BBA - General Subjects 1864 (2020) 129545

Contents lists available at ScienceDirect

BBA - General Subjects

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

Machine learning and ligand binding predictions: A review of data, methods, and obstacles

Sally R. Ellingson[a,b,*], Brian Davis[b], Jonathan Allen[c]

[a] College of Medicine, Division of Biomedical Informatics, University of Kentucky, Lexington, KY, United States of America
[b] Markey Cancer Center, Lexington, KY, United States of America
[c] Lawrence Livermore National Laboratory, Livermore, CA, United States of America

ARTICLE INFO

Keywords:
Machine learning
Drug discovery
Drug binding
Overfitting

ABSTRACT

Computational predictions of ligand binding is a difficult problem, with more accurate methods being extremely computationally expensive. The use of machine learning for drug binding predictions could possibly leverage the use of biomedical big data in exchange for time-intensive simulations. This paper reviews current trends in the use of machine learning for drug binding predictions, data sources to develop machine learning algorithms, and potential problems that may lead to overfitting and ungeneralizable models. A few popular datasets that can be used to develop virtual high-throughput screening models are characterized using spatial statistics to quantify potential biases. We can see from evaluating some common benchmarks that good performance correlates with models with high-predicted bias scores and models with low bias scores do not have much predictive power. A better understanding of the limits of available data sources and how to fix them will lead to more generalizable models that will lead to novel drug discovery.

1. Introduction

The ability to accurately predict protein-ligand interactions continues to be a difficult problem. Efficient calculations, such as molecular docking, that allow the exploration of large libraries of chemical compounds use simplified scoring functions and usually ignore things such as protein flexibility, entropic and de-solvation effects. Methods that use information from more computationally rigorous, and physically accurate calculations, such as coming from molecular dynamics simulations (MD), are limited due to their computational cost [1]. Even if the computational cost was not limiting, in order for these calculations to be reliable, there must be a high-quality structure of the protein, protonation states of the ligand and binding site must be known, the ligand binding mode must be correct, the protein cannot undergo substantial or slow changes, and the force fields must accurately describe the interactions [2].

Many fields have advanced using machine learning, including the use of machine learning to develop novel scoring functions for protein-ligand binding. Using machine learning in scoring functions offers a great advantage because they learn the parameters and model structure from data. This paper is a review of the types of models, types of features, and types of data that can be used for modeling protein-ligand interactions using machine learning. It also includes a review of some of the state-of-the-art models, problems that lead to overfitting and ungeneralizable models, and methods to quantify these problems.

2. Types of models

There are three distinct different ways models can be made to make binding predictions.

- Pose prediction. These models aim to accurately predict the correct binding conformation of the ligand. Input for these models would require three-dimensional experimental structures for protein-ligand complexes.
- Binding energy prediction. These models aim to correctly predict the experimental binding energy of a protein-ligand complex. Input for these models would require known binding energies of protein-ligand complexes.
- Virtual high-throughput screen. These models aim to separate potential binding ligands from a large library of ligands that contains mostly non-binding ligands. The success of these models is typically measured in dataset enrichment, i.e. the percent of actual binding ligands in the set of predicted binding ligands should be much higher than the percent of binding ligands in the entire dataset. This is often measured using the area under the curve (AUC) of a receiver

* Corresponding author at: College of Medicine, Division of Biomedical Informatics, University of Kentucky, Lexington, KY, United States of America.
E-mail address: sally@kcr.uky.edu (S.R. Ellingson).

https://doi.org/10.1016/j.bbagen.2020.129545
Received 3 November 2019; Received in revised form 21 December 2019; Accepted 30 January 2020
Available online 10 February 2020

operating characteristic curve. While these models can be built using known binding energies of protein-ligand complexes, models typically perform better if non-binding ligands are included in the input data. Since non-binding ligands do not have a binding energy, binary models can be built that just predict binding vs non-binding.

3. Examples of features to use in machine learning models

Features can be used to describe the drugs in the model, the proteins in the model, and also the three-dimensional structure of the drug-protein complex. A common drug feature is called a fingerprint. A fingerprint is a vectorized representation of a compound that can be used to assess the similarity of two compounds based on common substructures. Two common fingerprints are extended connectivity fingerprints (ECFP) and functional connectivity fingerprints (FCFP) and are also referred to as circular or Morgan fingerprints. Both of these fingerprints are followed by a number which indicates the diameter of connectivity. Extended fingerprints use atom features such as atomic number and charge while functional fingerprints use atom features related to binding such as number of hydrogen bond donors and acceptors. A wide variety of other descriptors can also be calculated for the drug compounds. As an example, the cheminformatics software Dragon calculates 5270 molecular descriptors which includes atom types, functional groups and fragment counts, topological and geometrical descriptors, three-dimensional descriptors, and estimations and alerts such as logP and Lipinski's alert.

