RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2024, **14**, 1341

Augmenting a training dataset of the generative diffusion model for molecular docking with artificial binding pockets†

Taras Voitsitskyi, *[ad] Volodymyr Bdzhola, [b] Roman Stratiichuk,[ae] Ihor Koleiev,[ad] Zakhar Ostrovsky,[a] Volodymyr Vozniak,[a] Ivan Khropachov,[a] Pavlo Henitsoi,[a] Leonid Popryho,[a] Roman Zhytar,[a] Semen Yesylevskyy, [acdf] Alan Nafiiev[a] and Serhii Starosyla [a]

This study introduces the PocketCFDM generative diffusion model, aimed at improving the prediction of small molecule poses in the protein binding pockets. The model utilizes a novel data augmentation technique, involving the creation of numerous artificial binding pockets that mimic the statistical patterns of non-bond interactions found in actual protein–ligand complexes. An algorithmic method was developed to assess and replicate these interaction patterns in the artificial binding pockets built around small molecule conformers. It is shown that the integration of artificial binding pockets into the training process significantly enhanced the model's performance. Notably, PocketCFDM surpassed DiffDock in terms of non-bond interaction and steric clash numbers, and the inference speed. Future developments and optimizations of the model are discussed. The inference code and final model weights of PocketCFDM are accessible publicly *via* the GitHub repository: https://github.com/vtarasv/pocket-cfdm.git.

Received 28th November 2023
Accepted 21st December 2023

DOI: 10.1039/d3ra08147h

rsc.li/rsc-advances

Introduction

Molecular docking plays a central role in modern computational drug discovery. Until recently docking was the only available method of predicting the poses of small molecules in the binding pockets of target proteins fast enough to be useful in large-throughput virtual screening projects. Despite its ultimate importance, there is a notable stagnation in improving the docking versatility, accuracy, computing cost, and predictive power.[1,2] Being based on inevitably simplified empirical potentials of intermolecular interactions and lacking explicit solvent, the docking scoring functions have arguably reached a plateau of practical accuracy. Despite a number of recent developments in the field, all of them are incremental improvements and domain-specific tuning rather than technological breakthroughs.

However, recent advancements in deep learning methodologies for predicting ligand poses in the protein binding pockets provide a promising alternative to docking algorithms. These techniques can be categorized into two primary groups.

The models from the first group utilize a one-shot inference regression-based approach.[3,4] They are developed with the aim of very fast inference, which is superior to traditional docking simulations. The geometric vector perceptrons (GVP) and equivariant graph neural networks (EGNN) are the most popular architectures for those types of models.[5–7] Such techniques as EquiBind[4] and TANKBind[3] demonstrated efficacy in predicting protein–ligand binding structure without prior knowledge about the binding pocket (also known as blind docking) while being faster than traditional docking techniques by several orders of magnitude. However, they still suffer from unrealistic ligand conformations and numerous sterical clashes in predicted complexes, which puts them behind the docking approaches in terms of structure quality and reliability. A possible reason for this is a mismatch between the objectives of molecular docking and the regression paradigm. Particularly, the accuracy metrics in molecular docking are based on structural similarity, rather than a regression loss.

The second and currently state-of-the-art approach uses generative AI models, which aim to be on par or better than

[a]*Receptor.AI Inc., 20-22 Wenlock Road, London N1 7GU, UK*
[b]*Institute of Molecular Biology and Genetics of The National Academy of Sciences of Ukraine, 150 Zabolotnogo Str., Kyiv 03143, Ukraine*
[c]*Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague 6 CZ-166 10, Czech Republic*
[d]*Department of Physics of Biological Systems, Institute of Physics of The National Academy of Sciences of Ukraine, 46 Nauky Ave., Kyiv 03038, Ukraine*
[e]*Department of Biophysics and Medical Informatics, Educational and Scientific Centre "Institute of Biology and Medicine", Taras Shevchenko Kyiv National University, 64 Volodymyrska Str., Kyiv 01601, Ukraine*
[f]*Department of Physical Chemistry, Faculty of Science, Palacký University Olomouc, 17 listopadu 12, Olomouc 771 46, Czech Republic*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3ra08147h

classic docking techniques in terms of accuracy and structure quality. The DiffDock,[8] a current leader in the field, demonstrates superior performance in comparison to some conventional docking techniques and the previous ML models in the blind docking scenarios. It produces much less steric clashes than its rivals and generates realistic ligand conformations. However, the enhanced precision of DiffDock comes at the cost of a substantial computational burden, which is on par with that of traditional docking methods. Since there is a significant potential for improvement in terms of inference speed, it makes generative models the most promising in the field at the moment of writing.

The major bottleneck, which hampers further improvement of the generative ligand pose prediction models, is the inherently limited amount of the training data. The overall number of experimentally determined protein–ligand complexes resolved by X-ray, Cryo-EM, or NMR techniques is now below 20 000. The PDBbind database,[9,10] which is a primary dataset for machine learning in protein binding site prediction[11,12] and ligand pose prediction,[3,4,8] contains 19 443 distinct protein–ligand complexes with the binding activity annotations (version 2020). Out of this, 2709 entries involve peptide ligands or multiple molecules in a single binding pocket; 3827 have an insufficient resolution (2.5 angstroms and more); 3523 contain weak binders ($K_d/K_i/IC_{50}$ > 10 μM) and 216 lack confident activity measurements. Thus there are only 10 270 high-quality complexes, which could be used for model training

Another issue of the experimental complexes is ligand data sparsity. There are 12 815 distinct small molecule ligands in PDBbind, of which only 1655 appear in two or more entries (Fig. 1). It indicates that ML-based approaches are mostly presented with a particular ligand bound to a single protein without any information about the possible variability of its binding modes. This might lead to overfitting to a single ligand pose, which wouldn't be the case when working with more dense data as was demonstrated for the affinity prediction models.[13]

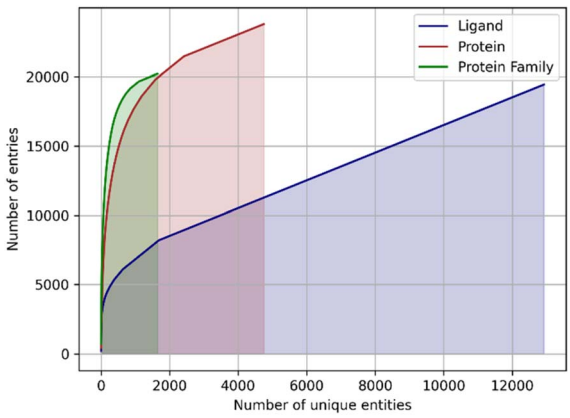

Fig. 1 The cumulative sums of entries in the PDBbind database assigned to a unique ligand, unique protein, or unique protein family. Note that the count of proteins and families may exceed the total number of complexes in PDBbind because a single PDB structure may be assigned to multiple entries.

In addition, the proteins are represented very unevenly in available data. The protein data bank (PDB) identifiers of 19 443 protein–ligand complexes are associated with 4749 unique UniProt[14] identifiers, but 1.500 most frequently occurring identifiers accounting for ∼80% of the total number of complexes (Fig. 1). The same is true for the protein families: a total of 1646 distinct protein families and superfamilies are present but the top 100 families account for almost 60% of complexes (Fig. 1). Thus, despite a significant overall variety of proteins, there is an obvious overrepresentation of some proteins and protein families, which may lead to model overfitting and imbalance.

There is little doubt that an insufficient overall number of samples, limited ligand diversity, and protein representation imbalance in the available data for model training are impairing the accuracy of ML-based approaches to the ligand pose prediction. These limitations are especially noticeable when comparing the amount and quality of the training data with a requirement for the model to operate on arbitrary protein targets and arbitrary ligands from an immense chemical space.[15]

It is clear that the dataset of experimentally resolved structures will not grow fast enough to satisfy the demands of the exploding field of ML ligand–protein binding prediction thus other approaches are needed for overcoming the lack of training data.

In this study, we develop an approach of augmenting the training set of the protein–ligand complexes with artificial data, which mimics real protein binding pockets in a number of structural characteristics. The statistical distributions of artificial pockets' parameters are fitted to the respective distributions of real protein–ligand structures so that both types of data could be used together seamlessly.

