E-26/210.935/2019; by PCI-LNCC grant numbers 300463/2019-7 and 312604/2016-5, by the French ANR agency (grant ToxME), by the INSERM institute and by University of Paris.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-82410-1.

Correspondence and requests for materials should be addressed to L.E.D. or M.A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.