Protein features can be limited to the assumed binding site or take in account the entire protein. These features can be based on sequence alone such as single, dipeptide, and tripeptide compositions, predicted properties based on sequence (such as predicted structure and likelihood of binding), composition of residue types (such as hydrophilic/hydrophobic/aliphatic). Several web tools can be used to generate such features [3–5].

Thirdly, input features can take into account the three-dimensional structure of a drug-protein complex. The atom coordinates can be featurized using a grid structure to represent atom and connections types in three-dimensional space. Several featurization methods can be found in the MoleculeNet/DeepChem [6] framework such as the Grid Featurizer that includes intermolecular interactions such as salt bridges and hydrogen bonding, as well as intra-ligand and intra-protein circular fingerprints. Another representation that can be used with graph based algorithms is Graph convolutions which is a feature vector of an atom's local environment with a list of neighboring atoms. There is a great interest in successfully using these representations with deep learning because of the promise of truly learning the interactions important for drug binding. However, it also limits the data to be used in the modeling to only those compounds with high-resolution experimental structures of the bound drug-protein complex or assuming that predicted bound structures from molecular docking simulations are accurate.

4. A review of some current machine learning methods in use to predict drug binding

A summary of some current trends in machine learning algorithms is given in Table 1. For each study, it is reported whether the prediction is binary (active vs decoy or binding vs non-binding), whether it has information about the protein in the model or only the ligand, whether the features come from 2-dimensional representations (such as ligand fingerprints and protein sequence), whether the features come from 3-dimensional representations of ligands bound in the binding pocket, whether the study has classical machine learning methods, whether the study has deep learning methods, and what the main datasets used in the study are. It can be seen that many studies still focus on a binary prediction. Making a continuous prediction limits datasets to those with experimental binding measurements, therefore do not train on the non-binding data. While these models may do well at predicting measurements of the active compounds, they do not perform as well for virtual screening when a larger number of non-binding compounds are in the compound libraries. There has been a push towards deep learning using 3-dimensional representations. However, as stated in the Atomic Convolutional Network study [12], these efforts still get better results using random forest with the 3-dimensional representation because advances are still needed in the data representation for deep learning. Even when protein features are considered in these models, it is usually the atomic coordinates of a bound structure, and models are protein specific to one protein. Many of these studies [8,10,11,13–15] do not consider potential biases in the datasets at all. Some studies create models using multiple proteins [7,9,14] and test on completely different datasets or left out proteins to reduce bias, but the potential biases between datasets are not quantified.

5. Obstacles in creating datasets for drug binding predictions

High-throughput screens (HTS) are commonly affected by experimental noise and artifacts. Organic chemicals can form aggregates and cause off-target and cytotoxic effects that disturb an assay's optical detection. Another problem with benchmark datasets is that experimentally measured binding affinities can have a large variation between methods, and this variation is often protein target and compound dependent [20]. These lead to fundamental errors in the input data to the models.

Models may lack interpretability and are often overfit to the data and are not generalizable to drug targets and chemotypes not in the training data. Benchmark datasets are prone to artificial enrichment and analogue bias due to the overrepresentation of certain scaffolds in experimentally determined active sets. Overoptimistic virtual screening results can be associated with dataset clumping, when the active class is easy to separate from the inactive class based on how the dataset was built, not what is guiding the binding [21,22]. A study showed that using only ECFP fingerprints of the ligands in the DUD-E dataset (described below) and a simple logistic regression (with 72 proteins in the training and 30 proteins in the test set), they achieved a mean AUC of 0.904. The model had no information on the targets [9].

5.1. Attempts to overcome obstacles

Datasets can be evaluated using spatial statistics [21,22] to quantify the dataset topology and better understand potential biases.

5.1.1. Quantifying bias in active to inactive datasets

The “nearest neighbor function” (G) gives the distance of any event to its nearest neighbor event, the distribution of intraset “active-to-active” distances,

$$G(t) = \sum_i \frac{I_t(i,j)}{n},\ i = 1,\ \ldots, n,$$

where $I_t(i,j) = 1$ if the distance between i and j (both actives) is less than the distance threshold t and n is the number of actives. The “empty space function” (F) gives the distance of a randomly chosen point to its nearest neighbor, the distribution of “active-to-background” distances,

$$F(t) = \sum_j \frac{I_t(j,i)}{m},\ j = 1,\ \ldots, m,$$

where $I_t(j,i) = 1$ if the distance between j (inactive) and i (active) is less than the distance threshold t and m is the number of inactives. A sum of the differences of these two measures, $S = (F - G)$, gives a quick and interpretable estimate of a datasets clumping or dispersion. Dataset clumping comprises a combination of self-similarity of actives and separation from decoys. The score described here was used to help develop the Maximum Unbiased Validation dataset described in Section

Table 1
A summary of current studies using machine learning for drug binding predictions.

