

RESEARCH

Open Access



# Exploring the ability of machine learning-based virtual screening models to identify the functional groups responsible for binding

Thomas E. Hadfield<sup>1</sup>, Jack Scantlebury<sup>1</sup> and Charlotte M. Deane<sup>1\*</sup>

## Abstract

Many recently proposed structure-based virtual screening models appear to be able to accurately distinguish high affinity binders from non-binders. However, several recent studies have shown that they often do so by exploiting ligand-specific biases in the dataset, rather than identifying favourable intermolecular interactions in the input protein-ligand complex. In this work we propose a novel approach for assessing the extent to which machine learning-based virtual screening models are able to identify the functional groups responsible for binding. To sidestep the difficulty in establishing the ground truth importance of each atom of a large scale set of protein-ligand complexes, we propose a protocol for generating synthetic data. Each ligand in the dataset is surrounded by a randomly sampled point cloud of pharmacophores, and the label assigned to the synthetic protein-ligand complex is determined by a 3-dimensional deterministic binding rule. This allows us to precisely quantify the ground truth importance of each atom and compare it to the model generated attributions. Using our generated datasets, we demonstrate that a recently proposed deep learning-based virtual screening model, PointVS, identified the most important functional groups with 39% more efficiency than a fingerprint-based random forest, suggesting that it would generalise more effectively to new examples. In addition, we found that ligand-specific biases, such as those present in widely used virtual screening datasets, substantially impaired the ability of all ML models to identify the most important functional groups. We have made our synthetic data generation framework available to facilitate the benchmarking of new virtual screening models. Code is available at <https://github.com/tomhadfield95/synthVS>.

**Keywords** Structure-based virtual screening, Machine learning, Interpretability

## Introduction

The drug discovery process is difficult and time consuming, with a recent study finding that the median time to develop a new drug was 8.3 years, at an average cost of \$985 million [1]. There is therefore a need to develop

novel techniques which can help design medicines more quickly and cheaply.

Fueled by their successes in a broad range of domains (e.g. [2–4]), there has been substantial recent interest in the development of machine learning (ML) algorithms to help accelerate the drug discovery process; recent examples include highly accurate protein structure prediction algorithms [4–6], generative models which propose target-specific libraries for compound design [7–11] and tools for automated synthesis planning for synthetically challenging molecules [12–14].

\*Correspondence:  
Charlotte M. Deane  
deane@stats.ox.ac.uk

<sup>1</sup> Oxford Protein Informatics Group, Department of Statistics, University of Oxford, Oxford, UK



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

The long-term hope for these algorithms is that it will one day be possible to build an end-to-end *in-silico* drug-discovery platform which, following the identification of a target, could produce a series of high affinity binders without human involvement. A key component of such a platform would be the ability to computationally predict whether a given ligand is likely to bind to the target with high affinity, as this would allow the prioritisation of promising compounds from a large library of molecules.

Traditional virtual screening models (e.g. [15, 16]) estimate the binding affinity of a protein-ligand complex as a weighted sum of physics-based terms (e.g. van der Waals contributions, hydrogen bond scores etc.), with solved protein-ligand complexes used to estimate the model weights. These methods are dependent on a set of hand-crafted features being fed into a model; subsequent approaches have sidestepped this requirement by converting an input molecule to a fingerprint representation and using machine learning algorithms to automatically approximate non-linear relationships between features, such as random forests [17] or neural networks [18].

Inspired by the remarkable ability of deep learning models to capture important spatial information in other fields, the most recent set of virtual screening models (e.g. [19–21]) have tended to use deep-learning architectures, representing the protein-ligand complex either as a graph or as a 3D image. It is hypothesized that such models are better able to identify important protein-ligand interactions and will be more generalizable to novel targets than fingerprint-based models or models which depend on a set of hand-picked features.

Despite their ability to capture important spatial information, recent studies [20, 21] have shown that fingerprint-based models (e.g. [18]) illustrated comparable or better predictive performance than deep learning models, calling into question whether deep learning models are actually able to use their learned representations to identify important intermolecular interactions. Moreover, a study by Chen et al. [22] raised concerns surrounding the susceptibility of deep learning-based models to ligand-specific biases, demonstrating that removing all protein information from a deep learning-based virtual screening model did not degrade its performance. Their findings indicated that the model was not capturing important spatial information but rather learning to classify examples based on ligand-specific biases. In a recent study by Volkov et al. [23], the authors trained a series of message passing neural networks with a variety of different inputs. They found that whilst the inclusion of protein-specific information aided the model's predictive performance compared to a model trained solely on ligand-specific features, the explicit featurisation of protein-ligand interactions did not improve performance

compared to the model trained on ligand-specific and protein-specific features. The authors argued that this suggested that the model was unable to learn to identify the underlying biophysical interactions responsible for binding and instead learned to classify examples based on distributional differences in the training data.

An approach to incentivise models to learn to use intermolecular interactions was described by Scantlebury et al. [24]. They proposed a data augmentation strategy where additional decoy examples were derived by taking an active protein-ligand complex from the training set and randomly rotating and translating the ligand. The augmentation procedure forced the convolutional neural network to use the protein-specific information when making decisions, illustrated by the degraded performance of the model when protein-specific information was removed.

To investigate the extent to which ML algorithms are able to accurately assign importance to individual atoms when making a prediction, several recent works [25–27] have proposed synthetic datasets which labelled ligands as 'active' if they contained a pre-defined molecular substructure. Using an attribution technique, such as Integrated Gradients [28], it is then possible to compare the model-assigned atom importances to the ground truth atom labels to assess whether the atoms which comprised the pre-defined substructure were identified as the most important atoms. While these studies provided valuable insights into the ability of ML algorithms to identify important functional groups, they did not assess whether the ML algorithms were able to capture important spatial information or identify intermolecular interactions.

Whilst several authors have used attribution techniques on real-world data to uncover important functional groups [24, 29, 30], it is often difficult to ascertain the precise contribution of each atom in an experimentally obtained protein-ligand complex. Combined with the difficulty in manually curating a large-scale test set, it is currently infeasible to objectively assess the attribution performance of ML algorithms on real-world virtual screening tasks.

To address this, we propose a protocol for generating a synthetic dataset which mimics the mechanics of protein-ligand binding. Each ligand in the dataset is surrounded by a randomly sampled point cloud of pharmacophores, "synthetic residues", which together comprise a "synthetic protein". Just as in real-world protein-ligand binding, where for a ligand to bind to a protein it must interact with a number of pharmacophores in the binding site, we define two simple deterministic binding rules, which decide whether a ligand "binds" to a synthetic protein based upon the relative position of ligand pharmacophores and synthetic residues with complementary types.

As the label of each example is determined by a known deterministic binding rule, we can precisely specify which functional groups, if any, in the ligand are responsible for binding. Although our deterministic binding rule is considerably simpler than the mechanics of real-world protein-ligand recognition, it allows us to assess whether an ML algorithm is able to capture important spatial information and use it when making predictions.

Using our synthetic dataset, we first quantified the ability of a fingerprint-based virtual screening model to correctly identify important functional groups, and investigated the effect of changing the model's parameters on its attribution performance. We then investigated the effect of ligand-specific biases on the fingerprint-based models, both in terms of predictive accuracy and attribution performance. Finally we compared the performance of the fingerprint-based models to a recently proposed Equivariant graph neural network, PointVS [30].

We found that although all models were able to accurately predict binding in the presence of ligand-specific biases, their ability to attribute binding to the correct functional groups was substantially degraded, indicating they were less able to generalise than models which were not susceptible to ligand-specific biases. We also found that the attribution performance of the fingerprint-based models was heavily dependent on the parameters used to define the fingerprint, and that they were less able to identify the most important functional groups compared to the EGNN method, PointVS. These findings illustrate the importance of investigating the reasons behind high predictive performance and the utility of our synthetic approach as a benchmark for model generalisability.

## Methods

It is not easily possible to precisely experimentally quantify the extent to which an intermolecular interaction contributes to protein-ligand binding on a large scale on real-world data. Therefore we propose two protocols for generating synthetic data where the contribution of any atom can be computed exactly. Whilst the synthetic data we generate gives only a coarse-grained approximation of real-world protein-ligand binding, it nevertheless allows us to assess the ability of virtual screening models to utilise important spatial information. We are also able to ensure that our synthetic datasets are free of ligand-specific bias and other extraneous factors which may inflate a model's predictive performance on a test set while degrading its generalizability to novel targets. This allows us to quantify the effect of such real-world factors on model generalizability.

