**RASPD+: Fast protein-ligand binding free energy prediction using simplified physicochemical features**

Stefan Holderbach,†‡ Lukas Adam,†‡ B. Jayaram,¶ Rebecca C. Wade,†§∥ and Goutam Mukherjee*,†§

†Molecular and Cellular Modelling Group, Heidelberg Institute of Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
‡Institute of Pharmacy and Molecular Biotechnology (IPMB), Heidelberg University, Im Neuenheimer Feld 364, 69120 Heidelberg, Germany
¶Supercomputing Facility for Bioinformatics & Computational Biology, Department of Chemistry, Kusuma School of Biological Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, 110016, India
§Center for Molecular Biology (ZMBH), DKFZ-ZMBH Alliance, Heidelberg University, Im Neuenheimer Feld 282, 69120 Heidelberg, Germany
∥Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

E-mail: goutam.mukherjee@h-its.org

**Abstract**

The virtual screening of large numbers of compounds against target protein binding sites has become an integral component of drug discovery workflows. This screening is often done by computationally docking ligands into a protein binding site of interest,
but this has the drawback that a large number of poses must be evaluated to obtain accurate estimates of protein-ligand binding affinity. We here introduce a fast prefiltering method for ligand prioritization that is based on a set of machine learning models and uses simple pose-invariant physicochemical descriptors of the ligands and the protein binding pocket. Our method, Rapid Screening with Physicochemical Descriptors + machine learning (RASPD+), is trained on PDBbind data and achieves a regression performance better than for the original RASPD method and comparable to traditional scoring functions on a range of different test sets without the need for generating ligand poses. Additionally, we use RASPD+ to identify molecular features important for binding affinity and assess the ability of RASPD+ to enrich active molecules from decoys.

**Introduction**

Virtual screening to assess in silico the binding of candidate ligands to a target protein is a key component of structure-based drug design procedures. Typically, screening is done by docking the ligands at many different positions or poses in the three-dimensional structure of the target protein. At every position, a scoring function is evaluated to approximate the binding free energy and this is used to rank the binding poses and to rank different candidate ligands for their ability to bind to the target protein. While correct docking poses are frequently generated, scoring functions often lack the accuracy necessary to correctly rank poses or ligands. Therefore, docking procedures are frequently supplemented by methods employing molecular dynamics simulations with the aim of computing more accurate binding affinities. However, both docking and molecular dynamics simulations often fail to provide predictions of binding free energy at the level of accuracy desired. Furthermore, they are demanding in terms of computational effort and expertise. Therefore, there is a need for quick approaches with robust predictive scoring functions to facilitate the screening and prioritization of large libraries of compounds prior to applying docking and simulation.
While the assessment of ligand properties, e.g., for drug-likeness, to filter ligand libraries is well established, we here address the need to filter and prioritize ligands based not only on ligand properties but also on the properties of the target protein. For this purpose, we previously developed a simple hybrid regression approach, called RASPD (Rapid Screening with Physicochemical Descriptors). In this linear regression model, the binding free energy $\Delta G$ was predicted using a minimal set of physicochemical descriptors for typical interactions. Hydrogen bonding was accounted for by counting potential donor and acceptor atoms. Van der Waals forces were approximated by the Wiener topology index and the molar refractivity, which describes the polarizability of a molecule. Additionally, the partition coefficient logP allowed for the estimation of the hydrophobic effect. While the descriptor values for the ligand are straightforward to compute, simplifying assumptions were made to obtain the physicochemical descriptors for the target protein. A sphere was centered on a known or assumed binding pocket position with a radius encompassing the maximum size of the ligand. This sphere was then used to select the amino acid residues for which descriptors were computed (Figure 1A). However, the linear regression model used has limited abilities to capture complex feature interactions compared to non-linear models. Since RASPD was first developed, more high-quality data sets on protein-ligand complexes with associated binding free energies have been made available and a large number of machine learning methods have been developed. Moreover, machine learning approaches have successfully been used to either replace or enhance the predictions of traditional scoring functions for protein-ligand binding.

Thus, we have here developed RASPD+, a new tool that improves on the conceptual framework of the original RASPD method by using: (i) a set of diverse machine learning methods to derive an ensemble prediction, (ii) additional and more fine-grained descriptors for the target proteins, and (iii) larger training sets of newer protein-ligand binding data. We here describe the training, testing, and application of RASPD+. We demonstrate the
capabilities of RASPD+ for binding free energy regression and compare its performance to established scoring functions. We also analyze the features contributing to the predictions to gain insights into the important features for binding affinity. Finally, we show that RASPD+ can enrich active molecules in tests with the Directory of Useful Decoys (DUD).  

Methods

The computational workflow and the training and validation procedure used for RASPD+ are illustrated in Figure 1.

Figure 1: The computational workflow of RASPD+ comprises featurization (A) and the training and evaluation of machine learning models (B). (A) For each ligand molecule, simple physicochemical descriptors are computed based on atomic contributions. Information about the target protein is gathered within a sphere around a putative binding position, whose size is determined by the radius of the ligand. For residues within this sphere, similar descriptors are computed. (B) Data from the PDBbind refined set featurized in this way served as the training data in a nested cross-validation strategy. To compare linear regression (LR), k-nearest neighbors (kNN), support vector regression (SVR), neural network (DNN), random forest (RF), and extremely random forest (eRF) models, test sets were split off in 10 replicates in an outer loop. In the inner loop, 6-fold cross-validation was used to select the best hyperparameters for the given model.
Datasets

The PDBbind refined data set (release 2018\textsuperscript{10,15}, containing 4463 protein-ligand crystal structures and experimentally measured binding affinities, served as the initial data set. As it contains high-quality structures of non-covalent protein-ligand interactions with a resolution better than 2.5 Å and no steric clashes, it is extensively used as a benchmark set for protein-ligand binding affinity prediction.\textsuperscript{10} We thus obtained structural information about each protein in the data set, the position and structure of the ligand binding to it, and the corresponding binding constant. As we considered modeling the coordination of metal ions to be beyond the scope of our approach, the structures were filtered to exclude cases with metal ions within 2.1 Å of the ligand. Dissociation and inhibition constants and $IC_{50}$ values were converted to binding free energies using the equation:

$$\Delta G = -RT \ln K$$ where $K \in \{K_d, K_i, IC_{50}\}$ assuming $T = 298.15 \text{ K}$ \hspace{1cm} (1)

This processing resulted in a set of 3925 protein-ligand complexes for training, validation, and testing.

For further testing, the following previously published benchmark sets served as external test sets: The Community Structure-Activity Resource (CSAR) NRC-HiQ 2010 selection,\textsuperscript{19,20} data sets from the CSAR 2012\textsuperscript{21} and CSAR 2014\textsuperscript{22} challenges, and a data set described by Wang et al.\textsuperscript{23}

The CSAR-NRC 2010 HiQ release\textsuperscript{19,20} contains two sets of protein-ligand complexes, with 55 and 49 docked complexes, respectively, as well as information about experimental binding affinities.

Another set of binding free energies and corresponding structures was assembled from the CSAR 2012\textsuperscript{21} and CSAR 2014\textsuperscript{22} data sets that are now curated by the Drug Design Data Resource (D3R) (drugdesigndata.org).\textsuperscript{9} For this set, which we refer to as the D3R data set, we downloaded the data for the proteins urokinase, cyclin-dependent kinase 2
(CDK2), checkpoint kinase 1 (CHK1), MAP kinase 1 (ERK2), LpxC deacetylase (LpxC), spleen tyrosine kinase (SYK), tRNA (m1G37) methyltransferase (tRMD), heat shock protein 90 (HSP90), and a CDK2-Cyclin A complex. The SMILES strings of 1271 active inhibitors of these proteins in the D3R data set were converted to 3D structures in PDB format using Open Babel. For HSP90, we excluded 46 compounds which were all assigned the same ΔG of −5.860 kcal/mol as this value likely represented a threshold value for the experimental measurements rather than the actual binding affinity of the ligands.

Wang et al. aggregated previous experimental results and PDB structures for 283 complexes of 7 different proteins: beta-secretase (BACE), CDK2, induced myeloid leukemia cell differentiation protein (Mcl-1), p38 MAP kinase, protein-tyrosine phosphatase 1B (PTP1B), thrombin, and tyrosine kinase 2 (TYK2). For this set, protein structures were retrieved from the RCSB protein data bank (http://www.rcsb.org) and hydrogen atoms were added to the protein structures with the tleap module of AMBER 14. This included additional ligands for Mcl-1 and TYK2 that were not used by Wang et al.. The structures of the 283 inhibitors were redrawn and verified in the MOE software (Chemical Computing Group, Montreal, QC).

Further details on the source of structures and experimental binding affinities are given in Table SI 1.

Generation of molecular descriptors

To model the non-covalent interactions, physicochemical molecular descriptors were computed using an improved pipeline based on that for the original RASPD procedure described in Ref. (Figure 1A). For each ligand, the molecular weight (here abbreviated as MASS), the number of hydrogen bond donors (D) and acceptors (A), an approximate octanol-water partition coefficient log P (logP), the molar refractivity (MR), and the Wiener topology index (W) were computed as described previously. Based on the ligand position in the pro-
tein structure, the most likely interacting amino acid residues were selected using a sphere, whose radius was derived from the maximum distance (maxD) between ligand atoms and the center of mass (Figure 1A). For the computation of the logP and MR descriptors, this sphere was extended by 0.9 Å over maxD and residues were selected based on their center of mass. To count hydrogen bond donors and acceptors, a sphere extending 3 Å beyond maxD was used to select atoms. Details regarding the protein pocket selection procedure and the choice of the cut-off radii are given in Ref. 6. To make the protein descriptors more fine-grained than in the previous RASPD procedure, we computed molar refractivity and log P for aromatic and non-aromatic residues separately (PMR(Arom), PMR(Non-Arom), PlogP(Arom), PlogP(Non-Arom)). Hydrogen bond donors were counted separately for the backbone amide group (PD(Amide-NH)) as well as for the following amino acid sets: Positively charged PD(K, R, HIP), neutral amino groups PD(K, N, Q), heteroaromatic donors PD(W, H), and hydroxyl-containing groups PD(T, S, Y, D, E). The number of hydrogen bond acceptors was determined for the backbone amide (PA(Amide-O)) and the following sets: negatively charged PA(D, E), neutral non-aromatic PA(N, Q, T, S, D-H, E-H), and aromatic acceptors PA(Y, H). The individual protein residue-derived descriptors were scaled by the ligand maxD. Additionally, the volume of the protein pocket (PVol) was computed using tools from the TRAPP software suite. Thus, in total, 6 ligand and 14 protein descriptors were computed per ligand-protein complex.