The idea of our approach is based on the assumption that the number of favorable interaction geometries between the protein amino acids and the chemical groups of the ligands is finite and is represented sufficiently well in the available experimental data. What is presumably lacking, is the sampling of all possible combinations of such interactions within the binding pocket. In other words, we assume that available experimental structures provide decent statistics about the preferable chemical identity of interacting atom pairs, distances between the atoms, and orientations of the corresponding chemical groups. However, only a very small fraction of all possible combinations of such pairs is observed in real proteins.

If one generates a large number of "artificial binding pockets", which follow the same statistical distribution of the interacting atom pairs as the real ones, but sample a much larger variety of their combinations, it might be possible to overcome the undersampling and to train the model on a more complete set of data.

Although we do not have a strong independent proof of our hypothesis, we decided to validate it experimentally by developing an algorithm for generating artificial pockets, compiling a dataset consisting of artificial and real pockets, and measuring the performance of the diffusion generative model,

which is inspired by DiffDock, and is trained on such augmented data – PocketCFDM (Pocket Conformation Fitting Diffusion Model). We show that PocketCFDM outperforms DiffDock, which is a recent breakthrough technology in the field of ML-based docking, in terms of generated ligand poses correctness (less steric clashes and more favorable non-covalent interactions). We also discuss future prospects of our methodology in terms of improving its predictive power and the speed of inference.

Methods

Protein and ligand preprocessing

The Python API of the RDKit v. 2021.03 was utilized for loading, processing, and feature generation of small molecules. A custom protein processing module was developed to extract protein data from the PDB files and to generate the necessary features for model training. This module utilizes the PDB atom names to obtain atom-level graph features, rather than relying on a third-party software to infer them. This approach decreases the exclusion rate for processed proteins due to inevitable inconsistencies in the PDB files.

In order to assess the protein–ligand interactions, we employed the SMILES arbitrary target specification (SMARTS) substructure search to classify the ligand atoms or chemical groups into the following categories: hydrophobic, aromatic, amide, donor, acceptor, cation, anion, or halogen. The protein atom assignment was conducted using a predefined mapping of the standard PDB atom names (Table S1†).

Prior to the assessment of non-covalent interactions, proteins and ligands were protonated (including any implicit hydrogens). The protein protonation was performed in a similar manner to EquiBind and DiffDock by using the reduce software in order to account for hydrogen bonding correctly. Additionally, we considered possible alternative positions of hydrogens, such as within hydroxyl or amine groups. In this work, we considered only amino acids as the entities interacting with the ligands, while the water molecules, ions and metal atoms, which are present in the experimental structures, are disregarded. This limitation could be addressed in the next versions of our technique as detailed in the Discussion. The source code of the preprocessing module is available: **https://github.com/vtarasv/rai-chem.git**.

The choice of non-bond interactions

The hydrophobic and electrostatic interactions as well as the hydrogen bonding (favorable non-bond interactions) were assessed between the protein and the ligand. We also accounted for unfavorable interactions, such as donor–donor atom pairs in close proximity, to improve the overall quality of artificial binding pockets by omitting such interactions. The summary of all used non-bond interactions is shown in Table S4.† The choice of included interactions is based on a compromise between the multiple approaches in the literature,[16–19] commonly used cheminformatics software[20–22] and widely-used molecular modeling tool BIOVIA Discovery Studio Visualizer.

Ligand–protein interaction statistics

The protein–ligand complexes with known 3D structures were taken from the PDBbind dataset.[9,10] Only the entries that satisfy the following criteria were used: the presence of a single small molecule ligand, resolution below 2.5 Å, and an activity/affinity less than 10 μM. A total of 10 270 protein–ligand complexes were selected. Among them 1805 ligand files were found to be unreadable, resulting in a final count of 8465 complexes that were used in this work.

The following statistical information was extracted:

- The probability of a specific ligand substructure (particular atom type, aromatic ring, or amide group) to participate in a protein–ligand interaction.
- The distribution of the number of the binding pocket substructures that are connected to ligand substructures through a specific interaction type.
- The distribution of amino acids involved in particular interaction types.
- The distributions of distances and angles involved in particular interaction types.

Artificial pockets generation

We utilized the PeptideBuilder package[23] to produce a collection of 20 amino acids in the PDB format, as well as 400 dipeptides representing each possible permutation of two amino acids. The amino acids and dipeptides were flanked by GLY residues on both sides and served as the basic building blocks for artificial pockets. The inclusion of flanking GLY residues helps in the generation of the correct peptide conformers, which are restrained by the peptide bonds on each side of the building blocks.

Artificial pocket generation is an iterative process of placing the building blocks around a small molecule in order to form a realistic network of non-covalent interactions. A total of 13 potential interactions, both directed and undirected, were taken into account during the pocket construction (Table 1).

Table 1 Non-bonded interactions taken into account during artificial pocket construction

#	Pocket feature	Ligand feature	Interaction type
1	Aromatic ring	Aromatic ring	Pi stacking
2	Amide group	Aromatic ring	Amide–pi
3	Aromatic ring	Amide group	Amide–pi
4	Aromatic ring	Cationic atom	Cation–pi
5	Hydrogen bond donor	Hydrogen bond acceptor	Hydrogen bond
6	Hydrogen bond acceptor	Hydrogen bond donor	Hydrogen bond
7	Hydrogen bond acceptor	Halogen atom	Halogen bond
8	Cationic atom	Anionic atom	Electrostatic
9	Anionic atom	Cationic atom	Electrostatic
10	Cationic atom	Aromatic ring	Cation–pi
11	C or S atom	F atom	Hydrophobic
12	C or S atom	Cl, Br or I atom	Hydrophobic
13	C or S atom	C or S atom	Hydrophobic

The pocket construction starts from the particular small molecule conformer (referred as ligand hereafter). For each of the 13 interactions listed in Table 1, the following steps are performed:

1. Find all the features of ligand L, which are compatible with the current interaction type i. Each such feature is denoted as F_{Li}.

2. Given experimental probability P of finding F_{Li} among all interactions of type i and the random number p, determine whether the new interaction should be added if $p < P$.

3. Select the peptide building block B to be placed by taking all building blocks with the features compatible with i and randomly selecting one of them according to the experimentally determined probability of the corresponding residue to participate in the interaction i.

4. For the chosen building block B, generate a random conformer taking into account the peptide bonds to flanking GLY residues. Then delete the flanking GLY residues.

5. Randomly sample the distance d between the F_{Li} and the matching feature of B from experimentally determined distributions and place the building block at a determined location.

6. Repeat the following steps until no steric clashes or unfavorable interactions are found between the building block and the ligand and between the building blocks:

a. Randomly rotate and translate the building block B preserving the distance d.

b. Determine whether the angular criteria of interaction i are met, if applicable.

c. If the maximal number of tries (2000 by default) is reached, the building block is skipped.

Using this algorithm 8465 artificial pockets were generated (one for each ligand from the PDBbind database). The distributions of the non-bond interactions of the generated pockets were computed and compared to the experimental distributions. Due to a good match of the distributions, no further tuning of the algorithm was required.

The examples of randomly selected artificial pockets can be found in the ESI (Fig. S1†).

Model training and testing datasets

In order to cover the maximal diversity of the ligands we employed the ZINC20 database of commercially available chemicals widely used for virtual screening.[24,25] The "In-Stock" category of chemicals was chosen, resulting in a collection of 13 million molecules represented by SMILES. The compounds were standardized using the ChEMBL Structure Pipeline[26] (**https://github.com/chembl/ChEMBL_Structure_Pipeline**). Particularly, we eliminated duplicates, compounds with less than seven heavy atoms, and large molecules with a molecular weight exceeding 750 g mol^{-1}. This resulted in 9 041 707 compounds. A subset of compounds (size depends on the model training settings) was randomly selected and used for artificial pockets generation. The pocket generation was repeated for each model training epoch.

In addition to the artificial pockets, the real binding pockets from the PDBbind database were added to the training set. These were defined as the residues with at least one heavy atom within a 5 Å from any heavy atom of the ligand. We compared the model training results achieved with the artificial pockets only, experimental pockets only, and a combination of both.

A subset of PDBbind complexes, which had been previously used in the analysis of DiffDock technique,[8] was used for model testing. We additionally removed all complexes containing the ligands with more than 100 heavy atoms in order to concentrate on the drug-like molecules.