Study	Binary/Continuous	Protein	2D	3D	Classic ML	Deep learning	Data
[7]	Binary	✓		✓		✓	DUD-E, ChEMBL
[8]	Binary		✓		✓	✓	PCBA, MUV, DUD-E, TOX21
[9]	Binary	✓		✓		✓	PDBBind, DUD-E, MUV
[10]	Binary	✓	✓			✓	DrugBank
[11]	Binary		✓		✓	✓	PubChem Bioassay
[12]	Continuous	✓	✓	✓	✓	✓	PDBBind
[13]	Binary	✓		✓		✓	DUD-E, ChEMBL
[14]	Continuous	✓		✓		✓	PDBBind, Astek diverse set
[15]	Continuous	✓	✓		✓		PDBBind, DUD-E
[16]	Binary		✓		✓	✓	ChEMBL
[17]	Continuous	✓	✓			✓	Kinase datasets [18,19]

6.3 by trying to minimize the sum of the differences of the nearest neighbor and empty space function, ΣS, or Σ *(F-G)*. Ideally *F* and *G* would have similar values for the range of distances, indicating that there is a similar spread of compounds within the active set as there is between the active and inactive sets. This can be evaluated by plotting *F(t)* and *G(t)* to see how much the curves overlap. Ideal curves would lead to similar ΣG and ΣF values and a ΣS near *0*. A high *G(t)* at low values of *t* means that the actives are all very similar and a low *F(t)* at high G(*t*) would mean that the inactives are very different from the actives. A big difference in ΣG and ΣF would indicate a big difference in the sets over a range of distances.

5.1.2. Quantifying bias with respect to training and validation sets

The Asymmetric validation embedding (AVE) bias [23] was specifically developed to address inflated metrics due to similarities between the validation (or testing) and training datasets when developing a machine learning model. Building on the spatial statistics described above, the AVE bias includes terms for how clumped the validation actives are among the training actives and also for the inactive clumping, as this class can be learned as well. Similar to the spatial statistics above, the AVE bias starts by defining a nearest neighbor function,

$$S(V,T,d) = \frac{1}{|V|}\sum_{v\in V} I_d(v,T),$$

where *V* is the set of validation molecules, *T* is the set of training molecules, *d* is a similarity distance threshold, and $I_d(v,T)$ is 1 if the distance between *v* and the nearest neighbor in *T* is smaller than *d*. The cumulative nearest neighbor function is then

$$H(V,T) = \frac{1}{|D|}\sum_{d\in D} S(V,T,d),$$

where D is a non-empty set of distance thresholds. Then the AVE bias score is defined as

$$B(V_a,V_i,T_a,T_i) = [H(V_a,T_a) - H(V_a,T_i)] + [H(V_i,T_i) - H(V_i,T_a)],$$

where V_a, V_i, T_a, and T_i are the sets of validation actives, validation inactives, training actives, and training inactives. $H(V_a,T_a)$-$H(V_a,T_i)$ is abbreviated as *(AA-AI)* and quantifies the active set clumping. $H(V_i,T_i)$-$H(V_i,T_a)$ is abbreviated as *(II-IA)* and quantifies the inactive set clumping. *(AA-AI)* negativity suggests validation actives are generally closer to training inactives than training actives, making the active set especially challenging to classify. *(II-IA)* positivity means inactives will be easier to classify. A bias *B* close to zero represents a less bias dataset split.

Major claims of this paper are that *k*-nearest neighbors ($k = 1$) can detect overfitting since the training data is memorized, that the AVE bias correlates with *k*-nearest neighbors ($k = 1$), the bias is a measure of the extent in which a dataset can be solved through overfitting and can be used to minimize this bias. The authors also provide an AVE minimization algorithm. It is a genetic algorithm with breeding operations: merge, add molecule, remove molecule, and swap subset. The algorithm first generates initial subsets through random sampling, measures the bias, and selects subsets with low biases for breeding. The algorithm repeats bias scoring, redundancy removal, and breeding until termination based on minimal bias or maximum iterations.

6. Description of existing benchmark datasets

There are many freely available data sources that can be used to build the models described above. A description of some of the popular sources and how they were constructed are given below. The first group of datasets describes data with experimental structures that can easily be used for pose predictions. The second group of datasets describes data with known active compounds and their binding energies that can be used for binding energy predictions. The third group of datasets cover diverse proteins and include both active and inactive compounds to enable building multi-protein models for high-throughput screens. These three datasets are described using the above mentioned spatial statistics. The spatial statistics from the MUV study use simple descriptors such as atom types etc. from the original MUV paper [22] and the Euclidean distance. The AVE bias analysis uses the defaults provided in the authors' script to minimize the bias between training and validation splits.

6.1. Binding pose prediction datasets

Datasets to build models that predict the correct binding pose need information on known experimental complexes. PDBBind [24] contains measured binding affinity data for protein-ligand complexes, protein-protein complexes, protein-nucleic acid complexes, and nucleic acid-ligand complexes with known structures in the PDB [25]. The database is updated every year and the 2017 release has a total of 17,900 entries, with 14,761; 2181; 837; and 121 entries per each class, respectively. This full set is called the “general set”. A “refined set”, containing 4154 entries, is also given and considered an acceptable dataset for molecular docking and scoring function studies. Not every year, but routinely a “core set” or CASF benchmark (Comparative Assessment of Scoring Functions) [26,27] is selected from the refined set using a systematic, non-redundant sampling procedure. These sets are named after the release year of the refined set in which they are built. Entries are selected for the core set by clustering similar proteins using their primary sequence and selecting a poor, medium, and good binder from each cluster. Entries with poor electron density and no structure factor data are filtered out.