General strategy for training and testing

To obtain a robust estimate of performance on the PDBbind data set, as well as the test sets, a nested cross-validation strategy was used (Figure 1B). For ten replicates, the PDBbind refined set was split into a test set covering 12.5% of the data and a set for cross-validation training. For each of these replicates, six-fold cross-validation training was performed to select the best hyperparameters for each replicate based on the Pearson correlation coefficient. Thus, for each replicate, 2860 complexes were used for training, 572 for cross-validation, and
The input features were robustly centered and scaled by the median and interquartile range (IQR) of the training set for each train-test split. All models obtained by the hyperparameter search were evaluated on the corresponding PDBbind test set as well as on the external test sets. We report the mean and standard deviation of the performance metrics.

**Evaluation metrics**

To assess model performance, the root-mean-squared error \( (RMSE) \), Pearson \( (r) \) and Spearman \( (\rho) \) correlation coefficients, and the coefficient of determination, \( R^2 \), were computed using the `sklearn.metrics` and `scipy.stats` Python modules. Additionally we report the \( Q^2_{F_3} \) metric (eqn. [2]43), as it is considered to be better suited for QSAR-like tasks than \( R^2 \)48.

\[
Q^2_{F_3} = 1 - \frac{\sum_{i}^{n_{test}} (\hat{y} - y_{test})^2}{\sum_{i}^{n_{test}} (\hat{y} - y_{train})^2} 
\]  \hspace{1cm} (2)

**Models and hyperparameters**

As part of this work, we evaluated different machine learning models. We considered linear regression (LR), as it was used in the previous RASPD approach,\(^6\) support vector regression\(^39\) (SVR), k-Nearest Neighbors (kNN), simple deep neural networks (DNN), random forests\(^40\) (RF), and a variant of the former, extremely random forests\(^41\) (eRF). The associated hyperparameters for each method were optimized by a grid search covering a typical space. Further details on each method and their associated hyperparameters are given in the Supporting Methods. A comprehensive list of tested hyperparameters is given in Table SI 2. All models except the neural networks were built using the `scikit-learn` Python package (version 0.20.2).\(^{42}\) For the neural networks, the Keras API (version 2.2.4)\(^{43}\) for TensorFlow (version 1.12) was used in conjunction with the `talos` package (version 0.4.6)\(^{44}\) for hyperparameter optimization.
Estimation of feature importance

To estimate the importance of individual input features, a simple permutation-based approach was used. After prediction on a real-world test set, the column of each feature in the data set was shuffled in 5 replicates, and the mean change in Pearson correlation coefficient was computed. Thereby, the model has to make a prediction based on a random sample from a distribution with the same mean and variance. A drop in predictive performance indicates that the prediction is dependent on this feature.

Enrichment analysis with decoy compounds from the DUD dataset

To evaluate the performance of RASPD+ for capturing active molecules from a pool of computationally generated decoys, 3D coordinates of active and decoy molecules were retrieved from the DUD data set. This set contains 40 targets, from which we considered 30 proteins for which cofactor binding or metal ions did not play an obvious role in the ligand-binding site of the provided reference protein structure. Information about the studied proteins and their PDB identifiers, as well as the number of active and decoy molecules, is given in Table SI 9. Enrichment was performed by selecting a given percentage of molecules that scored highest in the given method. For scoring, the predictions across the 6 cross-validation folds of a replicate were averaged. The enrichment factor was defined as the ratio of the fraction of active molecules in the enriched set divided by the fraction of the active molecules in the total set.
Results

Analysis of the descriptors and data sets

To confirm the usefulness of the chosen molecular descriptors, we performed correlation analysis on the PDBbind refined set (Figure 2, Table SI 3). The Spearman correlations with the binding free energy, $\Delta G$, were negative for most descriptors, as stronger binding is indicated by negative values of $\Delta G$. The strongest negative correlations were observed for the molar refractivity of the ligand molecule (MR, $-0.51$) and the abundance of peptide bond oxygen atoms (hydrogen bond acceptors) inside the protein binding pocket (PA(Amide-O), $-0.49$). The correlations with $\Delta G$ for the features for specific amino acids were lower than $0.25$ in magnitude, which is less than the corresponding correlation ($> 0.4$) obtained for the backbone (PA(Amide-O), PD(Amide-NH)) and non-aromatic amino acid (PlogP(Non-Arom), PMR(Non-Arom)) descriptors.

Figure 2: Correlation analysis on the PDBbind data set reveals that the experimental binding free energy has the strongest negative correlation with the ligand molar refractivity (MR, Spearman $\rho = -0.51$), and with the number of peptide bond oxygen atoms (hydrogen bond acceptors) present in the putative protein binding pocket (PA(Amide-O), Spearman $\rho = -0.49$). The value of the Spearman’s correlation coefficient is indicated by color.
We next analyzed the correlations among the descriptors, in particular, to check for possible biases for certain interactions in the protein-ligand complexes of the PDBbind data. Amongst the ligand descriptors, the strongest correlations were observed between molecular weight, molar refractivity, and Wiener index (MASS, MR, W). For the protein features, the strongest correlation was between the two descriptors for the aromatic amino acids, PlogP(Arom) and PMR(Arom). In addition, the backbone-based features, (PA(Amide-O) and PD(Amide-NH)), had a high correlation with the log P and molar refractivity values of the non-aromatic residues. Among the hydrogen bond contributions of the amino acids, we observed the strongest correlation, with $\rho = 0.66$, between PD(T+S+Y+DH+EH) and PA(N+Q+T+S+DH+EH). This correlation is expected because they share the highest number of amino acids.

A higher correlation between the ligand and the protein features was observed between ligand features that directly scale with the size of the molecule (MASS, W, MR) and the more general protein features, such as the backbone features and the log P and MR values of the non-aromatic residues. These protein features are expected to be related to the ligand size, and therefore, do not indicate any data set-specific bias of the PDBbind data set.

Comparing the distributions of binding free energies $\Delta G$ between the PDBbind data set used for training and validation and the CSAR 2012 and 2014\textsuperscript{21,22} and Wang et al.\textsuperscript{23} external data sets used for testing revealed that the PDBbind data set covers a wider range of binding free energies (Figure SI 1). In contrast, the 101 protein-ligand complexes from the CSAR NRC-HiQ release cover a wider $\Delta G$ range than PDBbind.

From the distribution of the individual descriptors, it is clear that the PDBbind data set encompasses the full range of descriptor values covered by the other data sets (Figure SI 2), even though there are differences in the mean values of the descriptors. For example, the average ligand molecular weight was lowest for the CSAR-NRC HiQ data and highest for the D3R data from CSAR 2012 and CSAR 2014.
Trained models – Random forests outperform neural networks

Initial tests revealed high variability in the performance metrics that depended on a random training and validation data split. We thus chose a nested cross-validation strategy to find the machine learning models best suited for the chosen descriptors (Figure 1B). Therefore, performance metrics are reported as the mean of sixty models resulting from 10 random data set draws and 6-fold cross-validation. The corresponding standard deviation enables the quantification of the uncertainty of the performance metrics. Apart from the baseline correlation values between the individual descriptors and the target variable $\Delta G$, we included a null model, which simply predicted the mean $\Delta G$ of the training data, to verify predictive power. The root-mean-squared error, $RMSE$, of $2.76 \pm 0.05$ kcal/mol measured for this null model is identical to the population standard deviation for the respective training folds (Table 1, Figure 3). The linear regression model derived by ordinary least squares fitting, similar to the original RASPD approach, achieved a $RMSE$ of $2.19 \pm 0.05$ kcal/mol on the test set. We tested six other methods and assessed whether they improved on this value.

SVR with a Gaussian radial basis function (RBF) kernel and a neural network with two hidden layers performed with $RMSE$ values of $2.04 \pm 0.05$ kcal/mol and $2.05 \pm 0.05$ kcal/mol, respectively, and similar to k-nearest neighbors with an $RMSE$ of $2.03 \pm 0.04$ kcal/mol. Superior performance in terms of both deviation, quantified by $RMSE$, and ranking, as measured by the Spearman correlation $\rho$, was achieved with the two random forest-based models. The eRF model had a $RMSE$ of $1.86 \pm 0.05$ kcal/mol and a Pearson correlation $r$ of $0.74 \pm 0.02$, and the RF model performed similarly (Table 1, Figure 3).

Therefore, we selected the resulting eRF models for further analysis. We note that these eRF regressors, which use 200 trees and have no limits on the number of samples per leaf, overfit the training set, despite showing better validation set performance compared to more strongly regularized variants (see Tables SI 4 and SI 5). Nevertheless, examination of the predictions of the eRF models on the PDBbind test data, shows that the general trends in the data are captured although the lowest $\Delta G$ values are overestimated and the highest $\Delta G$
Figure 3: Systematic evaluation of the predictive performance of the seven different machine learning methods on the PDBbind test set shows that, according to three metrics, the extremely random forests (eRF) model performs better than models derived by the other machine learning methods in predicting protein-ligand binding free energy. The error bars indicate the standard deviation for 10 replicates with 6-fold cross-validation.

Table 1: Comparison of the performance of the models derived with seven different machine learning methods for predicting the protein-ligand binding free energy for the PDBbind test set. The five metrics of performance are given as mean and standard deviation values computed by averaging from the 10 different random test set splits and 6 cross-validation folds. The RMSE is given in kcal/mol. NA : not applicable

| model   | RMSE    | r       | ρ       | $R^2$   | $Q_{F3}^2$ |
|---------|---------|---------|---------|---------|------------|
| null model | 2.76 ± 0.05 | 0.0 ± 0.0 | NA      | −0.00 ± 0.00 | −0.03 ± 0.05 |
| LR      | 2.19 ± 0.05 | 0.61 ± 0.02 | 0.60 ± 0.02 | 0.37 ± 0.02 | 0.35 ± 0.03 |
| kNN     | 2.03 ± 0.04 | 0.68 ± 0.02 | 0.67 ± 0.02 | 0.46 ± 0.03 | 0.44 ± 0.03 |
| lSVR    | 2.20 ± 0.05 | 0.61 ± 0.02 | 0.60 ± 0.02 | 0.37 ± 0.02 | 0.35 ± 0.03 |
| SVR     | 2.04 ± 0.05 | 0.68 ± 0.02 | 0.67 ± 0.02 | 0.45 ± 0.03 | 0.44 ± 0.03 |
| DNN     | 2.05 ± 0.05 | 0.67 ± 0.02 | 0.66 ± 0.02 | 0.45 ± 0.02 | 0.43 ± 0.03 |
| RF      | 1.88 ± 0.04 | 0.74 ± 0.02 | 0.73 ± 0.02 | 0.53 ± 0.02 | 0.52 ± 0.02 |
| eRF     | 1.86 ± 0.05 | 0.74 ± 0.02 | 0.74 ± 0.01 | 0.55 ± 0.02 | 0.54 ± 0.03 |
values are underestimated (Figure 4A). Thus, the greatest deviations from the experimental values are observed for those complexes with extremely low or high binding free energies (Figure 4B). There is, however, no clear relation between having a higher error value and the atom efficiency (Figure 4C).

