Journal of Advanced Research 67 (2025) 185–196

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Mathematics, Engineering, and Computer Sciences

Inactive-enriched machine-learning models exploiting patent data improve structure-based virtual screening for PDL1 dimerizers

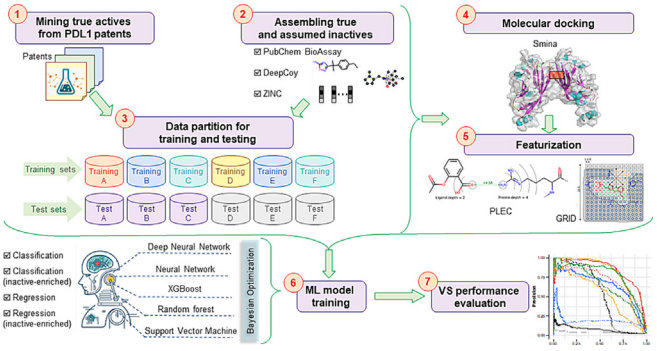
Pablo Gómez-Sacristán [a,1], Saw Simeon [a,1], Viet-Khoa Tran-Nguyen [a], Sachin Patil [b], Pedro J. Ballester [c,*]

[a] Centre de Recherche en Cancérologie de Marseille, Marseille 13009, France
[b] NanoBio Laboratory, Widener University, Chester, PA 19013, USA
[c] Department of Bioengineering, Imperial College London, London SW7 2AZ, UK

HIGHLIGHTS

- New machine-learning scoring functions were developed to predict PD1/PDL1 inhibitors.
- PDL1-specific regressors exploiting large volumes of inactive-enriched data and PLEC fingerprints as features were the most predictive for this important target.
- These new scoring functions are made available free of charge for prospective use.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 30 June 2023
Revised 1 December 2023
Accepted 21 January 2024
Available online 26 January 2024

Keywords:
PD1
PDL1
Immunotherapy
Virtual screening
Docking
Machine learning
Artificial intelligence

ABSTRACT

Introduction: Small-molecule Programmable Cell Death Protein 1/Programmable Death-Ligand 1 (PD1/PDL1) inhibition via PDL1 dimerization has the potential to lead to inexpensive drugs with better cancer patient outcomes and milder side effects. However, this therapeutic approach has proven challenging, with only one PDL1 dimerizer reaching early clinical trials so far. There is hence a need for fast and accurate methods to develop alternative PDL1 dimerizers.

Objectives: We aim to show that structure-based virtual screening (SBVS) based on PDL1-specific machine-learning (ML) scoring functions (SFs) is a powerful drug design tool for detecting PD1/PDL1 inhibitors via PDL1 dimerization.

Methods: By incorporating the latest MLSF advances, we generated and evaluated PDL1-specific MLSFs (classifiers and inactive-enriched regressors) on two demanding test sets.

Results: 60 PDL1-specific MLSFs (30 classifiers and 30 regressors) were generated. Our large-scale analysis provides highly predictive PDL1-specific MLSFs that benefitted from training with large volumes of docked inactives and enabling inactive-enriched regression.

Conclusion: PDL1-specific MLSFs strongly outperformed generic SFs of various types on this target and are released here without restrictions.

© 2022 The Authors. Published by Elsevier B.V. on behalf of Cairo University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

One of the most promising targets for cancer immunotherapy is the protein–protein interaction of Programmed Cell Death-1/Programmed Cell Death-Ligand 1 (PD1/PDL1) [1]. PD1 is an

* Corresponding author.
E-mail address: p.ballester@imperial.ac.uk (P.J. Ballester).
[1] joint first authors.

https://doi.org/10.1016/j.jare.2024.01.024
2090-1232/© 2022 The Authors. Published by Elsevier B.V. on behalf of Cairo University.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

endogenous immunoinhibitory receptor expressed on the surface of activated T cells, while the surface expression of its ligand PDL1 is enhanced in a variety of tumours, including non-small cell lung cancer, melanoma, epithelial ovarian cancer, breast cancer and gastrointestinal malignancies [2]. Under physiological conditions, the binding of PD1 with its ligand PDL1 leads to the negative regulation of immune response preventing its excessive stimulation. Under pathological conditions, however, the PD1/PDL1 immune checkpoint interaction leads to functional exhaustion and apoptosis of T cells, thus helping the tumours escape their immune destruction. Therefore, molecules preventing the PD1/PDL1 interaction result in boosting the activity of T cells, thereby strengthening the immune response against a broad spectrum of PDL1-presenting tumour cells [3]. The use of seven PD1/PDL1-targeting monoclonal antibody therapeutics that exhibit strong clinical benefits has been greenlit by the U.S. Food and Drug Administration (FDA). These include four binding to PD1 (dostarlimab, cemiplimab, nivolumab and pembrolizumab) and three binding to PDL1 receptors (atezolizumab, avelumab and durvalumab), leading to an effective PD1/PDL1 blockade [4]. These approved antibodies together with several other investigational anti-PD1/PDL1 immune checkpoint inhibitors are currently the subject of more than 1,100 active clinical trials (https://clinicaltrials.gov/; accessed on November 14, 2023), highlighting their significance in the cancer immunotherapy field.

In contrast to these peptide-based therapeutics including the FDA-approved antibody therapies, small-molecule PD1/PDL1 inhibitors for cancer immunotherapy have appropriate pharmacokinetic properties, tumour penetration, oral bioavailability, and acceptable half-lives [5]. Therefore, using these small-molecule inhibitors as cancer immunotherapies is likely to overcome the disadvantages of therapeutic antibodies (high price, poor bioavailability hampering oral administration, and may exert adverse effects on the immune system [6,7]. The direct targeting of the PDL1-binding site on PD1 has yielded small-molecule inhibitors with only modest activities in the μM range [8,9]. This may be attributed to the large PD1/PDL1 binding interface spanning over 1,970 $Å^2$ that is relatively featureless lacking any druggable binding pockets. In contrast, an innovative strategy to disrupt PD1/PDL1 is through the dimerization of PDL1 [10,11], where a small molecule binds to PDL1 [12] in a way that forms a PDL1 dimer, so that it can no longer interact with PD1 [13]. Such PDL1 dimerizers have shown impressive PD1/PDL1 inhibition with IC_{50} values in the pM-nM range [14]. They are, however, associated with acute cytotoxicity [15] and only one of them (INCB86550 from Incyte) has reached early clinical trial stages to date (NCT03762447, clinicaltrials.gov, accessed on November 14, 2023). Thus, there is a need for fast and accurate tools to identify alternative PDL1 dimerizers to mitigate this development hurdle. The design of novel small-molecule PD1/PDL1 inhibitors, including PDL1 dimerizers, by exploiting the generated data is a highly promising and largely unexplored therapeutic opportunity.

Structure-based (SB) virtual screening (VS) is a common approach to discovering drug leads by leveraging at least one atomic-resolution three-dimensional (3D) structure of the considered target [16,17]. In SBVS, a computational method is employed to screen a library of molecules for a subset enriched with a high proportion of potent actives (e.g. molecules with high pIC_{50}, or negative logarithm of the half-maximal inhibitory concentration, for the target). Docking [18] is a common type of SBVS methods, which generates 3D conformations of potential ligand molecules as bound to the considered target structure. Each docking method employs a classical scoring function (SF) to estimate the binding affinities of docked ligands. Classical SFs, whether force-field, knowledge-based or empirical, assume a linear relationship between the structure-derived features providing a numerical description of the protein–ligand complex and its binding affinity [19,20] (or its ligand-induced pIC_{50}). Machine learning (ML) approaches can be used to build SFs as an alternative to classical SFs. Such ML approaches capture binding interactions directly from training datasets without requiring limiting linearity assumptions [20]. Indeed, by not imposing a linear functional form for the SF, both linear and non-linear binding interactions can be directly inferred from protein structures and affinity/activity data.

MLSFs, now frequently outperforming classical SFs in benchmarks based on X-ray crystal structures of protein–ligand complexes by a large margin [21–29], are widely regarded as the state of the art in docking-based prediction of binding affinity. For example, Wei and co-workers showed that their convolutional neural network (CNN) allied with topological features was the top performer on many targets in the worldwide D3R competitions [25,26], with an independent analysis confirming that, in general, MLSFs work better than classical SFs on the targets selected for D3R [28]. MLSFs of this type are employed for problems such as drug lead optimisation [30] or drug repurposing [31]. On the other hand, MLSFs such as RF-Score-VS [32], CNN-Score [33] or DenseFS [34] have also outperformed classical SFs in the closely related SBVS problem (where the affinities of a much larger set of docked inactives have to be predicted too). Moreover, many studies have reported the discovery of potent actives using such MLSFs for a wide range of targets [35]: e.g. 460 nM in K_i for estrogen receptor alpha (ERα) using artificial neural networks (ANN) [36], 359 nM in IC_{50} for anaplastic lymphoma kinase (ALK) using support vector machines (SVM) [37], 400 nM in IC_{50} for aspartate *N*-acetyltransferase (ANAT) using CNN-based AtomNet [38], 173 nM in K_i for acetylcholinesterase (ACES) using extreme gradient boosting (XGBoost, or XGB for short) [24]. Interestingly, to our knowledge, MLSFs based on deep learning algorithms other than CNNs, including deep neural network (DNN) algorithms, have rarely been reported to be successful in prospective SBVS.

