A Proof for Proposition

Proof. of Proposition 1 In order to ensure SE(3) equivariance in the encoder architecture, we define an additional function to derive the relative distance matrix of two coordinate systems, c_x and c_y. Let $c_x = (x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_n, y_n, z_n)$ be the coordinates of system c_x, and let $c_y = (x'_1, y'_1, z'_1), (x'_2, y'_2, z'_2), \dots, (x'_n, y'_n, z'_n)$ be the coordinates of system c_y. Define a function $D(c_x, c_y) = (d_{11}, d_{12}, d_{13}), (d_{21}, d_{22}, d_{23}), \dots, (d_{n1}, d_{n2}, d_{n3})$ where $d_{ij} = \sqrt{(x_i - x'_j)^2 + (y_i - y'_j)^2 + (z_i - z'_j)^2}$ for $i, j = 1, 2, \dots, n$. Since most popular SE(3) frameworks [32, 4, 34] utilize the relative distance to represent coordinates, we replace all coordinates with this matrix representation.

The deviation between using accurate ligand coordinates and inaccurate ligand coordinates can be written as $s(\tilde{x}_m, x_p) - s(x_m, x_p)$. If we applied the Taylor expansion of the first order, the deviation becomes proportionate to the distance perturbation.

$$
\begin{aligned}
& s(\tilde{x}_m, x_p) - s(x_m, x_p) \\
&= f_\theta\left(D\left(c_m + \delta, c_m + \delta\right), h_m\right)^\top g\left(D\left(C_p, C_p\right), h_p\right) - f_\theta\left(D\left(c_m, c_m\right), h_m\right)^\top g\left(D\left(C_p, C_p\right), h_p\right) \\
&\approx \frac{\partial f_\theta}{\partial D(c_m, c_m)}\left(D(c_m, c_m), h_m\right) \cdot \left(D\left(c_m + \delta, c_m + \delta\right) - D\left(c_m, c_m\right)\right)
\end{aligned}
\tag{8}
$$

If the RDkit simulated conformation of the ligand is close enough to the protein-induced conformation, we can find the optimal Rotation R and translation t to fit the two conformations in 3D space that satisfies $(c_n + \delta)R^\top + t = c_n$, which further means $D\left(c_m + \delta, c_m + \delta\right) - D\left(c_m, c_m\right)) = 0$. Therefore, the deviation will be relatively small.

$$
\lim_{\delta \to 0} \{s(\tilde{x}_m, x_p) - s(x_m, x_p)\} = 0 \tag{9}
$$

However, when we applied the first-order Taylor expansion to the deviation of the Single-Tower model, we find out that the deviation is not proportional.

$$
\begin{aligned}
& k_\gamma\left(h_p, h_m, D\left(c_p, c_p\right), D\left(c_m + \delta, c_m + \delta\right), D\left(c_m + \delta, c_p\right)\right) \\
& - k_\gamma\left(h_p, h_m, D\left(c_p, c_p\right), D\left(c_m, c_m\right), D\left(c_m, c_p\right)\right) \\
\approx & \frac{\partial k_\gamma}{\partial D\left(c_m + \delta, c_m + \delta\right)}(\cdot)\dot{(}D\left(c_m + \delta, c_m + \delta\right) - D\left(c_m, c_m\right))) \\
& + \frac{\partial k_\gamma}{\partial D\left(c_m, c_p\right)}(\cdot)\dot{(}D\left(c_m + \delta, c_p\right) - D\left(c_m, c_p\right)))
\end{aligned}
\tag{10}
$$

Though $D\left(c_m + \delta, c_m + \delta\right) - D\left(c_m, c_m\right)) = 0$ can be quite small if highly accurate conformation is approximated by the simulation, the protein-molecule relative term $D\left(c_m + \delta, c_p\right) - D\left(c_m, c_p\right))$ have to be approximated by an additional molecule docking process. As a result, the supervised-learning based methods have to rely on molecule docking software to get the optimal rotation R and translation t. □

This mathematical derivation proves that our framework is more robust and will enjoy the advantages of introducing large amounts of noisy data for training.