Figure 4: (A) Binding free energies predicted by a single eRF model on unseen PDBbind test data. The dashed line indicates the ideal prediction. The absolute errors of the predictions shown in (A) are plotted against (B) the respective true $\Delta G$ and (C) the atom efficiency, which describes the $\Delta G$ contributed on average by each non-hydrogen atom (atom efficiency = $\Delta G/N_{\text{non-H-atoms}}$).

Results on external test sets

To compare our RASPD+ approach using eRF models with existing methods, we performed an evaluation on several external data sets from the literature\textsuperscript{19,21–23} that have different characteristics, as previously done by Jiménez et al.\textsuperscript{13} To compare to other methods for predicting protein-ligand binding free energy, we considered the previous RASPD approach\textsuperscript{6} as a method that does not rely on full docking, $K_{\text{DEEP}}$\textsuperscript{13} as a representative deep learning-based method, RF-Score\textsuperscript{45} as a method using random forests, and cyScore\textsuperscript{46} and X-Score\textsuperscript{47} as traditional docking scoring functions. Previously reported $RMSE$ values\textsuperscript{13} were transformed from errors in $pK$ values to errors in $\Delta G$ for the comparisons. With respect to the absolute deviation, measured by $RMSE$, the established scoring functions, RF-Score and X-Score
performed best (Table 2). Only on set 2 of the challenging CSAR-NRC HiQ release\textsuperscript{19} did RASPD+ with the eRF model have a lower RMSE, with a value of $2.23 \pm 0.04$ kcal/mol, than the existing methods. When considering the Pearson correlation as a proxy for the ranking performance, RASPD+ with eRF models not only achieved the best result on the CSAR-NRC HiQ set 2 ($r = 0.78 \pm 0.01$), but also achieved $r = 0.70 \pm 0.02$ on the data set curated by Wang et al. (Table 2).

Table 2: Comparison of the performance of RASPD+ using eRF models with five other methods to compute protein-ligand binding free energy. The $RMSE$ [kcal/mol] and Pearson correlation coefficients for predictions on five external test sets are given. The values for the other methods are taken from Jiménez et al.\textsuperscript{13} The values for the best performing models are shown in bold.

| Data set       | RASPD+  $\pm$ 0.04 | RASPD  3.43   | KDeep* 2.84 | RF-Score* 2.71 | CyScore* 3.18 | X-Score* 3.15 |
|----------------|-------------------|---------------|------------|----------------|-------------|-------------|
| CSAR-NRC HiQ 1 | 3.02              | 3.43          | 2.84       | **2.71**       | 3.18        | 3.15        |
| CSAR-NRC HiQ 2 | **2.23**          | 2.79          | 2.60       | 2.26           | 3.00        | 2.51        |
| CSAR12         | 1.50              | 1.93          | 2.17       | 1.36           | 2.84        | **1.27**   |
| CSAR14         | 1.36              | 2.05          | 2.39       | **1.19**       | 2.03        | 1.36        |
| Wang et al.    | 1.39              | 2.00          | 1.47       | **1.19**       | 5.74        | 1.49        |

| Data set       | RASPD+  $\pm$ 0.02 | RASPD  0.54   | KDeep* 0.72 | RF-Score* **0.77** | CyScore* 0.65 | X-Score* 0.60 |
|----------------|-------------------|---------------|------------|-------------------|-------------|-------------|
| CSAR-NRC HiQ 1 | 0.62              | 0.54          | 0.72       | **0.77**          | 0.65        | 0.60        |
| CSAR-NRC HiQ 2 | **0.78**          | 0.67          | 0.65       | 0.75              | 0.64        | 0.65        |
| CSAR12         | 0.40              | 0.29          | 0.37       | 0.46              | 0.26        | **0.48**   |
| CSAR14         | 0.55              | 0.32          | 0.61       | 0.80              | 0.67        | **0.82**   |
| Wang et al.    | **0.70**          | 0.55          | 0.29       | 0.24              | 0.27        | 0.25        |

*) $pK$ values reported by\textsuperscript{19} were converted to $\Delta G$ for comparison of $RMSE$ values.

The good performance of the RASPD+ eRF on the Wang et al. data set is also borne out in the distribution of predictions (Figure SI 3), which, compared to the results on CSAR-NRC HiQ (Figure SI 6), not only ranks but also faithfully captures the range of energies. On both the CSAR 2012 and CSAR 2014 data sets, clear failures of the RASPD+ eRF and most other methods can be observed. For some cases, the RASPD+ model predicts energies in a very narrow range around $-10.5$ kcal/mol (Figures SI 4, SI 5) but, interestingly, this value does not correspond to the mean $\Delta G$ value for the training data.
As the CSAR 2012\textsuperscript{21} and CSAR 2014\textsuperscript{22} releases and the data set from Wang et al.\textsuperscript{23} provided data for several ligands for each individual protein target, we analyzed the failure cases at the level of the individual proteins (Tables SI 6, SI 7, SI 8). The Pearson and Spearman correlations are below 0.3 for the BACE and CDK2 systems from the Wang et al.\textsuperscript{23} set and CHK1 and SYK in the CSAR sets. In contrast, the CDK2 complexes in the CSAR 2012 set\textsuperscript{21} achieved a Pearson correlation of $r = 0.50 \pm 0.05$. The highest correlations were observed for the PTP1B, Mcl-1, TYK2 systems in the Wang et al.\textsuperscript{23} data (Table SI 6) and for the CDK2-Cyclin A complex (Table SI 7) and TrmD on the CSAR data (Table SI 8). Strikingly, only for PTP1B, TYK2 and TrmD, was $R^2 > 0.3$ observed while all $Q^2_{F3}$ values were above 0.5.

**Feature importance analysis**

To assess which features contribute to accurate predictions, two strategies were chosen. By permutation feature importance, the contribution to the prediction was quantified by the change in the Pearson correlation coefficient after shuffling the values in the individual feature columns randomly. Three different model types – namely, linear regression, neural networks, and extremely random forests – showed different relative contributions of the individual features (Figure 5A). While LR assigned high contributions to a few features, the reduction in predictive performance for each shuffled feature was lower for eRF and the contribution signal was more evenly distributed among the different features. Molar refractivity (MR), which was the feature most strongly correlated with the target variable $\Delta G$, showed the strongest effect both in the LR and in the eRF models. For LR, randomizing MR almost completely removed the predictive power ($r < 0.2$). Among the protein features, all three methods showed high contributions for the general descriptor PlogP(Non-Arom) (Figure 5A). While both neural networks and eRF assigned high contributions to the PMR(Non-Arom), LR placed higher contributions on PlogP(Arom) and PMR(Arom) among the general protein features. PA(Amide-O), which had the second-highest correlation with $\Delta G$, showed a pronounced
signal only for the eRF model. The hydrogen bond acceptor count at the negatively charged amino acid residues (PA(D+E)) was informative for all three machine learning methods. In the eRF model, it had an importance value similar to the general protein features such as the residue log P values. This is especially surprising as no information on the ligand charge was provided and the count of positively charged amino acid hydrogen bond donors (PD(K+R+HIP)) did not contribute strongly to the predictions.

Additionally, we trained eRF models on subsets of the features and compared their performance to the full model (Figure 5B). Among the models trained on a single feature, the model trained on molar refractivity (MR) achieved better performance than that trained on molecular weight (MASS). Models trained on just the features of the protein pocket performed better than models using only ligand descriptors. In this case, the protein features still contained information about the ligand implicitly, as each protein descriptor is dependent on the size of the sphere surrounding the ligand. These reduced feature set models were also subjected to permutation feature importance analysis. For models with only ligand features, a very similar ranking of ligand features compared to the full training set was observed, illustrating the general preference for using those features for prediction (Figure SI 7).

When examining the feature importance for protein-only models, the backbone hydrogen bond acceptor (PA(Amide-O)) stands out compared to the feature importance on the full feature set (Figure SI 8). This could be partially explained by the fact that this feature showed a strong correlation with general ligand features (Figure 2) and thereby provides information related to general ligand size.

Enrichment of active molecules from the DUD data set

To assess the usefulness of our RASPD+ method, we simulated a drug discovery setting using 30 protein targets from the DUD data set, which contains several computationally generated decoys per active compound. For each of the seven machine learning models, we calculated enrichment factors (EF) to quantify how effective ranking by predicted binding free energies
Figure 5: Analysis of feature importance for the predictive performance of RASPD+ . (A) The average change in predictive performance as measured by Pearson correlation when the corresponding feature column was shuffled. Results are reported for linear regression (LR), artificial neural networks (DNN), and extremely random forests (eRF). (B) Ablation based analysis entailing training eRF models with different feature sets. No significant difference in performance was observed when adding the number of rotatable bonds. The models trained on just the values for molar refractivity (MR only) and molecular weight (MASS only) serve as lower bounds for measuring performance. The 'ligand only' and 'protein only' models trained on either only ligand or only protein features, respectively, perform better than the 'MR only' and 'MASS only' models but not as well as the models derived from all features. The protein features contain implicit information on the ligand size and this may indicate why the performance of the 'protein only' model is better than that of the 'ligand only' model. (See Figure 1 for details).
was at enriching active molecules from the whole data set (Table 3). We found that the linear LR and SVR models were the most effective when filtering to 1%, 5% and 10% of the samples, with EFs of $6.1 - 6.2 \pm 6$ when filtering down to 1% of the samples, followed by the random forest models eRF and RF with EF $5.0 - 5.4 \pm 5$. The high standard deviation in the mean EF resulted from high variability in the performance of different methods on individual proteins (Table SI 9). As methods that ranked on average less favorably provided the only acceptable enrichment on some of the systems, we chose a conservative approach to interpreting the results by combining the predictions of all the methods. We thus considered the union of the sets of top candidate molecules from all seven machine learning models. This combination outperformed the single models, as well as a mean ensemble model that averages the predicted binding free energy, with an EF of $7.3 \pm 5.9$ when enriching to 1% of the initial set. By excluding the kNN predictions from the union set the EF increased to $7.5 \pm 5.6$.