The SBVS performance provided by these generic SFs can be improved even further by using target-specific MLSFs (i.e. those using target-specific features and/or restricted to the target's most informative docked poses). This is shown retrospectively by Yasuo et al. [39] with random forest (RF)-based SFs and by Xiong et al. [21] with XGB-based SFs (these studies also reported hit rates up to 4 times higher than those from Glide's classical SFs). While target-specific SFs can only be generated when some actives are already known for the target, it is now a common situation owing to public bioactivity data resources [40,41]. However, this is only possible sometimes by mining patent data [42], which is a much less common route.

We hypothesize that applying the latest advances observed in studies based on other targets will lead to highly accurate target-specific MLSFs for PDL1. For instance, a large number of decoys (assumed inactives) in the training set were shown to boost the SBVS performance of MLSFs [32], but this has never been investigated for PDL1. Thus, it is not known if model training should be carried out with actives only, or by supplementing them with experimentally validated inactives [43], property-matched decoys [44] or random property-unmatched decoys [45]. Likewise, regression-based MLSFs are still to be applied to PDL1, despite the dependent variable to predict, pIC_{50}, being real-valued. This is probably due to the fact that the most popular SBVS benchmarks [43] do not provide or employ real-valued potency to evaluate screening performance, but only rely on sets of actives and decoys with binary classification metrics [46]. As a real-valued variable contains more information than any dichotomized version of that variable, it stands to reason that regression models should perform better than classification models, other things being equal. To explore both research opportunities, we will evaluate regression models that also exploit the information about the chemical diversity of

inactives, which we call inactive-enriched regression-based MLSFs. Another novel aspect of our study is investigating which combinations of featurization schemes and supervised learning algorithms are the most predictive for SBVS on PDL1.

The rest of the paper is organized as follows. First, the experimental design is described in detail, including the composition of our initial data (true actives, true inactives, decoys) and how they were assigned into the training set and the test set. VS performance on the test set of various PDL1-specific ML binary classifiers and regressors is next described, in comparison with that of three generic SFs. We then provide statistical evidence, based on our retrospective screening results, to illustrate the combinations of computational modelling choices that led to the most predictive models. Lastly, discussions and conclusions are presented, followed by detailed information on the materials and methodologies employed in this study.

Results

Experimental design

We started by analyzing the suitability of seven PDL1 dimeric structures (5 J89, 5J8O, 5N2D, 5N2F, 5NIU, 6NM8, and 6R3K) for docking and ultimately selected 6NM8 for this study (see Methods section). We then used 10 supervised learning algorithms that have led to predictive SFs for SBVS on other targets: ANN [47,48], SVM [37,49], RF[32,50], XGB [21,24], and DNN[51,52] (for each of them, its classification and regression variants were employed). Each of these algorithms exploited the same data instances and featurization schemes (protein–ligand extended connectivity fingerprints, PLEC [53]; and 3D grid-based features, GRID [54]). The performance of all resulting PDL1-specific MLSFs was next compared to that recently obtained with generic ones [55]: Smina [18], IFP [56,57], and CNN-Score [33]. These generic SFs were selected based on results in the literature and in preliminary non-systematic SBVS studies on a range of crystal structures, featurization schemes and generic SFs (see Methods section).

The experimental design of this study is illustrated in Fig. 1. Small-molecule PDL1 dimerizers, here the active molecules acting as positive instances, were mined from two patents published in the literature (step 1). Assumed inactives, including property-matched and property-unmatched decoys, were generated by DeepCoy [44] and selected at random from the ZINC database [58], respectively; while true inactives were retrieved from PubChem BioAssay [59]. Each set of inactives would act as a set of negative instances (step 2). Various training-test data partitions were made for subsequent model training and VS performance evaluation on the corresponding test set (step 3). Molecular docking of all these molecules was next carried out using the 6NM8 PDB structure (step 4), then the resulting docked ligand-receptor complexes were encoded as PLEC fingerprints and GRID features (step 5). PDL1-specific MLSFs (both classification and regression) were trained using the features computed previously on the training sets and 10 supervised learning algorithms (step 6), and were finally evaluated on the test sets to compare their performance with that of generic ones (step 7). Note that Bayesian optimization was applied to hyperparameter tuning during the model training process (supplementary information).

PDL1 benchmarks (Table 1) were constructed to investigate how well each SF ranks the same set of 297 docked actives (patent 1: WO2015034820A1) embedded in a much larger set of docked inactives (test set). Two test sets were employed in this study, each using a different set of docked inactives. The first test set is called TrueInactives, as its inactive instances were retrieved from the *in vitro* screening of 28,781 molecules (PubChem BioAssay AID 2316). It is uncommon to benefit from such a large number of true inactives to train target-specific MLSFs. Incidentally, this *in vitro* screen only identified a single active, suggesting that PD1/PDL1 is a hard target. The second test set is called DeepCoys, as its inactives are property-matched decoys generated by DeepCoy from each test active.

Concurrently, four training sets were built using the same set of 371 actives (patent 2: WO2015160641A2). Not only are there no molecules in common between any training and test set pair, but their chemical structures are also dissimilar (Figure S2). Each training set has a different set of docked inactives: DeepCoys (DeepCoy-generated decoys property-matched to each test active), RandomDecoys (a much larger number of property-unmatched decoys selected randomly from ZINC), TrueInactives (the same set of molecules used in the homonym test set), ActivesOnly (no training inactives). It is worth noting that, taking inspiration from the DUD-E data collection, whose active-to-inactive ratio is fixed at 1/50, we designed the TrueInactives set (for both training and testing) such that (i) it contains 18,550 inactive instances randomly chosen from experimental data of AID 2316, and (ii) the chosen inactives have physicochemical properties in the same ranges as those of training actives (Figure S3). Using ActivesOnly in the classification scenarios (Table 1) presents the issue that no inactives are considered while two classes are required. Thus, we followed the same approach as Wijewardhane et al. [42], in which two classes of actives (high and low potency) are defined according to their potency using an IC_{50} of 100 nM as the cutoff. SFs trained on RandomDecoys or ActivesOnly were later tested on both test sets (we did not include the two additional test sets employing either RandomDecoys or ActivesOnly because they would not be generally regarded as challenging). Another restriction is that the SFs trained on DeepCoys must not be tested on test sets also using DeepCoys (training inactives have to be assembled in a different way than test inactives to avoid performance overestimation due to decoy bias in the datasets[35]). Lastly, we only used TrueInactives either in the training set or in the test set, but not in the same partition to avoid having to split it and thus making it less challenging (a larger proportion of inactives makes it harder for VS models). Overall, this process gave rise to six VS scenarios corresponding to six pairs of training-test data for each type of SFs (classification or regression), as listed in Table 1.

Developing PDL1-specific classification-based SFs for SBVS

PDL1-specific classification-based SFs were trained and tested using each of the six training-test data partitions (Table 1). These ML models were built with two featurization schemes (PLEC, GRID) and five binary classification algorithms (RF, XGB, SVM, ANN and DNN). Three generic SFs were also evaluated on the same test sets. A precision-recall (PR) curve was traced using the scores output by each SF for PDL1-docked test-set molecules (Fig. 2). The corresponding PR area under the curve (AUC), or PR-AUC, values were also calculated (Table S1).

Despite the challenging test inactives, not including partitions likely to benefit from artificial decoy bias and the dissimilarity between training and test actives, some PDL1-specific MLSFs were able to discriminate very well between actives and inactives in the test set. We first analyze the models trained on inactive-enriched data. Among them, models employing RF, XGB and SVM as learning algorithms performed the best when trained on RandomDecoys from ZINC and tested on TrueInactives from PubChem, regardless of the featurization schemes (PR-AUCs $\geq$ 0.90 in one third of the cases and never lower than 0.75). While RF-based models gave consistent performance no matter which features were used and which training-test data partition was employed (PR-AUCs $\geq$ 0.80 most of the time and never below 0.72), the discrim-

