JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Perspective

Cite This: *J. Chem. Inf. Model.* 2019, 59, 947−961

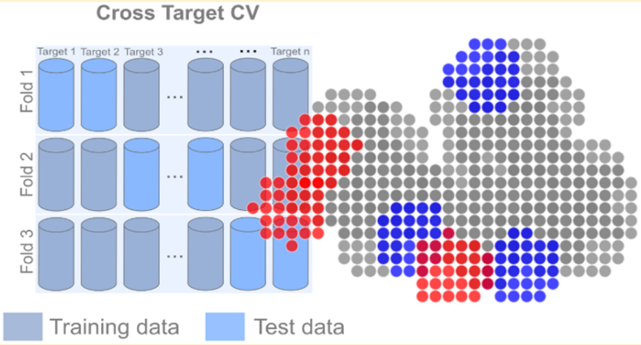
pubs.acs.org/jcim

In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening

Jochen Sieg, Florian Flachsenberg, and Matthias Rarey*

Universität Hamburg, ZBH - Center for Bioinformatics, Research Group for Computational Molecular Design, Bundesstraße 43, 20146 Hamburg, Germany

Supporting Information



ABSTRACT: Reports of successful applications of machine learning (ML) methods in structure-based virtual screening (SBVS) are increasing. ML methods such as convolutional neural networks show promising results and often outperform traditional methods such as empirical scoring functions in retrospective validation. However, trained ML models are often treated as black boxes and are not straightforwardly interpretable. In most cases, it is unknown which features in the data are decisive and whether a model's predictions are right for the right reason. Hence, we re-evaluated three widely used benchmark data sets in the context of ML methods and came to the conclusion that not every benchmark data set is suitable. Moreover, we demonstrate on two examples from current literature that bias is learned implicitly and unnoticed from standard benchmarks. On the basis of these results, we conclude that there is a need for eligible validation experiments and benchmark data sets suited to ML for more bias-controlled validation in ML-based SBVS. Therefore, we provide guidelines for setting up validation experiments and give a perspective on how new data sets could be generated.

1. INTRODUCTION

The basic task of virtual screening (VS) is to prioritize large *in silico* molecule libraries by the probability of the molecules to show activity against a particular protein target. A distinction can be made between ligand-based and structure-based VS. In ligand-based VS, new active molecules are predicted based on their similarity to known actives. In contrast, structure-based VS (SBVS) methods model the interactions between small molecules and the protein to predict actives.[1]

New VS methods are typically tested by retrospective validation on benchmark data sets.[2,3] Benchmark data sets contain molecules of known activity and are used for a standardized comparison of different methods to select the method best suited for a task.[4] Frequently used examples are the Directory of Useful Decoys (DUD),[5] the Directory of Useful Decoys - Enhanced (DUD-E),[6] Demanding Evaluation Kits for Objective *in silico* Screening (DEKOIS),[7,8] the Maximum Unbiased Validation Data set (MUV),[4] the Community Structure−Activity Resource (CSAR),[9] PDBbind,[10] and more.[11]

In recent years, machine learning (ML) methods have been trained and evaluated on these data sets.[12−17] The reported results show that ML methods outperform other methodologies such as empirical and knowledge-based scoring functions on these data sets.[12−17] However, the interpretability of many ML methods is not straightforward.[18] On the one hand, it is of great interest to understand the determinants of decision making of high-performing models to deduce the relationships potentially not captured otherwise. On the other hand, it is not recognizable whether a model's decisions are based on real signals in the data or on bias. We made two conspicuous observations in current literature that suggest that the latter is the case and bias is learned unnoticed from established data sets. We see the reason for this bias in the insufficiency of the current standard of validation experiment design, which is consistent with recent findings in a similar domain.[19]

In the following, an overview of benchmark data sets for VS and their advantages and disadvantages is given. Then it is evaluated if the unbiasing protocols of the examined data sets are suited for ML methods on the examples of the DUD, DUD-E and MUV data sets. Subsequently, our observations from the literature are described and analyzed. The results reveal that small molecule features dominate the predictions across dissimilar proteins when actually a structure-based descriptor is used, leading to biased models. Based on these results and observations, we propose guidelines for validation experiments to avoid bias and finally give an outlook on the generation of new data sets suitable for ML methods.

Special Issue: Machine Learning in Drug Discovery

Received: October 12, 2018
Published: March 5, 2019

DOI: 10.1021/acs.jcim.8b00712

1.1. Examples of Benchmark Data Sets in Virtual Screening. Benchmark data sets consist of sets of active and inactive molecules, each associated with a specific target. Often the actives are experimentally validated, but the documentation of experimentally validated inactive molecules is scarce. For this reason, assumed inactives, called decoys are frequently used.[20]

The first benchmark-like data sets were published in the early 2000s.[21−23] These data sets include randomly selected molecules as assumed inactives, and first approaches for picking samples while avoiding bias caused by the data set's compositions have been undertaken.[23]

Different biases have been identified by the community over time, which can either artificially increase or decrease prediction performance. Verdonk revealed that differences in distributions of basic physicochemical molecular properties of the active and inactive sets leads to artificial discrimination by those low dimensional features rather than features of higher dimensions.[24] An example is that many scoring functions favor molecules of larger size in docking because the potential number of interactions correlates with size.[24] This bias has been described by the term artificial enrichment and has been a problem when random molecules were selected as inactives. It can be counteracted by selecting inactives such that they are similar to the active molecules in terms of low dimensional properties.[24] In contrast to overestimations, the enrichment can be underestimated when utilizing experimentally non-validated inactives for which the assumption of inactivity turns out to be false.[5] This bias is termed false negative bias.

In 2006, Huang et al. introduced the DUD data set, which in its generation protocol addresses artificial enrichment and false negative bias.[5] DUD focuses on docking methods and comprises 40 different protein targets. The original DUD version contained 2950 actives and 95 326 assumed inactives. To circumvent the problem of the deficiency of experimentally validated inactive molecules, so-called decoys are selected *in silico* from the ZINC database.[25] Artificial enrichment has been addressed by selecting decoys such that they resemble the active molecules in their basic physicochemical properties. Those properties are molecular weight (MW), LogP, number of hydrogen bond acceptors and donors, as well as the number of rotatable bonds (see Table 1). It is worth mentioning that the presence of amine, amide, amidine and carboxylic acid has also been considered but with a lower priority. Simultaneously, to provide a higher confidence that decoys are actually inactive, the selection process ensures that each decoy is dissimilar to any of the active molecules with respect to CACTVS fingerprints[26] and a Tanimoto coefficient threshold of 0.9. The evaluation in a comparative docking study showed less artificial enrichment in DUD than in earlier data sets.[5] In fact, DUD has been considered the gold standard after its release and is still used today.[20]

Two years after the release of DUD, in 2008, analogue bias has been described by Good and Oprea.[27] Analogue bias is based on the observation that artificially improved enrichment can be achieved if a data set contains many analogue actives with the same chemotype. The activity of ligands in a cluster that share the same scaffold is easy to predict as soon as the activity of a single molecule of the cluster can be identified. Consequently, a common scaffold shared by actives but not present in the inactives leads to overestimations. This bias has been found in DUD.[27] A strategy to address this bias is to diversify the ligands by clustering actives by their scaffolds and selecting representatives.[27] Another limitation of DUD has been the chosen set of matched properties. Multiple groups reported that net charges are a strong discriminative feature in the data set,[28,29] which may lead to artificial enrichment in validation.

Table 1. List of Unbiased Features of DUD, DUD-E, and MUV

DUD[5]	DUD-E[6]	MUV[4]
molecular weight	molecular weight	
number of hydrogen bond acceptors	number of hydrogen bond acceptors	number of hydrogen bond acceptors
number of hydrogen bond donors	number of hydrogen bond donors	number of hydrogen bond donors
number of rotatable bonds	number of rotatable bonds	
logP	logP	logP
	net charge	
		number of all atoms
		number of heavy atoms
		number of boron atoms
		number of bromine atoms
		number of carbon atoms
		number of chlorine atoms
		number of fluorine atoms
		number of iodine atoms
		number of nitrogen atoms
		number of oxygen atoms
		number of phosphorus atoms
		number of sulfur atoms
		number of chiral centers
		number of ring systems
5 features	6 features	17 features

In 2012, after DUD had been analyzed in many studies and shortcomings had been identified, the DUD-Enhanced (DUD-E) data set was published.[6] The DUD-E compilation protocol addresses shortcomings of DUD and simultaneously extends the DUD data set to 22 886 actives and 1 411 214 decoys for 102 targets. The additional actives were retrieved from ChEMBL[30] and the inactives from the ZINC[25] database. To address analogue bias, active molecules were clustered by their Bemis-Murcko scaffolds.[31] To further reduce artificial enrichment bias, net charges were added to the matched properties between actives and decoys (see Table 1). Finally, a more stringent topology filter was employed during decoy selection to further reduce the probability of false negative inactives.[6]

Most of the benchmark data sets in VS focus on structure-based methodologies such as docking.[32] A popular example of a benchmark data set specifically designed for ligand-based methods is the maximum unbiased validation (MUV) data set collection.[4] MUV was published in 2009 and it comprises 17 separate data sets each associated with a target protein. Each data set contains 30 active and 15 000 inactive molecules, all retrieved from PubChem.[33] Note that MUV contains experimentally analyzed actives and inactives. Therefore, the probability that the inactives are in fact inactive is high. Samples of MUV were selected by a strategy addressing the data set's representation in a certain descriptor space (termed the data set's topology) with methods from spatial statistics.

The goal of MUV design was to reduce artificial enrichment and analogue bias by selecting samples such that a common spread between actives and other actives as well as actives and inactives is employed in a descriptor space of 17 simple features (see Table 1). The goal is a data set topology in simple descriptor space in which the probability that the nearest neighbor of each active is an active or an inactive is equal.[4] The MUV data sets have been developed with the focus on ligand-based methods, but the authors note its usability in SBVS as well,[4] which has been done in studies.[15]

Table 2 gives an overview of the DUD, DUD-E, and MUV data set.