### Generating a synthetic protein-ligand complex

We defined a “synthetic protein” to be the set  $\{(x_i, y_i, z_i, t_i) | i = 1, \dots, m\}$ , where each element, which we call a “synthetic residue”, comprises 3D coordinates  $(x_i, y_i, z_i)$  and an associated type,  $t_i$ . After specifying a ligand with a 3D conformation, we constructed a synthetic protein as follows:

- We first defined a box around the ligand. To obtain the  $x$ -axis of the box we identified the minimum and maximum  $x$ -coordinates,  $x_{min}$  and  $x_{max}$ , over all ligand atoms and defined the  $x$ -axis as  $[x_{min} - 5 \text{ \AA}, x_{max} + 5 \text{ \AA}]$ . The  $y$ -axis and  $z$ -axis were obtained in the same way.
- We then sampled a number of coordinates uniformly within the box, such that the density of points was invariant of the box volume. That is, we sampled  $m$  points, where  $m = a_{coef} \times b$  for box volume  $b$ .
- For each set of coordinates, we randomly sampled an associated type to create a synthetic residue.
- If any synthetic residue was within 2 Å of a ligand atom, it was deleted.
- The synthetic residues were filtered further so that no two synthetic residues are within 3 Å of each other.
- To reduce the risk of inducing ligand-specific bias dependent on the number of functional groups present in a ligand, we sampled the number of synthetic residues,  $n_{res}$ , as  $n_{res} = \text{floor}(n_{ops}/n_{lig})$ , where  $n_{ops}$  is a constant (50 for all experiments in this paper) and  $n_{lig}$  was the number of ligand functional groups which can interact with a protein (which varies according to the generative process used to generate the synthetic protein, see below).

**Polar generative process:** the type of each synthetic residue determines which ligand functional groups it is able to interact with. In the Polar dataset we restrict the synthetic residue types to “Hydrogen Bond Acceptor” (HBA) and “Hydrogen Bond Donor” (HBD). We used RDKit [31] to determine whether each ligand atom was a Hydrogen bond donor or acceptor and we define a ligand donor or acceptor to interact with a synthetic residue if:

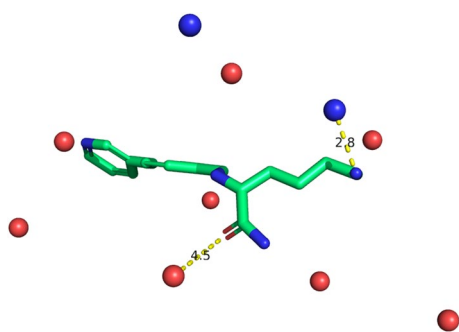
- Their types match: e.g. the synthetic residue has type HBD and the ligand atom is a Hydrogen Bond Acceptor, and
- The distance between the synthetic residue and ligand atom is below a specified threshold. For all experiments this was set at 4 Å.

For a given synthetic protein-ligand complex, if any ligand atom interacts with a synthetic residue, we say the complex is active, otherwise it is inactive. An example of

a synthetic protein-ligand complex generated using the Polar generative process is shown in Fig. 1.

### Contribution-based generative process

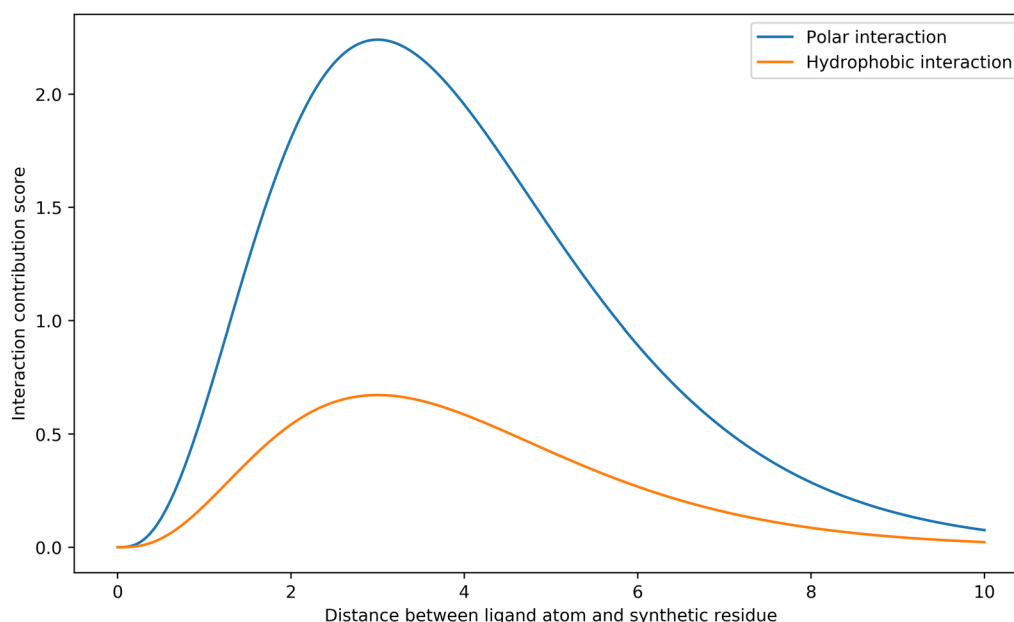
While the synthetic protein-ligand complexes generated by the Polar generative process only require a single



**Fig. 1** Example of synthetic protein-ligand complex. Green ligand atoms represent Carbons, red atoms denote Oxygens and blue atoms denote Nitrogens. The blue spheres represent synthetic residues with type ‘Hydrogen Bond Acceptor’ (HBA), whilst the red spheres represent synthetic residues with type ‘Hydrogen Bond Donor’ (HBD). The amine group is 2.8 Å away from a synthetic residue with type HBA; as the distance is less than the 4 Å specified by the deterministic binding rule, we consider this example to be active. While the carbonyl is 4.5 Å away from the nearest synthetic residue of with type HBD (and therefore does not interact with it), under the Polar generative process only a single interaction is needed for binding

interaction to be classified as active, in practice a ligand typically needs to make several different interactions in order to bind with high affinity. We therefore propose a second binding rule which assigns a score to each synthetic residue-ligand atom pair and classifies the complex depending on the cumulative score. In the Contribution dataset we extend the synthetic protein to include “Hydrophobic” synthetic residues as well as the HBA and HBD synthetic residues present in the Polar dataset. The synthetic protein is generated in the same way as when using the Polar generative process, with each synthetic residue type sampled at uniform from the set {HBA, HBD, Hydrophobic}.

To label a synthetic protein-ligand complex, we consider synthetic residue-ligand functional group pairs where the ligand functional group has type “Acceptor”, “Donor” or “Hydrophobe”, assigned by RDKit [31]. If the synthetic residue and ligand functional group do not have matching types, we define their interaction score as 0, otherwise their interaction score is calculated as a non-linear function of the interaction type and the euclidean distance between the synthetic residue and ligand atom (Fig. 2). For Hydrophobic interactions, we define the non-linear function as  $3 \times f(d)$ , where  $d$  is the distance between the ligand atom and synthetic residue, and  $f$  is the probability density function of the Gamma(4, 1) distribution. For Hydrogen Bond interactions, the non-linear function is defined as  $10 \times f(d)$ , meaning that Hydrogen Bonds are “stronger” than Hydrophobic



**Fig. 2** Non-linear functions used to determine the interaction score of a synthetic residue-ligand atom interaction; both functions are proportional to the Probability Density Function of a Gamma (4, 1) distribution. Separate functions are used for Polar interactions and Hydrophobic interactions, making it more difficult for the model to learn the deterministic binding rule

interactions. The sum of all pairwise interaction scores is calculated, and we label the example as “active” if the summed interaction score exceeds a pre-specified threshold. For all experiments in this work, we used a score threshold of 4.

### Machine learning algorithms

We used two different ML algorithms to classify examples as active or inactive. The first model was a random forest (RF), which took either Morgan fingerprints [32] or Protein-Ligand Extended Connectivity fingerprints (PLECs) [33] as input. We were able to represent our synthetic protein-ligand complexes as PLECs without needing to modify the Open Drug Discovery Toolkit [34] (oddt) codebase.

**Morgan fingerprints:** Morgan fingerprints are a vector representation of a molecule, calculated in an iterative fashion by first assigning an identifier to each atom and then updating the identifier to incorporate information from the identifiers of the atom's neighbours. This updating process is repeated a number of times so that information about atoms which are not immediate neighbours of an atom can be included in its identifier; in this work, all Morgan fingerprints are calculated using the RDKit [31] implementation with a radius of 2.

**Morgan fingerprints,** by design, do not incorporate any spatial information or any information about the target. In this work, the models trained using Morgan fingerprints serve as a baseline to assess whether a model can attain strong predictive accuracy using solely ligand-based features. Predictive performance that was substantially better than random would suggest that significant ligand-specific biases were present in the dataset, whereas close-to-random predictive performance would indicate that the active and inactive ligand sets were drawn from approximately the same population.