B Implementation details

B.1 Implementation of HomoAug

We propose a novel method called Homo-Aug, which utilizes the concept of homologous proteins in biology for data augmentation. Our core idea is to combine ligands from the PDBbind database with homologous proteins corresponding to their protein pockets, thereby generating new training data. Homologous sequences play a fundamental role in the domain of proteins, representing proteins that share a common ancestry in terms of evolutionary relationships. These homologous proteins exhibit certain resemblances in terms of their sequence, structure, and interactions with ligands. By incorporating homologous proteins alongside ligands, we introduce the noise of protein evolution, which can augment data while mitigates the risk of significant alterations in the binding properties of proteins and ligands.For our study, we opted to utilize the AlphaFold protein structure database [16, 40] as our search library for homologous proteins. This database leverages the AlphaFold2 [16] algorithm, enabling the prediction of protein structures for those lacking structural information but possessing sequence data. To ensure the reliability and integrity of the database, we implemented a series of stringent filtering operations.Specifically, we retained only instances exhibiting high structural confidence, as indicated by residues with plDDT values exceeding 0.7 accounting for more than 90% of the protein structure. This filtering criterion ensured that our database comprised instances with robust structural predictions.Furthermore, to enhance the diversity of our database, we employed the MMseqs [12] algorithm to cluster the data using a 50% identity threshold. This clustering process remove the very similar protein , promoting greater variation within the database.Through these rigorous filtering and clustering operations, we obtained a comprehensive homologous retrieval database comprising 8,449,772 protein sequences, each paired with its corresponding reliable protein structure. Utilizing the provided database, we have expanded and enriched the instances sourced from the PDBBind database. Our approach involved several steps to ensure the quality and diversity of the data. Initially, instances containing non-standard residues or pockets with multiple chains were excluded from the dataset. This step was undertaken due to the inherent difficulty in searching for homologous protein complexes. Next, for each protein's pocket-containing chain, we employed the Jackhmmer [14] Algorithm to conduct a search for homologous proteins. The top 200 homologous proteins identified in the Jackhmmer search results were retained for each instance, thereby augmenting the dataset and enhancing its diversity.To ensure ligand binding within the pocket of the homologous protein, we performed structure alignment between the homologous proteins and the original proteins using the TMalign [48] algorithm. This alignment process aimed to identify similarities between the overall protein structure and the pocket region. In order to ensure the quality of the newly generated protein-ligand pairs, we retained only those that exhibited a sufficient degree of structural similarity. Specifically, we imposed the condition that the TMscore should be equal to or greater than 0.4, indicating a significant structural similarity, and the alignment rate of the pocket region should be equal to or greater than 40%, denoting a substantial alignment of residues within the pocket region.Finally, we extracted the atoms of the homologous proteins located within a 6Å radius of the ligand, defining this extracted region as the new pocket. This step allowed us to precisely delineate the pocket for ligand binding and subsequent analysis.

By employing the data augmentation method described earlier, we have achieved significant success in obtaining 758,107 novel pocket-ligand pairs. This approach has resulted in the expansion of 51% of the original instances sourced from the PDBbind database. The implementation of the Homo-Aug method allows us to effectively harness the concept of homologous proteins and utilize it to augment our training data. Through a comprehensive set of filtering and alignment operations, we have successfully enhanced the diversity of the data. This augmentation process significantly broadens the foundation for the field of drug virtual screening, offering a more comprehensive and varied dataset for subsequent analyses and investigations.

B.2 Implementation of Fine-grained Atom Interaction

Besides aligning the representations of the global features from entire pockets and molecules, we also explore the usage of fine-grained features in our contrastive learning framework. When pretraining the 3D encoder, we also take the interactions between atoms into account. Specifically, we found out that in the complex structure, one single protein atom is only able to form strong interactions with a limited number of atoms from the binding molecule, and vice versa. From this biological intuition, we are able to propose an additional loss term that makes use of the fine-grained representation.

To define our training objective, we denote the atom-level representation of a molecule i as $[m_i^1, m_i^2, \cdots, m_i^N]$ and the atom-level representation of a pocket j as $[p_j^1, p_j^2, \cdots, p_j^M]$. To measure the alignment between the representations, we first employ a similarity metric as cosine similarity. Given an embedding m_i^u in m_i, we compute its similarity with all tokens in p_j and select the top K most similar tokens based on the similarity scores. We denote the set of indices of the selected tokens in p_j as $\mathbf{T}_{p_j}$.

Similarly, for each token embedding p_j^v, we find its K most similar tokens in m_i and represent the corresponding set of indices as $\mathbf{T}_{m_i}$.