Table 2. Overview of Three Benchmark Data Sets DUD, DUD-E, and MUV

	DUD	DUD-E	MUV
number of targets	40	102	17
targeted methodology	docking	docking	ligand-based similarity search
number of unbiased features	5	6	17
special design decision	2D dissimilarity	2D dissimilarity	experimental inactives
number of citations[34]	782	366	106

1.2. Bias in Chemical Data. The term bias has several connotations and is often not used uniformly. In essence, bias describes the distortion from a true underlying relationship. Available chemical data are biased because experiments are conducted with different intentions than sampling the chemical space uniformly.[35−37] Chemical space is infinite, but the pharmacologically relevant space is estimated to comprise about 10^{60} molecules.[38] The diversity of the synthesized subspace is biased due to known molecules and even *de novo* projects focus on molecules near the known active molecules.[36] There are legitimate reasons for excluding certain molecules from drug discovery projects for example costs, synthetic feasibility and availability in a library.[37] These reasons are comprehensible in drug discovery processes, but they prevent a uniform sampling of the chemical space. However, a nonuniform sampling does not mean that methods based on the available chemical data can not be used in practice, but it is important to consider the composition of the data in validation procedures and therefore in any benchmark data set. Otherwise it is not clear whether a method performs better because of a superior methodology or beneficial validation data.

Over time, several tendencies of bias in chemical data have been described. Cleves and Jain[35] presented general biases in chemical data as inductive bias. The authors showed that active ligands that are known today have been synthesized due to decisions of humans based on different assumptions, which may lead to advantageous performance of methods making the same assumptions. Those ligands are often synthesized based on their similarity to known ligands. They demonstrated that historically, known drugs for some targets show a noticeable 2D similarity in dependence of time, which they called 2D bias. Typically, actives for a specific target with high 2D similarity are patented in a narrow time span whereas more 2D dissimilar actives tend to be discovered years later. Consequently, for these ligands 2D methods have an artificial advantage over other methods.[35]

Another bias not specifically addressing the data composition has been described by Jain et al.[39] and is called confirmation bias. This bias is the tendency of a human to try to confirm a hypothesis by purely searching for a correlation with the outcome of the hypothesis. However, this correlation may not be physically founded and this approach can lead to false conclusions. For example, a model can be selected on the basis of correlation with some scoring function, but this scoring function might be based on assumptions that contradict the physical reality. An example from ligand-based VS would be the hypothesis that molecules similar to known active molecules are active as well, while ’activity cliffs’ are not considered.[39]

In summary, there are several bias specifications describing certain scenarios of distorted data composition in the literature that contain patterns or signals that should not be learned by a model, because they misrepresent the true underlying distribution.

We will introduce another specification related to confirmation bias that in our opinion describes the worst kind of bias in a data collection. In particular, we distinguish domain bias and noncausal bias. Both falsely present prospective predictivity. On the one hand, domain bias distorts a prediction because the distribution of the sampled population resembles easy test cases or less diverse samples. The bias is based on biological mechanisms, such that the model is right for the right reason, but the applicability domain is narrow. An example for domain bias is when the train and test samples are too similar, for example, when a common scaffold is shared by the actives in the training and test set. Therefore, the measured model performance is biased to a certain domain, which is acceptable if modeling this domain is the aim, but is insufficient for generalization.

On the other hand, noncausal bias describes the case in which there is correlation but no causation. In this case, good predictions can be achieved by patterns in the data that do not represent any biological mechanism relevant for binding, but exhibit pure correlation with the labeled outcome. This bias yields a good statistic, but on the basis of fallacious models not based on physical reality. Such models do not work in general but only on data that fits the bias pattern, which makes them unusable for prospective predictions. Interestingly, there are reports of successfully finding leads based on *in silico* predictions, for which it has been experimentally refuted that the molecule binds for the predicted reason.[40]

In Section 4, we will show in detail on ML-based scoring functions from the literature that noncausal bias has been learned implicitly and unnoticed from established benchmark data sets.

1.3. Review of Benchmark Data Sets in Context of Machine Learning. ML methods are increasingly used in SBVS,[12−17,41] but to our knowledge there is no data set specifically dedicated to ML. Therefore, we review the purpose of established data sets and their limitations regarding ML methods in this paper.

A general distinction between benchmark data sets for SBVS and ligand-based VS is often made.[20,32] However, this differentiation might be misleading. A more accurate and fine-grained differentiation would be to categorize these data sets according to the methodology they have been designed for. Exemplarily, DUD and DUD-E are generated for docking with conventional scoring functions. Therefore, those data sets have been tailored for the requirements defined by those

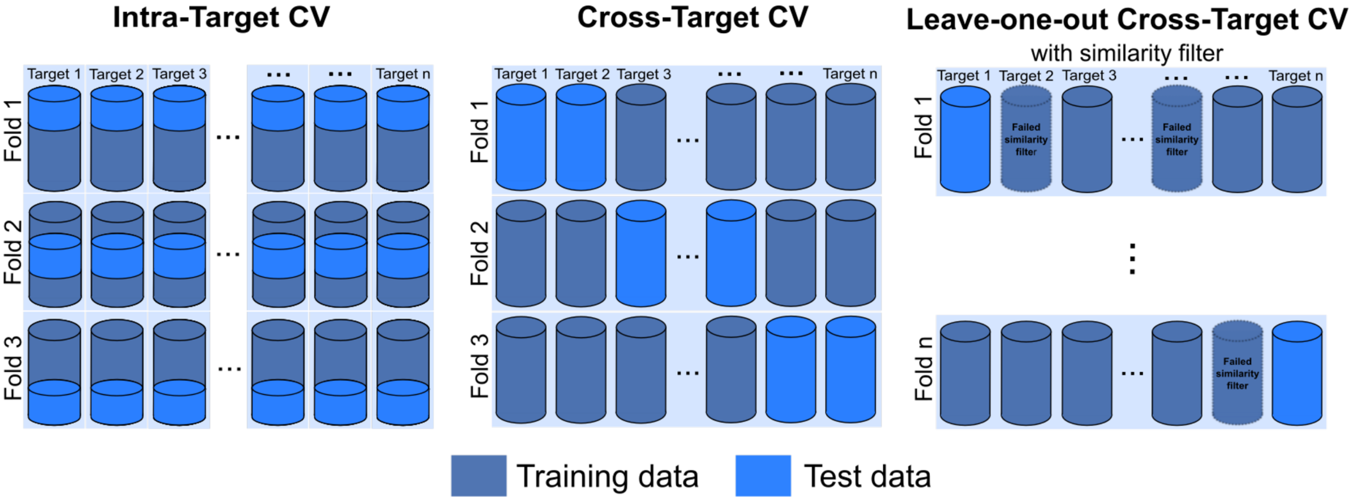


Figure 1. Three cross validation (CV) scenarios used in the experiments are depicted schematically, namely intra-target CV, cross-target CV, and leave-one-out cross-target CV with a protein similarity filter. In the case of the first two scenarios three folds are exemplarily depicted.

methods and their vulnerabilities. The 1D properties matched between actives and decoys by the DUD and DUD-E protocols are captured by scoring functions, for example the hydrogen bond donor count in hydrogen bonds terms. However, classification should not be driven by simple and unspecific molecular features such as donor and acceptor counts, but rather by the quality of interactions. Simultaneously, the 2D dissimilarity between actives and decoys can be employed, because conventional scoring functions do not capture 2D features like molecular topology. Similarly, the MUV data set is not compiled for general ligand-based VS methods but for nearest neighbor similarity search that starts with a query of actives and does not use the 17 simple features but rather more complex descriptors like MACCS structural keys. Consequently, different methods and descriptors not considered in a data set's compilation protocol might be unsuited for the data set even if they are also by definition ligand-based or structure-based.

An example illustrating this problem are net charges in the first version of the DUD data set, which have not been included in the matched properties between actives and inactives.[6] While 42% of the ligands were charged, only 15% of the decoys had a nonzero charge.[6] Therefore, it has been possible to artificially increase performance by just assigning charged molecules as active. After overoptimistic results were reported due to differing charge distributions, some updated versions of DUD have been released.[6,28,29] Accordingly, for validation it is important to compare the features considered in the compilation protocol with the descriptor to validate. It might be not straightforward to spot whether an improvement in score after the addition of a feature is due to bias.

Another example for a limitation of transferability of DUD and DUD-E is the employed 2D dissimilarity. It is known and stated by the authors[6] of DUD-E that 2D descriptors are inappropriate for use with their data set. Simultaneously, the same is obviously true for DUD. Still there are reports using these descriptors on those data sets. The extent of distinctness by 2D features has been analyzed by Bietz et al.[42] by mining the most discriminative SMARTS-patterns in the DUD data set. It could be shown that, for example, for the AMPC target 80% of all ligands contain a sulfur atom and only 10% of the decoys. There are other examples in which simple patterns such as the presence of single atoms can discriminate a noticeable portion of actives and inactives.[42] The 2D dissimilarity is expected, but it should be noted that the decoys can be easily distinguished according to very basic substructures, which might be relevant for the validation of novel descriptors.

In summary, data sets have a design purpose focusing on a specific methodology. Data set design decisions are made based on the goal of the data set to provide good test cases for the targeted methodology. This might lead to bias when the data set is used with methods or descriptors that are different than the targeted methodology. Special care has to be taken, because there is a repertoire of nonlinear ML algorithms that can be paired with a large variety of chemical descriptors. For this reason we evaluate whether ML methods can be validated on established data sets. Concretely, we will evaluate the unbiasing techniques of DUD, DUD-E, and MUV in Section 3 and analyze noncausal bias learned from a subset of these data sets in Section 4.