This allows us to quantify the extent to which a model might be susceptible to ligand-specific biases, as we would not expect models without access to information concerning the protein to be able to learn the deterministic binding rules and therefore better-than-random predictive accuracy would likely be due to biases in the training set.

**PLEC fingerprints:** PLEC fingerprints encode important spatial information as follows: First, all ligand atom-protein atom pairs which are closer together than a specified cutoff are identified, and integer radii,  $r_{lig}$  and  $r_{prot}$ , are specified for the ligand and protein respectively. (which need not be the same length). For each qualifying ligand atom-protein atom pair, substructures containing atoms which are up to  $r_{lig}$  or  $r_{prot}$  atoms away from the ligand or protein atom are identified and encoded in the fingerprint using a hashing algorithm. All PLEC

fingerprints were computed using the oddt [34] implementation with a ligand radius of 3 and a protein radius of 0 (as the synthetic residues are represented as a single, unconnected atom).

As ligand atom-protein atom pairs which are within a specified threshold are explicitly encoded within the fingerprint, we would expect models trained using PLECs to perform strongly on the Polar tasks when the PLEC distance cutoff closely matches the distance specified by the deterministic binding rule. However, as PLEC doesn't encode any more detailed spatial information, we would expect that it would be unable to learn the non-linear function of distance which calculates the contribution of a synthetic residue-ligand atom pair used by the Contribution generative process.

**Model naming:** we refer to models trained using a Morgan fingerprint as `RF_Morgan`. Models trained using PLEC fingerprints are referred to `RF_PLEC`. `RF_PLEC_n`, where  $n$  is a positive real number, refers to an RF model trained with a specific PLEC distance cutoff.

### Equivariant graph neural network (EGNN)

We compared the performance of the `RF_PLEC` models to a recently proposed deep learning-based approach [30], “PointVS”. PointVS is based on the E(n)-Equivariant Graph Neural Networks proposed by Satorras et al. [35]. In contrast to the fingerprint-based `RF_PLEC` models, PointVS takes as input the 3D coordinates of protein and ligand atoms, in addition to a one-hot encoding of the atom type. When constructing the input graph, where each node is an atom, two nodes are connected by an edge according to the following rules:

- Two ligand atoms are connected by an edge if they are within 2 Å of each other.
- Two protein atoms are connected by an edge if they are within 2 Å of each other.
- A protein atom and a ligand atom are connected by an edge if they are within 10 Å of each other.

For intra-molecular edges, the 2 Å cutoff connects atoms which are covalently bonded, whilst the inter-molecular edges connect atoms which might potentially interact. Scantlebury et al. [30] also applied a further distance cutoff, where any receptor atom which was not within 6 Å of any ligand atom was ignored; this reduced the dimensionality of the input graph by ignoring residues which were not part of the binding pocket. As we constrained each synthetic protein to be within a box defined as  $[x_{min} - 5, x_{max} + 5] \times [y_{min} - 5, y_{max} + 5] \times [z_{min} - 5, z_{max} + 5]$ , where  $x_{min}$  was the smallest ligand atom x-coordinate and the other values were defined similarly, the vast majority of synthetic residues would be within 6 Å of at



least one ligand atom and so this cutoff should have minimal impact on the performance of PointVS.

### "ZINC" dataset

Before exploring the effect of ligand-specific bias on model attribution performance, we constructed a baseline dataset to assess the ability of different ML algorithms to learn the deterministic binding rules outlined above. We obtained a set of 10k ligands from ZINC [36] and used the Polar and Contribution generative processes to generate two synthetic datasets ("Polar\_ZINC" and "Contribution\_ZINC"). The distribution of the number of Hydrogen Bond Acceptors/Donors in each ligand atom across the actives and decoys in the Polar\_ZINC set are shown in Fig. 3.

We sampled 500 examples from each dataset to serve as a test set, and used the remaining examples to train each model. We trained 8 different RF\_PLEC models, varying the PLEC distance cutoff by 0.5 Å from 2.5 Å to 6 Å. To train PointVS, we used the default hyperparameters outlined by Scantlebury et al. [30].

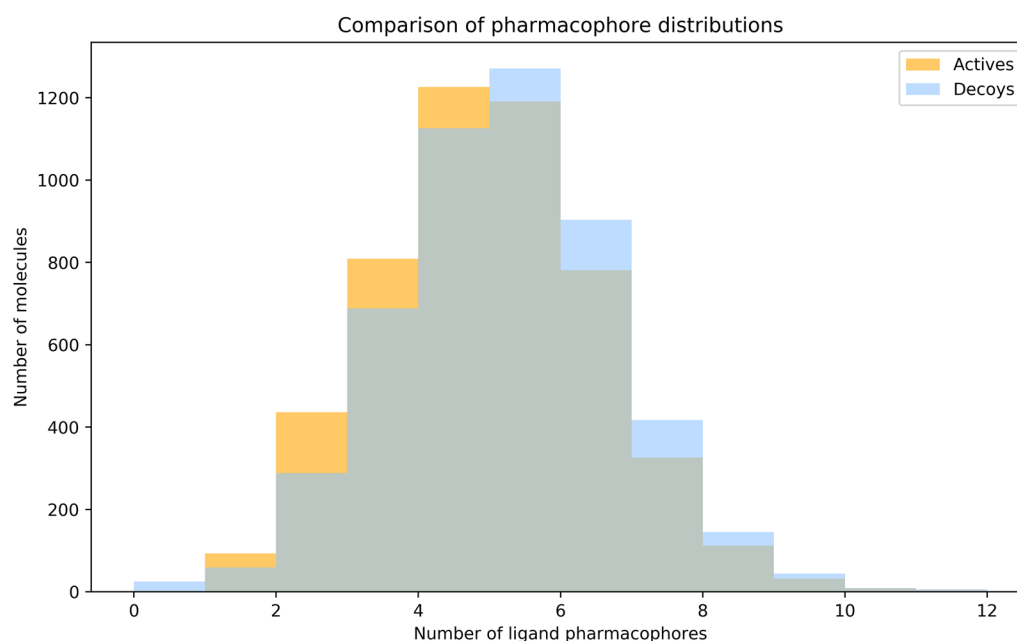
### Inducing ligand-specific bias

As mentioned above, Chen et al. [22] showed that the presence of ligand-specific biases allowed virtual screening models to disregard the provided protein-specific information and still achieve strong predictive accuracy. The models were able to do this by constructing a

decision rule which classified examples based on 2D ligand features rather than intermolecular interactions. We hypothesized that a dependence on ligand-specific biases would severely degrade the ability of a virtual screening model to identify important functional groups.

To explore the effect of ligand-specific bias on attribution performance, we constructed a set of synthetic protein-ligand complexes where each ligand was taken from the Directory of Useful Decoys—Enhanced (DUD-E) [37]. Sieg et al. [38] showed that ML algorithms were able to exploit distributional differences between the actives and inactives in DUD-E to achieve inflated predictive accuracy. We selected the five DUD-E targets containing the most ligands and used the ligands to construct five synthetic datasets. For each ligand, we extracted its true label from DUD-E and generated a synthetic protein using the Polar generation process outlined above. If the synthetic protein-ligand complex had a different label than the true label assigned to the ligand, we generated a new synthetic protein and recalculated the label until the true label and the label of the synthetic protein-ligand complex matched. If, after generating 100 synthetic proteins we were unable to attain the true label, we discarded the ligand. We used the same approach to derive an additional five synthetic datasets from the five LIT-PCBA [39] targets with the most ligands.

By constructing the synthetic virtual screening datasets in this way, an ML algorithm would be able



**Fig. 3** The distribution of the number of Hydrogen Bond Acceptor/Donors in each ligand atom for the Polar\_ZINC dataset, across the set of actives and decoys. The actives and decoys have almost identical distributions, suggesting that a virtual screening model would not be able to use the number of ligand pharmacophores to predict "binding"

to classify examples by using the spatial information in the synthetic protein-ligand complexes and/or by learning ligand-specific biases, mimicking the choice offered to an algorithm trained on the real DUD-E and LIT-PCBA datasets.