Model performance metrics

In a manner consistent with the EquiBind,[4] TANKBind,[3] and DiffDock[8] we used 25th, 50th, and 75th percentiles of symmetry-corrected[27] root mean square deviation (RMSD) between expected and predicted ligand pose, together with the percentage of predictions below 5 Å or 2 Å RMSD. The centroid distances between the expected and predicted positions of the ligands were tracked using the same metrics.

The following additional metrics of the non-bond interactions were used:

- The fraction of favorable contacts (F_{fav}) – total number of favorable interactions normalised by the number of ligand heavy atoms. The contribution of the hydrophobic interactions was accounted for with a weight of 1/6 due to their high relative abundance.
- The fraction of atoms involved in favorable contacts ($F_{fav\text{-}atoms}$) – a fraction of the heavy ligand atoms participating in at least one non-bond interaction.
- The fraction of unfavorable contacts (F_{unfav}) – total number of unfavorable interactions normalised by the number of ligand heavy atoms.

In order to evaluate the quality of predicted ligand poses, we determined the frequency of steric clashes between the ligand and protein atoms. The clash was defined as the distance between the atoms smaller than 70% of the sum of their respective van der Waals radii.

Model training and inference

The core architecture utilized in this study was the diffusion generative score model, which was adapted from DiffDock. This model is based on SE(3)-equivariant convolutional networks that operate on point clouds.[28,29] The model utilizes pocket and ligand graph representations and takes into account the spatial arrangement of the atoms. It produces SE(3)-equivariant vectors that describe the ligand's translations and rotations, along with SE(3)-invariant scalars per each compound's rotatable bond. The input position of a ligand is subsequently altered based on the model's output resulting in the final conformation of the molecule within a binding site. The torsion angles of the ligand are optimized as well as translations and rotations of the whole molecule. During the training phase, the position of each input ligand in a complex undergoes modifications caused by translational, rotational, and torsional noise, which can be referred to as forward diffusion. The model then learns how to reverse the diffusion process. This method enables the generation of numerous alternative ligand poses during the inference process.[8]

We adjust the DiffDock model inputs and architecture as follows:

- The atomic-level pocket graph is used instead of the residue-level graph of the whole protein.
- The categorical feature space of the ligand nodes was decreased significantly by narrowing down the number of atom types, number of neighbor heavy atoms, and atomic charges to those expected in the typical screening databases of small molecules.

The learning rate was set to 0.0125 based on initial tuning. During each epoch, the model is trained for 10 000 iterations with a batch size of 4. In our experimental conditions, the batch always consists of identical pocket and ligand components, whereas the level of diffusion noise varies between individual samples.

The models were trained for 25 epochs. During the training based on artificial data, 10 000 ligands are randomly sampled from the preprocessed ZINC dataset at each epoch. After that, the artificial pockets are created for each ligand. Thus, 250 000 unique artificial complexes were generated for the model training. Each epoch of training on experimental protein–ligand complexes is performed with 10 000 randomly sampled PDBbind entries (out of 16 379 complexes in the train split). The training on a combined dataset was performed with a 4 : 1 ratio of artificial and experimental data with 200 000 unique artificial complexes. The final production model was trained for 80 epochs using a combined dataset with 640 000 unique artificial complexes. The model training, which is a GPU-intensive task, and pocket generation, which is a CPU-intensive task, were separated into distinct parallel workflows.

Given the generative nature of the model, it is possible to produce an infinite number of alternate ligand poses. The real-world model performance is thus sensitive to the scoring function, which is used for the pose ranking and selection. The developers of DiffDock employed a confidence model that takes into account all protein atoms and produces a confidence score of the ligand pose. In contrast, we employed a non-covalent interaction score in the binding pocket, which reduces the inference cost significantly. Our scoring function S is computed as follows:

$$S = F_{\text{fav}} + F_{\text{fav-atoms}} - 2F_{\text{unfav}} - 2F_{\text{unfav-atoms}} - 10D_{\text{clash}}$$

where $F_{\text{unfav-atoms}}$ is a fraction of the ligand heavy atoms participating in at least one unfavorable interaction, D_{clash} is a sum of all distances, which are below the steric clashes threshold. The coefficients of the scoring function were adjusted empirically by visual inspection of the predicted protein–ligand complexes.

Results and discussion

Statistics of non-bond interactions in experimental and artificial protein–ligand complexes

Analysis of the high-quality PDBbind protein–ligand dataset revealed a total of 343 784 intermolecular non-covalent interactions. As anticipated, the hydrophobic contacts were the predominant kind of interaction, with a total of 267 774 atom pairs observed. The overwhelming majority of these interactions occurred between hydrophobic carbon or sulfur atoms. The hydrogen bonds were the second most prevalent form of interaction, occurring around once for every five hydrophobic pairs. The remaining contacts constituted less than 3% of the overall count (Fig. 2 and Table S2†).

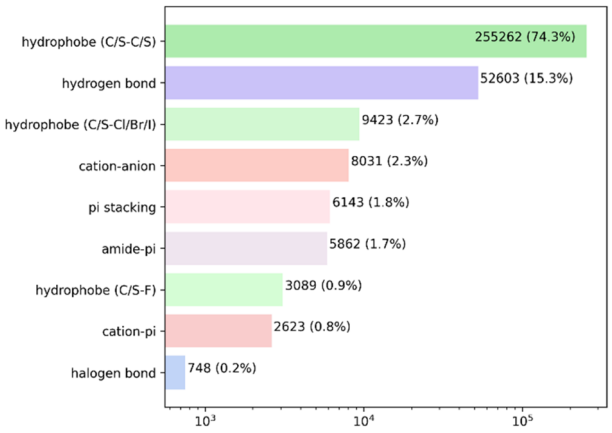

Fig. 2 Probability distribution of the non-bond protein–ligand interactions in the PDBbind complexes. Every bar is labeled with the total number of interactions and its fraction. Note the log scale of the X-axis.

The total number of non-bond interaction entries in the artificial pockets involving the same ligands amounted to 453 593. This significantly larger amount, in comparison to the real ones, is caused by formations of "unintended" interactions (mainly hydrophobic) during the placement of the pocket building blocks in the close proximity to the ligand. The pocket generation algorithm underestimates the solvent exposure contribution to the experimental interaction statistics because the ligand–water interactions are not taken into account during the pocket generation. This is also likely to produce additional interactions with the protein, which partially substitute the ligand–solvent interactions.

We compared the distributions of the parameters for each type of interaction between real and artificial pockets. For each interaction type, we also computed the probability of the involvement of particular amino acids in the binding pocket.

Fig. 3 shows the statistics of hydrophobic interactions in real and artificial binding pockets.

There is a reasonably good correspondence between the distributions, but the distributions of distances are systematically smoother and more monotonous for artificial pockets in comparison to real ones. This is an artifact caused by the frequent formation of "unintended" hydrophobic contacts while generating other types of interactions due to the abundance of hydrophobe atoms in both amino acids and small molecules. The distributions of interactions among amino acids are very similar except the notably increased involvement of PHE and TRP in artificial pockets. This is explained by the overlap between hydrophobic and pi stacking interactions, which is not checked in the algorithm for the sake of computational efficiency.