2. METHODS

Experiments are conducted in Python using RDKit[43] for reading molecules and calculating molecular features. For ML experiments the libraries scikit-learn,[44] Keras,[45] and TensorFlow[46] were utilized. The charge-corrected DUD, DUD-E, and MUV data sets were downloaded from the respective web pages.[47−49]

On the one hand, classification performance of predictors is evaluated with the area under the receiver operating characteristic curve (AUC). This metric provides a value between 0 and 1, where 0.5 indicates a random guess.[50] On the other hand, VS is an early recognition problem, which we assess with the enrichment factor at the top *x* percent of the ranked predictions, for example the top one percent (EF1%).[51] Both measures are highly used in the field of VS and therefore should make the results of our experiments comparable to published classifier performances.

In the following experiments, three different cross validation (CV) experiments are performed, which are illustrated in Figure 1. In the first validation scenario, the set of molecules belonging to a single target *t* is split into training and validation sets. This validation procedure will be called intra-target CV

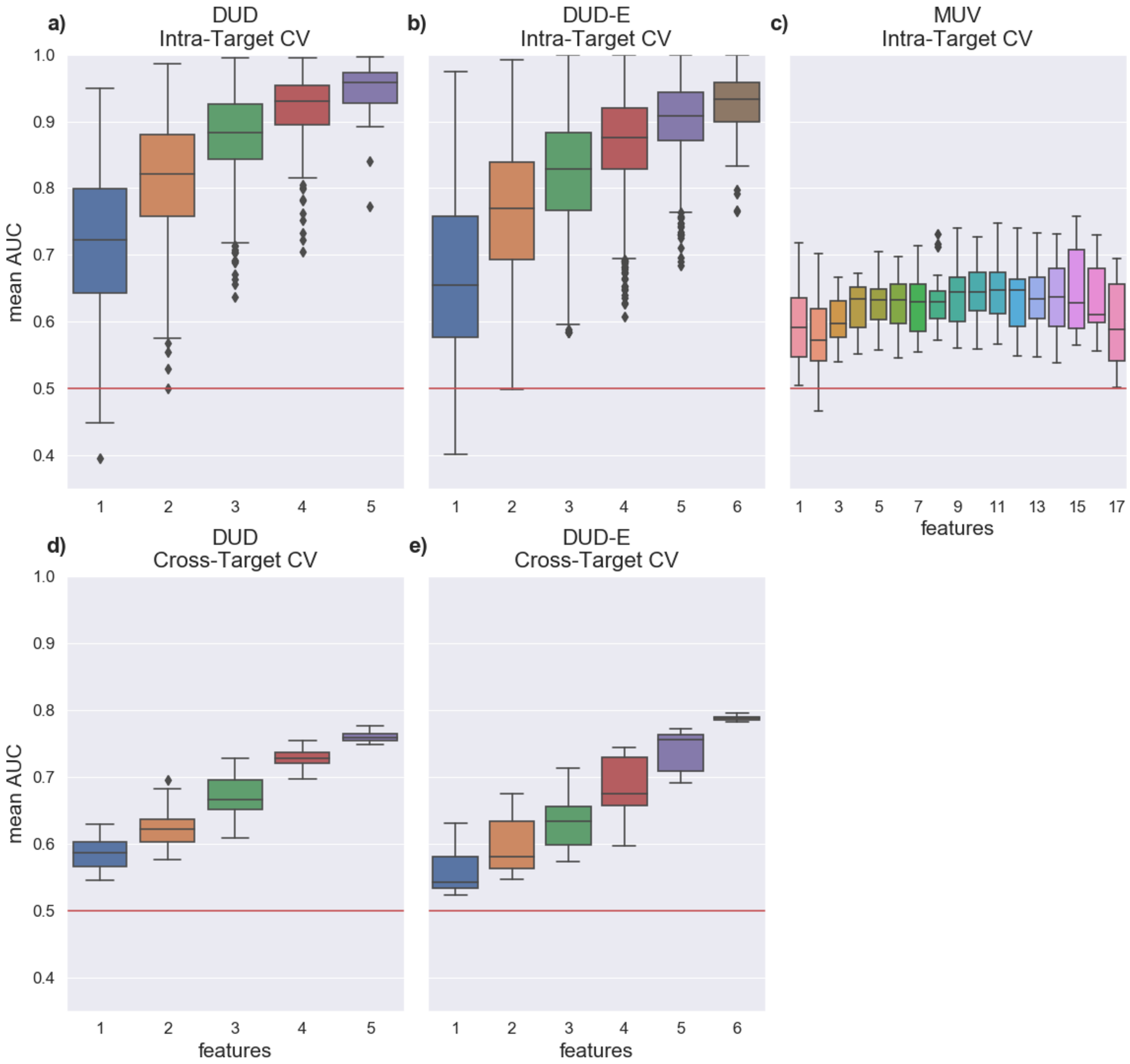


Figure 2. Results of the evaluation of unbiased features with AUC of DUD, DUD-E, and MUV with RF. The first row shows results of the intra-target CV for (a) DUD, (b) DUD-E, and (c) MUV. In the second row, results of cross-target CV are depicted for (d) DUD and (e) DUD-E.

and represents the task in VS of predicting further molecules that are active against a target based on known active and inactive molecules of this target. The second scenario is called cross-target CV. Here, the set of targets $T = \{t_1, ..t_n\}$ is split into disjunctive training set $T_{train} \subset T$ and test set $T_{test} \subset T$. With each target in T a set of molecules is associated. This CV represents a more challenging task for the method because the applicability domain is not restricted to a single target. Instead, the goal of this validation is to assess the predictive power across perhaps unrelated targets. There are different variations of cross-target CV. For example a more sophisticated version is leave-one-out cross-target CV with a similarity filter (see Figure 1). In n iterations each target $t \in T$ is once used as the test set. To aggravate predictions, a similarity filter is applied before training to remove targets from the training set, which are similar to the test target. The aim of this CV is to assess predictive power across dissimilar targets.

3. EVALUATING UNBIASING TECHNIQUES

The unbiasing techniques applied in the protocols of DUD, DUD-E, and MUV are evaluated for their consistency with machine learning (ML) methods. In all three data sets, the unbiasing comprises the reduction of the discriminative power of simple features (see Table 1 for a list of features) by the compilation protocols. Since protocols address those features, a reasonable assumption would be that those features barely contribute to predictions. To examine this unbiasing with ML, learning models were trained and tested while using only these features for predictions. Since ML methods are effective at capturing patterns across multiple features, it is also interesting to evaluate combinations of features. Accordingly, in this experiment we put to test whether the predictive power of these features is reduced in the data sets when ML is used for prediction.

3.1. Evaluation Setup. First, the unbiased features were calculated using RDKit. In the case of DUD all $\sum_{k=1}^{n=5}\binom{n}{k} = 31$

combinations of the 5 unbiased features were calculated. For DUD-E there are 63 feature sets and in the case of MUV 131 071 sets. Since for MUV the number of feature sets is too high, a greedy enumeration strategy was applied. A backward elimination[52] was employed to subsequently remove features from the whole unbiased feature set. Initially, the set of MUV features F_f, where f dedicates the number of features in the set, contains $f = 17$ features. The feature set F_{17} was used to train and evaluate models with cross validation. Then all possible sets of F_{f-1} were evaluated in the same way and the single feature contributing the least to the performance in terms of AUC in the cross validation was eliminated from the feature set. The process was iteratively repeated until $f = 1$ and the highest performing feature was determined. The strategy for feature subset enumeration is analogous to the enumeration used in feature selection tasks with wrapper methods.[52] In feature selection the goal is to reduce the number of features and select the best subset of features relevant to the learning task.[52] In contrast, the aim in this experiment is to identify whether and to what extent unbiased feature subsets of a benchmark data set perform with a given learning algorithm.

For evaluation, intra-target and cross-target CV were utilized (see Figure 1a,b). All CV experiments were conducted with 10 random folds. The folds of intra-target CVs were stratified to preserve the ratio of the samples for both classes in each fold.

As reference ML methods, Random Forest (RF) classifier[53] and logistic regression (LR) from scikit-learn were used with default parameters except for RF, for which the number of estimators was set to 400. LR was selected because it is a simpler linear model. In contrast RF is a nonlinear method and is considered comparatively robust against overfitting and easy to parametrize.[54] Furthermore, RF is a widely used method in the field of VS and drug discovery.

For these experiments, training and test sets were scaled column-wise on the basis of the respective training set by the formula $z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$, where i denotes the row and j the column in the feature matrix. Accordingly, from each single feature value x_{ij}, the mean μ_j of the column is subtracted and divided by the standard deviation σ_j.

3.2. Evaluation Results.

The results of intra-target CV with RF and LR evaluated with the AUC are presented in Figure 2a−c and S1a−c. Plots showing the performance assessed with the EF1% is shown in Figure S2a−c for RF and for LR in Figure S3a−c. Each boxplot shows the results of all targets of the respective data set. Hence, one point in the plot represents the mean performance of a method on a single target in 10-fold CV with a certain set of unbiased features. Consequently, a box in the plot shows the range of performance in AUC or EF1% over all targets when exactly x features are used for prediction.

The results of cross-target CV with AUC are depicted in Figure 2d and e for RF as well as in Figure S1d,e for LR. Results with EF1% are shown in Figure S2d,e for RF and Figure S3d,e for LR. In this experiment, random cross-target CV has been repeated ten times with different random splits to address empirical stability of the results. One point in these plots represents the mean performance of a method on a 10-fold cross-target CV with a certain set of unbiased features. Therefore, a box in these plots depicts the distribution of mean AUC or EF1% values of differently splitted cross-target CVs when x features are used.

We did not perform cross-target CV on MUV because of the high number of duplicates in the set of inactives of different targets. Of the 255 510 molecules in the whole MUV data set there are only 95 916 unique PubChem-compound-IDs. Deduplication would leave only 38% of samples and would yield an arbitrary composition not representing the MUV anymore. This redundancy is probably due to the fact that the experimentally analyzed space is generally small, but negative results are even less often reported.