#### “PDBBind” dataset

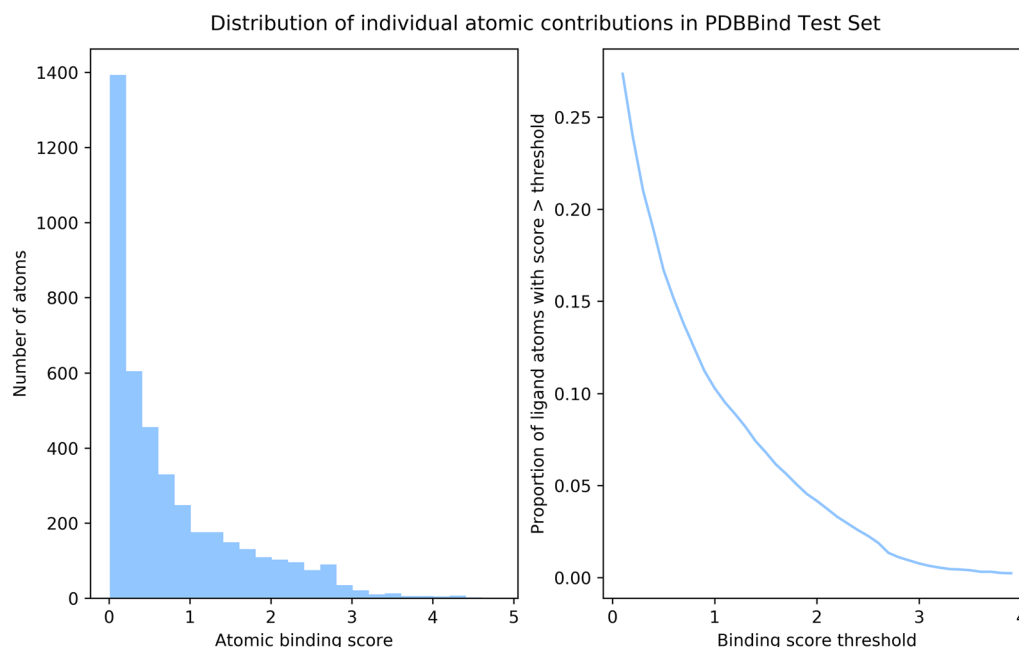
We used ligands from the PDBBind set [40] (v. 2019) to construct two external test sets, one using the Polar generative process and one using the Contribution generative process. To reduce the risk of inflated performance due to overfitting, any ligand with a Tanimoto similarity of more than 0.8 to any ligand in any of the above training sets was discarded and we also discarded any very small molecules (fewer than 15 heavy atoms). Of the remaining ligands, we randomly selected 500 and generated a synthetic protein-ligand complex from each. As the purpose of the test set was to assess the ability of the models to identify functional groups which were responsible for binding, we ensured that each example was “active” by resampling the synthetic protein for each ligand until it was active according to the deterministic binding rule. Figure 4 shows the distribution of the atomic scores contributed by each atom in the PDBBind\_Contribution dataset.

#### Ranking the importance of ligand atoms.

Following Hochuli et al. [29], we used atom masking to rank the importance assigned to a ligand atom by a particular model. For the  $i$ th atom in a molecule,  $m$ , we calculate the masking score as  $s_i = \text{score}(m) - \text{score}(m \setminus i)$ , where  $\text{score}(m)$  is the prediction given by the model for  $m$  and  $m \setminus i$  denotes the molecule  $m$  where the  $i$ th atom has been deleted. For the RF\_PLEC models, we delete an atom by replacing it with a dummy atom, and for PointVS we delete the corresponding node from the input graph. As higher scoring molecules are classified as active examples, masking assigns a high level of importance to atoms whose omission drastically reduces the model's confidence that an example has an active label. Whilst Scantlebury et al. [30] used an attention mechanism to score the relative importance of different atomic interactions, we used masking for the experiments in this paper as it can be used to generate attributions for any predictive model, allowing a closer comparison between different models.

#### Evaluation metrics

For datasets generated using the Polar process, where each atom either contributes to binding or is not involved at all, we would hope that an attribution method would give the highest rank to atoms involved in binding, allowing users to identify the most important atoms by their attribution scores. We propose an ‘Attribution AUC’;



**Fig. 4** The distribution of non-zero atomic contributions in the PDBBind test set. Atoms with a contribution of zero are not included to improve scaling; 7935 (62%) of the atoms in the test set have a ground-truth contribution of zero. **a** Distribution of atomic contribution scores. **b** Proportion of atoms with an above-threshold contribution. Approximately 9.5% of atoms have a score of more than 1 and 3.7% of atoms have a score of more than 2

where a ranking of atoms which places all binding ligand atoms at the top receives a score of 1, a ranking which places all binding ligand atoms at the bottom receives a score of 0, and all other rankings receive a score according to the following heuristic:

We define a ‘change’ to be the transposition of two adjacent rows in a dataframe sorted by the model attributions. We calculate the number of changes required to rank the ligand atoms correctly (all binding atoms ranked at the top), and the number of changes required to rank the ligand atoms correctly in the worst case scenario (all binding atoms ranked at the bottom).

$$1 - \frac{\text{\# changes needed}}{\text{worst case \# changes}}$$

For datasets generated using the Contribution process, we calculated the Spearman’s Rank Correlation Coefficient between the true ligand atom contributions and the model attributions for each example. To assess the extent to which the models were able to identify the most important atoms, we introduce two metrics:

- Mean Above-Threshold Ranking (MATR): for a specified score threshold,  $t$ , we compute the average rank (assigned by the model attributions) of all atoms whose true contribution was greater than  $t$ . A low MATR implies that a model can successfully identify important atoms; we compute the MATR for score thresholds between 0 and 3.
- Relative Efficiency of Ranking (RER): we compare the atomic rankings derived from the model attributions to the “perfect” ranking derived from the ground truth atomic contributions. We define:

$$RER(a, b) = \frac{MATR(a)}{MATR(b)}$$

where, for all experiments in this work,  $MATR(a)$  is the MATR score attained by the model  $a$ , and  $MATR(b)$  is the MATR score attained by the ground-truth atomic contributions. For the RER score we set the MATR score threshold to consider the ten percent of atoms with the highest score (i.e.  $t = 0.96$ )

In addition to the above metrics which assess the ability of a model to correctly identify which atoms were responsible for binding, we also computed the accuracy and area under the Precision-Recall curve (AU PRC) attained by the models on the held out test sets.

## Results and discussion

In order to quantify the extent to which different ML algorithms were able to accurately identify important intermolecular interactions, we tested the algorithms on

synthetic datasets using deterministic binding rules (see “Methods” section).

### Ligand-only model performance

In line with previous work [22, 24], we trained several models where no synthetic protein-specific information was provided to the model (RF\_Morgan, see “Methods” section). The performance of these models allowed us to assess whether a dataset exhibited any ligand-specific biases, as the deterministic binding rule was dependent on both ligand and synthetic-protein.

All RF\_Morgan models trained on DUD-E datasets attained substantially better-than-random predictive accuracy (Table 1). This was also true for all LIT-PCBA ligands if balanced numbers of actives and decoys were used (Table 1), although the balanced accuracy attained by the LIT-PCBA models was considerably lower than that obtained by the DUD-E models, suggesting that the LIT-PCBA ligands contained lower levels of ligand-specific bias. This suggests that RF\_Morgan models trained on DUD-E or LIT-PCBA datasets were able to exploit ligand-specific information to make accurate predictions. By contrast, when trained on the Polar\_ZINC dataset, the RF\_Morgan model attained a predictive accuracy of 0.52, indicating that without target-specific information a Random Forest using only ligand information was unable to accurately classify the examples (Table 2).

**Table 1** Performance of the RF\_Morgan model on different datasets

Dataset	Random Accuracy	Accuracy	AU-PRC	Balanced Accuracy	Balanced AU-PRC
ZINC	0.504	0.52	0.53	N/A	N/A
DUDE-AA2AR	0.93	1.0	1.0	0.984	0.996
DUDE-DRD3	0.972	0.998	1.0	0.978	0.995
DUDE-FA10	0.97	1.0	1.0	0.992	1.0
DUDE-MK14	0.976	0.998	1.0	0.994	1.0
DUDE-VGFR2	0.984	1.0	1.0	0.99	0.999
LIT-ALDH1	0.613	0.76	0.809	0.768	0.806
LIT-FEN1	0.956	0.958	0.584	0.778	0.883
LIT-MAPK1	0.964	0.964	0.292	0.692	0.797
LIT-PKM2	0.944	0.952	0.755	79	0.901
LIT-VDR	0.928	0.942	0.6	0.772	0.87

Predictive accuracy substantially better than random suggests that the datasets may suffer from ligand-specific bias

Accuracy denotes the proportion of correctly classified examples, whereas Random Accuracy denotes the accuracy that would have been obtained by assigning all examples the most common label (= max(% actives, % inactives). AU-PRC denotes the area under the Precision-Recall curve. Balanced Accuracy and Balanced AU-PRC denote the respective accuracy and area under the Precision-Recall curve when the model was trained using an equivalent number of actives and inactives