Mother of All Databases (MOAD) [28] is a source of experimentally resolved three-dimensional structures of ligand-protein complexes. In their 8th release in 2013, they had 23,269 complexes covering 6960 protein families and 11,173 unique ligands with an associated binding affinity for 35% of the complexes. The database also provides ligand annotation and protein classification and has some pre-computed docking files to use for scoring methods development on a subset of

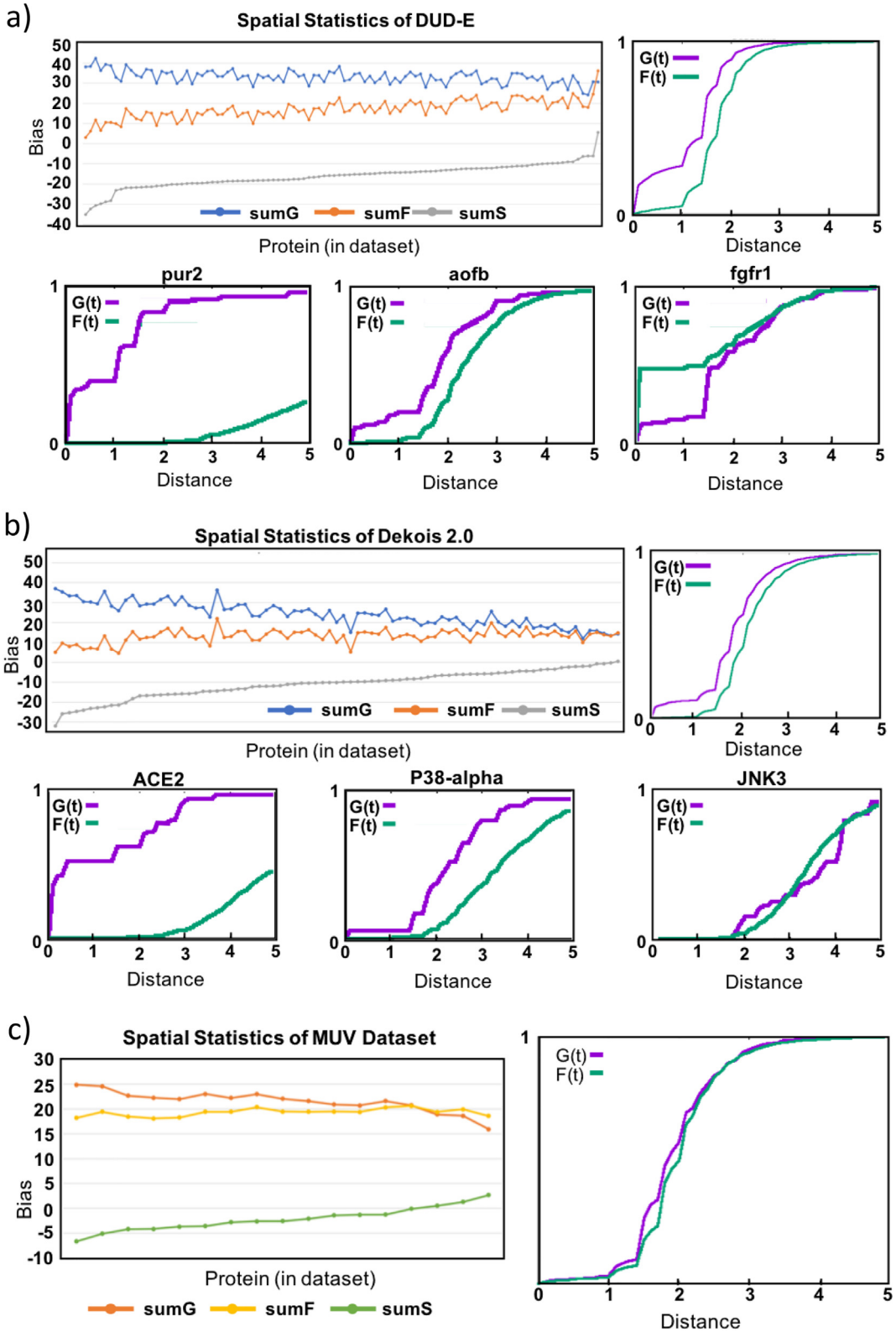

Fig. 1. Spatial statistics of a) DUD-E dataset, b) DEKOIS 2 dataset, and c) MUV dataset. The top left graph of each section shows ΣG, ΣF, and ΣS for each individual protein in the dataset ordered by increasing ΣS values. The top right graph of each section shows G(t) and F(t) for the combined dataset, comparing all active compounds to all decoy compounds. For a) and b) the bottom row, shows from left to right, the protein with the worst ΣS, the protein with the second to best ΣS, and the protein with the best ΣS.

data (CSAR described below).