Next, we defined the loss term as follows:

$$\mathcal{L}_{\text{topk-topk}} = \sum_{v \in \mathbf{T}_{m_i}} \sum_{u \in \mathbf{T}_{p_j}} \mathrm{s}\left(m_i^u, p_j^v\right) \tag{11}$$

By optimizing this topk-topk loss term, we encourage the model to focus on the most informative atom alignments, facilitating better representation on the fine-grained level. When implemented we add the topk-topk loss term as an auxiliary loss to the global-level contrastive learning objective as in Eq. 5. We also conduct experiments by extracting atom-level representations from different layers of the encoder to compare the difference. The experiment result for atom-level interaction is shown in section C.2.

B.3 Evaluation Metrics

There are several evaluation metrics we use in this paper for benchmarking virtual screening tasks. Here are the detailed explanations.

BEDROC incorporates exponential weights that assign greater importance to early rankings. In the context of virtual screening, the commonly used variant is BEDROC$_{85}$, where the top 2% of ranked candidates contribute to 80% of the BEDROC score (cite). The formal definition is:

$$\text{BEDROC}_\alpha = \frac{\sum_{i=1}^{\text{NTB}_t} e^{-\alpha r_i/N}}{R_\alpha\left(\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}\right)} \times \frac{R_\alpha \sinh(\alpha/2)}{\cosh(\alpha/2)-\cosh(\alpha/2-\alpha R_\alpha)} + \frac{1}{1-e^{\alpha(1-R_\alpha)}}. \tag{12}$$

Enrichment Factor(EF) is also a widely used metric, which is calculated as

$$\text{EF}_\alpha = \frac{\text{NTB}_\alpha}{\text{NTB}_t \times \alpha}, \tag{13}$$

where NTB_α is the number of true binders in the top $\alpha\%$ and NTB_t is the total number of binders in the entire screening pool.

We also adopted **ROC enrichment metric (RE)**, which is calculated as a ratio of the true positive rate to the false positive rate (FPR) at a given FPR threshold:

$$\text{RE}(x\%) = \frac{\text{TP} \times n}{\text{P} \times \text{FP}_{x\%}}, \tag{14}$$

where n is the total number of compounds, TP is the number of compounds that are correctly identified as active, P is the total number of active compounds, and $\text{FP}_{x\%}$ is the number of false positives predicted at a specified rate (e.g. 0.5%, 1%, etc.).

B.4 Encoder Pre-training

Our pre-training of the molecule and pocket encoders is based on the methodology proposed by UniMol [53]. Similar to BERT [5], we utilize a masked token prediction task. In the context of molecule or pocket data, this task involves predicting masked atom types. To augment the complexity of the pre-training task and extract valuable insights from 3D coordinates, we introduce an additional task called position denoising. Specifically, we add random uniform noise within the range of $[-1\text{Å}, 1\text{Å}]$ to 15% of the atom coordinates. Two tasks are incorporated to restore the original positions. Firstly, the model needs to predict the original distance between two corrupted atoms. Secondly, the model needs to estimate the original coordinates of a corrupted atom using the SE(3)-Equivariance coordinate system.

B.5 Contrastive Learning Training Details

We train our model using the Adam optimizer with a learning rate of 0.001. The other hyper-parameters are set to their default values. We have a batch size of 192, and we use 4 NVIDIA A100 GPU cards for acceleration. We train our model for a maximum of 200 epochs. To avoid overfitting, we use the CASF-2016 dataset as a validation set and select the epoch checkpoint with the best BEDROC_{85}. For more detailed training configurations, please refer to the code.

For the model used for human evaluation(DrugCLIP-L), we use dot product as the distance metric. For other models we use cosine similarity.

C Additional Experiments

C.1 Evaluation on Target Fishing

Since DrugCLIP has the ability to learn the matching between proteins and molecules, it could be also used for target fishing, another important task in drug discovery, which entails the identification of the target from a pool of candidate targets that have the potential to bind to a specific molecule. We establish a benchmark using the CASF-2016 dataset. For each molecule, we test whether the model can correctly find its corresponding pocket from all other pockets. As shown in Table 6, DrugCLIP exhibits superior accuracy in the top 1 to 5 predictions as compared to docking software, i.e. Glide, and Vina. Conversely, DrugBA performs much poorer, with results comparable to random guessing.

Note: In this benchmark, we are unable to use the CASF-2016 dataset as both the test set and the validation set. Therefore, we split our training set in a 9 to 1 ratio and allocate the latter portion as the validation set.

C.2 Global and Local interactions

As shown n Table 7, using atom embeddings from the last transformer layer yields worse performance. However, marginal improvement is observed when utilizing embeddings from the second last layer. Selecting the appropriate transformer layer is crucial for obtaining effective atom embeddings and enhancing model performance, and should be considered as future work.