**Table 2** Performance of different RF\_PLEC models when trained on the Polar\_ZINC dataset

Model	Accuracy	AU-PRC	Attribution AUC
RF_Morgan	0.52	0.53	0.47
RF_PLEC_2.5	0.62	0.66	0.57
RF_PLEC_3	0.70	0.78	0.65
RF_PLEC_3.5	0.79	0.89	0.77
RF_PLEC_4	0.95	0.99	0.89
RF_PLEC_4.5	0.81	0.86	0.86
RF_PLEC_5	0.79	0.80	0.78
RF_PLEC_5.5	0.75	0.79	0.73
RF_PLEC_6	0.72	0.77	0.70
PointVS	0.89	0.95	0.85

Accuracy denotes the proportion of correctly classified examples on the Polar\_ZINC test set, AU-PRC denotes the area under the Precision-Recall curve on the Polar\_ZINC test set, and Attribution AUC reflects the ability of a model to correctly identify the ligand atoms responsible for binding on the PDBBind test set (see "Methods" section)

The best performing model was RF\_PLEC\_4, which uses the same distance cutoff as the Polar deterministic binding rule

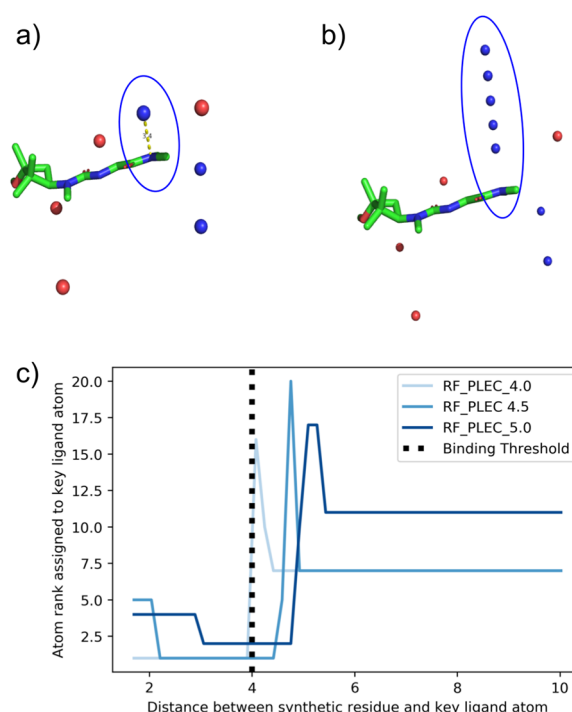
### Performance of RF\_PLEC models (ligand + protein fingerprints) on Polar\_ZINC dataset

Having validated that our Polar\_ZINC dataset did not contain trivial ligand-specific biases, we built RF\_PLEC models that include the synthetic proteins in the training process (see "Methods" section). To assess how sensitive the RF\_PLEC models were to the choice of PLEC distance cutoff, we trained eight distinct RF\_PLEC models, incrementing the PLEC distance cutoff by 0.5 Å between 2.5 Å and 6 Å. We found that all RF\_PLEC models attained a substantially better predictive accuracy than the RF\_Morgan model on the Polar\_ZINC dataset (Table 2), indicating that the inclusion of protein information into the model enabled the random forest to better approximate the deterministic binding rule. Unsurprisingly, the RF\_PLEC\_4 (the same cutoff as the deterministic binding rule) model obtained the highest predictive accuracy (Table 2).

The RF\_PLEC models exhibited a similar trend in terms of Attribution AUC (which assesses the ability of a model to assign high ranks to active ligand atoms), with the RF\_PLEC\_4 model most often correctly identifying the atoms responsible for binding (Table 2). We observed that performance degraded substantially as the PLEC distance cutoff diverged from 4 Å, suggesting that the RF\_PLEC models were highly dependent on the precise specification of the PLEC distance cutoff.

### Sensitivity of RF\_PLEC models to distance cutoff

To better understand the relationship between the PLEC distance cutoff and attribution performance, we



**Fig. 5** Case study illustrating the effect of changing the PLEC distance cutoff on model attributions. **a** The original synthetic protein-ligand complex generating using the Polar generative process, labelled as active as a result of the circled interaction. **b** Perturbed examples. We generated 50 synthetic protein-ligand complexes, which were all identical apart from the HBA synthetic residue contained in the circle, whose position was perturbed in relation to the ligand HBD with which it interacts. The 5 spheres within the circle illustrate 5 of the 50 positions occupied by the interacting residue. **c** Illustration of how the relative importance of the key ligand atom changes as the distance between it and the perturbed synthetic residue changes. Despite the synthetic residue and ligand atom no longer interacting when the distance between them is greater than 4 Å, the RF\_PLEC\_4.5 and RF\_PLEC\_5 models continued to rank the ligand atom highly until the synthetic residue-ligand atom distance was greater than the respective PLEC distance cutoff

consider a synthetic protein-ligand complex in detail. The original complex (shown in Fig. 5a) is labelled as active and has a single synthetic residue-ligand atom pair with complementary type and a distance below the deterministic binding threshold. We perturbed the synthetic residue to take a range of locations between 1.5 Å and 10 Å away from the matching ligand atom (Fig. 5b). We featurized each protein-ligand complex using PLEC and used masking to determine the importance of each ligand atom and calculated the rank of the ligand atom involved in binding. Figure 5c shows the relationship between synthetic residue-ligand atom distance and the rank assigned to the binding ligand atom by the highest performing RF\_PLEC models (RF\_PLEC\_4,

RF\_PLEC\_4.5 & RF\_PLEC\_5). Figure 5c shows that the RF\_PLEC models assign a high level of importance to the key ligand atom when the synthetic residue-ligand atom distance is less than the respective PLEC distance cutoff, and assigns a reduced level of importance when the distance is greater than the PLEC distance cutoff. In particular, when using a PLEC cutoff threshold of 4.5 or 5 Å, the ligand atom is assigned a high rank when the synthetic residue-ligand atom distance is greater than the true binding threshold but less than the PLEC distance cutoff. This demonstrates the sensitivity of the RF\_PLEC models to the exact specification of its distance cutoff; it is not able to encode a precise level of spatial information. This suggests that models trained using PLEC or based on distance-based cutoffs are unable to retain fine-grained spatial information, which may limit their usefulness for work with real-world protein-ligand complexes where different interactions can take place at different distances.

#### Performance of RF\_PLEC models on contribution dataset

We next examined the performance of the RF\_PLEC models on our Contribution\_ZINC dataset. The Contribution generative process more closely approximates real-world protein-ligand binding, as the strength of a synthetic residue-ligand atom interaction is a continuous function of the euclidean distance between them, and typically several high scoring interactions are required in order for a synthetic protein-ligand complex to be deemed active. This is significantly more challenging to learn than the Polar deterministic binding rule, which used a simple distance cutoff and only required a single synthetic residue-ligand atom pair to be 'active' for the complex to be active.

As with the dataset generated by the Polar generative process, we first fit an RF\_Morgan model on the Contribution\_ZINC dataset to assess whether the models were susceptible to ligand-specific bias. The RF\_Morgan model attained an accuracy of 0.47, suggesting that the structure-based models would have to use the protein to achieve strong predictive performance. All of the RF\_PLEC models attained a low Spearman's Rank Correlation Coefficient between the model assigned atom ranks and the ground truth atom ranks, ranging from 0.006 (RF\_PLEC\_2.5) to 0.093 ((RF\_PLEC\_6)). These values suggest that the RF\_PLEC models were unable to learn the deterministic binding rule. We next assessed whether the models were able to correctly assign a high rank to the most important functional groups. To do this, we computed the Relative Efficiency of Ranking (RER) attained by each RF\_PLEC model when compared to the "perfect" rankings obtained by ranking the atoms in a molecule by their ground truth contribution