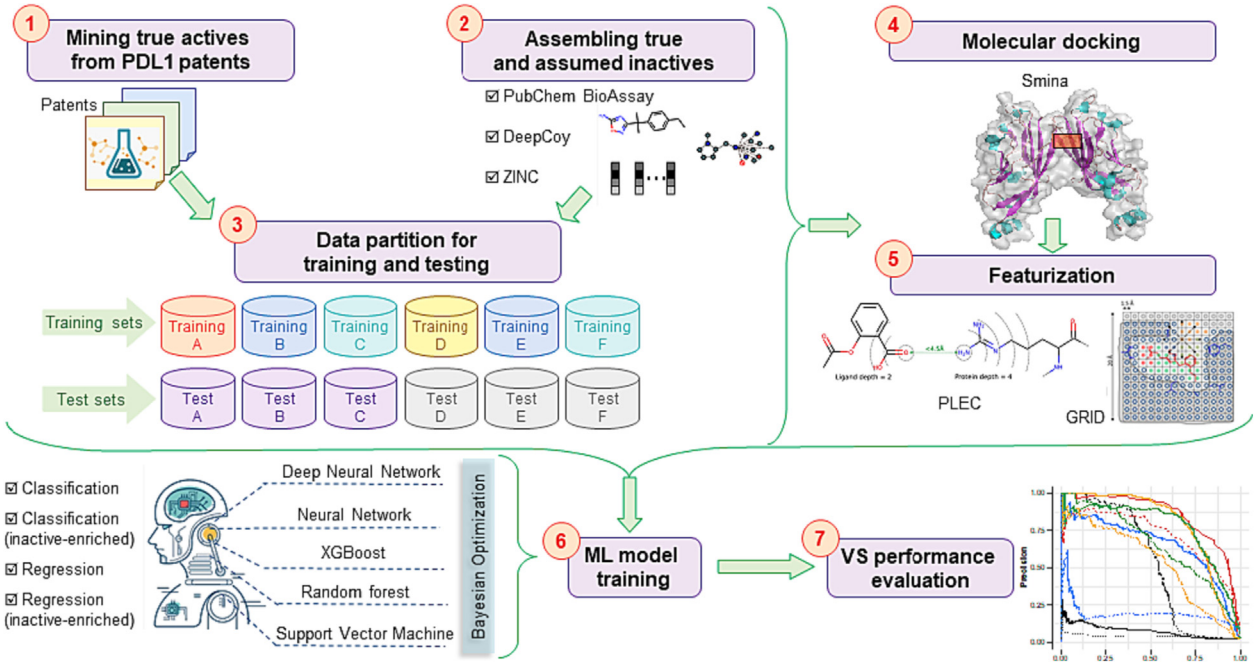


Fig. 1. Overview of the experimental design to generate and evaluate SFs for SBVS against PDL1. The experimental design involves several steps, including mining true actives from PDL1 patents, assembling true and assumed inactives, data partitioning for training and testing, molecular docking, featurization with PLEC and GRID, ML model training using various algorithms and performance evaluation.

Table 1
The 12 VS scenarios considered in this study employing six training-test data partitions (A-F). All training sets employ the same set of 371 actives (WO2015160641A2), but differ on the considered set of inactives and hence are uniquely identified by the latter (either TrueInactives, DeepCoys, RandomDecoys or ActivesOnly). Likewise, all test sets employ the same 297 actives (WO201503820A1), none of these molecules also included in the training set, but different sets of inactives (TrueInactives or DeepCoys). The number of successfully docked inactive instances of each dataset is provided between brackets: DeepCoys (14,700), TrueInactives (18,550), and RandomDecoys (337,194).

Partition ID	Training set	Test set	Model type
A	DeepCoys	TrueInactives	Classification
B	RandomDecoys	TrueInactives	Classification
C	ActivesOnly	TrueInactives	Classification
D	TrueInactives	DeepCoys	Classification
E	RandomDecoys	DeepCoys	Classification
F	ActivesOnly	DeepCoys	Classification
A	DeepCoys	TrueInactives	Regression
B	RandomDecoys	TrueInactives	Regression
C	ActivesOnly	TrueInactives	Regression
D	TrueInactives	DeepCoys	Regression
E	RandomDecoys	DeepCoys	Regression
F	ActivesOnly	DeepCoys	Regression

inatory power of XGB- and SVM-based ones depended on the combination of features and training-test scenarios (PLEC-based models performed here more consistently than GRID-based ones). By contrast, DNN's performance fluctuated within a wide range (PR-AUCs from 0.02 to 0.84), and was generally better when PLEC fingerprints were used as features. Lastly, ANN usually gave poor-to-modest performance (PR-AUCs between 0.3 and 0.7). The critical importance of using docked inactives for training PDL1-specific SFs is herewith demonstrated by the performance of all models worsening drastically (PR-AUCs never exceeding 0.04) when no inactives were included in the training sets (ActivesOnly).

The three generic SFs gave, at best, substantially worse PR-AUCs than PDL1-specific ML ones when distinguishing PDL1 actives in the test set from TrueInactives or DeepCoys (PR-AUCs below 0.1 in about half of the cases and never exceeding 0.53). The only ML-based SF among them (CNN-Score) outperformed the other two classical/interaction-based SFs (Smina, IFP) by a large margin.

Developing PDL1-specific regression-based SFs for SBVS

PDL1-specific regression-based SFs were also trained and tested using the same six training-test data partitions (Table 1). While those trained with ActivesOnly are common and generally unsuitable for SBVS[28], the regression algorithms using a training set with a much larger number of inactives led to inactive-enriched regression models instead (this requires assigning inactives a low pIC_{50}, whose value was set at 2 in this study without making any claim about the optimality of this choice). For each training set, two featurization schemes (PLEC, GRID) and five regression algorithms (RF, XGB, SVM, ANN and DNN) were employed and the resulting SFs evaluated. Three generic SFs were also evaluated on the same test sets. A PR curve was traced using the scores output by each SF for PDL1-docked test-set molecules (Fig. 3). The corresponding PR-AUC value was also calculated (Table S2).

PDL1-specific regression-based SFs generally outperformed their classification counterparts, especially when SVM and ANN were used as learning algorithms (Table S2, Fig. 3). SVM- and ANN-based models consistently gave outstanding performance regardless of the featurization schemes and training inactives (PR-AUCs above 0.9 in most cases and never lower than 0.86). Notably, the SVM algorithm, when using GRID features, trained on DeepCoys and tested on TrueInactives, gave the best performance overall that was not overshadowed by any other combination of learning algorithm-featurization scheme-training-test data, also by any aforementioned classification model (PR-AUC = 0.99). DNN also gave good performance, albeit not as consistent as SVM and ANN (PR-AUCs mostly > 0.8 and only < 0.7 in one VS scenario). RF-based models seemed to work better when trained on TrueInactives or RandomDecoys and using PLEC fingerprints as features. By contrast, XGB generally provided poorer performance than the other algorithms. Interestingly, models trained with more numerous RandomDecoys generally yielded better and more consistent