The AUC results of intra-target CV are similar for DUD and DUD-E in Figures 2 and S1, which is comprehensible since both data sets have been generated with a similar strategy. A noticeable observation is that the predictive performance achieved by many of the models is rather high. Comparing RF and LR in these experiments shows higher median and maximum values for the nonlinear RF, which was expected. The highest achieved mean AUC with RF is 1.0 for both DUD and DUD-E, which indicates perfect performance. With LR the maximum mean values are 0.95 and 0.99 for DUD and DUD-E, respectively, which also shows very good classification of molecules. An interesting observation is that AUC values increase when more features are included.

Evaluation with EF1% shows also very good performance on DUD and DUD-E. When AUC results are compared to EF1% a similar correlation of the number of features and performance can be observed for RF (see Figure S2a,b). In contrast, for LR this effect is only partly present on the DUD-E data set (see Figure S3a,b).

The predictive performance on MUV in the intra-target CV is substantially lower as on the other two data set, but still substantial values are achieved. The maximum AUC with RF is 0.76 when using 15 different features, while the best AUC with LR is 0.88 when seven features are used. In the case of MUV, combining features has no observable correlation with the performance.

When considering EF1% on MUV, the results show noticeable enrichment for both RF and LR. It is worth mentioning that for this experiment not all feature combinations were enumerated on MUV, but a backward elimination was employed, which was guided by the AUC metric.

Generally the performance with cross-target CV is lower than with intra-target CV. Still, a noteworthy separation can be achieved in this validation setup. Especially, RF achieves AUC values up to 0.78 and 0.80 on DUD and DUD-E, respectively. In contrast, when LR is utilized the best AUC values are 0.63 for DUD and 0.58 for DUD-E. The results of RF assessed with AUC and EF1% show the same correlation of the number of features with the performance as in the intra-target evaluation. This trend is also observable for most LR results, but not as prominent.

3.3. Evaluation Discussion.

3.3.1. DUD and DUD-E.

For the interpretation of the results of DUD and DUD-E it is important to consider models trained on a single feature and models trained on combinations of features separately. Since distributions of single features are matched between the classes by the compilation protocol by approximating mean and standard deviation it was expected that they contain not much information for discrimination and single feature-based predictions would be close to a random guess. However, the performance achieved with single features is far from an AUC of 0.5 with RF for most targets and also the LR performs very well. To analyze these results it is useful to examine the distributions of single features from both classes. In the Figures

DOI: 10.1021/acs.jcim.8b00712

S4 and S5 the distributions of the matched features for the DUD and DUD-E are depicted. These plots show the distributions over the whole data sets (over all targets). For both data sets, the histograms are mostly overlapping and the properties seem to be well matched, except for the molecular weight in DUD-E for which proportionally more actives are present from a molecular weight of approximately 500 Da. Consequently, it would be expected that molecular weight can be used to discriminate classes on DUD-E.

To explain the high performance with single features, we plotted the feature distribution of the target with the highest AUC in intra-target CV on DUD with RF. It can be seen at the distribution of LogP on the target PNP in Figure S6 (first row) that features are not as well matched as over the whole data set. In this example, a notable subset of actives has lower or higher LogP values than the decoy set. For further examination of results, we plotted the LogP histograms of the training and test sets of a single fold from intra-target CV in the second and third row of Figure S6. The fourth row of the plot shows the distribution of scores of the LR model in an one-dimensional contour plot over a range of LogP values. Test actives are marked by red triangles and test decoys by blue triangles. As can be seen from the blue color of the contour plot all test samples get scores near zero. This could be expected from the histograms since there is no LogP threshold, which could adequately separate the classes. In contrast, RF achieves almost perfect test performance with an AUC of 0.99 on this split. This can be explained with the last row of Figure S6, which shows the contour plot of scores from the RF model. There are many intervals with different scores, which was expected. It can be seen from the triangles marking the test actives that for each test active there exists an interval that corresponds to a higher than zero score. However, very low scores are predicted for almost all decoys of the test set. Therefore test actives get higher scores than the decoys, which is sufficient to classify them correctly. The performance of RF can be explained by the fact that the LogP values are matched in certain intervals only. Since predictions of RF are based on splitting the input space into intervals this unbiasing is ineffective, when the test set matches the intervals learned from the training set. In conclusion, the LogP values of actives and decoys in this example are well enough matched for removing bias for linear models, but not for a nonlinear model as RF. However, this does not explain the still very high AUC values achieved by LR models in Figure S1 when a single feature is used.

The high AUC values of LR model's when only a single feature is used can be explained by looking at Figure S7. In this example, the intra-target CV experiment on the target SAHH of DUD-E with the number of hydrogen bond acceptors as feature is shown. This feature is not well matched between classes. For this target, on average actives contain more acceptors than decoys. This distribution is still present after the random split into training and test set. For this reason, the LR model can learn a well separating threshold as depicted in the contour plot in Figure S7. For this example, LR and RF have the same performance in AUC because most molecules are easy to classify.

As mentioned before, an interesting observation is the correlation of the number of features and the performance, which increases until AUC values close to 1.0 are achieved for most experiments. This is not surprising since ML can capture synergies between subsets of features. However, separability is extremely high. This is probably because in the unbiasing one-dimensional feature spaces are matched, but not the multi-dimensional spaces. When considering a single target high performance with multiple features is nothing that needs to be avoided at any cost, because activity also depends on nonadditive molecular features.[39] However, when perfect performance is already reached with molecule features alone, a structure-based method might be strongly biased because it uses both the molecule and the protein. Therefore, in our opinion for DUD and DUD-E multidimensional unbiasing is necessary, if they are used for benchmarking of ML methods.

Interestingly, the same correlation can also be observed in the cross-target setup. This CV should be less prone to bias coming from molecular similar actives for example originating from the same molecular series. Correspondingly, the AUC values are lower as in the intra-target CV. However, the resulting AUCs are still reaching values of 0.78 and 0.80 for DUD and DUD-E, respectively, even though all predictions are based on molecular features only. An explanation for this is that we performed a random cross-target CV and the test and training actives of different targets might be similar because the targets could be related. This is further evaluated in Section 4 on the more stringent leave-one-out cross-target CV on DUD.

3.3.2. MUV. In contrast to DUD and DUD-E, the AUC values achieved for MUV are substantially lower (see Figure 2c). An explanation for this is that the embedding of actives among inactives in the 17 dimensional feature space with methods from spatial statistic removes more bias from the data set than approximating mean and standard deviation of single features of actives and inactives independently. This also explains that no synergistic effect by combining features is observed. In comparison with DUD and DUD-E, the unbiasing protocol of MUV seems to be more suited for ML methods, but still there are serious limitations. For example, the linear LR reaches higher maximum AUC values than RF. This is probably explained through overfitting of the RF models, which we did not investigated further. Interestingly, Wallach et al. showed for the ECFP that the suitability of MUV for ML is restricted by the MUV bias function, which considers the relation between actives to actives and actives to inactives in the feature space, but not inactives to actives and inactives to other inactives.[37] They proposed their own bias function called AVE[37] to overcome these limitations. However, it is comprehensible that MUV is not a perfect fit for ML because it was designed for similarity search starting from a query active rather than a training set of active and inactive molecules.

3.3.3. AVE Analysis. In an additional experiment, we evaluated the AVE-bias score proposed by Wallach et al.[37] on the intra-target CV experiments shown in Figures 2 and S1−S3. AVE is an extension of the MUV scoring function and assess the redundancy between training and test set. The results are shown in Figures S8−S10. AVE values different from zero indicate bias. For DUD, there is a notable correlation of AUC with AVE using RF but much less in LR and EF1% experiments. AVE bias in these experiment is not surprising because there was already substantial MUV bias reported for DUD.[4] For DUD-E, there is a notable correlation between AVE and AUC in RF and LR experiments and moderate correlation with EF1%. The correlation on MUV is weak for RF and moderate for LR experiments. Furthermore, the points in the plots are colored according to the number of features used (corresponding to the number of features on the *x*-axis in Figure 2). For DUD and DUD-E, it is observable that when more features are included also more AVE bias is exhibited. However, with most single features the performance is very high, but no AVE bias is present. In conclusion, there is a noteworthy correlation in DUD and DUD-E between performance and AVE, but in a substantial part of the experiments AVE can not explain the high performance.

3.3.4. Conclusion. In summary, DUD and DUD-E have been developed to evaluate conventional scoring functions for docking. With this goal, the unbiasing technique employed tries to approximate the distribution of single low dimensional features between actives and inactives. As our results demonstrate for a nonlinear ML method like RF it is not sufficient to select decoys to match mean and standard deviation of these low dimensional features. RF is able to accurately separate classes even on the basis of a single feature and especially when multiple low dimensional features are combined. Even LR is able to achieve impressive results, because for some targets features are not matched well enough. Moreover, experiments on MUV show also substantial performance when unbiased features are used with RF and LR, which is in accordance with others.[37]

Therefore, we can conclude that when the method is exchanged the unbiasing of the data set might be ineffective and molecular features alone might still play a major role in the validation of structure-based approaches, when ML is used. As a consequence, when using these data sets with ML methods, it should be kept in mind that low dimensional features are sufficient to achieve a notable performance in separating classes in the intra-target CV. Also a noteworthy separation can be achieved in the cross-target CV where predictions for targets dissimilar to the targets in the training set are made. These baseline performances should be considered when more complex descriptors are used which include those features. Finally, the performance should be compared to those baselines instead of random performance.

4. NON-CAUSAL BIAS IN LITERATURE

In the following, two examples of convolutional neural networks (CNN) for the scoring of protein–ligand complexes after docking are presented from literature. We made an observation that indicates that even after elaborate and conscientious validation by the authors bias might be a problem. In the subsequent examples, we first introduce the networks and their validation procedures from literature. Then the experiments we conducted to examine our observations are described.