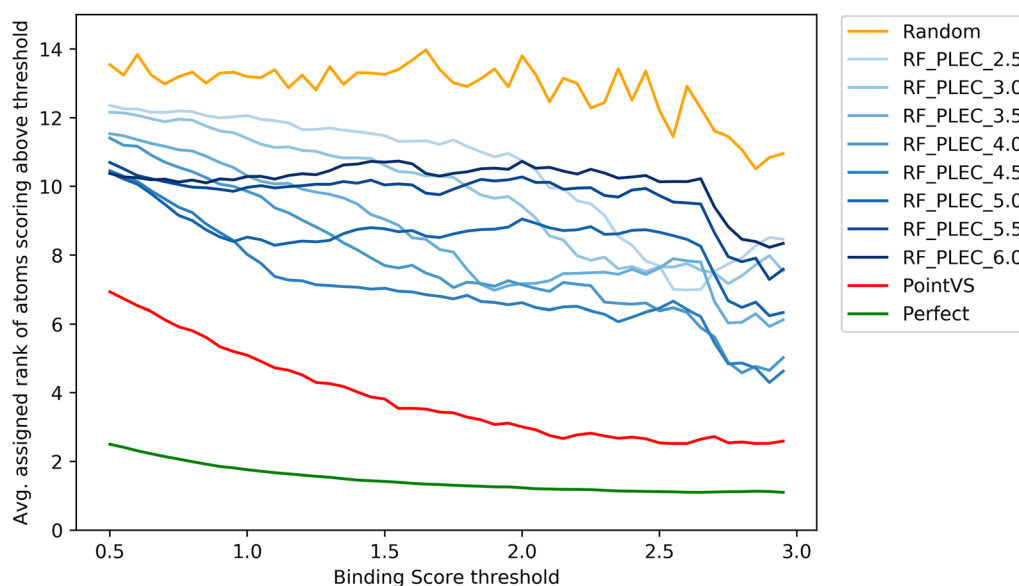
(see "Methods" section). The RER attained by the random baseline was 7.38, and the optimal attainable value for a model is 1.00. The RER scores attained by the RF\_PLEC models ranged between 4.66 (RF\_PLEC\_5) and 6.68 (RF\_PLEC\_2.5), suggesting that the models struggled to identify the most important functional groups when making predictions. The close-to-random performance of some of the RF\_PLEC models is also illustrated by Fig. 6, which shows the Mean Above Threshold Ranking (MATR) score (see "Methods" section) attained by each model for a variety of thresholds.

Our results suggest that whilst the PLEC fingerprints were often able to encode a sufficiently detailed level of spatial information for the simplistic Polar tasks, its representation of spatial information is inadequate to learn the more complicated Contribution binding rule. As the Contribution binding rule is itself considerably simpler than the rules which govern real-world protein-ligand binding, it is likely that PLEC (and other fingerprint-based methods which are based on a binary distance cutoff) are ill-equipped to ascertain which functional groups make a key contribution towards binding.

#### Quantification of ligand-specific bias on attribution performance

Although the above results suggest that fingerprint-based models struggle to learn complex binding rules, several previous studies (e.g. [21, 41]) have reported that fingerprint-based methods have attained strong predictive accuracy on real-world virtual screening datasets such as DUD-E [37]. However, recent studies (e.g. [22, 38]) have illustrated that virtual screening models are able to attain inflated performance by learning ligand-specific biases, with Sieg et al. [38] reporting that highly simplistic models (e.g. a model with the number of Hydrogen Bond Acceptors in a ligand as the only feature) served as highly accurate classifiers for certain DUD-E targets. Therefore, it is of interest to assess the extent to which real-world factors such as ligand-specific biases and labelling errors degrade the ability of virtual screening models to identify important ligand functional groups.

Although we found that the models trained on DUD-E and LIT-PCBA datasets attained strong predictive accuracy on their respective test sets (Table 3), overall we found that training the models on a dataset which exhibited ligand-specific bias degraded their attribution performance. Table 4 shows the Attribution AUC values attained by the best performing RF\_PLEC models when trained on a variety of different datasets and tested on the external PDBBind set, generated using the Polar generative process. When training on the DUD-E datasets, each of the RF\_PLEC\_4 models attained an attribution AUC value marginally below that obtained



**Fig. 6** The average rank assigned to all ligand atoms attaining a score above a specified threshold. Perfect is the curve attained when ranking atoms by the true atomic contribution, whilst random is the curve obtained when the atoms are ranked completely at random

**Table 3** Performance of the RF\_PLEC\_4 model on different datasets

Dataset	Random accuracy	Accuracy	AU-PRC	Balanced accuracy	Balanced AU-PRC
ZINC	0.504	0.948	0.988	N/A	N/A
DUDE-AA2AR	0.93	0.992	0.987	0.942	0.992
DUDE-DRD3	0.972	0.992	0.974	0.904	0.981
DUDE-FA10	0.97	0.992	0.971	0.934	0.993
DUDE-MK14	0.976	0.99	0.994	0.928	0.985
DUDE-VGFR2	0.984	0.992	0.92	0.926	0.987
LIT-ALDH1	0.613	0.948	0.987	0.946	0.988
LIT-FEN1	0.956	0.968	0.839	0.858	0.948
LIT-MAPK1	0.964	0.972	0.672	0.822	0.923
LIT-PKM2	0.944	0.974	0.976	0.916	0.984
LIT-VDR	0.928	0.98	0.931	0.928	0.979

Accuracy denotes the proportion of correctly classified examples, whereas Random Accuracy denotes the accuracy that would have been obtained by assigning all examples the most common label (= max(% actives, % inactives))

AU-PRC denotes the area under the Precision-Recall curve

Balanced Accuracy and Balanced AU-PRC denote the respective accuracy and area under the Precision-Recall curve when the model was trained using an equivalent number of actives and inactives

by the corresponding model trained on the ZINC dataset. However, we observed a considerably larger difference between the attribution AUC values obtained by the RF\_PLEC\_4.5 and RF\_PLEC\_5 models trained using the DUD-E datasets and the corresponding models trained using the ZINC ligands. Indeed, two of the RF\_PLEC\_5 models attained an Attribution AUC (FA10:

**Table 4** Attribution AUC obtained by the different RF\_PLEC models and PointVS on the different Polar datasets

Dataset	RF_PLEC_4	RF_PLEC_4.5	RF_PLEC_5	PointVS
ZINC	<b>0.89</b>	<b>0.86</b>	<b>0.78</b>	0.85
DUDE-AA2AR	0.86	0.76	0.73	0.84
DUDE-DRD3	0.83	0.68	0.63	0.69
DUDE-FA10	0.86	0.67	0.56	0.82
DUDE-MK14	0.82	0.63	0.61	0.73
DUDE-VGFR2	0.77	0.62	0.56	0.61
LIT-ALDH1	0.86	0.81	0.73	<b>0.93</b>
LIT-FEN1	0.83	0.69	0.62	0.87
LIT-MAPK1	0.81	0.61	0.59	0.70
LIT-PKM2	0.81	0.73	0.67	0.82
LIT-VDR	0.85	0.71	0.65	0.88
Random	0.503	0.503	0.503	0.503

The best performing dataset for each model is highlighted in bold

The RF\_PLEC models trained on the unbiased ZINC\_Polar dataset consistently attained a larger Attribution AUC than those trained on datasets susceptible to ligand-specific bias, suggesting that if a model learns to classify examples based on ligand-specific features, its ability to learn the true binding rule is impaired

By contrast, the highest Attribution AUC attained by PointVS was when training on the LIT-ALDH1 dataset, potentially illustrating that in some instances it was able to learn the deterministic binding rule in the presence of ligand-specific bias

0.555, VGFR2: 0.559) which was only marginally higher than the random baseline (0.503). We observed a similar trend for the RF\_PLEC models trained using LIT-PCBA ligands.

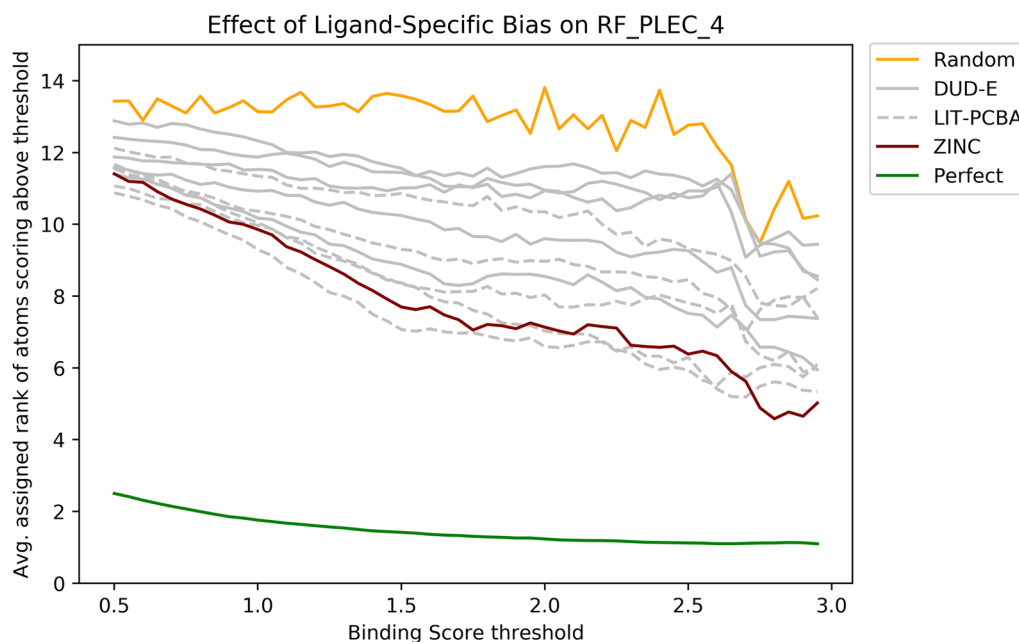
In contrast to the attribution performance attained on the Polar datasets, where the RF\_PLEC\_4 models were relatively robust to the introduction of ligand specific bias, the attribution performance attained on the Contribution datasets was substantially degraded in the presence of ligand-specific bias. Figure 7 compares the attribution performance of the RF\_PLEC\_4 model when trained on the Contribution\_ZINC dataset to the models trained on the DUD-E and LIT-PCBA datasets. Training on DUD-E or LIT-PCBA ligands degraded the attribution performance, with the average rank assigned to important ligand atoms decreasing substantially.