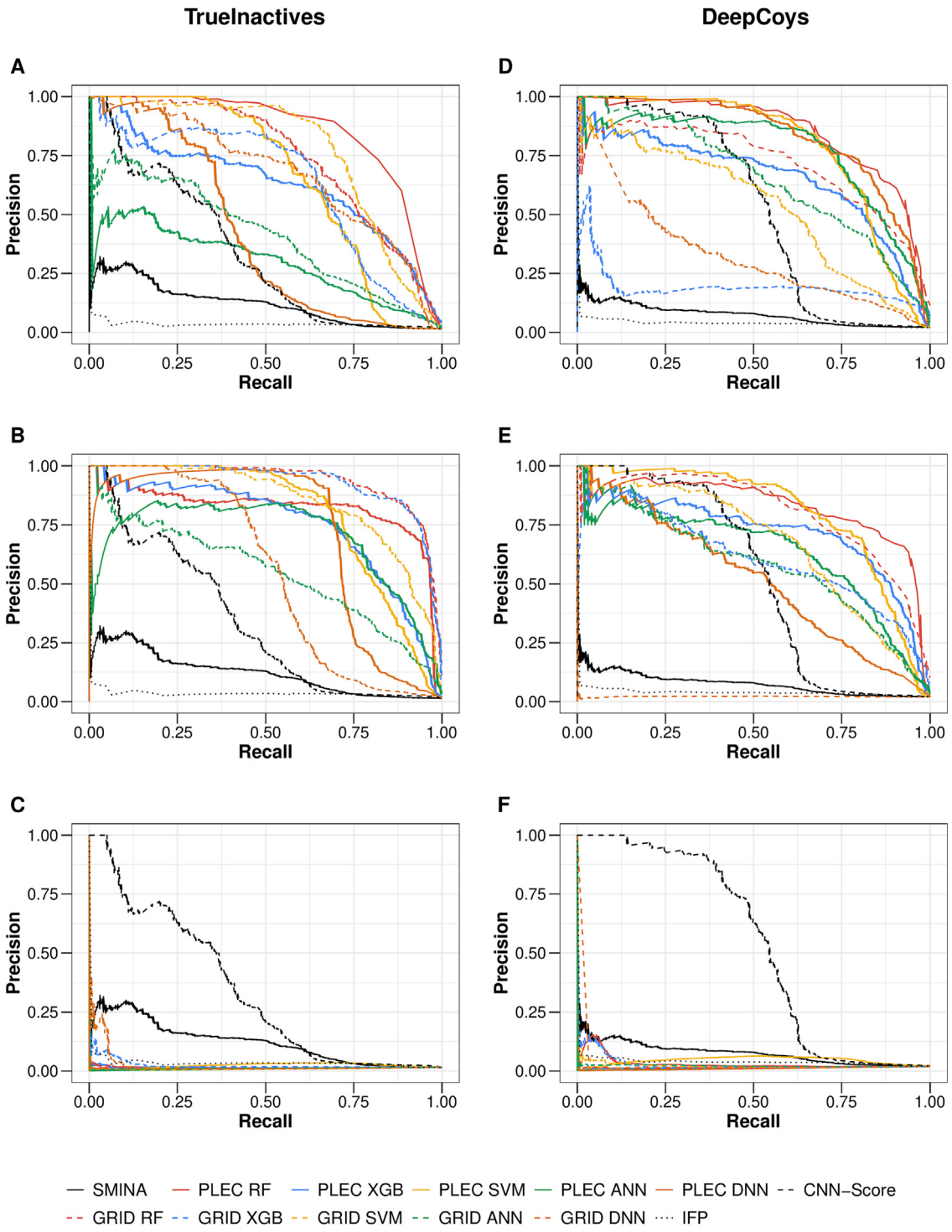


Fig. 2. VS performance on the test sets of generic SFs and PDL1-specific classification-based MLSFs, portrayed by PR curves. Each column of plots represents one test set, whereas rows represent training sets (top is TrueInactives (right) or DeepCoys (left), middle is RandomDecoys, and bottom is ActivesOnly), with A-F corresponding to the six training-test data partitions (Table 1). To generate PDL1-specific SFs, protein–ligand extended connectivity fingerprints (PLEC, solid lines) and 3D grid-based features (GRID, dashed lines) of PLD1-docked molecules were computed and the resulting sets of features were employed by each of the following classification algorithms: RF (red), XGB (blue), SVM (yellow), ANN (green), and DNN (orange). Generic SFs were also tested and their performance was used as the baseline. Their PR curves are all in black and represented as solid (Smina), dashed (CNN-Score) and dotted lines (IFP). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

performance than those trained with TrueInactives or DeepCoys (PR-AUCs > 0.9 in 75 % of the cases). Finally, regression models trained with ActivesOnly performed slightly better than their classification counterparts in Fig. 2, but still far worse than inactive-enriched models. This further demonstrates the critical importance of exploiting docked inactives to train PDL1-specific models. As the generic SFs were applied off-the-shelf to the same test sets, their PR curves are identical to those in Fig. 2.

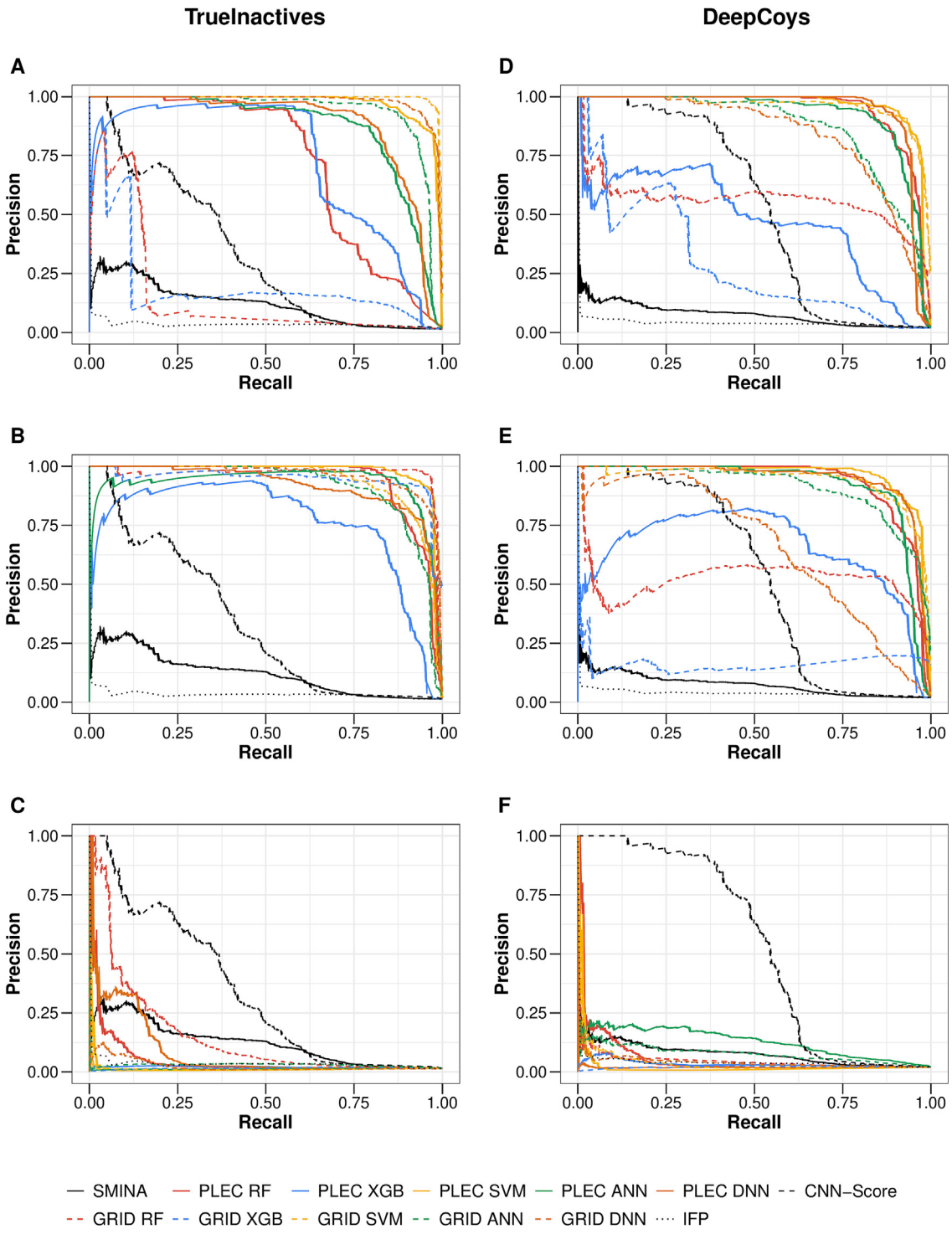


Fig. 3. VS performance on the test sets of generic SFs and PDL1-specific regression-based MLSFs, portrayed by PR curves. Each column of plots represents one test set, whereas rows represent training sets (top is TrueInactives (right) or DeepCoys (left), middle is RandomDecoys, and bottom is ActivesOnly), with A-F corresponding to the six training-test data partitions (Table 1). To generate PDL1-specific SFs, protein–ligand extended connectivity fingerprints (PLEC, solid lines) and 3D grid-based features (GRID, dashed lines) of PLD1-docked molecules were computed and the resulting sets of features were employed by each of the following regression algorithms: RF (red), XGB (blue), SVM (yellow), ANN (green), and DNN (orange). Generic SFs were also tested and their performance was used as the baseline. Their PR curves are all in black and represented as solid (Smina), dashed (CNN-Score) and dotted lines (IFP). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The impact of computational modelling choices on SBVS performance

The boxplots in Fig. 4 summarize the distribution of the plotted PR-AUCs obtained on the test sets according to (i) the nature of SFs (ML or non-ML, generic or PDL1-specific, classification or regression), (ii) the featurization schemes (PLEC or GRID), (iii) the presence or absence of inactive instances in the training data, and (iv) the source of inactive molecules in the training and test sets (TrueInactives, DeepCoys, or RandomDecoys).