The Community Structure Activity Resource (CSAR) [29] is a dataset that has 6 protein targets with 82 associated crystal structures (59 with extra stringent criteria) and 647 compounds with biological binding affinities. CSAR in concerned with collecting two types of data. 1) Comprehensive sets: targets that have known active compounds that have a span of 3–8 orders of magnitude in binding with some known inactive compounds and a highly-confident crystal structure. 2) Activity cliffs: pairs of compounds that have minimal changes but large changes in binding affinity, hopefully in interaction with a target in the comprehensive set. CSAR also hopes to elucidate the experimental variance of experimental binding measurements, and for their in-house sets, they include several experimental measurements over three different binding measurement techniques (isothermal titrating calorimetry, ThermoFluor, and Octet RED). CSAR has stringent criteria for protein and small molecules structures and they collaborated primarily with pharmaceutical companies to develop well validated data. All of their included crystal structures are set up for pose prediction docking tests. They ran redocking experiments with DOCK [30] and provide one near native structure (< 1 A RMSD) and a set of pose decoys that span the binding pocket and are more than 2 Å RMSD from the native pose.

6.2. Binding energy prediction datasets

While subsets of the other datasets can also be used for binding energy predictions, these datasets focus on known binding compounds and do not necessarily have experimental structures of the complexes associated with every entry. BindingDB [31] is a database of experimental protein-small molecule interactions that is collected from scientific literature and patents and coming from nearly 500,000 molecules and thousands of proteins. The database collects other experimental conditions of the measurements, such as temperature, pH, and buffer composition. Their collection of information from US patents is a valuable resource of data not necessarily in the literature. They invite investigator direct deposition of data and curate all data. The database also includes some simulated data. They have potential 3D conformations of ligands generated with Vconf [32]. The Therapeutic Target Database (TTB) [33] has a recent update in which they include, along with much more data relevant to protein-drug binding research, drug resistance mutations in 83 targets and regulatory genes that are resistant to 228 drugs targeting 63 unique diseases, differential expression profiles in disease relevant and drug-targeted tissues and in healthy individuals. DrugBank [34] is a rich database first established in 2006 and still being updated. It contains drug, drug-target, drug action, and drug interaction data for FDA approved and experimental drugs. The Drug Target Commons (DTC) maintains a community-driven effort to curate bioactivity data, which is available for bulk download and user derived queries. The DTC provides access to experimental binding affinity measurements for a wide range of compounds and targets [35,36].

6.3. Virtual high-throughput screening datasets

In order to make a model that correctly filters out non-binding compounds, it is important to have inactive (or decoy) molecules in the dataset. Directory of Useful Decoys – Enhanced (DUD-E) [37] provides 102 datasets of unique proteins, with an average of 124 active compounds per target and 50 decoys per active compound. The active sets are clustered by Bemis-Murcko atomic frameworks to reduce chemotype bias in the active sets. Decoy compounds are generated using ZINC [38] and property matching to the actives based on molecular weight, MiLogP, number of rotatable bonds, hydrogen bond acceptors, hydrogen bond donors, and net charge. Potential active compounds are filtered out of the decoy set using the ECFP4 fingerprint and removing 75% of the decoys most similar to the active set. Only 9219 decoys are experimentally confirmed. Since a ΣS near zero is best according to the spatial statistics work, we can see from Fig. 1-a that none of the individual protein datasets in DUD-E are very good by this measure. Looking at the dataset that performed the worst by this measure, pur2, we can see that nearly all of the active compounds have a nearest neighbor less than a distance of 2 ($G(t)$), while none of them have a nearest decoy within this distance ($F(t)$). This gap gets smaller with *aofb,* but still exists. Strangely, the trend switches for *fgfr1*, where more actives have a decoy nearest neighbor than another active, which may be an error in the dataset itself. Another study pointed out bias in DUD-E when it showed that using only ECFP fingerprints of the ligands and a simple logistic regression (with 72 proteins in the training and 30 proteins in the test set), it achieved a mean AUC of 0.904, with no information on the targets [9].

DEKOIS 2 [39] provides 81 benchmark datasets of 80 unique proteins, and one that has separate datasets for two different known binding pockets in the same protein. There are four proteins that overlap with DUD-E. The active sets are extracted from BindingDB [31]. Weak binders are excluded, and 40 distinct actives are selected by clustering by FCFP6 Tanimoto similarities. Three datasets are extended by selecting up to 5 actives from each structurally diverse cluster. The decoy set is generated using ZINC [38] and property matching to the actives based on molecular weight, octanol-water partition coefficient (logP), hydrogen bond acceptors and donors, number of rotatable bonds, positive charge, negative charge, and aromatic rings. Possible latent actives in the decoy set are removed using a score based on the FCFP6 and the size of the matching substructures and any decoy that contained a complete active structure as a substructure is removed. The majority of the DEKOIS 2 proteins have a ΣS greater than −10 while only a few of the DUD-E protein sets do (Fig. 1). The sets with the worst score are similar to the ones in DUD-E, but improve quickly compared to DUD-E sets.