4.1. Example 1: DeepVS. DeepVS is a CNN inspired by natural language processing.[14] The structure-based approach uses a novel descriptor to featurize and vectorize 3D protein-molecule complexes after docking. The aim is to learn general protein–ligand interactions from basic features in the local neighborhood of atoms of the small molecule in the 3D complex, called atom contexts.[14]

The DeepVS descriptor is depicted in Figure 3. For each atom a in the ligand molecule, the local neighborhood is considered. The neighborhood is described by the distances, atom types, atomic partial charges, and associated protein residues of the k_c atoms in the ligand molecule nearest to a and the k_p atoms in the protein nearest to a. For example, in Figure 3, k_c is three and includes N3 (atom a itself), H, and C2, as indicated by the red circle. In the same example, k_p is 2 and includes the protein atoms CD and OE. These discrete values are transformed into real-valued vectors, which constitute the first hidden layer of the network. The network consists of the first hidden layer, a convolutional layer (second hidden layer), a third hidden layer and an output layer with a softmax classifier.[14] In the final DeepVS network, the hyperparameters k_c and k_p have been set to 6 and 2, respectively. The training was performed on minibatches of size 20 with stochastic gradient descent (SGD), negative log-likelihood as loss function and backpropagation.[14]

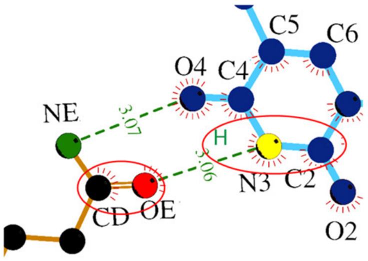


Figure 3. This figure, taken from Pereira et al.,[14] illustrates the DeepVS descriptor on the example of the local atom context of atom N3 (yellow) of thymidine in complex with a thymidine kinase (PDB-ID:1KIM). The parameters k_c = 3 and k_p = 2 are indicated by the two large red circles.

DeepVS has been validated on the DUD data set in a leave-one-out cross-target CV (LOO−CV) with a similarity filter as illustrated in Figure 1. In each iteration of the LOO−CV, one of the n = 40 DUD targets is left out as a test set. From the remaining 39 proteins, all that are similar to the selected one are discarded and the rest form the training set. Similarity of proteins has been described as sharing the same protein class or showing a positive cross-target enrichment in the original DUD paper.[5] Each trained model has been used to make predictions for the respective test protein. The validation revealed a mean AUC of 0.81 and an enrichment factor at 2% (EF2%) of 6.62 in the LOO−CV, which outperformed several other scoring functions.[14]

One experiment in the validation of DeepVS raised our attention. It is reported by the authors that setting the parameter k_p to 0 yields an AUC of 0.80 and EF2% of 6.95 (Table 6 in Pereira et al.[14]).[14] By setting k_p = 0, all explicit protein information is removed from the descriptor. The remaining descriptor considers only small molecule atoms, but the resulting AUC drops only by 0.01 and the EF2% marginally increases by 0.33, therefore, it seems DeepVS is invariant to the information from protein atoms. The descriptor with k_p = 0 corresponds roughly to a ligand-based approach. With the exception of the molecule conformation which has been generated through docking, no information from the protein is contained in the descriptor. Ligand-based approaches are based on finding active molecules due to their similarity to known ligands. Since targets similar to the test target are removed before training, it is not expected that any ligands of the training targets are structurally similar to the test ligands. Therefore, the predictability based on ligand similarity should be low. However, the achieved prediction performance is almost unchanged, which indicates noncausal bias.

4.2. Evaluating DeepVS. To understand these results, we reimplemented the DeepVS network in TensorFlow and performed the same validation experiment with altered input data. Since the input to the original DeepVS was docking output, the small molecules have a binding site specific conformation. To remove this protein-dependent information from the experiment, we generated small molecules con-

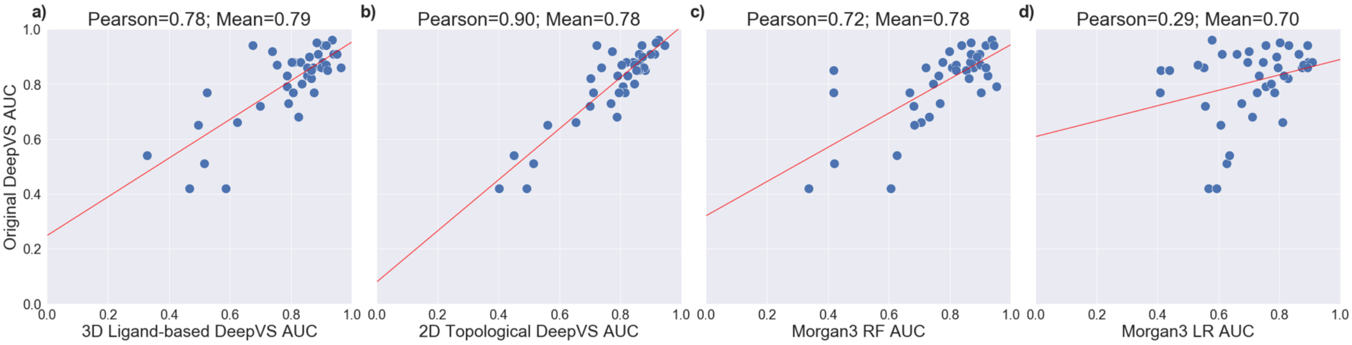


Figure 4. Correlation plots of AUC values of the structure-based original DeepVS (values taken from ref 14) and the four other approaches. (a) Performance of our 3D ligand-based reimplementation. (b) Correlation of the reimplementation using the topological distance on the molecular graph instead of 3D distances. Finally, the performance of (c) RF and (d) LR with Morgan3 fingerprints is plotted against the original DeepVS.

formations with RDKit, not considering the individual protein structure. This yielded a purely ligand-based descriptor.

The results of our ligand-based reimplementation are plotted against the reported folds of the original DeepVS in Figures 4 and S11. The achieved mean AUC of the reimplementation is 0.79, as shown in Figure 4a and the mean EF2% is 5.03 (see Figure S11a). ROC curves are provided in Figure S12. On the basis of these results, it seems that the overall performance of the DeepVS approach is in fact mostly invariant to protein information. Although, for both scores the ligand-based results correlate with the original structure-based DeepVS, it can be observed that the results in AUC are more similar than the achieved EF2% values. For single folds the simple molecule conformations of the ligand-based version suffice to achieve higher early enrichment while for other folds using docking poses and protein information is advantageous. However, the high performance of the ligand-based version in the LOO−CV implies noncausal bias.

Since small molecules only are sufficient to discriminate actives from inactives across dissimilar targets, there must be a discriminative noncausal molecular property across the whole DUD data set. To ensure that decoys are not actually active molecules, a certain minimum topological distance from every decoy to every ligand in the whole data set was required when the data set was designed.[5] For this reason, 2D molecular features should be discriminatory across all DUD targets.

By examining the DeepVS descriptor, it becomes obvious that the descriptor is able to capture 2D topology. In Figure 3, the red circle on the right containing the small molecule atoms shows the local atom context of N3 in 3D space. Indeed, the red circle also marks the local substructure around atom N3, because the nearest atoms in 3D space often correspond to the nearest atoms in the molecular graph.

To compare the local neighborhood of an atom in 3D space with the topological neighborhood, we conducted another experiment. Using our ligand-based reimplementation, instead of considering the k_c nearest atoms in 3D space we considered the k_c = 6 nearest atoms in the molecule graph, yielding a topological version of the ligand-based DeepVS. The resulting mean AUC and EF2% over all 40 folds of the topological DeepVS is 0.78 and 5.41, respectively (see Figures 4b and S11b). ROC curves for the 40 folds are shown in Figure S13. A comparison of AUC and EF2% values between our 3D and 2D reimplementations is shown in Figure S14, which indicates a strong correlation in terms of AUC and a good correlation for EF2% values. To further evaluate whether the 3D descriptor captures the 2D information, we examined if the same actives were enriched by both methods. In the first 5% of the ranked lists, 81% of predicted actives are identical between both methods over all 40-folds with a standard deviation of 18%. A full distribution of the active identity is depicted in Figure S15a. These experiments demonstrate that similar results can be achieved when using exclusively 2D information.

Another open question is to determine to which extent the usage of a CNN contributes to the performance and how standard ML approaches would perform. To examine the baseline performance of a 2D descriptor, we applied Morgan3 fingerprints folded to 4096 bits with RF and LR in the same validation setup as used for DeepVS. The resulting mean AUC is 0.78 for RF and 0.70 for LR (see Figure 4c,d), while the resulting mean EF2% are 5.52 and 5.47 for RF and LR, respectively (see Figure S11c,d). On the one hand, these results show that also other nonlinear ML methods such as RF are sufficient to achieve a comparable performance and using a CNN does not improve the prediction results significantly. On the other hand, when using a simple linear LR model the mean AUC is already quite high with 0.70, which shows that correct classification can be achieved with a linear function for many test cases. As in the previouse case, we compare the hit lists to see whether the same actives are enriched by conventional ML methods as with the CNN. At the first 5% of the hit lists on average 46% (28% standard deviation) of actives are identical (see Figure S15b for details). Therefore for most targets (folds) there is a difference between the information captured and used by RF with Morgan3 fingerprints and the 3D DeepVS model but still a comparable performance can be achieved and ligand features suffice to make predictions across dissimilar targets.

4.3. Example 2: Grid-Based 3D CNN.

For our second example, we wanted to examine a currently frequent approach of ML-based scoring functions using a CNN inspired by image recognition. In these approaches the 3D complex of the protein and ligand is discretized with a grid around the binding site. There are multiple examples, such as AtomNet,[13] an unnamed network by Ragoza et al.[15] and K_{DEEP}.[17] As in the case of image recognition, the intention of these networks is to automatically and hierarchically abstract high-level features holding information for binding from low-level features of the complex. Because of the workload required to rebuild and re-evaluate these networks, we decided to examine the CNN by Ragoza et al. only because the authors provided an elaborate validation from different perspectives and shared their code.