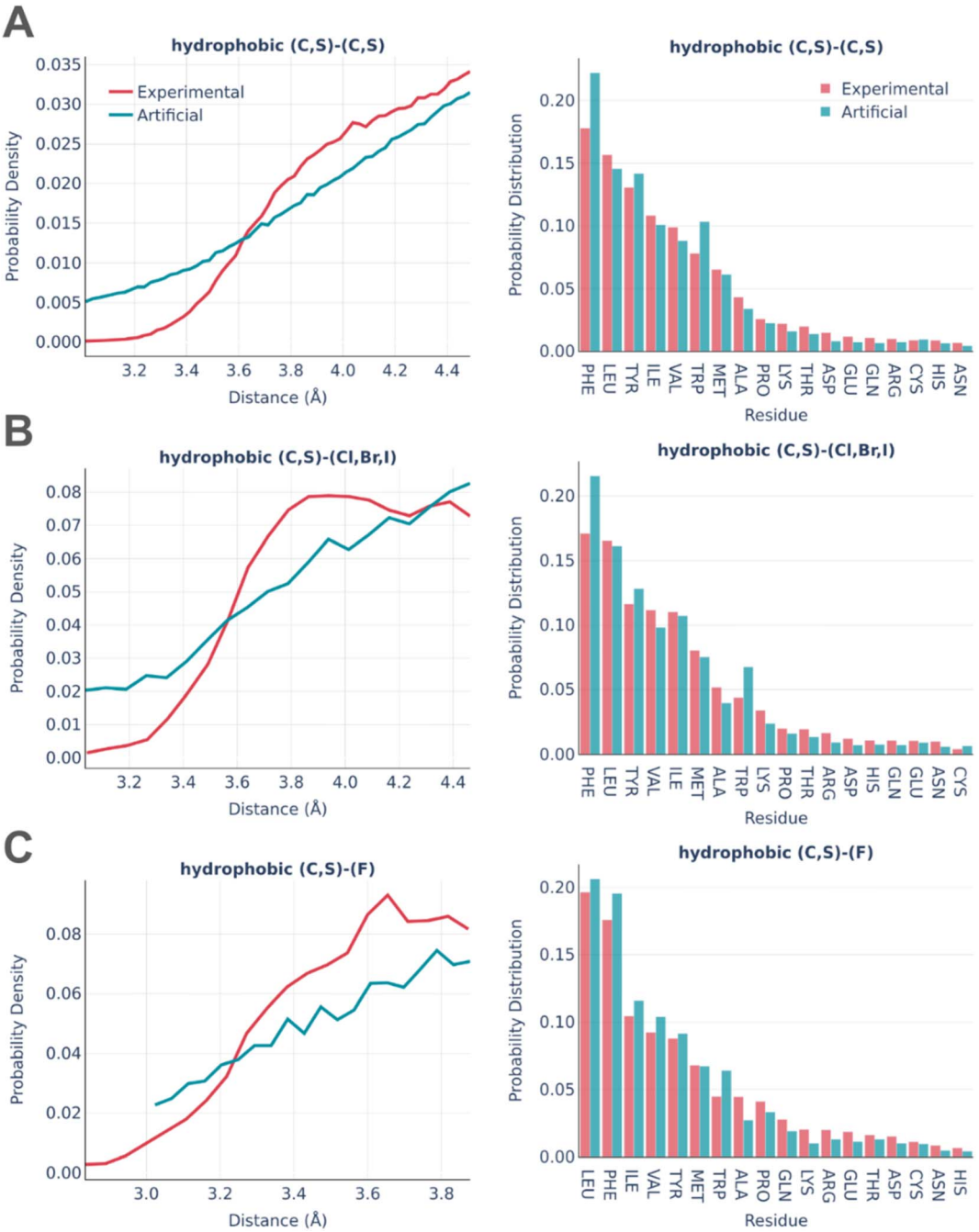

Fig. 3 Statistics of hydrophobic interactions in the PDBbind and artificial pockets involving (A) C or S atoms of the protein and the ligand (B) Br, Cl or I atoms of the ligand (C) F atoms of the ligand. The distance distributions are shown in the left columns and the pocket residues occurrence ratio is shown in the right column.

Fig. 4 shows the statistics of pi stacking interactions. There are two types of these interactions: parallel (true pi stacking) and T-shaped (aromatic–pi interactions). The distance distributions of both types of interactions are remarkably similar in real and artificial pockets as well as the involvements of different aromatic amino acids. However, the theta angle distributions of parallel interactions for artificial pockets are shifted toward 90° by 10–15° in comparison to experimental ones since the generation algorithm only checks if the aromatic ring has an angle within a given range.

Fig. 5 shows the statistics of amide–pi interactions, which could also be classified into parallel and T-shaped.

Similarly to pi stacking interactions, the distance distributions are very similar between real and artificial pockets, while the theta angles in artificial pockets are somewhat shifted towards 90°. The primary residues serving as donors of the amide group were found to be glycine, which exposes the peptide bond due to the lack of a side chain, as well as asparagine and glutamine, which possess an amide group at the terminus of their side chains. The predominant residue bearing

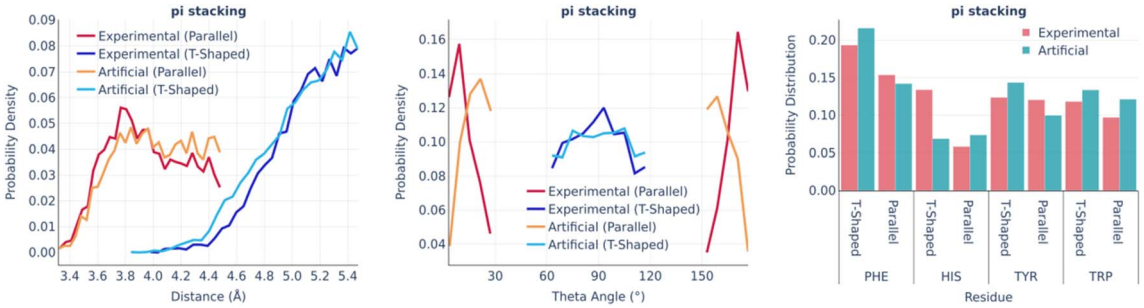

Fig. 4 Pi stacking: the distance distribution (left), theta angle distribution (middle), and pocket residues frequency (right) in the PDBbind and artificial pockets.

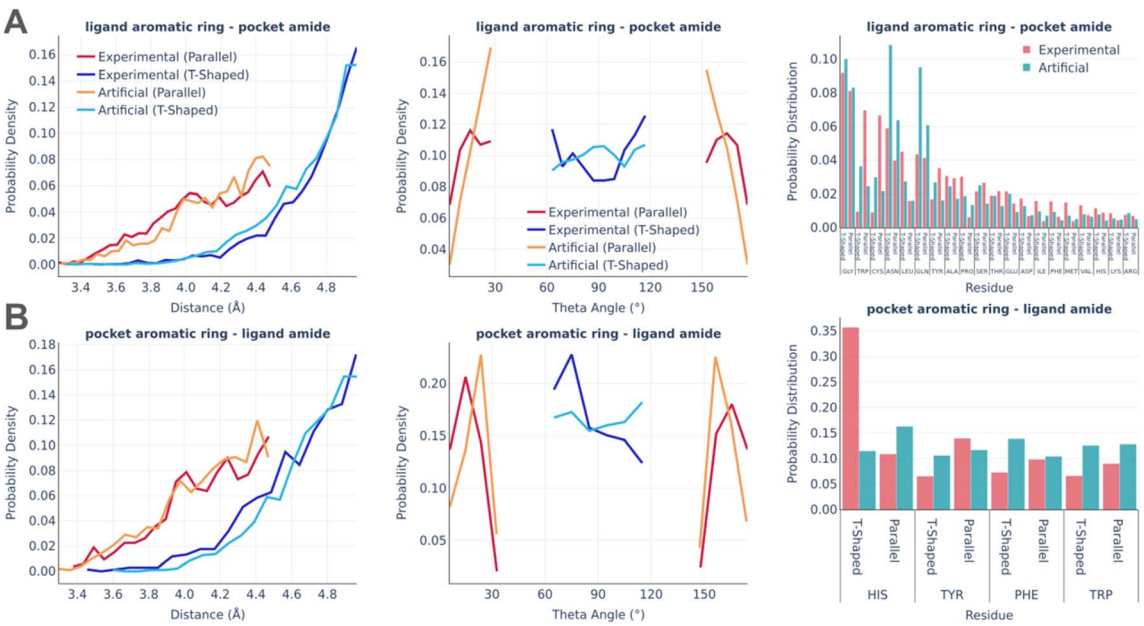

Fig. 5 Amide–pi interactions: the distance distribution (left), theta angle distribution (middle), and pocket residues frequency (right) in the PDBbind and artificial pockets for the ligand aromatic ring – pocket amide contacts (A) and pocket aromatic ring – ligand amide contacts (B).

an aromatic ring in the amide–pi linkages is HIS, which accounts for the majority of experimentally observed T-shaped interactions. There are significantly less HIS contacts in artificial pockets because unfavorable donor–donor contacts between the nitrogens of the HIS ring and amide group were omitted during the pocket generation.