The Maximum Unbiased Validation (MUV) dataset [22] is built using PubChem Bioassays [40] employing the spatial statistics described in Section 5.1 to ensure an unbiased dataset. The MUV dataset comes from 18 pairs of assays. To be considered for the dataset a target must have a HTS and a confirmatory screen (low-throughput dose-response experiment). HTS screens are commonly affected by experimental noise and artifacts, so all potential active compounds were confirmed with a confirmatory bioassay with a lower error rate. All non-binders from the primary HTS are kept as potential decoys and all binders from the secondary confirmatory screen are kept as potential actives. Even though the potential actives were verified by a secondary screen, these are still prone to artifacts. Three assay artifact filters were used: Hill slope filter (detects aggregates), frequency of hits filter (detects promiscuous binders), and the autofluorescence and luciferase inhibition filter. Potential false decoys were tested by doing a similarity search using the simple descriptors. Next, to ensure that all actives were within the chemical space needed to create an unbiased dataset, a chemical space embedding filter was derived using spatial statistics. The final dataset was established by selecting 30 actives with a common level of spread. It is no surprise that the spatial statistics for MUV look much better than the previous examples since this dataset was generated using them (Figure 2). However, it often receives poor metrics in machine learning models because of how difficult the active set is to classify and the small number of actives for training data [9].

It can be seen in the AVE bias study of DUD-E and DEKOIS 2 (Fig. 2 and Table 2), that the highest correlation between AUC and the bias score is with the k-nearest neighbors, as expected by the claims in their study. It can also be seen that the lowest AUC values with the k-nearest neighbors correspond to the protein datasets that have the best bias score (close to zero). It can also be seen that not many of the proteins in DUD-E get a bias score close to zero which is no surprise based on the ΣS values of the dataset and DEKOIS 2 does better. It can also be seen that there is less correlation with MUV because it is inherently already a hard dataset to classify. A full list of AUC and bias scores for each individual protein in the three different benchmark datasets are given in