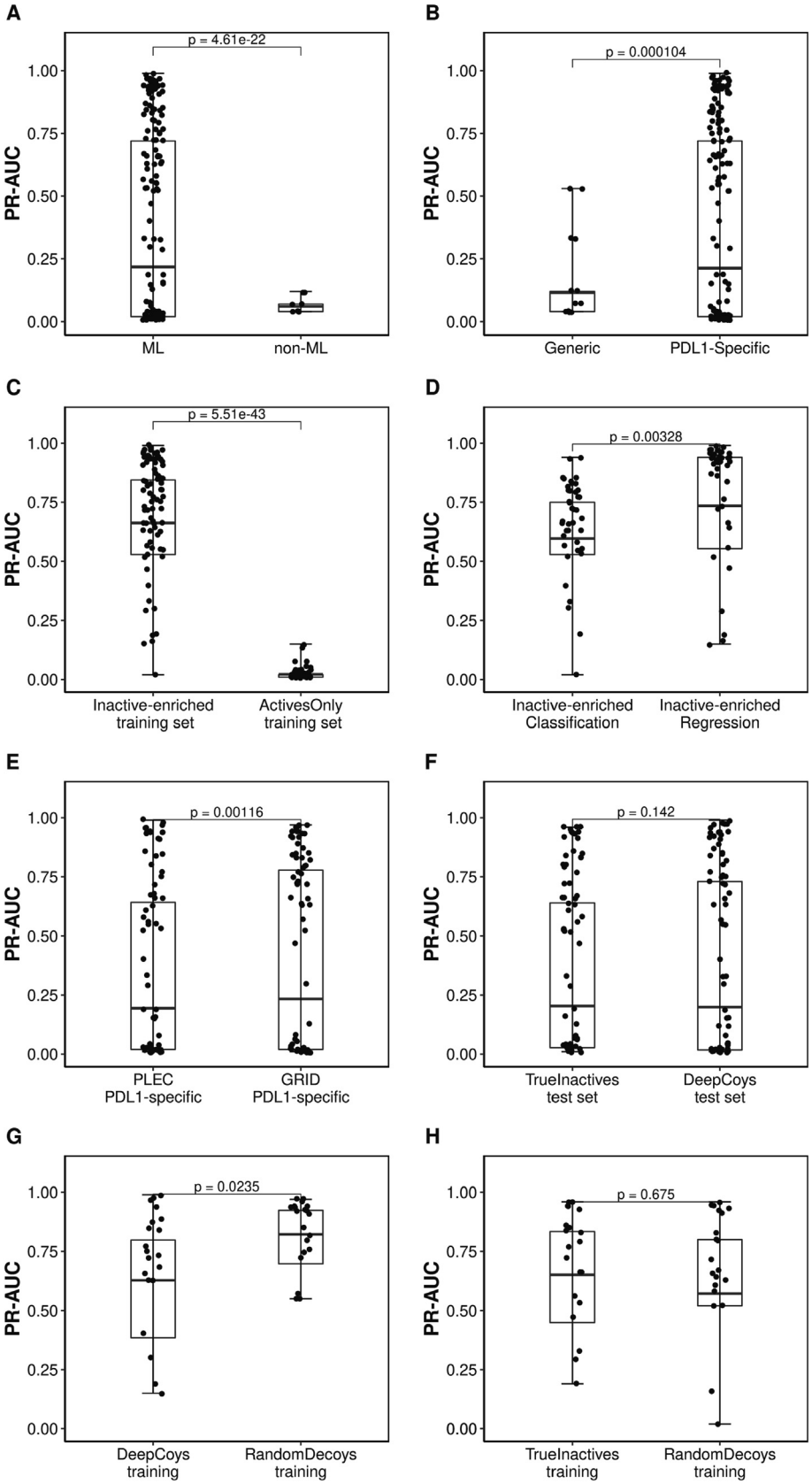


Fig. 4. The impact of computational design on test set performance. (A) ML or not. (B) generic SFs or PDL1-specific SFs. (C) inactive-enriched or not (classification and regression SFs). (D) inactive-enriched classification or inactive-enriched regression. (E) PLEC or GRID features (PDL1-specific SFs). (F) TrueInactives or DeepCoys in the test set. (G) DeepCoys or RandomDecoys as training data when TrueInactives are used in the test set. (H) TrueInactives or RandomDecoys as training data when DeepCoys are used in the test set. For each boxplot, the median value is represented by the horizontal line inside the box. Welch's t-tests were carried out for statistical analyses: paired tests in case the samples have the same size (as in boxplots D, E, F, G, H), and unpaired tests when the samples have different sizes (as in boxplots A, B, C).

These pairs of PR-AUC distributions and their associated p-values show the level of generality and strength of each result from this study. First, there is a strong advantage in using MLSFs over classical/interaction-based ones (Fig. 4A), although this would not be observed if an MLSF trained on ActivesOnly was chosen for the comparison (Fig. 4C). Second, among MLSFs, the PDL1-specific ones are generally more predictive than generic ones (Fig. 4B). Third, the most important factor for gaining predictive SBVS performance is training ML models with inactive-enriched data (Fig. 4C). Also, within the latter SFs, regression models perform substantially better than classification ones (this is particularly noticeable when the upper-quartile PR-AUCs are compared, Fig. 4D). GRID features generally led to better-performing PDL1-specific SFs than PLEC fingerprints (Fig. 4E). In case there are not sufficient TrueInactives retrieved from PubChem BioAssay to test SFs, property-matched DeepCoys can be used as an alternative offering a similar level of benchmark difficulty (Fig. 4F). On the other hand, as molecules in the DeepCoys set are property-matched to test actives and these actives are dissimilar to those in the training set (Figure S2), like RandomDecoys, these DeepCoys are property-unmatched to training actives. Therefore, given that RandomDecoys are almost 23 times more numerous than DeepCoys, it follows that using a larger number of such property-unmatched decoys as training data tends to improve SBVS performance (Fig. 4G). Lastly, no significant performance difference is observed between PDL1-specific MLSFs trained on TrueInactives and those trained on much more numerous RandomDecoys (Fig. 4H). Note that the right-hand-side boxplots in Fig. 4G and 4H were created from the same PDL1-specific MLSFs trained on the same actives and the same RandomDecoys, but different PR-AUC distributions are observed because they were evaluated on different test sets (TrueInactives and DeepCoys, respectively).

Discussion and conclusions

This large-scale study has investigated SBVS for PDL1 dimerizers using state-of-the-art approaches. We mined patents to gather actives (chemical structures and their target activities) and almost 20,000 experimentally-determined inactives, which is both beneficial and unusual for building target-specific MLSFs. In a broader yet less detailed preliminary stage (data not shown), we explored seven PDB structures of this target, three generic MLSFs and seven featurization schemes. This was followed by a systematic analysis focusing on the modelling choices that performed the best in the preliminary stage: 6NM8 as the PDL1 template structure, CNN-Score as the generic MLSF, and PLEC and GRID as the featurization schemes. We also employed three types of inactives, 10 supervised learning algorithms (five for binary classification and five for regression), and a leading Bayesian optimization method for tuning their hyperparameters. We also evaluated two non-MLSFs as performance baselines. This process has resulted in highly promising SFs for this important target (e.g. GRID SVM gave a PR-AUC of at least 0.94 with four different data partitions on two different test sets).

The performance of these PDL1-specific MLSFs was boosted by training sets enriched with a large proportion of inactives (between 35 and 909 per active). Regarding which type of inactives is preferable, we observed that model training with experimentally-determined inactives did not yield significantly better-performing SFs than training with assumed inactives (Fig. 4H), although the latter set is much larger (over 18 times in this study). Also, using 23 times more assumed inactives as training data led to a significant SBVS improvement (Fig. 4G) on a test set whose inactive instances are experimentally confirmed. These results encourage the use of larger sets of docked decoys in training SFs for SBVS of this and other targets.

We have also found that SFs based on inactive-enriched regression are more predictive than those based on inactive-enriched classification (Fig. 4D). By assigning low potency to the inactives (pIC_{50} = 2), the use of such molecules to train regression models is enabled, thus allowing better learning to identify true inactives (i.e. reducing false positives). This is also beneficial for learning from the actives, as the potency differences between actives are now taken into account, while they are neglected when classifiers are built (the most common type of SFs for SBVS).

Practically all discriminatory power of these SFs vanished following the removal of inactives from the training sets (Fig. 4C). Thus, the fact that an SF trained only on actives has already managed to discover a potent PDL1 dimerizer makes inactive-enriched SFs even more promising for this target. Indeed, Wijewardhane et al. [42] trained an ensemble of Energy-Graph ANN classifiers with 762 PDL1 dimerizers docked into the PD1 pocket of another PDL1 homodimer crystal structure (5N2F). To enable classifier training, each active molecule was categorised as either "high potency" or "low potency" (this is effectively the same approach followed by the classifiers trained on ActivesOnly in Fig. 2). The resulting PDL1-specific SF was employed to screen a small in-house library of molecules, leading to the discovery of a new small-molecule inhibitor of the PD1/PDL1 interaction with an IC_{50} of 340 nM without requiring any potency optimization. Having just seen that inactive-enriched regression is much more predictive than actives-only classification on this target, it is reasonable to expect that the released SFs can directly discover even better PDL1 dimerizers than those in the aforementioned study of Wijewardhane et al. [42].

The availability of true PDL1 inactives retrieved from PubChem BioAssay benefited the study in various ways. First, it gave the most realistic training and test sets, as the lack of activity of these compounds was not assumed. Second, a much higher number of property-unmatched decoys could be identified as an effective substitute of true inactives in the training set (Fig. 4H). Third, the consistency of the results across two different test sets could be established. Lastly, property-matched DeepCoys could be identified as an alternative to true inactives in the test set, as they posed a similar level of difficulty (Fig. 4F).

Regarding ML algorithms, there is no strong advantage of using one over another. As it has been observed in other targets [23,29] and problems [60,61], deep learning algorithms are not always better than other algorithms on average across cases. By contrast, the featurization scheme, or representation, affected performance much strongly, as schemes other than GRID or PLEC were much less predictive. GRID-based SFs were significantly more accurate than PLEC-based SFs, but the former has the disadvantage of taking around 100 times more computing time than the latter (for feature extraction and model training altogether). The vast majority of research in this area presents the application of either a new algorithm or new representation. Interestingly, our results suggest that the way synthetic training data (docked actives and inactives) are generated and used is of higher importance.

The PDL1-specific MLSFs strongly outperformed generic ML ones (Fig. 4B). Thus, it is worthwhile to spend the resources required to build SFs tailored to PDL1 SBVS, which might include mining patent data as it was the case here. These MLSFs were in turn much more predictive than non-MLSFs (Fig. 4A). In particular, IFP obtained the worst performance despite being found to be more predictive than two actives-only generic MLSFs on other targets [57] (this study did not consider CNN-Score or target-specific MLSFs). This outcome has also been observed when comparing IFP to MLSFs on other targets [62].