Table 3: Average enrichment factors with corresponding standard deviations for the top 1%, 5%, and 10% of the data selected for each of the 30 DUD targets analyzed. Union is the enrichment achieved by selecting the non-redundant set of candidate compounds obtained by combining the selections of each method.

| Method | Enrichment Factor (EF) |
|--------|------------------------|
|        | 1%   | 5%    | 10%   |
| eRF    | 5.0 $\pm$ 5.2         | 2.0 $\pm$ 1.7 | 1.3 $\pm$ 1.0 |
| RF     | 5.4 $\pm$ 5.3         | 2.1 $\pm$ 1.6 | 1.4 $\pm$ 0.9 |
| DNN    | 4.0 $\pm$ 4.2         | 1.6 $\pm$ 1.3 | 1.2 $\pm$ 0.9 |
| kNN    | 1.5 $\pm$ 3.1         | 1.0 $\pm$ 1.4 | 0.9 $\pm$ 0.9 |
| lSVR   | 6.1 $\pm$ 5.7         | 1.9 $\pm$ 1.6 | 1.3 $\pm$ 0.9 |
| SVR    | 4.4 $\pm$ 4.2         | 1.7 $\pm$ 1.4 | 1.3 $\pm$ 0.9 |
| LR     | 6.2 $\pm$ 5.9         | 1.9 $\pm$ 1.7 | 1.0 $\pm$ 1.0 |
| Mean ensemble | 5.4 $\pm$ 4.8 | 1.8 $\pm$ 1.5 | 1.2 $\pm$ 0.9 |
| Union  | 7.3 $\pm$ 5.9         | 2.6 $\pm$ 1.7 | 1.7 $\pm$ 0.9 |
| Union w/o kNN | 7.5 $\pm$ 5.6 | 2.6 $\pm$ 1.7 | 1.7 $\pm$ 1.0 |
Discussion

As the global health crisis surrounding the SARS-CoV-2 pandemic has demonstrated, there is a need for fast computational tools to accelerate drug design and development processes. The method we present here, RASPD+, is able to perform virtual screening of large libraries of compounds at a fraction of the time typically required for protein-ligand docking methods. This enables quick prioritization of candidates for a follow-up with more accurate yet computationally more demanding methods, such as docking. We achieved the speed up by training machine learning models on simple pose-invariant ligand and protein descriptors. With this simplified approach, we achieved results comparable to existing scoring functions when predicting the binding free energy $\Delta G$ on several data sets. By splitting the PDBbind training, testing, and validation data in a nested cross-validation setup, we were able to assess reliably that random forest models, and in particular the extremely random forest model, performed best on this type of data. While this splitting strategy increases confidence in the comparison of learning methods and feature importance analysis within the study, other data set splitting strategies, which explicitly control how similar proteins or ligands are between training and test sets, may be more appropriate to assess performance on completely different ligands or proteins directly. We accounted for this deficiency by not only testing the regression performance on different external test sets but also by assessing the ability of the RASPD+ models to enrich active molecules from a set of inactive decoys. Although the achievable enrichment factors were not as high as state-of-the-art docking or free energy prediction methods, RASPD+ still displayed appreciable enrichment of active molecules on the DUD data set. This is remarkable for two reasons. First, the training set only includes molecules displaying binding to their specific target protein. Secondly, 4 of the 6 physicochemical descriptors (molecular weight, hydrogen bond donor and acceptor count, and logP value), used to describe the ligand molecule, were initially used to select decoys similar to the active molecules for the DUD data set. This makes the task of distinguishing active and inactive molecules particularly difficult for our
models that employ only basic ligand descriptors. Notably, however, molar refractivity, which was not used for the creation of the DUD decoys, was not only a powerful predictor on its own ($r > 0.5$) but was also consistently assigned the highest feature importance among the ligand features. The high importance of MR agrees with results from a recent study that used ligand descriptors to enhance the performance of a common docking scoring function.

When analyzing cases where predictions failed, we were not able to identify a particular protein class for which enrichment or regression was particularly unreliable. Yet, it was clear from the individual enrichment results that different machine learning methods provided predictions of varying quality for the different protein systems. Random forest methods, which were best suited for the $\Delta G$ regression on known binders, were for some proteins outperformed by the simpler linear regression methods. This observation might support the recent finding that random forest methods, in particular, benefit from highly similar training molecules. Thus, considering the strengths and weaknesses of the different machine learning methods, we recommend that for applications of RASPD+, the results of the seven different machine learning methods are combined by picking top candidates from the rankings produced by each method.

If this approach is applied to pick the top 10% of RASPD+ candidates, this can provide a 10-fold reduction in the time spent for docking. Notably, we achieved computation times for RASPD+ that were over one hundred times faster than Glide SP docking (Schrödinger Release 2019-4: Glide, Schrödinger, LLC, New York, NY) on a laptop grade CPU (data not shown), meaning that computation times for RASPD+ screening are negligible compared to times for docking and molecular dynamics simulation.

Thus, the use of RASPD+ is clearly beneficial in time-critical applications of virtual screening of large compound libraries against individual protein targets. Moreover, higher structure-based screening throughput could also enable more effective inverse virtual screening of protein databases to assess the specificity and potential side-effects of candidate molecules.
Supporting Information Available

The supporting information contains Tables S1-10, Figures S1-9, and a description of the machine learning methods used. The code used for creating the models and figures in this paper, and scripts to run the RASPD+ pipeline are available on GitHub: https://github.com/HITS-MCM/RASPDplus Precomputed descriptor values and the model weights can be downloaded from https://doi.org/10.5281/zenodo.3937425

Author contributions

G.M. developed the descriptor pipeline and performed enrichment analysis. L.A. and S.H. trained and evaluated machine learning models. S.H. performed feature importance analysis. S.H., L.A., G.M., and R.C.W. analyzed data. B.J. provided advice for the development of the descriptor pipeline. R.C.W. provided guidance and supervised the work with G.M. S.H. wrote the manuscript with input from all authors.

Acknowledgement

This work was supported by the Klaus Tschira Foundation. We thank Dr. Daria Kokh and Jui-Hung Yuan for providing code from the TRAPP software \cite{TRAPP1,TRAPP2} for the calculation of binding pocket volume and Dr. Stefan Richter for technical assistance with the software and setting up the computing environment for the calculations. We thank Dr. Abhilash Jayaraj for help with program compilation. We thank Dr. Satyanarayana Rao and Vidhu S. Pandey for providing helpful utility code. We thank all those who made data on protein-ligand binding available in the databases that we have used.
Conflict of interest statement

The authors declare no competing interest.

References

(1) Torres, P. H. M.; Sodero, A. C. R.; Jofily, P.; Silva-Jr, F. P. Key Topics in Molecular Docking for Drug Design. *International Journal of Molecular Sciences* **2019**, *20*, 4574, DOI: 10.3390/ijms20184574.

(2) Wang, Z.; Sun, H.; Shen, C.; Hu, X.; Gao, J.; Li, D.; Cao, D.; Hou, T. Combined strategies in structure-based virtual screening. *Physical Chemistry Chemical Physics* **2020**, *22*, 3149–3159, DOI: 10.1039/C9CP06303J.

(3) Li, J.; Fu, A.; Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdisciplinary Sciences: Computational Life Sciences* **2019**, *11*, 320–328, DOI: 10.1007/s12539-019-00327-w.

(4) Willems, H.; De Cesco, S.; Svensson, F. Computational Chemistry on a Budget: Supporting Drug Discovery with Limited Resources. *Journal of Medicinal Chemistry* **2020**, *acs.jmedchem.9b02126*, DOI: 10.1021/acs.jmedchem.9b02126.

(5) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **2001**, *46*, 3–26, DOI: 10.1016/S0169-409X(00)00129-0.

(6) Mukherjee, G.; Jayaram, B. A rapid identification of hit molecules for target proteins via physico-chemical descriptors. *Physical Chemistry Chemical Physics* **2013**, *15*, 9107, DOI: 10.1039/c3cp44697b.
(7) Wiener, H. Structural Determination of Paraffin Boiling Points. *Journal of the American Chemical Society* **1947**, *69*, 17–20, DOI: 10.1021/ja01193a005.

(8) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *Journal of Chemical Information and Modeling* **1987**, *27*, 21–35, DOI: 10.1021/ci00053a005.

(9) Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H.; Stuckey, J. A.; Kang, Y. N.; Delproposto, J.; Kubish, G.; Dunbar, J. B.; Carlson, H. A.; Burley, S. K.; Walters, W. P.; Amaro, R. E.; Feher, V. A.; Gilson, M. K. D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 651–668, DOI: 10.1007/s10822-016-9946-8.

(10) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015**, *31*, 405–412, DOI: 10.1093/bioinformatics/btu626.

(11) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chemical Reviews* **2019**, *119*, 10520–10594, DOI: 10.1021/acs.chemrev.8b00728.

(12) Gomes, J.; Ramsundar, B.; Feinberg, E. N.; Pande, V. S. Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. *arXiv preprint* **2017**.

(13) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. K DEEP : Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **2018**, *58*, 287–296, DOI: 10.1021/acs.jcim.7b00650.

(14) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.;
Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Central Science* **2018**, *4*, 1520–1530, DOI: 10.1021/acscentsci.8b00507.

(15) Pei, J.; Zheng, Z.; Kim, H.; Song, L. F.; Walworth, S.; Merz, M. R.; Merz, K. M. Random Forest Refinement of Pairwise Potentials for Protein–Ligand Decoy Detection. *Journal of Chemical Information and Modeling* **2019**, *59*, 3305–3315, DOI: 10.1021/acs.jcim.9b00356.

(16) Boyles, F.; Deane, C. M.; Morris, G. M. Learning From The Ligand: Using Ligand-Based Features To Improve Binding Affinity Prediction. *Bioinformatics* **2020**, *36*, 758–764, DOI: 10.1093/bioinformatics/btz665.