The distributions of the hydrogen bonds are shown in Fig. 6. Hydrogen bonding imposes the biggest challenge in terms of artificial pocket generation. It is clearly seen there are the biggest discrepancies between real and artificial pockets in terms of the hydrogen bonds' parameters, which are especially visible for the angle distributions. In general, our algorithm tends to underestimate the D–H–A angles (the significance of the angle distribution was deliberately reduced to speed up the algorithm) and doesn't capture the distance peak at 2.8–2.9 Å, which is observed in real pockets, while generating significantly more long h-bonds (short h-bonds tend to be discarded more often during the generation as they often lead to steric clashes between the atoms not participating in the interaction). These compromises allow us to keep the algorithm fast enough for routine practical usage. At the same time, the involvement of different amino acids is reproduced remarkably well in artificial pockets.

The halogen bonds are the least frequent type of interaction in real binding pockets. Their distributions are shown in Fig. 7. Taking into account the small number of observed interactions of this kind, the experimental and artificial distributions are sufficiently similar to each other.

The electrostatic interactions were analyzed separately for the “ligand anion – pocket cation” and “pocket anion – ligand cation” pairs (Fig. 8).

The experimentally observed involvement of charged amino acids is reproduced almost ideally in artificial pockets, while the distance distributions are somewhat different in real and artificial pockets. Artificial pockets possess more iterating pairs at larger distances than real ones, which could be explained by the

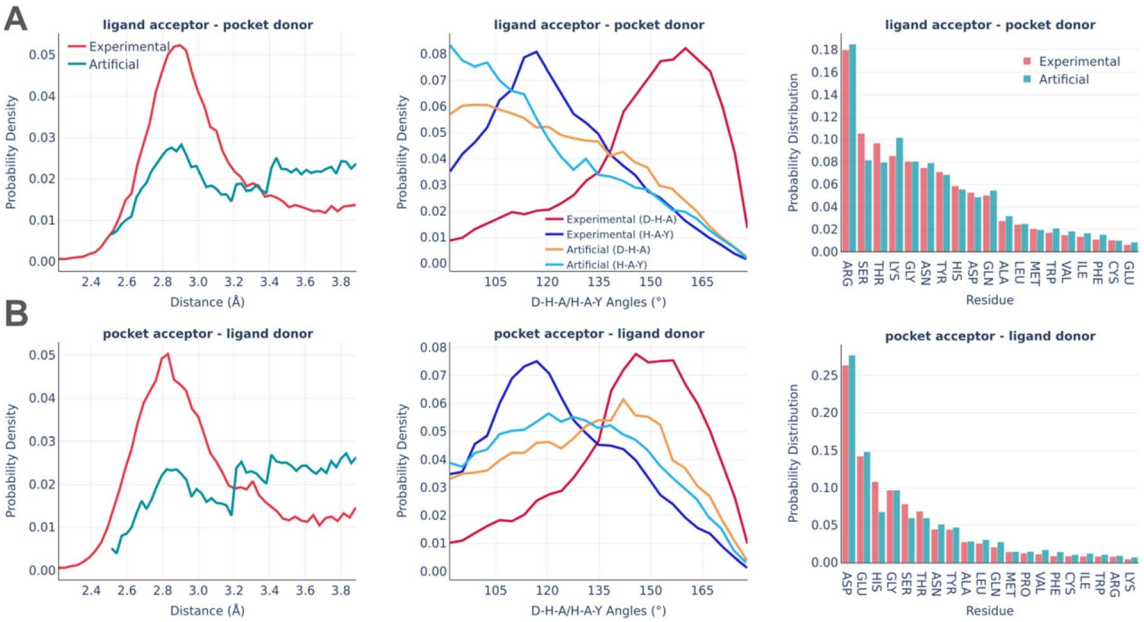

Fig. 6 Hydrogen bonds: the distance distribution (left), D–H–A/H–A–Y angles distribution (middle), and pocket residues frequency (right) in the PDBbind and artificial pockets for the ligand acceptor – pocket donor contacts (A) and pocket acceptor – ligand donor contacts (B). In the D–H–A/H–A–Y angles, D is the donor atom covalently bound to the hydrogen; H is a hydrogen atom; A is an acceptor atom; Y is a heavy covalently bound neighbor of the acceptor atom.

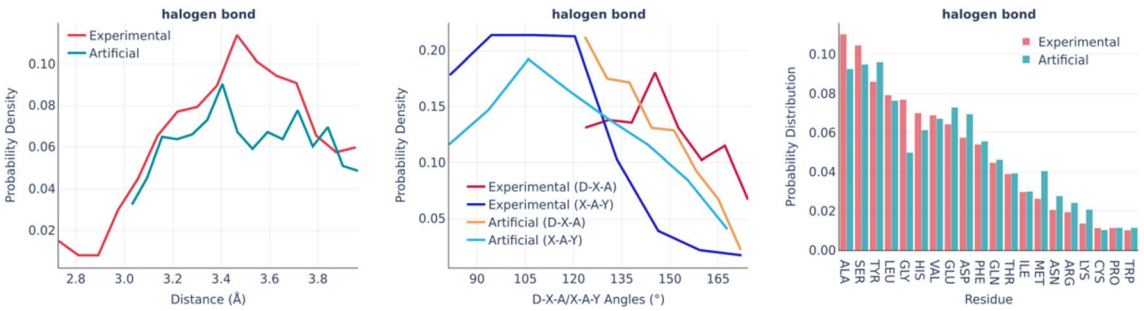

Fig. 7 Halogen bonds: the distance distribution (left), D–X–A/X–A–Y angles distribution (middle), and pocket residue frequencies (right) in the PDBbind and artificial pockets. In the D–X–A/X–A–Y angles, D is the donor atom covalently bound to the halogen; X is a halogen atom; A is the halogen bond acceptor atom; Y is a heavy covalently bound neighbor of the acceptor atom.

increased complexity of fitting pocket residues in the close proximity of a small molecule. Such placement often leads to steric clashes and thus is often discarded by the algorithm.

The last type of interaction is cation–pi pairs (Fig. 9). They are rather minor and the distributions of their parameters are very similar in real and artificial pockets without any significant feature worth commenting on.

Impact of artificial data on model performance

To evaluate the potential influence of data augmentation using artificial protein–ligand complexes on model performance, we compared three models. The first model was exclusively trained on experimental protein–ligand complexes, the second model was trained on the artificially generated pockets with corresponding small molecules, and the third model was trained on the combination of both. The metrics are reported for the top-1 predicted complex and the best of the top-5 predicted complexes based on custom scoring function S (see the Methods section).

Table 2 illustrates that the inclusion of artificial or combined data in the training process led to enhancements in both the RMSD and centroid distance metrics, as compared to the model trained exclusively on experimental complexes. Unexpectedly, the training of the artificial dataset is superior to the combined dataset in certain metrics, particularly when the generation of 40 poses for each pocket was used. We hypothesize that this might be caused by ignoring the solvent exposure contribution in the pocket generation algorithm, which results in an exaggerated number of the protein–ligand interactions and the tighter spatial constraints in the final ligand pose. In addition, we noticed multiple experimental complexes with the steric clashes (*e.g.* 1gj4, 5yr6) or even covalent bonds (*e.g.* 3s3q, 6eyz)