Ragoza et al. discretize the protein−ligand complex with a uniform 3D grid of 24 Å centered around the binding site. The

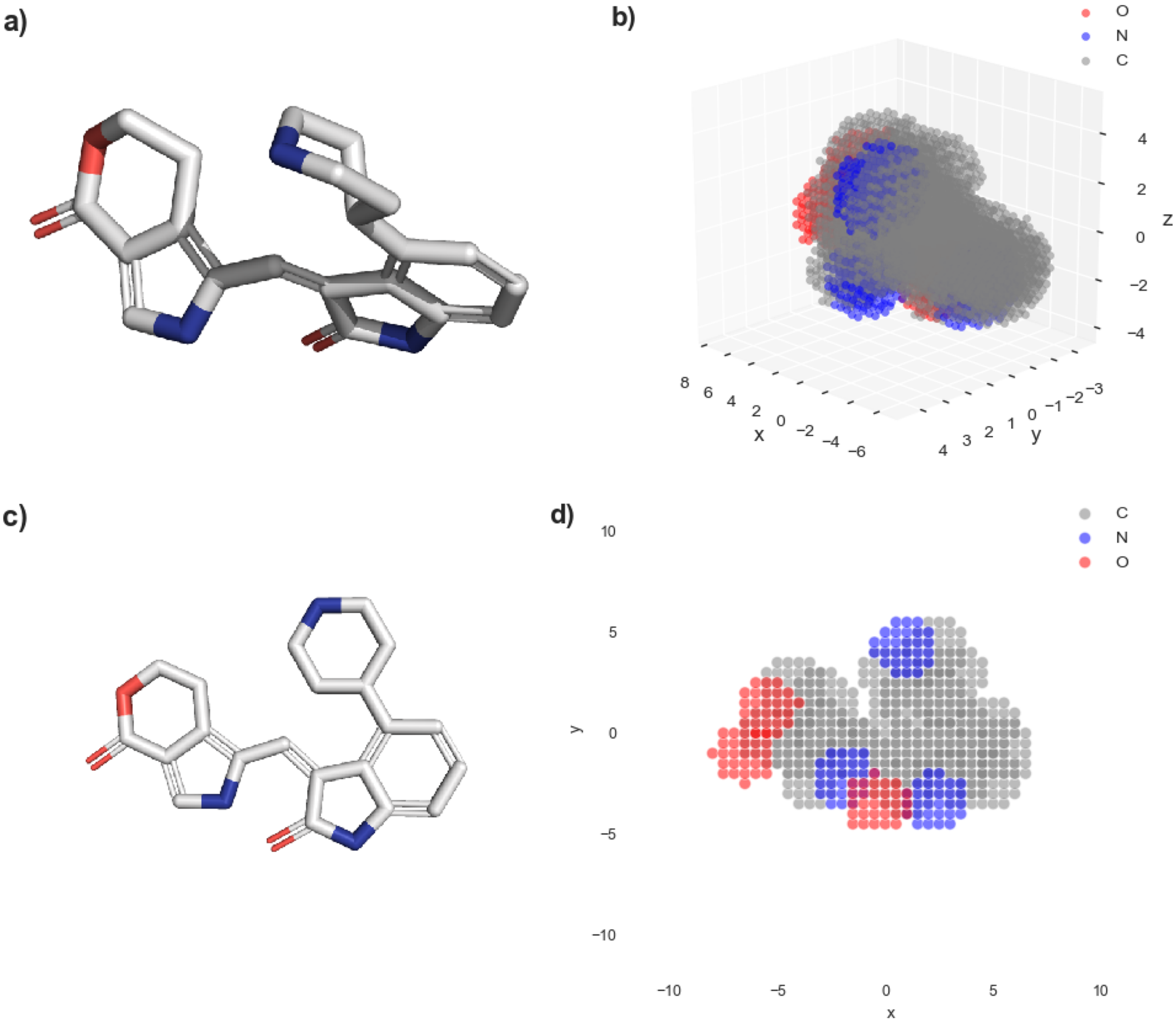


Figure 5. Illustration of the two variants of the grid descriptor on the example of compound CHEMBL58224. The first row depicts a (a) 3D structure of the molecule and (b) corresponding representation of the 3D grid descriptor. In the second row, the (c) 2D conformation of the molecule and (d) according 2D grid representation are illustrated. In panels b and d, the coloring of the boolean-valued grid points is overlaid for the illustration of the channels of the grid.

resolution is set to 0.5 Å. A grid point stores information about heavy atom types represented by a continuous density function depending on the distance of an atom to the grid point and the respective van der Waals radius. Each channel of the grid resembles a different atom type (like RGB channels in images). Atom types for protein atoms and ligand atoms are considered separately in different channels. In total 34 distinct atom types are considered. These include simple elements of atoms, such as nitrogen, oxygen or sulfur. In addition aliphatic and aromatic properties of carbons are considered as well as nitrogen and oxygen atoms acting as hydrogen bond acceptors or donors.[15]

The architecture of the CNN consists of three subsequent pooling and convolutional layers and a final output layer. The input to the network is the featurized 3D grid of the binding pocket. The final scores are provided with the softmax function.[15] See Ragoza et al.[15] for a more detailed description of the architecture. Ragoza et al. performed the training of the network with SGD and backpropagation while minimizing the multinomial logistic loss. Oversampled batches of size 10 have been utilized such that each batch is balanced according to the number of actives and inactives. In addition, training data have been augmented by random rotation and translation.[15]

The described CNN architecture has been evaluated comprehensively for the tasks of pose prediction and VS on CSAR, DUD-E and independent test sets.[15] CSAR and DUD-E have been used in cross-target three-fold cluster cross validations (cCCV). For each of the two data sets, all targets were clustered into three-folds to ensure that targets with sequence identity greater 90% for CSAR and 80% for DUD-E are in the same fold. This should prevent the training targets from being too similar to test targets.[15]

The achieved results for pose prediction and VS differ. Pose prediction with cCCV on CSAR yielded a mean AUC of 0.815 outperforming AutoDock Vina (Vina). In contrast, in an intra-target validation, Vina outperformed the CNN or is almost as good. For the VS task, the best reported mean AUC with the CNN in the cCCV on DUD-E is 0.86. On the test, Vina achieved an AUC of only 0.68. Interestingly, Ragoza et al. showed that a model trained with DUD-E performed rather poor on CSAR test data and vice versa. The authors evaluated models trained on combinations of DUD-E and CSAR data. The combined model exhibits an AUC of 0.83 on the VS task and 0.79 at the pose prediction task, showing slightly worse performance as the networks trained for a single task. For this reason Ragoza et al. conclude that the results demonstrate that the data sets generated for different tasks prevent models from learning a similar scoring function.[15]

4.4. Evaluating Grid-Based 3D CNN. As in the first example from literature, a CNN was trained for the scoring of docking output. The evaluation was performed on the DUD-E data set, which is compiled with a comparable strategy as the DUD data set. Interestingly, Ragoza et al. noticed that learning with data sets for different tasks leads to models with differing scoring functions.[15] A reason for this might be bias learned from the DUD-E data set. Ragoza et al. already suspected overly optimiztic predictions in the case of their DUD-E experiments, which should be mitigated through cCCV.[15] The intention in utilizing the cCCV was that the targets used for training are not similar to the test targets. As in the case of DeepVS, the performance of the CNN could be due to small molecule information only.

To examine this possibility, we reimplemented the described CNN in Keras with some adaptions to the descriptor and replicated the cCCV experiments described by Ragoza et al.[15]. First, as in the reimplementation of DeepVS, protein information was left out completely. No docking was performed. Instead, conformations of the small molecules in DUD-E were generated using RDKit. The molecules are then put into a 3D grid of 48^3 grid points with a spacing of 0.5 Å as the original descriptor. The second adaptation was that for the reimplemented descriptor not all 34 atom types were used. For computational efficiency only the elements bromine, carbon, chlorine, fluorine, iodine, nitrogen, oxygen, phosphorus and sulfur were used. This reduced the channels of each grid from 34 to 9. The third adaptation made to the descriptor was that no density function representing the atoms was utilized. Instead, a boolean function was applied, which sets a grid point to 1 if this grid point is in the van der Waals radius of an atom. This adapted descriptor is a simpler version, capturing only a subset of the features of the original descriptor. The descriptor is illustrated in Figure 5. The 3D conformation of the molecule (Figure 5a) is discretized by the grid descriptor. The descriptor for the depicted conformation is shown in Figure 5b, where the different channels (O, N, and C) are overlaid.

With the modified descriptor, the reimplemented CNN was trained and evaluated. The three-folds for the cCCV on DUD-E were generated using the Python script[55] provided by Ragoza et al.[15] The resulting AUC values achieved with the simpler descriptor and the reimplemented CNN are 0.82, 0.84, and 0.85 for the three-folds, which results in a mean of 0.84. Therefore, the difference to the original CNN model trained on DUD-E is 0.02, which is very similar to the performance of the CNN of the original publication using docking. In terms of early enrichment EF2% values of 12.60, 14.87, and 13.98 were achieved for the three folds. ROC curves are shown in Figure S16. The experiment here is ligand-based; therefore, predictions across dissimilar targets should not be possible on the basis of ligand similarity. The still high AUC values strongly indicate that a similar bias is learned as in the example of DeepVS. As with DUD, it is possible for DUD-E to learn activity prediction across targets with small molecule information only.

As in the case of DUD, molecular topology is a discriminative feature across the DUD-E data set.[6] To examine if this is the reason for the ligand-based results for this CNN, we conducted a 2D version of the original 3D experiment using exclusively a 2D description of the molecules. Instead of generating 3D molecule conformations, we generated only 2D depictions (see Figure 5c) using a constant atom radius of 1.5 Å for all atoms and calculated their respective 2D grid representation as shown in Figure 5d. The 2D grid builds the input to a 2D version of the CNN. We performed the same cCCV and achieved AUC values of 0.82, 0.85, and 0.84 for the three-folds, yielding a mean of 0.84. The achieved EF2% values are 11.70, 15.41, and 11.31. ROC curves are shown in Figure S17. To compare whether the same actives were enriched with the 3D and 2D version of the CNN, we compared the hit lists from both methods regarding the actives. On average, 73% (1% standard deviation) of the actives are identical at the first 5% of the hitlists. The full distribution of identical actives can be found in Figure S18. These results indicate that a substantial part of the performance of the original grid-based CNN is, such as in the case of DeepVS, based on the difference across all DUD-E targets of lower dimensional molecular features included in the molecules' topology.