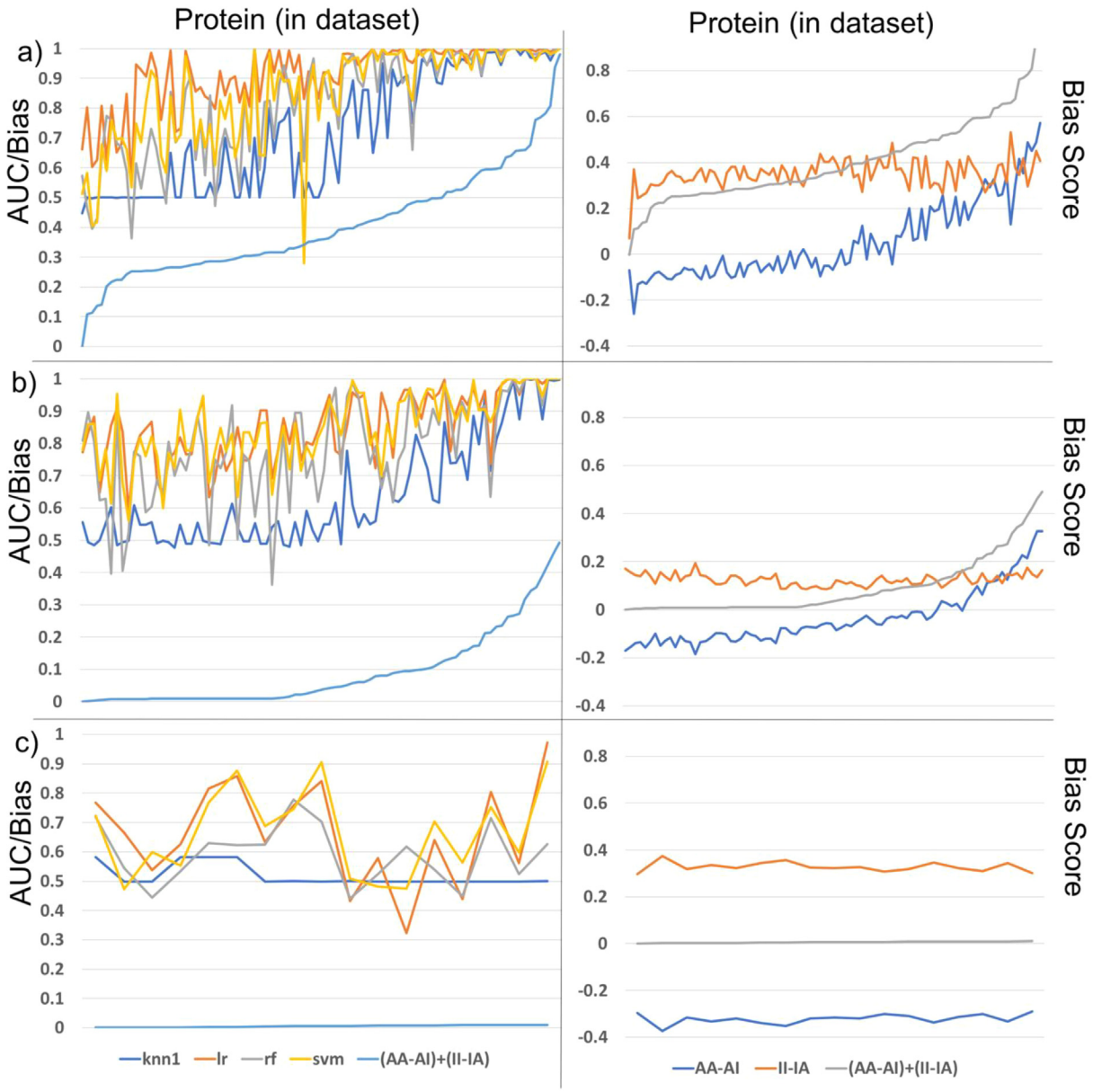

Fig. 2. The left-hand side gives the AUC of a AVE split using k-nearest neighbors (k = 1), linear regression, random forest, and support vector machine and also the AVE bias, (AA-AI) + (II-IA). The x-axis is each protein in the dataset, order by lowest to highest AVE bias from left to right. The right-hand side gives the breakdown of the active set clumping (AA-AI) and the inactive set clumping (II-IA) and the overall bias, again ordered by the AVE bias. This information is reported for the a) DUD-E dataset, b) DEKOIS 2 dataset, and 3) MUV dataset.

the Supporting Material.

7. Conclusions and future directions for the field

This paper provides an overview of types of models for drug binding predictions, sources of data for the models, some current trends in modeling, and also highlights some of the problems with the data and current strategies to quantify it. The summary from previous work is that knn (n = 1) can detect the potential for overfitting since it memorizes the training data, the AVE Bias correlates with the AUC for KNN

Table 2
Average AUC over all the proteins in each dataset using k-nearest neighbors (k = 1), linear regression, random forest, and support vector machine and the correlation of the AUC with the AVE bias.

	DUD-E		DEKOIS 2		MUV	
	Pearson correlation	Average AUC	Pearson correlation	Average AUC	Pearson correlation	Average AUC
knn1	0.85	0.72	0.91	0.64	−0.58	0.52
lr	0.65	0.92	0.63	0.86	−0.13	0.66
rf	0.73	0.82	0.58	0.79	−0.06	0.59
svm	0.69	0.85	0.64	0.85	0.06	0.67

($n = 1$) and therefore can be a measure for the potential for overfitting and used to attempt to debias datasets. However, we can see from evaluating some common benchmarks that good performance correlates with models with high-predicted bias scores and models with low bias scores do not have much predictive power. Our current and future interests include using weighed metrics that do not limit what the training data may be to keep predictive power but still have a better understanding of the potential overfitting. We are also interested in developing bias scores based on the protein-ligand complexes and visualizing this space to better understand the data missing from our models and the domain in which we can make useful predictions.

Acknowledgements

This research was supported by the Cancer Research Informatics Shared Resource Facility of the University of Kentucky Markey Cancer Center (P30CA177558), the University of Kentucky CCTS KL2TR000116 and 1KL2TR001996-01 grants, and the Markey Women Strong – philanthropy grant.

Declaration of Competing Interests

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagen.2020.129545.