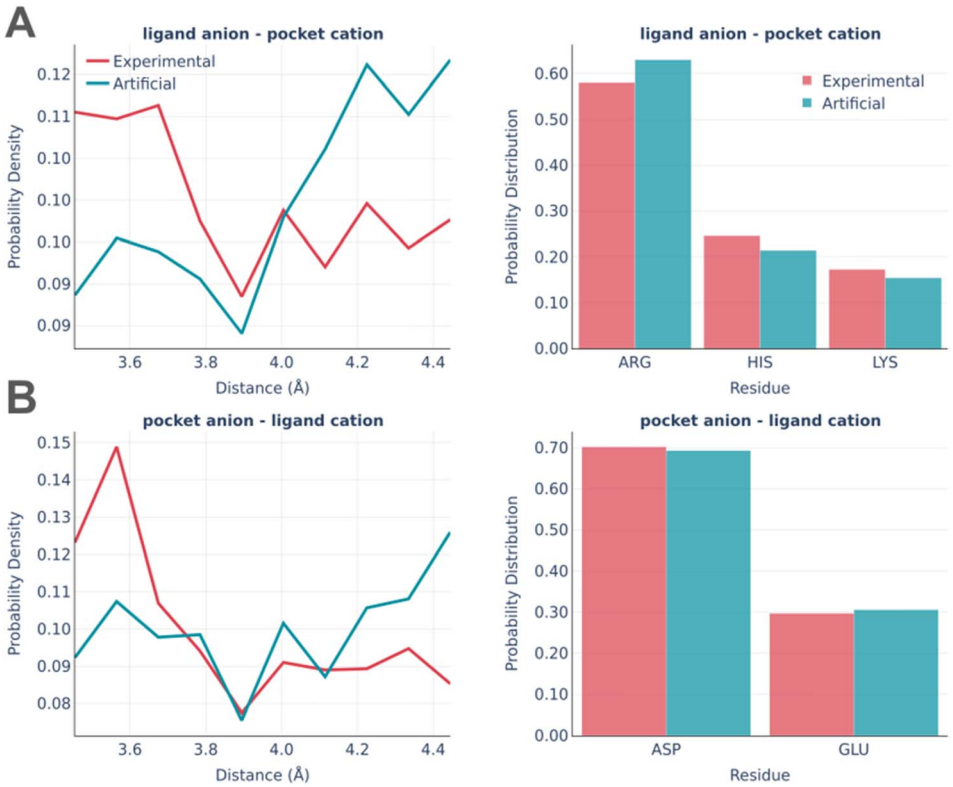

Fig. 8 Electrostatic interactions: the distance distribution (left) and pocket residues frequency (right) in the PDBbind and artificial pockets for the ligand anion – pocket cation (A) and pocket anion – ligand cation (B).

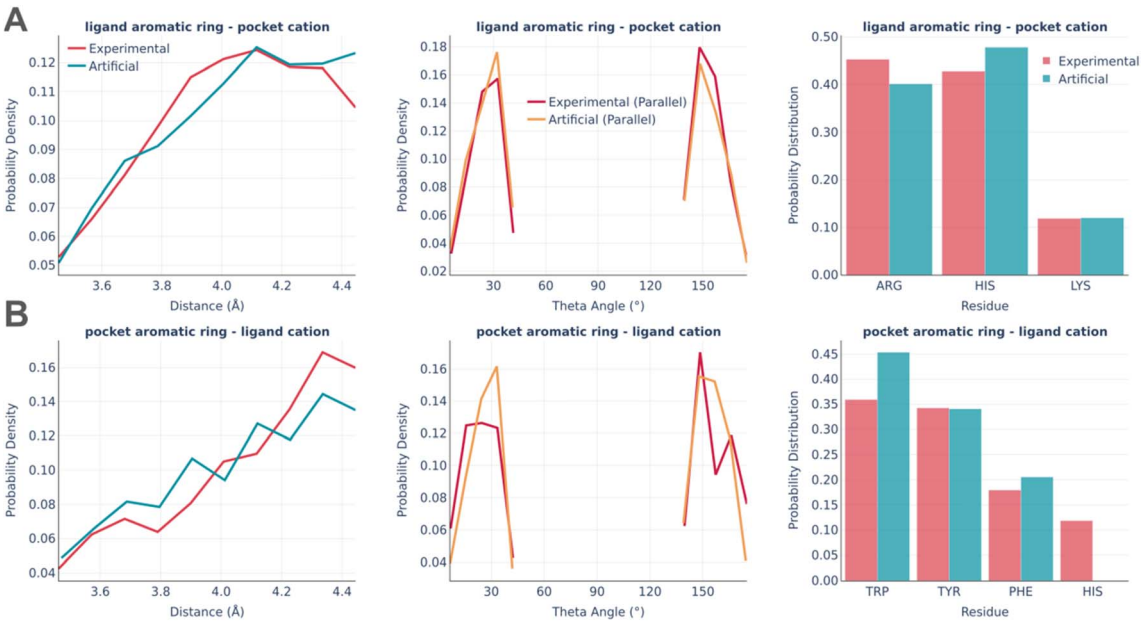

Fig. 9 Cation–pi interactions: the distance distribution (left), theta angle distribution (middle), and pocket residues frequency (right) in the PDBbind and artificial pockets for the ligand aromatic ring – pocket cation (A) and pocket aromatic ring – ligand cation (B).

between the protein and the ligand. Such artifacts constitute nearly 5% of the PDBBind training data, which may contribute to confusing the model concerning valid physical distance constraints.

The combined dataset of protein–ligand complexes exhibited superior performance compared to the other two datasets in terms of both favorable and unfavorable non-covalent interactions, as illustrated in Table 3. Additionally, it was observed that

Table 2 The symmetry-corrected root mean square deviation (RMSD) and centroid distance metrics between predicted and real ligand positions in the PDBbind test complexes. The models generated either 10 or 40 ligand poses. The best of top 5 poses is picked based on the lowest RMSD. ↑ – means the higher the better; ↓ – means the lower the better

	RMSD					Centroid distance				
	Percentiles ↓			% Below threshold ↑		Percentiles ↓			% Below threshold ↑	
Dataset	25th	50th	75th	5 Å	2 Å	25th	50th	75th	5 Å	2 Å
10 samples, top 1										
Experimental	4.08	5.51	7.45	40.93	5.02	1.02	1.75	2.7	94.59	57.92
Artificial	3.69	5.75	7.39	40.15	5.02	1	1.46	2.19	97.68	69.88
Combined	3.76	5.13	7.12	45.95	7.34	1.05	1.69	2.33	98.07	63.32
10 samples, top 5										
Experimental	2.68	3.72	4.64	82.24	11.2	0.95	1.37	1.97	99.61	75.29
Artificial	2.49	3.47	4.6	81.08	16.99	0.78	1.24	1.86	98.84	80.31
Combined	2.4	3.31	4.55	85.33	17.37	0.81	1.29	1.88	99.23	77.99
40 samples, top 1										
Experimental	3.89	5.54	7.36	40.93	4.63	1.01	1.73	2.48	96.14	57.92
Artificial	3.47	5.46	7.25	47.1	7.72	1	1.39	2.09	97.68	72.59
Combined	3.86	5.2	7.16	45.95	6.18	1.05	1.57	2.42	97.68	64.09
40 samples, top 5										
Experimental	2.37	3.35	4.6	82.24	15.06	0.85	1.19	1.69	99.61	81.08
Artificial	1.99	3.08	4.3	85.33	25.1	0.73	1.11	1.69	99.61	82.63
Combined	2.09	3.3	4.6	82.24	23.94	0.8	1.26	1.99	98.84	75.29

Table 3 The favorable rate, favorable rate uniq, and unfavorable rate for the predicted PDBbind test complexes. The models generated either 10 or 40 samples. The best of the top 5 candidates are picked based on the lowest RMSD. ↑ – means the higher the better; ↓ – means the lower the better

	Favorable rate[a]			Favorable rate uniq[b]			Unfavorable rate[c]		
	Percentiles ↑			Percentiles ↑			Percentiles ↓		
Dataset	25th	50th	75th	25th	50th	75th	25th	50th	75th
10 samples, top 1									
Experimental	0.27	0.37	0.51	0.38	0.53	0.67	0	0.08	0.17
Artificial	0.27	0.39	0.53	0.38	0.53	0.67	0	0.07	0.15
Combined	0.27	0.38	0.54	0.41	0.54	0.68	0	0.07	0.15
10 samples, top 5									
Experimental	0.28	0.38	0.52	0.35	0.53	0.66	0.04	0.1	0.23
Artificial	0.3	0.39	0.52	0.38	0.52	0.68	0.04	0.1	0.21
Combined	0.3	0.4	0.56	0.39	0.55	0.67	0.03	0.09	0.18
40 samples, top 1									
Experimental	0.29	0.37	0.54	0.4	0.54	0.71	0	0.06	0.14
Artificial	0.29	0.41	0.58	0.42	0.58	0.71	0	0.04	0.12
Combined	0.31	0.43	0.59	0.42	0.58	0.71	0	0.05	0.12
40 samples, top 5									
Experimental	0.26	0.4	0.54	0.39	0.53	0.69	0.03	0.09	0.18
Artificial	0.29	0.41	0.55	0.39	0.54	0.7	0	0.07	0.14
Combined	0.32	0.43	0.55	0.41	0.56	0.71	0	0.06	0.14

[a] Total number of favorable interactions normalized by the number of ligand heavy atoms. [b] A fraction of the heavy ligand atoms participating in at least one non-bond interaction. [c] Total number of unfavorable interactions normalized by the number of ligand heavy atoms.

the integration of the combined data for training purposes led to a decrease in the proportion of samples exhibiting steric clashes, as shown in Table 4.