(17) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry* **2006**, *49*, 6789–6801, DOI: 10.1021/jm0608356.

(18) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry* **2004**, *47*, 2977–2980, DOI: 10.1021/jm0305801.

(19) Dunbar, J. B.; Smith, R. D.; Yang, C.-Y.; Ung, P. M.-U.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes. *Journal of Chemical Information and Modeling* **2011**, *51*, 2036–2046, DOI: 10.1021/ci200082t.

(20) Smith, R. D.; Dunbar, J. B.; Ung, P. M.-U.; Esposito, E. X.; Yang, C.-Y.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. *Journal of Chemical Information and Modeling* **2011**, *51*, 2115–2131, DOI: 10.1021/ci200269q.

(21) Dunbar, J. B.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y.-N.; Kubish, G.; Gestwicki, J. E.;
Stuckey, J. A.; Carlson, H. A. CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. *Journal of Chemical Information and Modeling* 2013, 53, 1842–1852, DOI: 10.1021/ci4000486.

(22) 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.; Dunbar, J. B. CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *Journal of Chemical Information and Modeling* 2016, 56, 1063–1077, DOI: 10.1021/acs.jcim.5b00523.

(23) Wang, L. et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society* 2015, 137, 2695–2703, DOI: 10.1021/ja512751q.

(24) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* 2011, 3, 33, DOI: 10.1186/1758-2946-3-33.

(25) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* 2005, 26, 1668–1688, DOI: 10.1002/jcc.20290.

(26) Cumming, J. N. et al. Structure based design of iminohydantoin BACE1 inhibitors: Identification of an orally available, centrally active BACE1 inhibitor. *Bioorganic & Medicinal Chemistry Letters* 2012, 22, 2444–2449, DOI: https://doi.org/10.1016/j.bmcl.2012.02.013.

(27) Wang, L.; Deng, Y.; Knight, J. L.; Wu, Y.; Kim, B.; Sherman, W.; Shelley, J. C.; Lin, T.; Abel, R. Modeling Local Structural Rearrangements Using FEP/REST: Application to
Relative Binding Affinity Predictions of CDK2 Inhibitors. *Journal of Chemical Theory and Computation* **2013**, *9*, 1282–1293, DOI: 10.1021/ct300911a.

(28) Friberg, A.; Vigil, D.; Zhao, B.; Daniels, R. N.; Burke, J. P.; Garcia-Barrantes, P. M.; Camper, D.; Chauder, B. A.; Lee, T.; Olejniczak, E. T.; Fesik, S. W. Discovery of Potent Myeloid Cell Leukemia 1 (Mcl-1) Inhibitors Using Fragment-Based Methods and Structure-Based Design. *Journal of Medicinal Chemistry* **2013**, *56*, 15–30, DOI: 10.1021/jm301448p.

(29) Goldstein, D. M. et al. Discovery of 6-(2,4-difluorophenoxy)-2-[3-hydroxy-1-(2-hydroxyethyl) propylamino]-8-methyl-8*H*-pyrido[2,3-*d*]pyrimidin-7-one (pamapimod) and 6-(2,4-difluorophenoxy)-8-methyl-2-(tetrahydro-2*H*-pyran-4-ylamino)pyrido[2,3-*d*]pyrimidin-7(8*H*)-one (R1487). *Journal of Medicinal Chemistry* **2011**, *54*, 2255–2265, DOI: 10.1021/jm101423y.

(30) Wilson, D. P. et al. Structure-Based Optimization of Protein Tyrosine Phosphatase 1B Inhibitors: From the Active Site to the Second Phosphotyrosine Binding Site. *Journal of Medicinal Chemistry* **2007**, *50*, 4681–4698, DOI: 10.1021/jm0702478.

(31) Baum, B.; Mohamed, M.; Zayed, M.; Gerlach, C.; Heine, A.; Hangauer, D.; Klebe, G. More than a Simple Lipophilic Contact: A Detailed Thermodynamic Analysis of Non-basic Residues in the S1 Pocket of Thrombin. *Journal of Molecular Biology* **2009**, *390*, 56–69, DOI: https://doi.org/10.1016/j.jmb.2009.04.051.

(32) Liang, J. et al. Lead identification of novel and selective TYK2 inhibitors. *European Journal of Medicinal Chemistry* **2013**, *67*, 175–187, DOI: https://doi.org/10.1016/j.ejmech.2013.03.070.

(33) Liang, J. et al. Lead Optimization of a 4-Aminopyridine Benzamide Scaffold To Identify Potent, Selective, and Orally Bioavailable TYK2 Inhibitors. *Journal of Medicinal Chemistry* **2013**, *56*, 4521–4536, DOI: 10.1021/jm400266t.
(34) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 868–873, DOI: 10.1021/ci9903071.

(35) Kokh, D. B.; Richter, S.; Henrich, S.; Czodrowski, P.; Rippmann, F.; Wade, R. C. TRAPP: A Tool for Analysis of Transient Binding Pockets in Proteins. *Journal of Chemical Information and Modeling* **2013**, *53*, 1235–1252, DOI: 10.1021/ci4000294.

(36) Yuan, J.-H.; Han, S. B.; Richter, S.; Wade, R. C.; Kokh, D. B. Druggability Assessment in TRAPP Using Machine Learning Approaches. *Journal of Chemical Information and Modeling* **2020**, acs.jcim.9b01185, DOI: 10.1021/acs.jcim.9b01185.

(37) Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q2 parameter for QSAR validation. *Journal of Chemical Information and Modeling* **2009**, *49*, 1669–1678, DOI: 10.1021/ci900115y.

(38) Todeschini, R.; Ballabio, D.; Grisoni, F. Beware of Unreliable Q2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models. *Journal of Chemical Information and Modeling* **2016**, *56*, 1905–1913, DOI: 10.1021/acs.jcim.6b00277.

(39) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference. 1997; pp 155–161.

(40) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32, DOI: 10.1023/A:1010933404324.

(41) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Machine Learning* **2006**, *63*, 3–42, DOI: 10.1007/s10994-006-6226-1.
(42) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011, 12, 2825–2830.

(43) Chollet, F.; Others, Keras. https://keras.io, 2015.

(44) Kotila, M. Talos. 2018; https://github.com/autonomio/talos.

(45) Ballester, P. J.; Mitchell, J. B. O. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* 2010, 26, 1169–1175, DOI: 10.1093/bioinformatics/btq112.

(46) Cao, Y.; Li, L. Improved protein–ligand binding affinity prediction by using a curvature-dependent surface-area model. *Bioinformatics* 2014, 30, 1674–1680, DOI: 10.1093/bioinformatics/btu104.

(47) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design* 2002, 16, 11–26.

(48) Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* 2020, 579, 265–269, DOI: 10.1038/s41586-020-2008-3.

(49) Irwin, J. J.; Shoichet, B. K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 2005, 45, 177–182, DOI: 10.1021/CI049714+

(50) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* 2006, 34, D668–72, DOI: 10.1093/nar/gkj067.

(51) 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 Depended on the Training
Set? *Journal of Chemical Information and Modeling* 2020, acs.jcim.9b00714, DOI: 10.1021/acs.jcim.9b00714.

(52) Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling* 2019, 59, 947–961, DOI: 10.1021/acs.jcim.8b00712.

(53) Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *Journal of Chemical Information and Modeling* 2014, 54, 1717–1736, DOI: 10.1021/ci500081m.

(54) Lagarde, N.; Zagury, J.-F.; Montes, M. Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives. *Journal of Chemical Information and Modeling* 2015, 55, 1297–1307, DOI: 10.1021/acs.jcim.5b00090.

(55) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry* 2004, 47, 1739–49, DOI: 10.1021/jm0306430.

(56) Nair, V.; Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. Proceedings of the 27th International Conference on International Conference on Machine Learning. USA, 2010; pp 807–814.

(57) Clevert, D. A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (ELUs). 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings. 2016.

(58) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 2014, 15, 1929–1958.
(59) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536, DOI: 10.1038/323533a0.
Supporting Information

Supplementary methods

Details regarding the different machine learning methods used in this study.

Linear Regression (LR)

In the previous RASPD approach, the simplifying assumption was made that the dependent variable \( y \), binding free energy, follows the physicochemical features \( x \) with a linear relationship. Therefore, all contributions behave purely additively. The parameters \( w \) and \( b \) in such a linear regression model (eqn. 3) can be fit to the data with the ordinary least squares algorithm.

\[
\hat{y} = w^T \cdot x + b = \sum_{i}^{n_{\text{features}}} w_i \cdot x_i + b
\]

(3)

Thus, if the features are linearly independent and possess a linear relationship with the output value, the coefficients \( w \) are informative about the relationship of the features to the predicted value in the model.

k-Nearest Neighbor (kNN)

A naive way to capture nonlinear relationships between the input features \( x \) and the output value \( y \) is to rely on a nearest-neighbor based approach to assign the binding free energy values of the closest training examples in feature space. As Euclidean distance is most commonly used, no explicit prioritization of features is enforced but dependent on the separation of training data points in the feature space. A tunable hyperparameter controls how many of the \( k \) closest neighbors are averaged to obtain a prediction \( \hat{y} \).

Support Vector Regression (SVR)

Another approach to achieve separation based on distance in the feature space is support vector regression. Here, a regression function is described based on the inner product of...
the input features and a set of the training examples, the support vectors. Regularization is applied to minimize the contribution of support vectors within an error margin $\epsilon$ allowing to tune the bias-variance trade-off with an additional regularization parameter $C$. The necessary inner product can be computed not only linearly on the feature space but also in a high dimensional space using a kernel function as a similarity measure, thus allowing for nonlinear models. The most common kernel is the radial basis function ($rbf$) which decays with distance in feature space according to a Gaussian density function.

**Neural Networks – Deep Learning (DNN)**

Neural networks introduce nonlinearity and the ability to capture complex relationships by stacking multiple layers of linear matrix multiplication operations interleaved with a nonlinear function $f$.

\[
\begin{align*}
    h_0 &= x \\
    \text{for } n \text{ layers} & \\
    h_i &= f(w_i h_{i-1} + b_i) \\
    y &= h_n
\end{align*}
\]

Such a nonlinear activation function $f$ could simply be a Rectified Linear Unit (ReLU) or an Exponential Linear Unit (ELU), which has the additional benefit of being continuous and smooth.

\[
ReLU(x) = \max(0, x)
\]

\[
ELU(x) = \begin{cases}
    x & \text{if } x \geq 0 \\
    e^x - 1 & \text{otherwise}
\end{cases}
\]
The number and size of these layers have to be chosen according to the complexity of the problem. To avoid overfitting, dropout regularization with an additional dropout probability hyperparameter can be adopted. Training of the neural network models was performed using mini-batch stochastic gradient descent with mean squared error as the loss function. This leads to the additional training hyperparameters of batch size, learning rate, and the number of epochs (full iterations over the training set).