Methods

Mining chemical and bioactivity data from patents

A data set of PDL1 dimerizers and their activity values were obtained from the patent 1 (WO2015034820A1) and patent 2 (WO2015160641A2), consisting of 297 and 371 compounds tested against PDL1, respectively. For each patent, bioactivity data were retrieved using an in-house R script complemented with manual curation. First, the *read_html* function from the R package *xml2* was used to read the html page [63]. The *html_nodes* and *html_text* functions from the R package *rvest* were employed to select the html nodes containing the IUPAC (International Union of Pure and Applied Chemistry) name of each compound corresponding to a bioactivity and to extract texts from the html page, respectively [64]. IUPAC names of compounds were manually inspected and curated. Second, SureCHEMBL searches on the patents WO2015034820A1 and WO2015160641A2 were performed to obtain the SMILES (simplified molecular input line entry specification) strings that correspond to the IUPAC names retrieved above [65]. Third, the *left_join* function from the R package *dplyr* was applied to link each measured bioactivity range to its corresponding chemical structure, by IUPAC names. As a result, the information gathered for each molecule includes its SMILES string, its IUPAC name, its compound ID and its potency. The activity of the molecule, measured with a homogeneous time-resolved fluorescence (HTRF) binding assay, was reported as one of the following active concentration ranges: 6–100 nM, 110–1000 nM, 1010–10,000 nM for the actives of patent 1; 0.6–10 nM, 10.01–100 nM, 100.01–20,000 nM for those of patent 2. For each range, the value equal to 50 % of the upper limit was chosen as its representative; therefore, the actives of patent 1 have active concentrations reported at either 50 nM, 500 nM, or 5,000 nM; and those of patent 2 at either 5 nM, 50 nM, or 10,000 nM.

Preparing sets of inactives (RandomDecoys, DeepCoys, TrueInactives)

Three types of inactives were employed. First, RandomDecoys are property-unmatched decoys selected randomly from a pool of 265,450,385 molecules from the ZINC database [58] (1000 decoys per active were selected, each as a ready-to-dock 3D conformer, at reference pH only, not reactive according to filters). Second, DeepCoys are property-matched decoys that were generated using DeepCoy, a novel method based on graph-generative neural network (NN) models [44]. Briefly, this method takes an active molecule as an input and generates new structurally-dissimilar molecules with similar physicochemical properties (listed in the supplementary information). To achieve this goal, new molecules are created in an iterative and bond-by-bond manner from a pool of atoms, while ensuring the chemical validity of the generated molecules by incorporating a set of permitted atom types and complying to basic atomic valency rules. The generative models convert the graphs of active molecules into property-matched decoys under an augmented variational autoencoder setting, employing standard gated-graph NN in both the encoder and the decoder. Third, TrueInactives are molecules retrieved from PubChem BioAssay[59] (accession number AID 2316). The data originated from an *in vitro* screen of a chemical library of 28,781 compounds against the PD1/PDL1 interaction, which turned out to be all inactive but one, and hence constitute true inactive molecules. To compare the physicochemical property space of the TrueInactives and the patent actives, we analyzed their Lipinski's rule-of-five descriptors (supplementary information). Figure S3 shows a substantial overlap between TrueInactives and the 668 actives in the spanned physicochemical space, thus precluding the possibility of trivial active-inactive discrimination.

Crystal structures of PDL1 in complexes with small-molecule ligands

A total of 33 crystal structures of PDL1 were downloaded from the Protein Data Bank in Europe (PDBe) [66] and superimposed. PDB structures without a ligand or with monoclonal antibodies were removed. Only the PDB crystal structures containing small molecules binding with PDL1 in a non-covalent manner and resolved in X-ray diffraction studies were further considered. Thus, seven structures associated with seven dimerizers reported in BMS (Bristol-Myers Squibb) patents 1 (WO2015034820A1) and 2 (WO2015160641A2) were retained, including PDB IDs 5J8O, 5 J89, 5N2D, 5N2F, 5NIU, 6NM8, and 6R3K, with resolutions (Å) at 2.30, 2.20, 2.35, 1.70, 2.01, 2.79, and 2.20, respectively. A recent analysis concluded that these structures are promising to support SBVS efforts against the unconventional binding pocket of the PDL1 dimer [11]. Here we focused on 6NM8, which was the structure selected for cross-docking [11] and yielded here the best performance in this preliminary assessment using Smina (see supplementary information).

Docking all molecules into the target structures

The preparation of each molecule (smi format) was carried out using the ChemAxon suite (version 21.18.0) [67] according to the following steps: (i) inorganics and mixtures were removed, (ii) the SMILES strings were converted into optimal 3D structures, (iii) specific chemotypes were normalized, and (iv) duplicates were removed; thanks to the options *Strip Salts*, *Aromatize*, *Clean3D*, *Tautomerize at* pH *7.4*, *Neutralize* and *Remove Explicit Hydrogens* [68]. At the end of these steps, a 3D structure (in sdf) was returned for each ligand. From this, mol2 files were next generated using the Open Babel software (version 2.3.1)[69] and the integrated MMFF94 force field (*--gen3D* option). The final output files were ready for the subsequent docking procedure.

The PDL1 homodimer structures (in the pdb file format) were downloaded from the Protein Data Bank and were prepared using the DockPrep tool of Chimera (version 1.15) [70], according to recommended parameters [71] and similarly to a previously described workflow [72].

The Smina software (version 2019–10-15), a fork of AutoDock Vina (version 1.1.2), was used to carry out molecular docking [18]. The search space was centered on the position of the ligand structure bound to the PD1 binding site (*--autobox_ligand* option), and its volume was set at 30 x 30 x 30 ($Å^3$) so that the ligands would have sufficient space to rotate. The active site of the protein structure was defined using the coordinates of the bound ligand [73]. The Smina docking code was executed using the following additional options: number of restarts for the conformational search algorithm (*--exhaustiveness* = 1) and number of poses to output per docked molecule (*--num_modes* = 1). All output poses were saved as multi-mol2 files. The binding free energy of each docked molecule was computed by the Smina generic SF.

Calculating features for each docked molecule

Each docked target-ligand complex must be numerically encoded as features to enable supervised learning [74,75]. We initially considered the following featurization schemes: interaction fingerprints (IFP) [56], RF-Score version 2 features (RF-Score-v2) [74], protein–ligand extended connectivity fingerprints (PLEC) [53], and 3D grid-based features from *RDKitGridFeaturizer* (GRID) [54]. Our preliminary tests (data not shown) revealed that PLEC

and GRID were clearly the most predictive for SBVS in PDL1 and thus we focused on these two schemes alone.

PLEC [53] fingerprints represent a sub-type of extended-connectivity fingerprints (ECFPs), where only atoms in contact with another molecule are considered. For a pair of protein–ligand atoms whose distance in 3D space does not exceed a user-defined threshold (referred to as "interacting atoms"), one or more "environments" are identified for each of them (the "root" atom). The aforementioned distance threshold is defined using the *distance_cutoff* parameter of the ODDT [76] (https://github.com/oddt/oddt) *PLEC* function, and was set at 4.5 Å in this study. Each environment covers the "root" and its neighbors found within a diameter of i bonds, i being an integer value such that $0 \leq i \leq n$, where n defines the largest environment of an atom and is also user-supplied, using the *depth_protein* parameter if the "root" is part of the protein, or the *depth_ligand* parameter if otherwise (set at 5 and 1 in this study, respectively). Atomic features of the "root" atom are registered, including (i) its atomic number, (ii) the number of neighboring heavy atoms, (iii) the number of attached hydrogens (implicit and explicit), (iv) the atomic mass, (v) the formal charge, (vi) ring membership and aromaticity. Each previously identified protein environment is then paired with a ligand environment having the same depth (protein environments with a depth larger than the maximal depth of ligand ones are paired with the largest ligand environment, or vice versa). All possible environment pairings (for an interacting protein–ligand atom pair) are hashed, using the Python *hash* function, and folded to a certain fingerprint size determined by users (*size* parameter, 4092 in this case).

The GRID featurization scheme imposes, for each protein–ligand complex, a box whose center is placed on the centroid of the bound/docked ligand. The size of the box is defined by the *box_width* parameter of the *RDKitGridFeaturizer* from the *deepchem* Python package (left as default, i.e. 16 Å, in this study). The box is then subdivided by a grid into voxels of the same size, which is user-defined by the *voxel_width* parameter (set at 16 Å in this case, as observed in the literature [51,52]. For each voxel, intra-ligand, intra-protein, and intermolecular interaction fingerprints would then be generated, taking into account all intra-voxel ligand atoms and intra-voxel protein atoms which are found within 4.5 Å from any ligand atom. Intra-ligand and intra-protein fingerprints are computed using ECFPs of radius r and 2^n bits (r determined by the *ecfp_degree* parameter, set at 2 in this study; n determined by the *ecfp_power* parameter, herein set at 9), then hashed together to form a vector of 512 bits. Intermolecular interactions in each voxel are then enumerated, distributing all intra-voxel protein–ligand atom pairs into three bins according to their intra-pair distance: 0–2 Å, 2–3 Å, 3–4.5 Å. For each pair, the structural protein–ligand interaction fingerprints (SPLIFs) of 2^m bits (m determined by the *splif_power* parameter, set at 9) are registered, and the fingerprints of each distance bin are then hashed into a vector of 512 bits (i.e. 1536 bits for three bins). The number of hydrogen bonds whose lengths fall into each of the three ranges 2.2–2.5 Å, 2.5–3.2 Å, 3.2–4.0 Å is counted, along with the number of salt bridges, for each voxel. All above-mentioned intra-voxel features are finally concatenated, yielding a vector of 2052 features (i.e. a 2052-bit fingerprint) per voxel in the end.