5. TOWARD BIAS-CONTROLLED VALIDATION

A comprehensive and elaborated validation for a VS method is the basis for reliability, acceptance, and usage of this method in the scientific community. Elaborate efforts are made to validate methods, but as we showed in the previous section on two examples from literature, it is possible to achieve comparable results when a ligand-based version of a structure-based descriptor is used. Simultaneously, the performance in those experiments is noncausal bias, because ligand similarity should not suffice to make predictions across dissimilar targets.

A reason this bias remained unnoticed is the non-transparency of the used ML models. In particular, deep learning models such as CNNs are difficult to interpret and are often treated as black boxes. This lack of interpretability increases the effort required for validation as well as the necessity for a comprehensive validation. In the two examples, elaborated validation experiments have been conducted by the authors.[14,15] Still, as we have shown, substantial noncausal bias remained unrecognized.

We demonstrated in detail for DeepVS that molecular topology is a discriminative feature in the used validation setup and is captured implicitly by the DeepVS descriptor. Our results strongly indicate the same reason for the bias influencing the performance measuring of the grid-based 3D CNN. In both cases, discriminative lower dimensional features of the small molecules were contained in an nontransparent structure-based 3D descriptor. This is difficult to spot, and an apparently reasonable model is actually learning noncausal bias, especially, when difficult-to-interpret neural networks are employed. Our experiments with standard ML methods showed that the difference in performance between RF and CNNs is very small, which makes it debatable whether benefits of CNNs outweigh the additional effort and the lack of interpretability that is associated with deep learning models. This also shows the importance of performing baseline experiments and comparing complex methods with simple and interpretable ones.

That lower dimensional features are biasing higher dimensional descriptors is an recurring problem that has already been addressed in the context of simple properties and conventional scoring functions in DUD and DUD-E as well as for similarity search in MUV. However, the increasing trend of applying ML methods in SBVS comes with a variety of novel descriptors and expressive methods that can abstract higher dimensional features from low dimensional ones. This holds the risk of learning bias, because the currently established data sets focus on different methodologies. When comparing the ML methodology on the basis of DUD and DUD-E it is evident that those data sets have not been designed with ML methods in mind. Conventional scoring functions such as empirical scoring functions are typically weighted sums over physically motivated descriptors.[41] Current ML methods operate differently. They are strongly data-driven, make use of a large number of free variables, and are able to derive nonlinear relationships. While this is an advantage in general, it leads to a higher risk of learning noncausal bias from data sets. In particular, the protocols of DUD and DUD-E assume that simple 1D properties might cause bias. However, as the results of Section 3 show, in contrast to MUV, the combination of 1D features for example by a ML method has not been considered in the unbiasing protocol of DUD and DUD-E and it is arguable if combinations of these features should have a strong influence on structure-based descriptors. In addition, the design decision to employ a 2D dissimilarity of each decoy to any active of the whole data set generates a strongly discriminating feature and could be employed because 2D features are not captured by conventional scoring functions. The authors of DUD-E therefore state that their data set is not suited for topological or 2D methods because "Through its construction, ligands light up against DUD-E decoys using these 2-D similarity methods, which create an artificially favorable enrichment bias for them.".[6] This design decision restricts the applicability of DUD and DUD-E for 2D methods, but not in general. However, what should be more emphasized is that the 2D dissimilarity is employed across all targets of the data sets, which can artificially enable the differentiation of actives and inactives on the basis of molecules from biologically unrelated targets.

In conclusion, whether bias occurs in the validation of a method depends on the composition of the data set used as well as the method and descriptor used on the data set (and other factors such as the performance metric and the strategy for splitting into train/test sets).

In our studies, we focused on three well established data sets in VS, but there are other established and also more recent data sets, which should be evaluated for their suitability with ML[8−11,56] as well before similar problems arise from different data sets.

This discloses two current problems in VS with ML. First, method developers need to evaluate the appropriateness of a benchmark data set for their method-descriptor combinations. Second, there is a need for a benchmark data set suited for ML methods in VS. To both, we provide some first guidelines to foster the discussion in the scientific community.

6. GUIDELINES FOR VALIDATION EXPERIMENTS

We suggest five guidelines for the setup of validation experiments for ML in VS, which are not necessarily restricted to these methods. This includes determining if a data set is suitable for validating a particular method and descriptor, as well as identifying implicit bias.

(i) Validation domain of a data set: If a new method is in place, a data set is selected for validation. In general, it is very good to choose established data sets since they were assembled independently from the method development. It is not sufficient, however, to just consider the application problem. The data set must be suited for the method applied. Some authors of data sets already give hints in their publications. For example, DUD and DUD-E are known to be not suitable for methods capturing topological descriptors.

(ii) Method and descriptor design: More detailed evaluations can be conducted if the method and descriptor are modular. Particularly, components of the descriptor and method should be easily controllable in their information content and ideally exchangeable. This is beneficial to evaluate different descriptors and method variations. A positive example for a modular descriptor is the DeepVS descriptor because it allows to control the extent of information from the protein and from the small molecule.

(iii) Data set's unbiasing strategy: As observed from the compilation protocols of DUD, DUD-E, and MUV, each data set applies some unbiasing techniques to reduce the predictive power of certain features. These features are considered biologically irrelevant in that they should have not much influence on the distinction of actives from inactives. Whether an unbiasing technique of a data set compilation protocol works with a certain method needs to be validated. Evaluating these unbiased features and their combinations with a particular method, as in the experiment in Section 3, is important because perceptible predictive power in this experiment indicates that the data set's unbiasing strategy and the employed method are not compatible. If the data set with the unbiased features is still used, it is necessary to compare the method's performance not against random predictions, but to the baseline performance given by the unbiased features. If, however, the unbiased features are not used explicitly or implicitly by the method, the data set can be used.

(iv) Baseline definition: Comparing a method's performance to random predictions is not enough to recognize bias or test cases which are too easy. A novel and complex method should always be compared to standard methods and simple approaches in the same validation scenario. Perhaps simple linear classifiers or nearest neighbor searches are as good as a more complex nonlinear method. The importance of weak-but-realistic baselines was also recognized by others.[37] Beyond that, we showed when a structure-based method identifies most active molecules of a single target in a test set the developer might assume that the performance achieved is due to relationships derived from protein and small molecule features. However, without performing a baseline experiment this conclusion can not be taken for granted. A simple ligand-based similarity search achieving a similar performance on the test set indicates that the test cases might be too simple. Furthermore, it is worthwhile to also perform ligand-based baseline experiments when the validation scenario would be that ligand-based methods have no predictive power. As described in Section 4, unexpected good results strongly indicate implicit and noncausal bias.

(v) Detecting noncausal bias: Our final guideline deals with evaluating a model with multiple sanity tests in the sense of negative controls to identify noncausal bias. This corresponds to our approach in Section 4, where we evaluated feature subsets of the descriptor where no causal relationship exists and expected to see no performance, for example, the descriptor can be systematically decomposed into features that are expected to be insufficient for activity prediction in the specific validation scenario. This yields a set of negative controls that the model needs to pass to validate against noncausal bias and should finally reveal which feature subsets are decisive. The first step in the validation of a structure-based method would be to remove the protein from the experiment and examine the performance without protein information. Furthermore, one can replace the target protein by a biologically unrelated protein and see whether the performance drops. The hard part in this systematic bias detection is to actually look at the descriptor and validation experiment and come up with a subset of features to evaluate.

Gabel et al.[57] also proposed two guidelines for better validation experiments for ML-based scoring functions for affinity prediction, which are complementary to ours. These guidelines describe testing the sensitivity of the scoring function to the ligand pose and the ability of the scoring function to discriminate actives from inactives in a VS task.[57]

7. ON GENERATING NEW DATA SETS

We see a need for data sets appropriate for ML methods in SBVS. A basic concept for designing a data set without ML-specific bias would be to define a bias scoring function to optimize the selection of data points for a new data set. For ligand-based VS, Wallach et al.[37] proposed a variation of the MUV function for ML called AVE. This scoring function assess bias by equating bias with a one-nearest neighbor predictor, a simple learning method. A similar concept was applied during the development of the original MUV data set, for which bias was measured by the performance in a simple descriptor space. This essentially resembles a negative control. When there is good performance in an insufficient experimental setup, then there is bias. This concept could be extended to a wide range of possible biases. For example, to remove the samples introducing the bias in the experiment of DeepVS, the performance in the LOO−CV with similarity filter could be used as a bias scoring function for optimization. Therefore, it is possible to generate specifically tailored data sets for certain methodologies.

The ideal data set would be a data set of uniformly sampled data points from the chemical space without bias for any methodology. Such a data set would enable the comparison of different methodologies without restrictions. However, the explored chemical space is not sampled uniformly. This most likely leads to a trade-off between bias reduction and the comparability of methodologies. Dissimilar methodologies might have different vulnerabilities to bias and tailored data sets for conventional scoring functions might be unsuited to ML methods. For this reason, there is a need for data sets that are suitable for both ML and conventional scoring functions.

Ultimately, an open problem is the quality and quantity of the available data. Some problems could be fixed if more experimentally validated molecules would be published, especially if inactivity will be more frequently reported to the community. However, it is unforeseeable when the level of an adequate number of data points will be reached. There are, however, ongoing efforts to further refine decoy selection protocols.[58] Finally, an important step toward better data sets would be a more open sharing of negative results, for example, from large industrial screening campaigns.[3,59]

8. CONCLUSION

This perspective draws attention to the problem of bias in current ML-based scoring functions for SBVS. We showed that bias, such as artificial enrichment, is still a problem in established data sets when the context of methodology is changed. ML functions are susceptible to unnoticed bias because they tend to be black boxes and in addition are validated on data sets that are not compiled for ML. More complex methods and descriptors can disguise decisive lower dimensional features. The current popularity of difficult-to-interpret deep learning models covers up bias even more, which can lead to models that instead of protein−molecule interactions learn only molecular features from a structure-based descriptor. To recognize this and other biases, we proposed practical guidelines, which should aid the validation process and avoid fallacious models. The guidelines (ii), (iv), and (v) encourage to design novel methods and descriptors modular to compare variations of them to multiple baselines and validate the model against different sanity tests, which enhances the understanding of the model and reveals biased predictors.