A positive correlation was identified between the amount of small molecule heavy atoms and the metrics such as RMSD and steric clashes. Additionally, a negative correlation was found between the count of atoms and favorable non-bond interaction rates, as depicted in Fig. S2.† This indicates that the model performance decreases with the increase of ligand size.

The final production PocketCFDM model was trained for 80 epochs using high-quality experimental data (used for the non-bond interactions statistics retrieval) and artificial samples, similar to the aforementioned combined dataset settings. A significant reduction in the occurrence of protein–ligand steric clashes was observed, with percentages of 19.31% and 13.90% for the top-1 and best of top-5 samples, respectively. This is a nearly 10% decrease compared to the most optimal model trained for 25 epochs (Table 4). There has been no substantial improvement in other metrics.

Table 4 The percentage of samples within the predicted PDBbind test complexes exhibiting at least one steric clash. The models generated 40 ligand poses. The best of top-5 candidates are picked based on the lowest number of clashes

	Steric clashes	
Dataset	Top 1	Top 5
Experimental	33.59%	30.12%
Artificial	30.50%	26.64%
Combined	28.19%	23.94%

Table 5 Comparison between the DiffDock and PocketCFDM on the test PDBbind dataset

Metric	PocketCFDM	DiffDock
Avg. inference time (s); 40 samples	49	90
RMSD; median; 40 samples; top-1	5.14	2.8
RMSD; median; 40 samples; top-5	3.02	2.17
Centroid distance; median; 40 samples; top-1	1.4	1.01
Centroid distance; median; 40 samples; top-5	1.08	0.81
Favourable rate; median; 40 samples; top-1	0.43	0.37
Favourable rate; median; 40 samples; top-5	0.42	0.39
Favourable rate uniq; median; 40 samples; top-1	0.58	0.47
Favourable rate uniq; median; 40 samples; top-5	0.56	0.5
Unfavourable rate; median; 40 samples; top-1	0.02	0.07
Unfavourable rate; median; 40 samples; top-5	0.07	0.08
Steric clashes; 40 samples; top-1	19.31%	46.51%
Steric clashes; 40 samples; top-5	13.90%	25.97%

Comparison with other methods

We also performed a detailed comparison between PocketCFDM and DiffDock, which is currently considered as the most accurate AI technique for predicting ligand binding poses (Table 5).

DiffDock exhibited superior accuracy in terms of RMSD and centroid distance. In contrast, PocketCFDM exhibited superior outcomes in terms of favorable and unfavorable non-covalent interactions, as well as a notably lower incidence of steric clashes. It was expected due to the divergence in the scoring algorithms employed to evaluate and rank the generated samples. The DiffDock approach utilized a confidence model that was trained to prioritize samples with lower RMSD to the actual ligand pose. However, in our context, the scoring approach was more focused on the number of non-covalent contacts and the absence of steric clashes, without considering the specific conformation of the ligand pose. Another notable difference pertaining to the disparity in scoring functions is the diversity observed within anticipated poses. The median RMSD values between the alternative top-5 samples in the PocketCFDM and DiffDock methods were 2.59 and 1.29 respectively. Additionally, it was noted that the mean inference time for the PocketCFDM was approximately 1.8 times quicker. This is attributed mostly to the reduced size of the protein graph and the decrease in the graph's node feature space.

PocketCFDM differs from DiffDock in both the scoring function and the training dataset. DiffDock is trained on the residue-level full protein graphs, while PocketCFDM utilizes the atomic-level graphs limited to the binding pockets (real or artificial, see the "Model training and inference" section above). The training dataset for PocketCFDM is augmented by artificially generated pockets, while DiffDock is trained on the experimentally resolved protein–ligand complexes only. Thus observed better performance of PocketCFDM originates from both of these factors.

Additionally, we compared PocketCFDM performance to conventional molecular docking methods such as Autodock Vina,[30] QuickVina-W,[31] GNINA,[32] SMINA,[33] and GLIDE.[34] The results of traditional docking programs on the PDBbind dataset were reported by the DiffDock authors.[8] According to Table 6, PocketCFDM performs comparably to the conventional docking methods. It is worth mentioning that the ML-based methods don't guarantee the absence of inter-/intramolecular clashes in

Table 6 Comparison between the PocketCFDM, DiffDock, and conventional docking techniques on the test PDBbind dataset. The ML-based method results are reported for the best of the top 5 pose candidates among 40 predicted samples. ↑ – means the higher the better; ↓ – means the lower the better

	RMSD					Centroid distance				
	Percentiles ↓			% Below threshold ↑		Percentiles ↓			% Below threshold ↑	
Method	25th	50th	75th	5 Å	2 Å	25th	50th	75th	5 Å	2 Å
Autodock Vina (top-1)	5.7	10.7	21.4	21.2	5.5	1.9	6.2	20.1	47.1	26.5
QuickVina-W (top-1)	2.5	7.7	23.7	40.2	20.9	0.9	3.7	22.9	54.6	41
GNINA (top-1)	2.4	7.7	17.9	40.8	22.9	0.8	3.7	23.1	53.6	40.2
SMINA (top-1)	3.1	7.1	17.9	38	18.7	1	2.6	16.1	59.8	41.6
GLIDE (top-1)	2.6	9.3	28.1	33.6	21.8	0.8	5.6	26.9	48.7	36.1
GNINA (top-5)	1.6	4.5	11.8	52.8	29.3	0.6	2	8.2	66.8	49.7
SMINA (top-5)	1.7	4.6	9.7	53.1	29.3	0.6	1.85	6.2	72.9	50.8
DiffDock	1.2	2.4	5	75.5	44.7	0.4	0.9	1.9	88	76.7
PocketCFDM	2.1	3.3	4.6	82.2	23.9	0.8	1.3	2.0	98.8	75.3

the docked complexes while the traditional approaches mostly provide physically valid poses.

Limitations and perspectives

The primary area of enhancement of the PocketCFDM model lies in inference time, which is currently still too large for efficient practical deployment. The mean time required to predict 40 ligand poses is around ~50 seconds at 24 GB NVIDIA L4 GPU. The inference time could be improved significantly by replacing computationally intensive EGNN blocks with more efficient alternatives like GVP. The inclusion of the artificial pockets into the training of even simpler regression-based techniques, such as EquiBind and TANKBind, could also be beneficial since these architectures are generally more sensitive to the number of distinct training samples. It is possible to explore the usage of multiple augmentations of the same input instead of directly including specific symmetries and equivariances into graph neural network (GNN) designs. The utilization of such augmentation is popular in convolutional neural networks (CNNs). Moreover, it demonstrated promising outcomes in the domain of geometric graph learning.[35] This strategy may potentially increase the model training time while resulting in reduced inference time.

Another notable drawback of the present proof-of-the-principle implementation is the possibility of ligand self-intersections (intramolecular clashes), which have to be filtered out during the post-processing steps. Incorporating intramolecular and intermolecular clashes into the loss function during the training process could potentially address this problem.

Also, we believe that incorporating larger ligands into the training process will address the challenges encountered with relatively large compounds. The ZINC20 dataset, which was utilized in this work, has only 2.5% of molecules larger than 40 heavy atoms, which makes them underrepresented during the model training. Although the median size of the ligands in this dataset is 25 heavy atoms, which is quite common for the datasets of drug-like molecules, a higher percentage of large ligands may enhance the inference capabilities for larger compounds while maintaining the same level of performance for smaller molecules.

It should be pointed out that the current iteration of the pocket generation algorithm doesn't consider water molecules, ions, and metal atoms, which are known to be important mediators of interactions in a significant amount of protein–ligand complexes. Thus currently our technique should be used with caution in the cases when the involvement of water, ions, and the metal atoms in the ligand binding is anticipated. We assume that the inclusion of these components into the artificial pockets in the next iterations of our technique should further improve the model performance and universality.