### Performance of PointVS

We next examined the performance of an EGNN-based [35] method, PointVS (see "Methods" section). When trained using the Polar\_ZINC dataset, the PointVS attained an accuracy of 0.886, a AU-PRC of 0.948 and an Attribution AUC of 0.851 (Table 2). Whilst PointVS performed slightly worse than the RF\_PLEC\_4 and RF\_PLEC\_4.5 models, those models were able to exploit the fact that the PLEC featurisation explicitly encoded information about which synthetic residue-ligand atom pairs were near- or below the deterministic binding threshold. By contrast, PointVS was only provided with the unprocessed atomic coordinates, making it more challenging to learn the deterministic binding rule.

When trained and tested on datasets using the Contribution generative process we found that PointVS comfortably outperformed all of the RF\_PLEC models. Whilst the Spearman's rank correlation (0.18) obtained by PointVS was not particularly strong, it was higher than any correlation attained by the RF\_PLEC models (max: RF\_PLEC\_6, 0.09). As with the RF\_PLEC models, we computed the RER score (see "Methods" and "Evaluation metrics" section) to compare PointVS with the ground truth attributions. PointVS attained an RER score of 2.86, considerably better than the top-performing RF\_PLEC model (RF\_PLEC\_5, 4.66). Fig. 6 illustrates the Mean Above Threshold Ranking (MATR) score attained by each model for a variety of thresholds, and shows that PointVS uniformly outperformed all RF\_PLEC models for all thresholds. These results are perhaps not surprising, as the RF\_PLEC models do not capture a precise encoding of each atom's position, only recording whether the distance between two atoms is below a pre-specified threshold, making it difficult for the RF\_PLEC models to approximate the non-linear deterministic binding rule. By contrast, PointVS is not constrained by a particular featurization and can learn the deterministic binding rule from the data.

When trained on the DUD-E or LIT-PCBA datasets, PointVS demonstrated a degree of robustness to ligand-specific bias under the Polar generative process (Table 4),



**Fig. 7** The average rank assigned to all ligand atoms attaining a score above a specified threshold on the PDBBind test set. Each line corresponds to the performance obtained by an RF\_PLEC\_4 model trained on a different training set. The model trained on the unbiased ZINC dataset outperforms the models trained on the real-world DUD-E and LIT-PCBA datasets, illustrating that ligand-specific biases hamper the ability of virtual screening models to identify the most important functional groups

although in several cases the Attribution was substantially lower than the value obtained for the Polar\_ZINC data. Under the Contribution generative process, consistent with the results obtained using the RF\_PLEC\_4 model, PointVS' attribution performance was considerably worse when trained with the DUD-E or LIT-PCBA datasets compared to the Contribution\_ZINC dataset (Fig. 8). As both the PointVS and RF\_PLEC\_4 models struggle to learn complex binding rules in the presence of ligand-specific bias, it is likely that the models would be unable to identify the functional groups responsible for binding on real-world datasets if similar biases were present.

## Conclusion

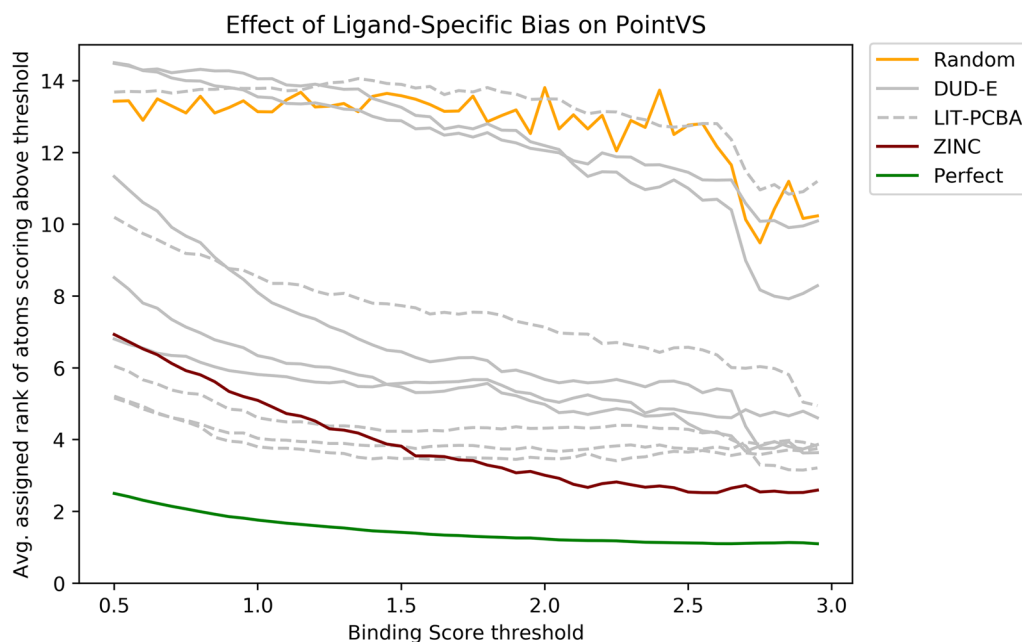
We have proposed a framework for assessing the ability of different ML algorithms to attribute binding at the atomic level. By simulating synthetic datasets we are able to define a ground truth for which atoms in a synthetic protein-ligand complex are responsible for binding, something which is not currently possible with real-world data. Despite several recent studies suggesting that fingerprint-based machine learning algorithms achieve comparable performance with current deep learning methods, we found that a deep learning-based EGNN was better able to elucidate which atoms were responsible for 'binding' when the training datasets were not

susceptible to ligand-specific biases. This suggests that deep learning methods should have a greater ability to generalize to novel targets.

When we trained our models on datasets containing ligand-specific bias, we found that a model's ability to accurately assign atomic importance was diminished, consistent with the notion that dataset-specific bias degrades model generalisability.

However, for the Polar datasets, we found that the drop in attribution performance attained by PointVS was far smaller than the corresponding drop in performance observed for the RF\_PLEC\_4.5 and RF\_PLEC\_5 models and similar to that observed for the RF\_PLEC\_4 model, suggesting that models which can more easily learn the true binding rule are less susceptible to the effects of bias. The development of models which are able to capture and apply important spatial information should therefore continue to be a priority.

Our analysis has a number of limitations. The deterministic binding rules used in both the Polar and Contribution generative processes are significantly simplified compared to real-world protein-ligand binding. However, our objective was to assess the extent to which different ML algorithms were able to capture and apply relevant spatial information and our relatively simple generative processes were sufficient to highlight the difficulties of explicitly encoding a precise level of



**Fig. 8** The average rank assigned to all ligand atoms attaining a score above a specified threshold on the PDBBind test set. Each line corresponds to the performance obtained by a PointVS model trained on a different training set. The model trained on the unbiased ZINC dataset outperforms the models trained on the real-world DUD-E and LIT-PCBA datasets, illustrating that ligand-specific biases hamper the ability of virtual screening models to identify the most important functional groups



3D information into a fingerprint which uses a binary distance cutoff to summarise information about the binding site. An interesting avenue for further work would be to benchmark the performance of a wider variety of virtual screening models than the two presented in this paper, including fingerprints such as NNScore [18], which incorporate distance-based terms from the scoring function of Autodock Vina [15] into their fingerprint, and convolutional neural networks (e.g. [20, 21]).

Whilst there is significant interest in developing models which can predict protein-ligand binding with high accuracy, there is a strong need for models which can identify the functional groups which contribute most heavily towards binding; the identification of key functional groups would allow easier iterative optimisation of lead compounds. Moreover, models which demonstrated an understanding of biophysical rules would be more likely to be accepted by human experts. In addition to the development of machine learning architectures which can better capture and apply important spatial information, avenues of research which might better enable models to identify important functional groups include data augmentation and multi-task learning. An alternative strategy which can improve the quality of virtual screening models is the incorporation of experimentally verified misses into the training data [42].