Training-test data partitions

As the selected patents contain compounds chemically different from each other (Figure S2), we used patent 1 (WO2015034820A1) actives in the test set and patent 2 actives (WO2015160641A2) in the training set. This is aimed at evaluating the ability of the models to identify not only unseen active molecules, but furthermore those with novel scaffolds. For actives-only classifiers, the potencies of 366 molecules were labeled as high and the remaining 302 as low depending on whether their pIC_{50} was above or below 7 (100 nM), respectively. To enable regression, we assigned RandomDecoys, DeepCoys, and TrueInactives a low pIC_{50} value of 2. Table 1 specifies the sets of inactives employed in each partition.

Re-scoring docked molecules with generic SFs

To complement the Smina classical SF, we considered several other generic SFs. First, the similarity of docked molecules, in terms of IFPs with the protein, to the co-crystallized ligand of the X-ray structure employed for docking was found to be the most predictive in SBVS against other targets [57]. Second, CNN-Score [77] gave better VS performance on PDL1 in preliminary tests than other generic MLSFs (RF-Score-VS v2 [32] and DenseFS [34]). We only tested CNN-Score on all test sets to avoid overcrowding our PRC plots. In this manner, we had as baselines one generic SF from each class (classical, interaction-based, ML).

The IFP module from the IChem package (version 5.2.9) was used to compute the Tanimoto score between the IFP of each docked pose and that of the crystal 6NM8 ligand (HET code: KSD). Molecules whose docked poses were more similar to the crystal template, in terms of IFPs (higher Tanimoto scores), were regarded as more likely to bind according to IChem developers. The module requires, as input files, the 3D structure (mol2) of the receptor as well as the 3D coordinates (mol2) of the crystal ligand and of each docked pose to calculate the corresponding IFPs, on the basis of all receptor residues engaging in direct interactions with the small molecules. The IFPs take into account seven features per ligand (crystal or docked) as follows: hydrophobic, aromatic (face to face), aromatic (edge to face), hydrogen bond (protein atom as donor), hydrogen bond (protein atom as acceptor), electrostatic (positively charged protein atom), and electrostatic (negatively charged protein atom).

CNN-Score re-scores docked molecules using the *score_only* and *cnn_scoring* options of the *gnina* function in GNINA 1.0, a deep learning framework for molecular docking [77]. This generic MLSF is an ensemble of five CNN models selected to balance pose prediction, VS performance and runtime: dense (1,429,790 training instances), general_default2018 (11,324 training instances), default2017 (1,429,790 training instances), crossdock_default2018 (13,839 training instances) and redock_default2018 (13,780 training instances). Each of the models was trained using a different training set and a different model architecture, as previously described [55,77]. CNNs are the state of the art in 2D image recognition. CNN-Score treats the full 3D "images" of protein–ligand complexes as a generalization of 2D images. Each voxel represents a cube divided by a regular 3D grid with features describing its atomic contents. The entire set of voxel features from the complex is then employed to rescore docked poses. However, the dimensionality of a 3D voxel is much higher than that of a 2D pixel (each pixel usually has three features, each specifying the intensity of a primary colour), which makes it a harder problem for the CNNs.

Developing PDL1-specific MLSFs

These SFs were implemented for each data partition and featurization scheme using the following supervised learning algorithms: random forest (RF) [78], extreme gradient boosting (XGB) [79], support vector machine (SVM) [80], deep neural network (DNN) [81], and artificial neural network (ANN). Implementation and hyperparameter tuning details can be found in supplementary information.

Measuring SBVS performance

Precision-recall (PR) curves are useful for assessing VS performance when the classes are highly imbalanced, as in this case. The PR curve here depicts the trade-off between precision (a measure of result relevance) and recall (a measure of completeness) in the test molecules ranked by the score of a given model (a higher score indicates that the molecule is more likely to have the activity of interest). The PR area under the curve (PR-AUC) of each model-test set pair was computed using the scikit-learn toolkit (version 1.0.2).

Ethics statement

No experiment in this study involved animals or human patients.

Code availability

Code to evaluate and run the best scoring functions for PDL1 are freely available at https://github.com/sawsimeon/MLSF-PDL1.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are thankful for the financial support from the Fondation ARC pour la Recherche sur le Cancer (ARC), the Indo-French Centre for the Promotion of Advanced Research (CEFIPRA) and the Wolfson Foundation and the Royal Society for a Royal Society Wolfson Fellowship (RSWF\R1\221005). We are also thankful to Sachin Vishwakarma for assistance in tuning the DNN algorithm, Muhammad Junaid for comments and Chayanit Piyawajanusorn for assistance in designing Fig. 1. We also thankfully acknowledge the use of software packages (ChemAxon, OpenBabel, Gnina, IChem, Chimera, Scikit-learn).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2024.01.024.

References

[1] Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer 2012;12:252–64.
[2] Wu, Y., Chen, W., Xu, Z. P. & Gu, W. PD-L1 distribution and perspective for cancer immunotherapy— blockade, knockdown, or inhibition. *Frontiers in Immunology* vol. 10 2022 Preprint at https://doi.org/10.3389/fimmu.2019.02022 (2019).
[3] Tang J et al. The clinical trial landscape for PD1/PDL1 immune checkpoint inhibitors. Nat Rev Drug Discov 2018;17:854–5.
[4] Upadhaya S, Neftelinov ST, Hodge J, Campbell J. Challenges and opportunities in the PD1/PDL1 inhibitor clinical trial landscape. Nat Rev Drug Discov 2022. doi: https://doi.org/10.1038/d41573-022-00030-4.
[5] Zhan, M. M. *et al.* From monoclonal antibodies to small molecules: The development of inhibitors targeting the PD-1/PD-L1 pathway. *Drug Discovery Today* vol. 21 1027–1036 Preprint at https://doi.org/10.1016/j.drudis.2016.04.011 (2016).
[6] Qin M et al. Discovery of [1,2,4]Triazolo[4,3-a]pyridines as Potent Inhibitors Targeting the Programmed Cell Death-1/Programmed Cell Death-Ligand 1 Interaction. J Med Chem 2019;62:4703–15.
[7] Konieczny M et al. Di-bromo-Based Small-Molecule Inhibitors of the PD-1/PD-L1 Immune Checkpoint. J Med Chem 2020;63:11271–85.
[8] Patil SP et al. Macrocyclic Compounds from Ansamycin Antibiotic Class as Inhibitors of PD1–PDL1 Protein-Protein Interaction. Chem Pharm Bull (Tokyo) 2018;66:773–8.
[9] Patil SP et al. Identification of Small-Molecule Inhibitors of PD-1/PD-L1 Protein-Protein Interaction. ChemistrySelect 2018;3:2185–9.
[10] Shi D et al. Computational Insight Into the Small Molecule Intervening PD-L1 Dimerization and the Potential Structure-Activity Relationship. Front Chem 2019;7.
[11] Kuang Z et al. Partial Least-Squares Discriminant Analysis and Ensemble-Based Flexible Docking of PD-1/PD-L1 Inhibitors: A Pilot Study. ACS Omega 2020;5:26914–23.
[12] Kythreotou, A., Siddique, A., Mauri, F. A., Bower, M. & Pinato, D. J. PD-L1. *Journal of clinical pathology* vol. 71 189–194 Preprint at https://doi.org/10.1136/jclinpath-2017-204853 (2018).
[13] Zak KM et al. Structural basis for small molecule targeting of the programmed death ligand 1 (PD-L1). Oncotarget 2016;7:30323–35.
[14] Guzik K et al. Development of the Inhibitors That Target the PD-1/PD-L1 Interaction—A Brief Look at Progress on Small Molecules, Peptides and Macrocycles. Molecules 2019;24:2071.
[15] Ganesan A et al. Comprehensive in vitro characterization of PD-L1 small molecule inhibitors. Sci Rep 2019;9:1–19.
[16] Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. AAPS J 2012;14:133–41.
[17] Batool, M., Ahmad, B. & Choi, S. A Structure-Based Drug Discovery Paradigm. *International journal of molecular sciences* vol. 20 2783 Preprint at https://doi.org/10.3390/ijms20112783 (2019).
[18] Koes DR, Baumgartner MP, Camacho CJ. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. J Chem Inf Model 2013;53:1893–904.
[19] Ain QU, Aleksandrova A, Roessler FD, Ballester PJ. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. WIREs Comput Mol Sci 2015;5:405–24.
[20] Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics 2010;26:1169–75.
[21] Xiong G-L et al. Improving structure-based virtual screening performance via learning from scoring function components. Brief Bioinform bbaa094 2020. doi: https://doi.org/10.1093/bib/bbaa094.
[22] Fresnais L, Ballester PJ. The impact of compound library size on the performance of scoring functions for structure-based virtual screening. *Brief Bioinform* bbaa095 2020. doi: https://doi.org/10.1093/bib/bbaa095.
[23] Li H, Sze K-H, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based virtual screening. In: WIREs Computational Molecular Science e1478. doi: https://doi.org/10.1002/wcms.1478.
[24] Adeshina YO, Deeds EJ, Karanicolas J. Machine learning classification can reduce false positives in structure-based virtual screening. Proc Natl Acad Sci 2020;117:18477–88.
[25] Nguyen DD et al. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. J Comput Aided Mol Des 2019;33:71–82.
[26] Nguyen DD, Gao K, Wang M, Wei GW. MathDL: mathematical deep learning for D3R Grand Challenge 4. J Comput Aided Mol Des 2020;34:131–47.
[27] Li H et al. Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. Bioinformatics 2019;35:3989–95.
[28] Li H, Sze K-H, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based drug lead optimization. Wiley Interdiscip Rev Comput Mol Sci 2020;10:e1465.
[29] Shen C et al. From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. Wiley Interdiscip Rev Comput Mol Sci e1429 2019. doi: https://doi.org/10.1002/wcms.1429.
[30] Jiménez-Luna J et al. DeltaDelta neural networks for lead optimization of small molecule potency. Chem Sci 2019;10:10911–8.
[31] Ziehm M et al. Drug repurposing for aging research using model organisms. Aging Cell 2017;16:1006–15.
[32] Wójcikowski M, Ballester PJ, Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. Sci Rep 2017;7:46710.
[33] Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-Ligand Scoring with Convolutional Neural Networks. J Chem Inf Model 2017;57:942–57.
[34] Imrie F, Bradley AR, van der Schaar M, Deane CM. Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data. J Chem Inf Model 2018;58:2319–30.
[35] Ghislat, G., Rahman, T. & Ballester, P. J. Recent progress on the prospective application of machine learning to structure-based virtual screening. *Current Opinion in Chemical Biology* vol. 65 28–34 Preprint at https://doi.org/10.1016/j.cbpa.2021.04.009 (2021).
[36] Durrant JD et al. Neural-Network Scoring Functions Identify Structurally Novel Estrogen-Receptor Ligands. J Chem Inf Model 2015;55:1953–61.
[37] Sun H et al. Constructing and Validating High-Performance MIEC-SVM Models in Virtual Screening for Kinases: A Better Way for Actives Discovery. Sci Rep 2016;6:24817.
[38] Stecula A, Hussain MS, Viola RE. Discovery of novel inhibitors of a critical brain enzyme using a homology model and a deep convolutional neural network. J Med Chem 2020;63:8867–75.
[39] Yasuo N, Sekijima M. An Improved Method of Structure-based Virtual Screening via Interaction-energy-based Learning. J Chem Inf Model 2019;59:1050–61.
[40] Peón A et al. MolTarPred: A web tool for comprehensive target prediction with reliability estimation. Chem Biol Drug Des 2019;94:1390–401.
[41] Mathai, N., Chen, Y. & Kirchmair, J. Validation strategies for target prediction methods. *Briefings in Bioinformatics* vol. 21 791–802 Preprint at https://doi.org/10.1093/bib/bbz026 (2020).
[42] Wijewardhane PR, Jethava KP, Fine JA, Chopra G. Graph Neural Networks Bootstrapped for Synthetic Selection and Validation of Small Molecule