**Random Forests – Decision Tree Ensembles**

Decision trees try to separate data by recursively finding features and thresholds that split the data into two groups with the most dissimilar output value, which are on the other hand as similar as possible within each group. As single decision trees tend to overfit the training data, several weak decision tree regressors are combined in their predictions by bagging resulting in random forests (RF). To generate independent regressors based on the same training data, two strategies are used: In the original random forest implementation, bootstrap samples of the training data for each tree increase diversity. Another approach termed extremely random forests (eRF) or extra trees sets the decision boundary for a given feature at random. Thus, the algorithm only picks the best-separating feature given the random boundary rather than additionally computing the best boundary. In both methods, only a random subset of the available features is used at each branch point when searching for the best-separating feature. To control the bias-variance trade-off, several hyperparameters can be set: The number of trees controls the achievable bias and regularizes the variance until the number of trees exceeds the number of truly independent tree samples. The method by which subsets of features are chosen determines how diverse or random the resulting trees are. Additionally, the number of splits can be controlled by a limit to how many samples can be placed in a leaf node.
Supplementary tables
Table SI 1: Content and origin of the datasets of protein-ligand complexes from D3R ([drugdesigndata.org](https://drugdesigndata.org)) CSAR12 and CSAR14 - and from Wang et al.

| Data set | Protein target | PDB ID | # ligands | Source of ligand structures and experimental affinities |
|----------|----------------|--------|-----------|--------------------------------------------------------|
| CSAR12   | Urokinase      | 5YC6   | 35        | [https://drugdesigndata.org/php/file-download.php?type=extended&id=76](https://drugdesigndata.org/php/file-download.php?type=extended&id=76) |
|          | CDK2-Kinase    | 1H1Q   | 25        | [https://drugdesigndata.org/php/file-download.php?type=extended&id=99](https://drugdesigndata.org/php/file-download.php?type=extended&id=99) |
|          | CDK2-CyclinA   | 4GCJ   | 23        | [https://drugdesigndata.org/php/file-download.php?type=extended&id=111](https://drugdesigndata.org/php/file-download.php?type=extended&id=111) |
|          | CHK1-Kinase    | 2YEX   | 110       | [https://drugdesigndata.org/php/file-download.php?type=extended&id=70](https://drugdesigndata.org/php/file-download.php?type=extended&id=70) |
|          | ERK2           | 4ZZN   | 298       | [https://drugdesigndata.org/php/file-download.php?type=extended&id=71](https://drugdesigndata.org/php/file-download.php?type=extended&id=71) |
|          | LpXc           | 3UHM   | 20        | [https://drugdesigndata.org/php/file-download.php?type=extended&id=73](https://drugdesigndata.org/php/file-download.php?type=extended&id=73) |
| CSAR14   | SYK            | 5LMA   | 583       | [https://drugdesigndata.org/php/file-download.php?type=extended&id=74](https://drugdesigndata.org/php/file-download.php?type=extended&id=74) |
|          | tRMD           | 4YQD   | 31        | [https://drugdesigndata.org/php/file-download.php?type=extended&id=75](https://drugdesigndata.org/php/file-download.php?type=extended&id=75) |
|          | HSP90          | 4YKU   | 146       | [https://drugdesigndata.org/php/file-download.php?type=extended&id=100](https://drugdesigndata.org/php/file-download.php?type=extended&id=100) |
| Wang et al. | BACE        | 4DJW   | 36        | Ligand 2D-structure; experimental parameters: SI of [23] |
|          | CDK2           | 1H1Q   | 16        | Ligand 2D-structure; experimental parameters: SI of [23] |
|          | MCL1           | 4HW3   | 67        | Ligand 2D-structure; experimental parameters: SI of [23] |
|          | p38            | 3FLY   | 34        | Ligand 2D-structure; experimental parameters: SI of [23] |
|          | PTP1B          | 2QBS   | 48        | Ligand 2D-structure; experimental parameters: SI of [23] |
|          | Thrombin       | 2ZFF   | 11        | Ligand 2D-structure; experimental parameters: SI of [23] |
|          | TYK2           | 4GIH   | 71        | Ligand 2D-structure; experimental parameters: SI of [23] |
Table SI 2: Hyperparameters evaluated for the different machine learning methodologies. The combinatorial space was explored by a naive grid search over all reasonable combinations. The most frequently chosen hyperparameter configuration is indicated in bold font.

| Model | Parameter | Values |
|-------|-----------|--------|
| (l)SVR | $C$ | $0.01, 0.1, 1, 10$ |
|       | $\epsilon$ | $0, 0.1$ |
|       | Kernel | linear, gaussian rbf |
|       | kernel parameter $\gamma$ | scikit-learn auto |
| kNN | $k$ | $1, 5, 10, 20$ |
|       | distance metric | **Euclidean**($L2$) |
| RF | Number of trees | $20, 100, 200$ |
|     | Ratio of features | $log_2, sqrt$ |
|     | min. number of samples per leaf | $1, 3, 5$ |
| eRF | Number of trees | $20, 100, 200$ |
|     | Ratio of features | $log_2, sqrt$ |
|     | min. number of samples per leaf | $1, 3, 5$ |
| DNN | Number of hidden layers | $2, 3$ |
|     | Activation Function | ReLU, ELU |
|     | Dropout Probability | $0, 0.1, 0.2$ |
|     | Hidden layer size | $10, 20, [layer 1: 20, layer 2: 10]$ |
|     | Optimizer | SGD, Adam |
|     | Learning rate | $0.01, 0.003, 0.001$ |
|     | Loss function | MSE |
|     | Initializer | variance scaling ($\sqrt{a/fan-in}$), $a = 1, 2$, normal/uniform |
Table SI 3: Values of the Spearman correlation between the protein and ligand descriptors and $\Delta G$ in the PDBbind refined data set. The highest correlation is observed for MR and shown in bold.

| Descriptor | $\rho(\Delta G, \text{Descriptor})$ |
|------------|------------------------------------|
| PA(D+E)    | -0.240                             |
| PA(N+Q+T+S+DH+EH) | -0.150                     |
| PA(Y+H)    | 0.061                              |
| PD(K+R+HIP) | -0.120                             |
| PD(LYN+N+Q) | 0.060                              |
| PD(T+S+Y+DH+EH) | -0.087                        |
| PD(W+H)    | 0.053                              |
| PA(Amide-O) | -0.490                             |
| PD(Amide-NH) | -0.447                           |
| PlogP(Arom) | -0.190                             |
| PlogP(Non-Arom) | -0.450                          |
| PMR(Arom)  | -0.187                             |
| PMR(Non-Arom) | -0.460                        |
| PVol       | -0.224                             |
| A          | -0.187                             |
| D          | -0.091                             |
| logP       | -0.364                             |
| W          | -0.474                             |
| MR         | **-0.508**                         |
| MASS       | -0.454                             |
Table SI 4: Comparison of the performance of different models on the PDBbind training set using five different metrics. Results for the different random test set splits and cross-validation folds were averaged and the mean and standard deviation are reported. The RMSE is given in kcal/mol. NA: Not applicable

| Model   | RMSE   | r      | ρ      | $R^2$   | $Q^2_{F3}$ |
|---------|--------|--------|--------|--------|------------|
| null model | 2.72 ± 0.02 | 0.00 ± 0.00 | NA     | 0.00 ± 0.00 | 0.00 ± 0.00 |
| LR      | 2.17 ± 0.01 | 0.60 ± 0.01 | 0.60 ± 0.01 | 0.36 ± 0.01 | 0.36 ± 0.01 |
| kNN     | 1.65 ± 0.05 | 0.80 ± 0.02 | 0.79 ± 0.02 | 0.63 ± 0.02 | 0.63 ± 0.02 |
| lSVR    | 2.19 ± 0.02 | 0.60 ± 0.01 | 0.59 ± 0.01 | 0.36 ± 0.01 | 0.36 ± 0.01 |
| SVR     | 1.74 ± 0.01 | 0.77 ± 0.00 | 0.77 ± 0.00 | 0.59 ± 0.01 | 0.59 ± 0.01 |
| DNN     | 1.86 ± 0.03 | 0.73 ± 0.01 | 0.73 ± 0.01 | 0.53 ± 0.02 | 0.53 ± 0.02 |
| RF      | 0.70 ± 0.01 | 0.98 ± 0.00 | 0.98 ± 0.00 | 0.93 ± 0.00 | 0.93 ± 0.00 |
| eRF     | 0.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |

Table SI 5: Comparison of the performance of different models on the PDBbind validation set using five different metrics. Results for the different random test set splits and cross-validation folds were averaged and the mean and standard deviation are reported. The RMSE is given in kcal/mol. NA: Not applicable

| model   | RMSE   | r      | ρ      | $R^2$   | $Q^2_{F3}$ |
|---------|--------|--------|--------|--------|------------|
| null model | 2.72 ± 0.07 | 0.00 ± 0.00 | NA     | −0.00 ± 0.00 | −0.00 ± 0.06 |
| LR      | 2.19 ± 0.06 | 0.59 ± 0.02 | 0.59 ± 0.03 | 0.35 ± 0.03 | 0.35 ± 0.04 |
| kNN     | 2.02 ± 0.05 | 0.68 ± 0.02 | 0.67 ± 0.02 | 0.45 ± 0.03 | 0.45 ± 0.03 |
| lSVR    | 2.20 ± 0.07 | 0.59 ± 0.02 | 0.59 ± 0.03 | 0.35 ± 0.03 | 0.35 ± 0.04 |
| SVR     | 2.03 ± 0.06 | 0.67 ± 0.02 | 0.66 ± 0.02 | 0.45 ± 0.03 | 0.45 ± 0.04 |
| DNN     | 2.02 ± 0.05 | 0.67 ± 0.02 | 0.66 ± 0.02 | 0.45 ± 0.03 | 0.45 ± 0.03 |
| RF      | 1.88 ± 0.05 | 0.73 ± 0.02 | 0.72 ± 0.02 | 0.52 ± 0.02 | 0.52 ± 0.03 |
| eRF     | 1.85 ± 0.05 | 0.74 ± 0.02 | 0.73 ± 0.02 | 0.54 ± 0.02 | 0.54 ± 0.03 |
Table SI 6: Performance of the model for the individual protein targets provided by Wang et al.\textsuperscript{23}

| data set            | N   | RMSE    | r      | $\rho$     | $R^2$   | $Q_{F3}^2$ |
|---------------------|-----|---------|--------|------------|---------|-----------|
| BACE                | 36  | 0.96 ± 0.06 | -0.1 ± 0.1 | -0.1 ± 0.1 | -0.5 ± 0.2 | 0.87 ± 0.02 |
| CDK2                | 16  | 1.8 ± 0.1   | -0.2 ± 0.2 | 0.0 ± 0.2  | -1.4 ± 0.3 | 0.55 ± 0.05  |
| Mcl-1               | 67  | 1.86 ± 0.08 | 0.70 ± 0.02 | 0.64 ± 0.07 | -0.5 ± 0.1 | 0.53 ± 0.04  |
| PTP1B               | 48  | 1.22 ± 0.09 | 0.78 ± 0.02 | 0.72 ± 0.03 | 0.48 ± 0.08 | 0.80 ± 0.03  |
| TYK2                | 71  | 1.13 ± 0.03 | 0.67 ± 0.01 | 0.54 ± 0.03 | 0.39 ± 0.03 | 0.83 ± 0.01  |
| Thrombin            | 11  | 1.7 ± 0.1   | 0.4 ± 0.1  | 0.34 ± 0.09 | -2.6 ± 0.5 | 0.61 ± 0.06  |
| p38                 | 34  | 0.98 ± 0.04 | 0.3 ± 0.1  | 0.3 ± 0.1  | 0.03 ± 0.07 | 0.87 ± 0.01  |

Table SI 7: Performance of the model for the individual protein targets in the CSAR 2012 data set.\textsuperscript{21}

| data set            | N   | RMSE    | r      | $\rho$     | $R^2$   | $Q_{F3}^2$ |
|---------------------|-----|---------|--------|------------|---------|-----------|
| CDK2                | 25  | 1.9 ± 0.1 | 0.50 ± 0.05 | 0.50 ± 0.07 | -1.8 ± 0.3 | 0.53 ± 0.06 |
| CDK2-Cyclin A       | 23  | 1.09 ± 0.09 | 0.65 ± 0.03 | 0.47 ± 0.04 | -0.1 ± 0.2 | 0.84 ± 0.03 |
| CHK1                | 110 | 1.71 ± 0.04 | 0.12 ± 0.03 | 0.05 ± 0.03 | -0.33 ± 0.06 | 0.60 ± 0.02 |
| ERK2                | 298 | 1.34 ± 0.02 | 0.36 ± 0.05 | 0.33 ± 0.05 | 0.08 ± 0.03 | 0.76 ± 0.01 |
| LpxC                | 20  | 2.1 ± 0.1  | 0.38 ± 0.03 | 0.30 ± 0.04 | -0.6 ± 0.2 | 0.39 ± 0.06 |
| Urokinase           | 35  | 1.60 ± 0.06 | 0.46 ± 0.02 | 0.46 ± 0.03 | -0.4 ± 0.1 | 0.66 ± 0.03 |

Table SI 8: Performance of the model for the individual protein targets in the CSAR 2014 data set.\textsuperscript{23}

| data set            | N   | RMSE    | r      | $\rho$     | $R^2$   | $Q_{F3}^2$ |
|---------------------|-----|---------|--------|------------|---------|-----------|
| HSP90               | 146 | 1.81 ± 0.04 | 0.33 ± 0.02 | 0.32 ± 0.02 | -0.40 ± 0.06 | 0.56 ± 0.02 |
| SYK                 | 583 | 1.23 ± 0.04 | 0.0 ± 0.1  | 0.0 ± 0.1  | -0.10 ± 0.07 | 0.79 ± 0.01 |
| TrmD                | 31  | 1.19 ± 0.06 | 0.63 ± 0.03 | 0.46 ± 0.04 | 0.35 ± 0.07 | 0.81 ± 0.02 |
Table SI 9: Enrichment factors for protein targets in the processed DUD data set for different fractions of the total data selected. PDB IDs of the used structures in parentheses. Union describes the combination of all machine learning methods, where redundant molecules are removed.