We hope that our synthetic framework will prove useful to researchers seeking to benchmark the ability of different virtual screening models to capture and apply important spatial information. We have made the datasets used in this study available, alongside the code to generate synthetic protein-ligand complexes from a set of ligands.

#### Abbreviations

Å	Angstrom
DUD-E	Directory of useful decoys—enhanced
EGNN	Equivariant Graph Neural Network
HBA	Hydrogen Bond Acceptor
HBD	Hydrogen Bond Donor
MATR	Mean Above-Threshold Ranking
ML	Machine learning
ODDT	Open Drug Discovery Toolkit
PLEC	Protein-Ligand Extended Connectivity
RER	Relative Efficiency of Ranking
RF	Random forest
ZINC	ZINC is not commercial

#### Acknowledgements

The authors would like to thank Garrett M. Morris, Fergus Boyles, Kris Birchall and Andy Merritt for helpful discussions.

#### Author contributions

TEH and CMD conceived and designed the study; TEH and JS implemented the methods; TEH conducted the experiments and analysed the results. All authors contributed to the writing, reviewing and editing of the manuscript.

#### Funding

T.E.H. was supported by funding from the Engineering and Physical Sciences Research Council (EPSRC), LifeArc, F. Hoffmann-La Roche AG, and UCB Pharma (Reference: EP/L016044/1). J.S. is supported by funding from the Biotechnology and Biosciences Research Council (BBSRC) BB/S507611/1 and BenevolentAI.

#### Availability of data and materials

Code is available at <https://github.com/tomhadfield95/synthVS>. Datasets are available at [https://opig.stats.ox.ac.uk/data/downloads/synthVS\\_data\\_for\\_release.tar.gz](https://opig.stats.ox.ac.uk/data/downloads/synthVS_data_for_release.tar.gz).

#### Declarations

##### Competing interests

T.E.H. is an employee of AstraZeneca PLC; all work was done whilst a doctoral student at the University of Oxford. C.M.D. is an employee of Exscientia PLC.

Received: 2 May 2023 Accepted: 25 August 2023

Published online: 19 September 2023

#### References

- Wouters OJ, McKee M, Luyten J (2020) Estimated Research And Development Investment Needed To Bring A New Medicine To Market, 2009–2018. *J Am Med Assoc* 323(9):844–853
- Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A et al (2017) Mastering the game of go without human knowledge. *Nature* 550(7676):354–359
- Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. *Proc Adv Neural Inf Process Syst* 33:1877–1901
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A et al (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583–589
- Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373(6557):871–876
- Du Z, Su H, Wang W, Ye L, Wei H, Peng Z, Anishchenko I, Baker D, Yang J (2021) The trRosetta server for fast and accurate protein structure prediction. *Nat Prot* 16(12):5634–5651
- Skalic M, Sabbadin D, Sattarov B, Sciabola S, De Fabritiis G (2019) From target to drug: generative modeling for the multimodal structure-based ligand design. *Mol Pharm* 16(10):4282–4291
- Ragoza M, Masuda T, Koes DR (2022) Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chem Sci* 13(9):2701–2713
- Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. *J Cheminform* 9(1):1–14
- Imrie F, Bradley AR, van der Schaar M, Deane CM (2020) Deep generative models for 3d linker design. *J Chem Inf Model* 60(4):1983–1995
- Hadfield TE, Imrie F, Merritt A, Birchall K, Deane CM (2022) Incorporating target-specific pharmacophoric information into deep generative models for fragment elaboration. *J Chem Inf Model* 62(10):2280–2292
- Genheden S, Thakkar A, Chadimová V, Reymond J-L, Engkvist O, Bjerrum E (2020) AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J Cheminform* 12(1):1–9
- Ishida S, Terayama K, Kojima R, Takasu K, Okuno Y (2022) Ai-driven synthetic route design incorporated with retrosynthesis knowledge. *J Chem Inf Model* 62(6):1357–1367
- Dai H, Li C, Coley C, Dai B, Song L (2019) Retrosynthesis prediction with conditional graph logic network. *Proc Adv Neural Inf Process Syst*. Vol 32
- Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comp Chem* 31(2):455–461
- Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein-ligand docking using GOLD. *Proteins* 52(4):609–623

17. Ballester PJ, Mitchell JB (2010) A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* 26(9):1169–1175
18. Durrant JD, McCammon JA (2011) NNScore 2.0: a neural-network receptor-ligand scoring function. *J Chem Inf Model* 51(11):2897–2903
19. Pereira JC, Caffarena ER, Dos Santos CN (2016) Boosting docking-based virtual screening with deep learning. *J Chem Inf Model* 56(12):2495–2506
20. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR (2017) Protein-ligand scoring with convolutional neural networks. *J Chem Inf Model* 57(4):942–957
21. Imrie F, Bradley AR, van der Schaar M, Deane CM (2018) Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. *J Chem Inf Model* 58(11):2319–2330
22. Chen L, Cruz A, Ramsey S, Dickson CJ, Duca JS, Hornak V, Koes DR, Kurtzman T (2019) Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS ONE* 14(8):0220113
23. Volkov M, Turk J-A, Drizard N, Martin N, Hoffmann B, Gaston-Mathé Y, Rognan D (2022) On the frustration to predict binding affinities from protein-ligand structures with deep neural networks. *J Med Chem* 65(11):7946–7958
24. Scantlebury J, Brown N, Von Delft F, Deane CM (2020) Data set augmentation allows deep learning-based virtual screening to better generalize to unseen target classes and highlight important binding interactions. *J Chem Inf Model* 60(8):3722–3730
25. McCloskey K, Taly A, Monti F, Brenner MP, Colwell LJ (2019) Using attribution to decode binding mechanism in neural network models for chemistry. *Proc Natl Acad Sci* 116(24):11624–11629
26. Sundar V, Colwell L (2020) Attribution methods reveal flaws in fingerprint-based virtual screening. *arXiv*. <https://doi.org/10.48550/arXiv.2007.01436>
27. Matveeva M, Polishchuk P (2021) Benchmarks For interpretation Of QSAR models. *J Cheminform* 13(1):1–20
28. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Sundararajan M (ed) Proceedings of 34th international conference on machine learning. Proceedings of Machine Learning Research, Pittsburgh, pp 3319–3328
29. Hochuli J, Helbling A, Skaist T, Ragoza M, Koes DR (2018) Visualizing convolutional neural network protein-ligand scoring. *J Mol Graph Model* 84:96–108
30. Scantlebury J, Vost L, Carbery A, Hadfield TE, Turnbull OM, Brown N, Chenthamarakshan V, Das P, Grosjean H, von Delft F et al (2023) A small step toward generalizability: training a machine learning scoring function for structure-based virtual screening. *J Chem Inform Model*. <https://doi.org/10.1021/acs.jcim.3c00322>
31. Landrum G (2006) RDKit: Open-Source Cheminformatics
32. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inform Model* 50(5):742–754
33. Wójcikowski M, Kukiela M, Stepniewska-Dziubinska MM, Siedlecki P (2019) Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* 35(8):1334–1341
34. Wójcikowski M, Zielenkiewicz P, Siedlecki P (2015) Open drug discovery toolkit (ODDT): a new open-source player in the drug discovery field. *J Cheminform* 7(1):1–6
35. Satorras VG, Hoogeboom E, Welling M (2021) E (n) equivariant graph neural networks. In: Satorras VG (ed) Proceedings of the 38th international conference on machine learning. Proceedings Machine Learning Research, Pittsburgh, pp 9323–9332
36. Sterling T, Irwin JJ (2015) ZINC 15-ligand discovery for everyone. *J Chem Inform Model* 55(11):2324–2337
37. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* 55(14):6582–6594
38. Sieg J, Flachsenberg F, Rarey M (2019) In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J Chem Inform Model* 59(3):947–961
39. Tran-Nguyen V-K, Jacquemard C, Rognan D (2020) LIT-PCBA: an unbiased data set for machine learning and virtual screening. *J Chem Inform Model* 60(9):4263–4273
40. Liu Z, Su M, Han L, Liu J, Yang Q, Li Y, Wang R (2017) Forging the basis for developing protein-ligand interaction scoring functions. *Acc Chem Res* 50(2):302–309
41. Wójcikowski M, Ballester PJ, Siedlecki P (2017) Performance of machine-learning scoring functions in structure-based virtual screening. *Sci Rep* 7(1):1–10
42. Poelking C, Chessari G, Murray CW, Hall RJ, Colwell L, Verdonk M (2022) Meaningful machine learning models and machine-learned pharmacophores from fragment screening campaigns. *arXiv*. <https://doi.org/10.48550/arXiv.2204.06348>

## Publisher's Note

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

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