Another shortcoming of the current pocket generation algorithm is the frequent formation of "unintended" contacts (mainly hydrophobic) while generating other types of interactions, which is evident from the disbalance between the number of non-bond interactions found in the experimental and artificial pockets. When the pocket building block is added, additional interactions could be formed accidently apart from the intended contact pair. Additional checks could be added in future versions of the algorithm to minimize the amount of such unwanted random interactions.

Although the present work concentrates on predicting the ligand poses, it is necessary to note that the pose prediction is only a part of the accurate prediction of the protein–ligand binding. Two other important components are the accurate scoring of the binding poses in terms of affinity or activity and the accounting for protein flexibility and dynamics. The former is currently being addressed not only by traditional docking force fields but also by ML-based pose rescoring techniques.[36] The latter problem could be tackled by multiple approaches including flexible and ensemble docking. Accounting for the protein flexibility and the ensembles of protein conformations are among the future directions of improvement for our data augmentation technique.

Conclusions

In this work, we introduced the PocketCFDM generative diffusion model for predicting the poses of small molecules in the protein binding pockets. The model is trained using an innovative approach of data augmentation, which involves the construction of a large number of artificial binding pockets that follow the same statistical patterns of non-bond interactions as the real ones. In order to construct such artificial pockets we thoroughly evaluated the statistical characteristics of non-bond interactions in the real protein–ligand complexes and designed an algorithmic approach that reproduces them in artificial pockets, which are built around given small molecule conformers. The performance of the models trained on experimental data only, artificial data only, and a combination of both was evaluated and compared to the currently most promising ML model for binding pose prediction DiffDock. It is shown that the inclusion of artificial binding pockets into the model training resulted in a significant increase of model performance. Particularly, PocketCFDM outperforms DiffDock in terms of the non-bond interaction counts, the number of steric clashes, and the inference speed. The prospects of further improvements of PocketCFDM are discussed. The inference code, final model weights, and model prediction examples are publicly available in the GitHub repository (**https://github.com/vtarasv/pocket-cfdm.git**).

Author contributions

VB, SS and AN designed the study and supervised data quality of generated pocket as well as model development and testing. SY coordinated the work and participated in results interpretation. TV developed the modules for pocket and features generation, model training and testing. TV, IK, and RS researched existing methods for machine-learning-based molecular docking. LP, ZO, and RZ participated in consultations about the best practices and strategies for model tuning and efficient training. IK,

PH, and VV performed and supervised the tuning, training and testing process. The manuscript was written by TV and SY.

Conflicts of interest

All authors but VB are employees of Receptor. AI INC. SS, AN and SY have shares in Receptor.AI INC.

Acknowledgements

SY was supported through the MSCA4Ukraine project, which is funded by the European Union.

References

1 X. Li, Y. Li, T. Cheng, Z. Liu and R. Wang, *J. Comput. Chem.*, 2010, **31**, 2109–2125.
2 J. B. Ghasemi, A. Abdolmaleki and F. Shiri, in *Pharmaceutical Sciences: Breakthroughs in Research and Practice*, IGI Global, 2017, pp. 770–794.
3 W. Lu, Q. Wu, J. Zhang, J. Rao, C. Li and S. Zheng, *bioRxiv*, 2022, preprint, DOI: **10.1101/2022.06.06.495043**.
4 H. Stärk, O.-E. Ganea, L. Pattanaik, R. Barzilay and T. Jaakkola, *arXiv*, 2022, preprint, arXiv:2202.05146, DOI: **10.48550/arXiv.2202.05146**.
5 B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend and R. Dror, *arXiv*, 2021, preprint, arXiv:2009.01411, DOI: **10.48550/arXiv.2009.01411**.
6 B. Jing, S. Eismann, P. N. Soni and R. O. Dror, *arXiv*, 2021, preprint, arXiv:2106.03843, DOI: **10.48550/arXiv.2106.03843**.
7 O.-E. Ganea, X. Huang, C. Bunne, Y. Bian, R. Barzilay, T. Jaakkola and A. Krause, *arXiv*, 2022, preprint, arXiv:2111.07786, DOI: **10.48550/arXiv.2111.07786**.
8 G. Corso, H. Stärk, B. Jing, R. Barzilay and T. Jaakkola, *arXiv*, 2023, preprint, arXiv:2210.01776, DOI: **10.48550/arXiv.2210.01776**.
9 R. Wang, X. Fang, Y. Lu and S. Wang, *J. Med. Chem.*, 2004, **47**, 2977–2980.
10 Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu and R. Wang, *Bioinformatics*, 2015, **31**, 405–412.
11 Ž. H. Petrovski, B. Hribar-Lee and Z. Bosnić, *Pharmaceutics*, 2022, **15**, 119.
12 J. Kandel, H. Tayara and K. T. Chong, *J. Cheminf.*, 2021, **13**, 65.
13 M. Volkov, J.-A. Turk, N. Drizard, N. Martin, B. Hoffmann, Y. Gaston-Mathé and D. Rognan, *J. Med. Chem.*, 2022, **65**, 7946–7958.
14 The UniProt Consortium, *Nucleic Acids Res.*, 2019, **47**, D506–D515.
15 J.-L. Reymond, R. Van Deursen, L. C. Blum and L. Ruddigkeit, *MedChemComm*, 2010, **1**, 30.
16 C. Bissantz, B. Kuhn and M. Stahl, *J. Med. Chem.*, 2010, **53**, 5061–5084.
17 R. Ferreira De Freitas and M. Schapira, *MedChemComm*, 2017, **8**, 1970–1981.
18 R. Wilcken, M. O. Zimmermann, A. Lange, A. C. Joerger and F. M. Boeckler, *J. Med. Chem.*, 2013, **56**, 1363–1388.
19 B. Kuhn, E. Gilberg, R. Taylor, J. Cole and O. Korb, *J. Med. Chem.*, 2019, **62**, 10441–10455.
20 M. Wójcikowski, P. Zielenkiewicz and P. Siedlecki, *J. Cheminf.*, 2015, **7**, 26.
21 H. C. Jubb, A. P. Higueruelo, B. Ochoa-Montaño, W. R. Pitt, D. B. Ascher and T. L. Blundell, *J. Mol. Biol.*, 2017, **429**, 365–371.
22 M. F. Adasme, K. L. Linnemann, S. N. Bolz, F. Kaiser, S. Salentin, V. J. Haupt and M. Schroeder, *Nucleic Acids Res.*, 2021, **49**, W530–W534.
23 M. Z. Tien, D. K. Sydykova, A. G. Meyer and C. O. Wilke, *PeerJ*, 2013, **1**, e80.
24 J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *J. Chem. Inf. Model.*, 2020, **60**, 6065–6073.
25 T. Sterling and J. J. Irwin, *J. Chem. Inf. Model.*, 2015, **55**, 2324–2337.
26 A. P. Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L. J. Bellis, M. De Veij and A. R. Leach, *J. Cheminf.*, 2020, **12**, 51.
27 R. Meli and P. C. Biggin, *J. Cheminf.*, 2020, **12**, 49.
28 N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff and P. Riley, *arXiv*, 2018, preprint, arXiv:1802.08219, DOI: **10.48550/arXiv.1802.08219**.
29 M. Geiger and T. Smidt, *arXiv*, 2022, preprint, arXiv.2207.09453, DOI: **10.48550/arXiv.2207.09453**.
30 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**, 455–461.
31 N. M. Hassan, A. A. Alhossary, Y. Mu and C.-K. Kwoh, *Sci. Rep.*, 2017, **7**, 15451.
32 A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri and D. R. Koes, *J. Cheminf.*, 2021, **13**, 43.
33 D. R. Koes, M. P. Baumgartner and C. J. Camacho, *J. Chem. Inf. Model.*, 2013, **53**, 1893–1904.
34 T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard and J. L. Banks, *J. Med. Chem.*, 2004, **47**, 1750–1759.
35 FAENet, *Frame Averaging Equivariant GNN for Materials Modeling | OpenReview*, **https://openreview.net/forum?id=HRDRZNxQXc**, accessed August 16, 2023.
36 C. Yang, E. A. Chen and Y. Zhang, *Molecules*, 2022, **27**, 4568.