| Target     | Fraction (%) | eRF | RF  | DNN | kNN | lSVR | SVR  | LR  | Union |
|------------|--------------|-----|-----|-----|-----|------|------|-----|-------|
| ACE (1O86) | 1            | 13.8| 7.6 | 8.5 | 0.0 | 6.2  | 7.2  | 6.8 | 11.8  |
|            | Active: 48   | 5   | 5.0 | 4.4 | 4.4 | 0.0  | 2.9  | 4.1 | 2.5   | 5.5   |
|            | Decoys: 1786 | 10  | 3.0 | 2.8 | 3.1 | 0.2  | 1.5  | 2.7 | 1.4   | 3.1   |
| ACHE (1EVE)| 1            | 5.9 | 7.8 | 4.0 | 0.9 | 6.8  | 5.4  | 6.9 | 9.6   |
|            | Active: 103  | 5   | 3.1 | 3.3 | 1.9 | 1.8  | 2.6  | 2.4 | 2.2   | 2.7   |
|            | Decoys: 3848 | 10  | 2.2 | 2.0 | 1.8 | 1.5  | 2.1  | 1.7 | 1.9   | 1.9   |
| ADA (1NDW)| 1            | 3.0 | 4.9 | 3.6 | 0.3 | 8.2  | 2.5  | 8.2 | 8.3   |
|            | Active: 39   | 5   | 1.5 | 1.6 | 2.0 | 1.0  | 1.6  | 1.4 | 1.6   | 1.7   |
|            | Decoys: 920  | 10  | 0.8 | 0.8 | 1.6 | 0.7  | 1.2  | 1.4 | 1.1   | 1.0   |
| AMPC (1XGJ)| 1            | 1.1 | 2.7 | 0.0 | 0.0 | 0.0  | 0.0  | 0.0 | 0.0   |
|            | Active: 21   | 5   | 1.6 | 1.9 | 0.5 | 0.7  | 0.0  | 0.0 | 0.0   | 1.5   |
|            | Decoys: 778  | 10  | 1.0 | 1.5 | 1.7 | 0.7  | 0.3  | 0.0 | 0.1   | 1.2   |
| AR (1XQ2) | 1            | 10.6| 8.9 | 5.2 | 0.6 | 9.3  | 6.3  | 9.6 | 15.3  |
|            | Active: 79   | 5   | 4.3 | 3.7 | 1.4 | 0.2  | 2.9  | 1.9 | 2.0   | 4.4   |
|            | Decoys: 2824 | 10  | 2.9 | 3.2 | 1.1 | 0.3  | 1.7  | 1.3 | 1.4   | 2.6   |
| CDK2 (1CKP)| 1            | 3.6 | 4.0 | 8.8 | 0.0 | 6.9  | 6.5  | 7.2 | 9.0   |
|            | Active: 70   | 5   | 1.4 | 1.8 | 2.7 | 0.6  | 2.8  | 2.7 | 2.6   | 3.0   |
|            | Decoys: 2055 | 10  | 1.1 | 1.3 | 1.6 | 0.6  | 2.1  | 1.6 | 1.4   | 1.6   |
| COX2 (5IKR)| 1            | 1.7 | 2.6 | 2.6 | 0.0 | 2.4  | 1.8  | 1.4 | 3.2   |
|            | Active: 424  | 5   | 0.6 | 0.8 | 1.4 | 0.0  | 1.0  | 1.9 | 0.6   | 1.4   |
|            | Decoys: 13098| 10  | 0.4 | 0.5 | 1.3 | 0.0  | 1.0  | 1.5 | 0.5   | 1.0   |
| DHFR (3DFR)| 1            | 0.8 | 0.4 | 0.0 | 0.6 | 1.6  | 0.4  | 2.4 | 1.6   |
|            | Active: 409  | 5   | 0.5 | 0.6 | 0.1 | 1.5  | 0.7  | 0.3 | 1.5   | 1.0   |
| Protein          | ID    | Decoys | 0.4 | 0.5 | 0.2 | 1.5 | 0.8 | 0.3 | 1.2 | 1.1 |
|------------------|-------|--------|-----|-----|-----|-----|-----|-----|-----|-----|
| EGFR (1M17)      |       | 8216   |     |     |     |     |     |     |     |     |
| Active: 470      |       |        | 13.7| 12.0| 10.0| 1.2 | 13.2| 8.6 | 13.9| 12.3|
| Decoys: 15849    |       | 10     | 2.6 | 2.3 | 2.1 | 1.0 | 2.7 | 2.2 | 2.7 | 2.6 |
| ER (2P15)        |       | 15849  | 3.5 | 0.5 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 |
| Active: 39       |       |        | 2.1 | 1.8 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.9 |
| Decoys: 1435     |       | 10     | 1.4 | 1.3 | 0.0 | 1.1 | 0.0 | 0.0 | 0.0 | 0.9 |
| ER-agonist (2P15)|       | 1435   | 2.3 | 2.9 | 2.4 | 2.1 | 2.3 | 1.5 | 1.5 | 2.8 |
| Active: 67       |       |        | 1.0 | 1.5 | 0.9 | 0.8 | 0.6 | 0.9 | 0.6 | 0.9 |
| Decoys: 2557     |       | 10     | 0.9 | 1.2 | 0.8 | 0.7 | 0.9 | 0.8 | 0.7 | 0.9 |
| FGFR1 (1AGW)     |       | 4482   | 0.7 | 1.5 | 0.8 | 0.0 | 18.9| 1.8 | 20.3| 8.1 |
| Active: 120      |       |        | 0.4 | 0.9 | 0.5 | 0.1 | 5.9 | 1.0 | 6.3 | 4.3 |
| Decoys: 2132     |       | 10     | 0.4 | 0.9 | 0.4 | 0.2 | 3.5 | 0.7 | 3.5 | 2.8 |
| FXA (1EZQ)       |       | 5678   | 0.6 | 0.5 | 0.0 | 0.1 | 0.7 | 0.0 | 0.7 | 0.7 |
| Active: 144      |       |        | 0.2 | 0.3 | 0.0 | 0.0 | 0.3 | 0.2 | 0.3 | 0.3 |
| Decoys: 2915     |       | 10     | 0.2 | 0.3 | 0.1 | 0.1 | 0.3 | 0.3 | 0.4 | 0.2 |
| GPB (1A8I)       |       | 2915   | 3.4 | 5.8 | 1.6 | 2.0 | 4.0 | 0.2 | 4.0 | 6.8 |
| Active: 77       |       |        | 1.8 | 2.1 | 0.9 | 1.2 | 1.3 | 0.6 | 0.9 | 2.2 |
| Decoys: 2132     |       | 10     | 1.6 | 1.3 | 1.2 | 1.3 | 0.9 | 1.0 | 0.9 | 1.5 |
| GR (1M2Z)        |       | 2132   | 0.4 | 1.6 | 2.7 | 0.8 | 0.0 | 2.7 | 0.0 | 2.5 |
| Active: 77       |       |        | 0.4 | 0.8 | 1.4 | 1.1 | 0.5 | 1.6 | 0.5 | 1.3 |
| Decoys: 2915     |       | 10     | 0.4 | 0.5 | 1.3 | 1.4 | 0.4 | 1.5 | 0.4 | 1.1 |
| HIVPR (1HPX)     |       | 2915   | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Active: 62       |       |        | 0.0 | 0.1 | 0.3 | 0.2 | 0.3 | 0.0 | 0.0 | 0.2 |
| Decoys: 2012     |       | 10     | 0.1 | 0.2 | 0.3 | 0.2 | 0.2 | 0.1 | 0.1 | 0.3 |
| HIVRT (1RT1)     |       | 2012   | 5.1 | 10.1| 9.9 | 0.2 | 4.8 | 12.6| 4.8 | 8.7 |
| Active: 43       |       |        | 2.5 | 3.3 | 3.2 | 0.3 | 2.1 | 3.3 | 2.5 | 3.8 |
| Protein  | Decoys | 1 | 1.5 | 1.7 | 2.2 | 0.2 | 1.6 | 2.2 | 1.6 | 2.2 |
|----------|--------|---|-----|-----|-----|-----|-----|-----|-----|-----|
| Decoys: 1514 | 10 | 3.8 | 4.9 | 12.0 | 0.8 | 0.0 | 7.4 | 0.0 | 7.4 |
| HSP90 (1UY6) | 1 | 1.1 | 1.2 | 3.2 | 1.3 | 1.1 | 3.6 | 11.1 | 3.2 |
| Active: 37 | 5 | 0.6 | 0.8 | 1.9 | 0.9 | 0.5 | 2.3 | 0.5 | 1.9 |
| Decoys: 973 | 10 | 10.5 | 15.9 | 4.3 | 1.3 | 9.3 | 8.3 | 9.4 | 8.9 |
| InhA (1P44) | 1 | 3.5 | 4.6 | 2.2 | 0.6 | 2.3 | 2.8 | 2.4 | 4.0 |
| Active: 86 | 5 | 2.1 | 2.5 | 1.4 | 0.4 | 1.7 | 1.8 | 1.7 | 2.3 |
| Decoys: 3217 | 10 | 7.1 | 7.1 | 7.1 | 10.0 | 7.1 | 7.1 | 7.1 | 19.6 |
| MR (2AA2) | 1 | 5.9 | 5.0 | 2.1 | 2.7 | 2.1 | 2.5 | 1.7 | 5.4 |
| Active: 15 | 5 | 1.6 | 2.1 | 1.7 | 3.9 | 1.4 | 0.7 | 1.1 | 3.2 |
| Decoys: 627 | 10 | 17.6 | 16.6 | 5.2 | 2.1 | 8.4 | 7.5 | 8.4 | 16.1 |
| NA (1A4G) | 1 | 5.9 | 5.0 | 2.1 | 2.7 | 2.1 | 2.5 | 1.7 | 5.4 |
| Active: 48 | 5 | 3.5 | 2.8 | 1.3 | 2.4 | 1.3 | 1.6 | 1.4 | 3.0 |
| Decoys: 1864 | 10 | 1.0 | 1.9 | 1.3 | 0.7 | 3.5 | 2.8 | 4.7 | 3.0 |
| P38 (1KV2) | 1 | 0.1 | 0.1 | 3.8 | 0.0 | 3.6 | 4.5 | 3.6 | 5.3 |
| Active: 454 | 5 | 0.4 | 0.4 | 1.5 | 0.1 | 0.1 | 1.6 | 3.0 | 1.9 | 2.1 |
| Decoys: 9041 | 10 | 0.1 | 0.1 | 3.8 | 0.0 | 3.6 | 4.5 | 3.6 | 5.3 |
| PDE5 (1XP0) | 1 | 0.9 | 0.4 | 0.1 | 1.1 | 0.0 | 0.3 | 0.0 | 0.4 |
| Active: 86 | 5 | 0.4 | 0.4 | 1.5 | 0.1 | 0.1 | 1.6 | 3.0 | 1.9 | 2.1 |
| Decoys: 1950 | 10 | 0.4 | 0.4 | 1.5 | 0.1 | 0.1 | 1.6 | 3.0 | 1.9 | 2.1 |
| PPARg (1FM9_D) | 1 | 0.2 | 0.0 | 0.0 | 1.6 | 0.0 | 0.1 | 0.0 | 0.8 |
| Active: 85 | 5 | 0.9 | 0.4 | 0.1 | 1.1 | 0.0 | 0.3 | 0.0 | 0.4 |
| Decoys: 3095 | 10 | 0.7 | 0.4 | 0.1 | 1.2 | 0.2 | 0.2 | 0.2 | 0.5 |
| PR (1SR7) | 1 | 3.9 | 6.2 | 3.9 | 1.2 | 7.8 | 6.2 | 7.8 | 9.1 |
| Active: 27 | 5 | 1.7 | 1.5 | 1.1 | 0.9 | 1.7 | 1.4 | 1.5 | 1.6 |
| Decoys: 1027 | 10 | 1.2 | 1.0 | 0.7 | 0.6 | 1.0 | 0.9 | 1.0 | 0.9 |
| SRC (2SRC) | 1 | 10.5 | 11.8 | 3.4 | 2.2 | 19.8 | 7.3 | 20.4 | 13.5 |
| Active: 158 | 5 | 5.1 | 5.6 | 2.0 | 1.4 | 6.2 | 3.4 | 6.1 | 5.4 |
| Protein     | Decoys | 3.1 | 3.2 | 1.6 | 1.2 | 3.5 | 2.4 | 3.7 | 3.3 |
|-------------|--------|-----|-----|-----|-----|-----|-----|-----|-----|
| Thrombin (1BA8) | 6235   | 10  | 2.5 | 1.8 | 0.0 | 0.0 | 1.4 | 0.0 | 1.4 | 2.0 |
| Active: 72  |        | 5   | 1.5 | 0.7 | 0.0 | 0.2 | 0.4 | 0.0 | 0.4 | 0.9 |
| Decoys: 2439|        | 10  | 1.1 | 0.6 | 0.0 | 0.3 | 0.3 | 0.2 | 0.3 | 0.7 |
| TK (1KIM)  | 2439   | 10  | 13.7| 13.7| 12.8| 11.5| 13.7| 13.7| 13.7| 14.9|
| Active: 22  |        | 5   | 3.4 | 3.6 | 2.7 | 2.7 | 2.7 | 2.7 | 2.7 | 3.4 |
| Decoys: 885 |        | 10  | 1.9 | 2.1 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.8 |
| Trypsin (1BJU)| 1649   | 10  | 0.4 | 0.3 | 2.6 | 0.2 | 2.4 | 2.7 | 3.9 | 1.6 |
| Active: 49  |        | 5   | 0.5 | 0.5 | 2.5 | 0.1 | 2.1 | 2.7 | 3.8 | 1.6 |
| Decoys: 1649|        | 10  | 0.4 | 0.3 | 2.6 | 0.2 | 2.4 | 2.7 | 3.9 | 1.6 |
| VEGF2 (1VR2)| 2873   | 10  | 1.4 | 1.6 | 1.0 | 1.5 | 2.5 | 1.1 | 2.5 | 2.5 |
Supplementary figures

Figure SI 1: Distribution of binding free energy ($\Delta G$) in the protein-ligand complexes in the training and external data sets. D3R contains protein-ligand complexes from both the CSAR 2012 and 2014 releases.\cite{21,22}
Dataset

Figure SI 2: Distribution of the protein and ligand descriptor values in the training set and the external test sets shown colored by data set. The vertical lines indicate the mean values for the respective data sets. D3R contains protein-ligand complexes from both the CSAR 2012 and 2014 releases.21,22
Figure SI 3: Individual predictions on the external data set from Wang et al. [3] (A) Predicted $\Delta G$ against experimental $\Delta G$. (B) Error for each prediction against the experimental $\Delta G$. (C) Error for each prediction against the atom efficiency ($\Delta G / N_{\text{non-H-atoms}}$).

Figure SI 4: Individual predictions on the external CSAR 2012 data set. (A) Predicted $\Delta G$ against experimental $\Delta G$. (B) Error for each prediction against the experimental $\Delta G$. (C) Error for each prediction against the atom efficiency ($\Delta G / N_{\text{non-H-atoms}}$).
Figure SI 5: Individual predictions on the external CSAR 2014 data set. (A) Predicted $\Delta G$ against experimental $\Delta G$. (B) Error for each prediction against the experimental $\Delta G$. (C) Error for each prediction against the atom efficiency ($\Delta G/N_{\text{non-H-atoms}}$).

Figure SI 6: Individual predictions on the external CSAR-NSR HiQ data set. (A) Predicted $\Delta G$ against experimental $\Delta G$. (B) Error for each prediction against the experimental $\Delta G$. (C) Error for each prediction against the atom efficiency ($\Delta G/N_{\text{non-H-atoms}}$).
Figure SI 7: Feature importance for models trained with only the ligand descriptors. The feature importance is computed as the average change in Pearson correlation coefficient for five permutations of the respective feature column by shuffling the data (see Methods for details.)

Figure SI 8: Feature importance for models trained with only the protein descriptors. Feature importance is computed as the average change in Pearson correlation coefficient for five permutations of the respective feature column by shuffling the data (see Methods for details.)
Graphical TOC Entry

Rapid screening with physicochemical descriptors + machine learning

Simple descriptors
- H-bonds
- logP
- MW
- ...

Protein Ligand

RASPD+

RF
DNN
SVR
LR kNN