References

[1] F. Gräter, et al., Protein/ligand binding free energies calculated with quantum mechanics/molecular mechanics, J. Phys. Chem. B 109 (20) (2005) 10474–10483.
[2] D.L. Mobley, M.K. Gilson, Predicting binding free energies: frontiers and benchmarks, Annu. Rev. Biophys. 46 (2017) 531–558.
[3] C. Mirabello, G. Pollastri, Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility, Bioinformatics 29 (16) (2013) 2056–2058.
[4] E. Gasteiger, et al., Protein identification and analysis tools on the ExPASy server, The proteomics protocols handbook, Springer, 2005, pp. 571–607.
[5] P. Zhang, et al., A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks, Brief. Bioinform. 18 (6) (2016) 1057–1070.
[6] Z. Wu, et al., MoleculeNet: a benchmark for molecular machine learning, Chem. Sci. 9 (2) (2018) 513–530.
[7] I. Wallach, M. Dzamba, A. Heifets, AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint, arXiv:1510.02855 (2015).
[8] B. Ramsundar, et al., Massively multitask networks for drug discovery, arXiv Preprint (2015) arXiv:1502.02072.
[9] A. Gonczarek, et al., Interaction prediction in structure-based virtual screening using deep learning, Comput. Biol. Med. 100 (2018) 253–258 Elsevier.
[10] M. Wen, et al., Deep-learning-based drug–target interaction prediction, J. Proteome Res. 16 (4) (2017) 1401–1409.
[11] A. Korotcov, et al., Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets, Mol. Pharm. 14 (12) (2017) 4462–4475.
[12] J. Gomes, et al., Atomic convolutional networks for predicting protein-ligand binding affinity, arXiv preprint (2017) (arXiv:1703.10603).
[13] M. Ragoza, et al., Protein–ligand scoring with convolutional neural networks, J. Chem. Inf. Model. 57 (4) (2017) 942–957.
[14] M.M. Stepniewska-Dziubinska, P. Zielenkiewicz, P. Siedlecki, Development and evaluation of a deep learning model for protein-ligand binding affinity prediction, Bioinformatics 1 (2018) 9.
[15] I. Kundu, G. Paul, R. Banerjee, A machine learning approach towards the prediction of protein–ligand binding affinity based on fundamental molecular properties, RSC Adv. 8 (22) (2018) 12127–12137.
[16] A. Mayr, et al., Large-scale comparison of machine learning methods for drug target prediction on ChEMBL, Chem. Sci. 9 (24) (2018) 5441–5451 Royal Society of Chemistry.
[17] H. Öztürk, E. Ozkirimli, A. Özgür, DeepDTA: deep drug-target binding affinity prediction, Bioinformatics 34 (17) (2018) i821–i829 Oxford University Press.
[18] M.I. Davis, et al., Comprehensive analysis of kinase inhibitor selectivity, Nat. Biotechnol. 29 (11) (2011) 1046.
[19] J. Tang, et al., Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis, J. Chem. Inf. Model. 54 (3) (2014) 735–743.
[20] T.R. Stouch, The errors of our ways: taking account of error in computer-aided drug design to build confidence intervals for our next 25 years, J. Comput. Aided Mol. Des. 26 (1) (2012) 125–134.
[21] S.G. Rohrer, K. Baumann, Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics, J. Chem. Inf. Model. 48 (4) (2008) 704–718.
[22] S.G. Rohrer, K. Baumann, Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data, J. Chem. Inf. Model. 49 (2) (2009) 169–184.
[23] I. Wallach, A. Heifets, Most ligand-based classification benchmarks reward memorization rather than generalization, J. Chem. Inf. Model. 58 (5) (2018) 916–932.
[24] Z. Liu, et al., Forging the basis for developing protein–ligand interaction scoring functions, Acc. Chem. Res. 50 (2) (2017) 302–309.
[25] H.M. Berman, et al., The protein data bank, Nucleic Acids Res. 28 (1) (2000) 235–242.
[26] Y. Li, et al., Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results, J. Chem. Inf. Model. 54 (6) (2014) 1717–1736.
[27] Y. Li, et al., Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set, J. Chem. Inf. Model. 54 (6) (2014) 1700–1716.
[28] A. Ahmed, et al., Recent improvements to Binding MOAD: a resource for protein–ligand binding affinities and structures, Nucleic Acids Res. 43 (D1) (2014) D465–D469.
[29] J.B. Dunbar Jr.et al., CSAR data set release 2012: ligands, affinities, complexes, and docking decoys, J. Chem. Inf. Model. 53 (8) (2013) 1842–1852.
[30] D.T. Moustakas, et al., Development and validation of a modular, extensible docking program: DOCK 5, J. Comput. Aided Mol. Des. 20 (10 − 11) (2006) 601–619.
[31] M.K. Gilson, et al., BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic Acids Res. 44 (D1) (2015) D1045–D1053.
[32] C.E. Chang, M.K. Gilson, Tork: conformational analysis method for molecules and complexes, J. Comput. Chem. 24 (16) (2003) 1987–1998.
[33] Y.H. Li, et al., Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics, Nucleic Acids Res. 46 (D1) (2017) D1121–D1127.
[34] D.S. Wishart, et al., *DrugBank 5.0:* A major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (D1) (2017) D1074–D1082.
[35] J. Tang, et al., Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions, Cell Chem. Biol. **25** (2) (2018) 224–229 (e2).
[36] Z. Tanoli, et al., Drug Target Commons 2.0: a community platform for systematic analysis of drug–target interaction profiles, Database (2018) 2018.
[37] M.M. Mysinger, et al., Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking, J. Med. Chem. 55 (14) (2012) 6582–6594.
[38] J.J. Irwin, et al., ZINC: a free tool to discover chemistry for biology, J. Chem. Inf. Model. 52 (7) (2012) 1757–1768.
[39] M.R. Bauer, et al., Evaluation and optimization of virtual screening workflows with DEKOIS 2.0–a public library of challenging docking benchmark sets, J. Chem. Inf. Model. 53 (6) (2013) 1447–1462.
[40] Y. Wang, et al., PubChem BioAssay: 2017 update, Nucleic Acids Res. 45 (D1) (2016) D955–D963.