Moreover, we see the need for new data sets suited for ML, because established data sets have not been compiled with the nature of ML methods in mind. This makes current benchmark data sets unsuited for differently operating novel methods and descriptors. Therefore, the proposed guidelines (i) and (iii) also include the verification of the validation domain and unbiasing strategy of a data set because not every data set is suited for every method.

To demonstrate the issues raised, we investigated the behavior of two recently published CNN-based scoring functions. Both scoring functions are well-designed and reflect state-of-the-art methodologies. The utilized methods, that is, docking and the CNNs, do not pose a problem, but the validation experiments do pose a problem. The reported performances change little when essential protein information is removed from the descriptors, especially in terms of AUC. Therefore, the validation does not reveal much about the true practical performance on new targets and compound classes. Since experimental prospective studies can be performed only on rather limited scale, they are not appropriate for measuring method performance in general. As a consequence, it is necessary for the method developer to not only validate the model's predictive capabilities in certain applicability domains, as with intra-target validation and cross-target validation, but also to verify that the chosen validation setup is valid for the model under consideration. As part of ML-based research, it is an important challenge for the near future to come up with reasonable validation schemes. Only then will we be able to exploit the full potential of modern machine learning for drug discovery.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00712.

Boxplots of AUC and EF1% values achieved with logistic regression and random forest in evaluation of unbiasing techniques of DUD, DUD-E, and MUV; distribution of matched properties of DUD and DUD-E; analysis of model performance when using single unbiased features; AVE analysis of DUD, DUD-E, and MUV; correlation of EF2% of the original DeepVS and our reimplementations as well as random forest and logistic regression; ROC curves of our 2D and 3D reimplemations of DeepVS; AUC and EF2% correlation plots of folds of our 2D and 3D reimplementation of DeepVS; analysis of actives identity between our 2D and 3D reimplementation of DeepVS as well as random forest; ROC curves for 3D and 2D reimplementation of grid-based CNN; analysis of actives identity between 2D and 3D reimplementation of grid-based CNN (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

ORCID

Jochen Sieg: 0000-0001-5343-7255
Florian Flachsenberg: 0000-0001-7051-8719
Matthias Rarey: 0000-0002-9553-6531

Notes

The authors declare no competing financial interest.

Software availability: The software containing the reimplemented networks and descriptors as described in section 4.2 and 4.4 are available open source via github from https://github.com/rareylab/MLValidation.

ACKNOWLEDGMENTS

The authors thank D. Koes for providing additional information and scripts related to the 3D CNN approach to us. This work was orally presented at the International Conference on Chemical Structures (ICCS), Noordwijkerhout in June 2018. We are grateful for all the comments and suggestions we received from numerous participants of the conference.

REFERENCES

(1) *Virtual Screening in Drug Discovery*; Alvarez, J., Shoichet, B., Eds.; CRC Press, 2005; Vol. *1*.

(2) Special Issue: A snapshot in time: Docking Challenge. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 675−799.

(3) Carlson, H. A.; Smith, R. D.; Damm-Ganamet, K. L.; Stuckey, J. A.; Ahmed, A.; Convery, M. A.; Somers, D. O.; Kranz, M.; Elkins, P. A.; Cui, G.; Peishoff, C. E.; Lambert, M. H.; Dunbar, J. B. CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *J. Chem. Inf. Model.* **2016**, *56*, 1063−1077.

(4) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49*, 169−184.

(5) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(6) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.

(7) Vogel, S. M.; Bauer, M. R.; Boeckler, F. M. DEKOIS: Demanding Evaluation Kits for Objective in Silico Screening - A Versatile Tool for Benchmarking Docking Programs and Scoring Functions. *J. Chem. Inf. Model.* **2011**, *51*, 2650−2665.

(8) Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 - A Public Library of Challenging Docking Benchmark Sets. *J. Chem. Inf. Model.* **2013**, *53*, 1447−1462.

(9) Dunbar, J. B.; Smith, R. D.; Yang, C.-Y.; Ung, P. M.-U.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Selection of the Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2036−2046.

(10) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **2017**, *50*, 302−309.

(11) Wallach, I.; Lilien, R. Virtual Decoy Sets for Molecular Docking Benchmarks. *J. Chem. Inf. Model.* **2011**, *51*, 196−202.

(12) Ballester, P. J.; Mitchell, J. B. O. A machine learning approach to predicting proteinligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.

(13) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *arXiv preprint arXiv:1510.02855*, **2015**.

(14) Pereira, J. C.; Caffarena, E. R.; dos Santos, C. N. Boosting Docking-Based Virtual Screening with Deep Learning. *J. Chem. Inf. Model.* **2016**, *56*, 2495−2506.

(15) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−957.

(16) Wójcikowski, M.; Ballester, P. J.; Siedlecki, P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.* **2017**, *7*, 46710.

(17) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287−296.

(18) Polishchuk, P. Interpretation of Quantitative Structure-Activity Relationship Models: Past, Present, and Future. *J. Chem. Inf. Model.* **2017**, *57*, 2618−2639.

(19) Chuang, K. V.; Keiser, M. J. Comment on "Predicting reaction performance in C-N cross-coupling using machine learning". *Science* **2018**, *362*, eaat8603.

(20) Réau, M.; Langenfeld, F.; Zagury, J.-F.; Lagarde, N.; Montes, M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Front. Pharmacol.* **2018**, *9*, 11.

(21) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(22) McGovern, S. L.; Shoichet, B. K. Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes. *J. Med. Chem.* **2003**, *46*, 2895−2907.

(23) Diller, D. J.; Li, R. Kinases, Homology Models, and High Throughput Docking. *J. Med. Chem.* **2003**, *46*, 4638−4647.

(24) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.

(25) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(26) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility. *J. Chem. Inf. Model.* **1994**, *34*, 109−116.

(27) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169−178.

(28) Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Moretti, L.; Cooper, R. I.; Richards, W. G. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 789−801.

(29) Mysinger, M. M.; Shoichet, B. K. Rapid Context-Dependent Ligand Desolvation in Molecular Docking. *J. Chem. Inf. Model.* **2010**, *50*, 1561−1573.

(30) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani,

B.; Overington, J. P. ChEMBL: a largescale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(31) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(32) Xia, J.; Tilahun, E. L.; Reid, T.-E.; Zhang, L.; Wang, X. S. Benchmarking methods and data sets for ligand enrichment assessment in virtual screening. *Methods* **2015**, *71*, 146−157.

(33) Wheeler, D. L.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2007**, *36*, D13−21.

(34) *Web of Knowledge*; Clarivate, 2018. www.webofknowledge.com (accessed Sept 7, 2018).

(35) Cleves, A. E.; Jain, A. N. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 147−159.

(36) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.

(37) Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **2018**, *58*, 916−932.

(38) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.

(39) Jain, A. N.; Cleves, A. E. Does your model weigh the same as a Duck? *J. Comput.-Aided Mol. Des.* **2012**, *26*, 57−67.

(40) Shoichet, B. K.; Stroud, R. M.; Santi, D. V.; Kuntz, I. D.; Perry, K. M. Structure-based discovery of inhibitors of thymidylate synthase. *Science* **1993**, *259*, 1445−1450.

(41) Liu, J.; Wang, R. Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, *55*, 475−482.

(42) Bietz, S.; Schomburg, K. T.; Hilbig, M.; Rarey, M. Discriminative Chemical Patterns: Automatic and Interactive Design. *J. Chem. Inf. Model.* **2015**, *55*, 1535−1546.

(43) *RDKit: Open-source cheminformatics*. Version: 2017.09.3.0; RDKit, 2018. http://www.rdkit.org (accessed Jan 21, 2018).

(44) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(45) Chollet, F. *Keras*. Version: 2.1.5; Keras, 2015. https://keras.io (accessed Mar 19, 2018).

(46) Abadi, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Version: 1.7.0; TensorFlow, 2015; https://www.tensorflow.org/ (accessed Apr 22, 2018).

(47) *Directory of Useful Decoys website - Partial Charges for DUD Molecules recalculated by Inhibox. AM1*; Docking.org, 2017 (accessed May 12, 2017).

(48) *Directory of Useful Decoys Enhanced website*; Docking.org, 2017 (accessed May 21, 2017).

(49) *Maximum Unbiased Validation (MUV) Data Sets website*; Technische Universitat Braunschweig, 2017. https://www.tu-braunschweig.de/pharmchem/forschung/baumann/muv (accessed Jun 5, 2017).

(50) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861−874.

(51) Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligandbased virtual screening. *J. Cheminf.* **2013**, *5*, 26 DOI: 10.1186/1758-2946-5-26.

(52) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157−1182.

(53) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(54) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(55) *Clustering Script*; GitHub, 2018. https://github.com/gnina/scripts/blob/9cf8892010873b672f370a122e32aa8bc496a5e1/clustering.py (accessed May 8, 2018).

(56) Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.

(57) Gabel, J.; Desaphy, J.; Rognan, D. Beware of Machine Learning-Based Scoring Functions-On the Danger of Developing Black Boxes. *J. Chem. Inf. Model.* **2014**, *54*, 2807−2815.

(58) Xia, J.; Reid, T.-E.; Wu, S.; Zhang, L.; Wang, X. S. Maximal Unbiased Benchmarking Data Sets for Human Chemokine Receptors and Comparative Analysis. *J. Chem. Inf. Model.* **2018**, *58*, 1104−1120.

(59) Gaieb, Z.; Liu, S.; Gathiaka, S.; Chiu, M.; Yang, H.; Shao, C.; Feher, V. A.; Walters, W. P.; Kuhn, B.; Rudolph, M. G.; Burley, S. K.; Gilson, M. K.; Amaro, R. E. D3R Grand Challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1−20.