Immunomodulators. ChemRxiv Preprint 2021. doi: https://doi.org/10.33774/chemrxiv-2021-r4xnx-v2.
[43] Réau M, Langenfeld F, Zagury J-F, Lagarde N, Montes M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. Front Pharmacol 2018;9:11.
[44] Imrie F, Bradley AR, Deane CM. Generating Property-Matched Decoy Molecules Using Deep Learning. Bioinformatics 2021. doi: https://doi.org/10.1093/bioinformatics/btab080.
[45] Irwin JJ et al. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. J Chem Inf Model 2020. doi: https://doi.org/10.1021/acs.jcim.0c00675.
[46] Ballester, P. J. Selecting machine-learning scoring functions for structure-based virtual screening. *Drug Discovery Today: Technologies* vols 32–33 81–87 Preprint at https://doi.org/10.1016/j.ddtec.2020.09.001 (2020).
[47] Durrant JD, McCammon JA. NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein−Ligand Complexes. J Chem Inf Model 2010;50:1865–71.
[48] Durrant JD, McCammon JA. NNScore 2.0: A Neural-Network Receptor-Ligand Scoring Function. J Chem Inf Model 2011;51:2897–903.
[49] Li L et al. Target-Specific Support Vector Machine Scoring in Structure-Based Virtual Screening: Computational Validation, In Vitro Testing in Kinases, and Effects on Lung Cancer Cell Proliferation. J Chem Inf Model 2011;51:755–9.
[50] Ballester PJ et al. Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification. Journal of the Royal Society, Interface / the Royal Society 2012;9:3196–207.
[51] Wang D et al. Improving the Virtual Screening Ability of Target-Specific Scoring Functions Using Deep Learning Methods. Front Pharmacol 2019;10:924.
[52] Ashtawy HM, Mahapatra NR. Task-Specific Scoring Functions for Predicting Ligand Binding Poses and Affinity and for Screening Enrichment. J Chem Inf Model 2018;58:119–33.
[53] Wójcikowski M, Kukiełka M, Stepniewska-Dziubinska MM, Siedlecki P. Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. Bioinformatics 2019;35:1334–41.
[54] Wu Z et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci 2018;9:513–30.
[55] Tran-Nguyen VK, Junaid M, Simeon S, Ballester PJ. A practical guide to machine-learning scoring for structure-based virtual screening. Nature Protocols 2023;18(11):3460–511.
[56] Radifar M, Yuniarti N, Istyastono EP. PyPLIF: Python-based Protein-Ligand Interaction Fingerprinting. Bioinformation 2013;9:325–8.
[57] Tran-Nguyen VK, Bret G, Rognan D. True Accuracy of Fast Scoring Functions to Predict High-Throughput Screening Data from Docking Poses: The Simpler the Better. J Chem Inf Model 2021;61:2788–97.
[58] Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A Free Tool to Discover Chemistry for Biology. J Chem Inf Model 2012;52:1757–68.
[59] Kim S et al. PubChem Substance and Compound databases. Nucleic Acids Res 2016;44:D1202–13.
[60] Chen D et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit Med 2019;2:43.
[61] Borisov V et al. Deep Neural Networks and Tabular Data. A Survey 2021.
[62] Tran-Nguyen VK, Ballester PJ. Beware of Simple Methods for Structure-Based Virtual Screening: The Critical Importance of Broader Comparisons. J Chem Inf Model 2023;63:25.
[63] Wickham H, Hester J, Ooms J. xml2: Parse XML. R package version 2018;1.
[64] Wickham, H. & Wickham, M. H. Package 'rvest'. *URL: https://cran. r-project. org/web/packages/rvest/rvest. pdf* (2016).
[65] Papadatos G et al. SureChEMBL: a large-scale, chemically annotated patent document database. Nucleic Acids Res 2016;44:D1220–8.
[66] Gutmanas A et al. PDBe: Protein Data Bank in Europe. Nucleic Acids Res 2014;42:D285–91.
[67] ChemAxon v21.18.0. *ChemAxon, Ltd: Budapest, Hungary* Preprint at (2021).
[68] Simeon S et al. Unraveling the origin of splice switching activity of hemoglobin β-globin gene modulators via QSAR modeling. Chemom Intel Lab Syst 2016;151.
[69] O'Boyle NM et al. Open Babel: An open chemical toolbox. J Cheminform 2011;3:33.
[70] Pettersen EF et al. UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605–12.
[71] dos Santos RN, Ferreira LG, Andricopulo AD. Practices in molecular docking and structure-based virtual screening. in *Methods in Molecular Biology*. (Humana Press Inc.; 2018.
[72] Tran-Nguyen VK, Simeon S, Junaid M, Ballester PJ. Structure-based virtual screening for PDL1 dimerizers: Evaluating generic scoring functions. Curr Res Struct Biol 2022;4:206–10.
[73] Perry E et al. Fragment-based screening of programmed death ligand 1 (PD-L1). Bioorg Med Chem Lett 2019;29:786–90.
[74] Ballester PJ, Schreyer A, Blundell TL. Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? J Chem Inf Model 2014;54:944–55.
[75] Xiong G et al. Featurization strategies for protein–ligand interactions and their applications in scoring function development. In: *WIREs Computational Molecular Science* e1567. doi: https://doi.org/10.1002/wcms.1567.
[76] Wójcikowski M, Zielenkiewicz P, Siedlecki P. Open Drug Discovery Toolkit (ODDT): A new open-source player in the drug discovery field. J Cheminform 2015;7:1–6.
[77] McNutt AT et al. GNINA 1.0: molecular docking with deep learning. J Cheminform 2021;13:43.
[78] Breiman L. Random Forests. Mach Learn 2001;45:5–32.
[79] Chen T, Guestrin C. In: XGBoost: A Scalable Tree Boosting System. (ACM Press; 2016. doi: https://doi.org/10.1145/2939672.2939785.
[80] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.
[81] Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw 2015;61:85–117.