Table 6: Result of Target Fishing Task on CASF-2016 dataset

	Accuracy				
	@1	@2	@3	@4	@5
Vina [39]	3.38	5.26	7.52	9.02	10.15
Glide [11]	14.98	22.85	30.34	35.58	39.33
DrugBA	0.37	0.74	1.11	2.22	2.22
DrugCLIP	24.07	42.96	51.11	59.26	62.59

Table 7: Performance Comparison on DUD-E and LIT-PCBA Datasets by adding atom-level interactions

	DUD-E		
	AUROC %	BEDROC %	EF@1%
Global only	80.93	50.52	31.89
with last	78.87	44.72	28.65
with second	82.79	50.57	32.45

C.3 GPCR

In this section, we demonstrate the ability of our model to pair all known human GPCR proteins with 31,422 human metabolites using AlphaFold2 predicted models. We aim to identify unrevealed GPCR ligands to facilitate functional studies, as certain GPCR proteins may have unexpected functions. For example, hOF17-4, an olfactory receptor, locates on sperms and contributes to egg localization. To achieve this, we utilized Fpocket for ligand-binding pocket detection on GPCR protein surfaces and obtained 17,702 pockets. Evaluating more than 5×10^8 pocket-ligand pairs would typically take around one CPU year with cutting-edge active-learning-assisted docking; however, our model can rank these pairs within minutes.

We manually evaluated top-ranked pairs and predicted their binding poses using commercialized docking software GLIDE in the Schrodinger Suite. Our findings revealed several particularly interesting pairs, including three kidney-enriched olfactory GPCRs, OR2T5, OR2T11, and OR4C3, which were predicted to bind known metabolic wastes. The kidney-expressed olfactory system has long been known to influence urine production. Additionally, the presence of olfactory G protein, G_{olf},and olfactory-related adenylate cyclase AC3 was detected in the distal convoluted tubule. When olfactory signaling was blocked via AC3 knock-out, creatinines accumulated in the blood, indicating defective renal function.

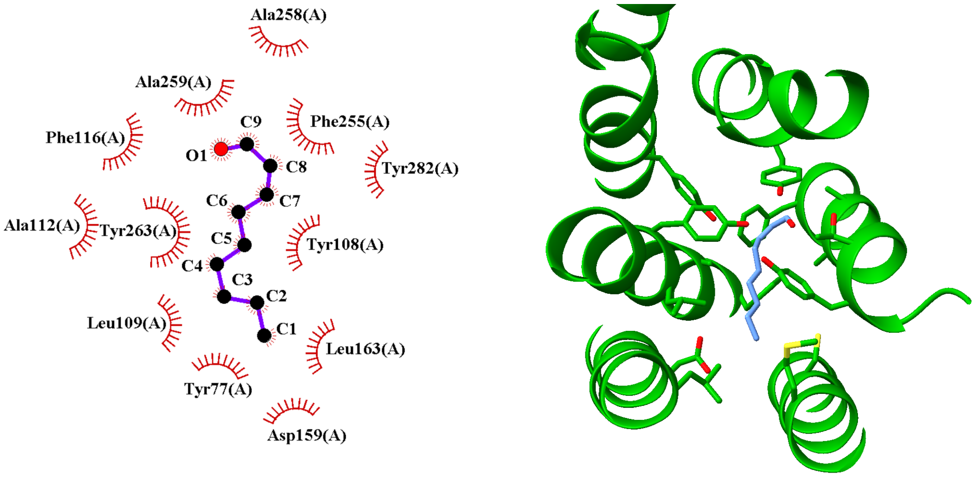
Our model identified OR2T5 paired with 2-nonenal, OR2T11 paired with p-cresol, and OR4C3 paired with D-lactic acid. Docking poses revealed potential hydrophobic interactions, hydrogen bonds, and $\pi - \pi$ interactions between pockets and ligands. As previous studies reported, 2-nonenal is a uremic toxin; p-cresol is an intermediate of tyrosine metabolism; and D-lactic acid is a widely distributed waste product. These molecules are highly toxic and require timely cleaning/recycling by either the excretory system or cellular processes. Our findings suggest that olfactory receptors in the kidney can sense metabolic wastes and regulate the excretion process as a feedback loop. Visualizations are shown in Figure 6,7,8.

D Limitations

The major limitation of our paper lies pertains to its interpretability. Although our model demonstrates enhanced effectiveness and efficiency, it falls short in terms of interpretability compared to traditional docking methods. These conventional approaches offer visualizations that elucidate the binding mechanism between a pocket and a molecule, providing clear explanations.

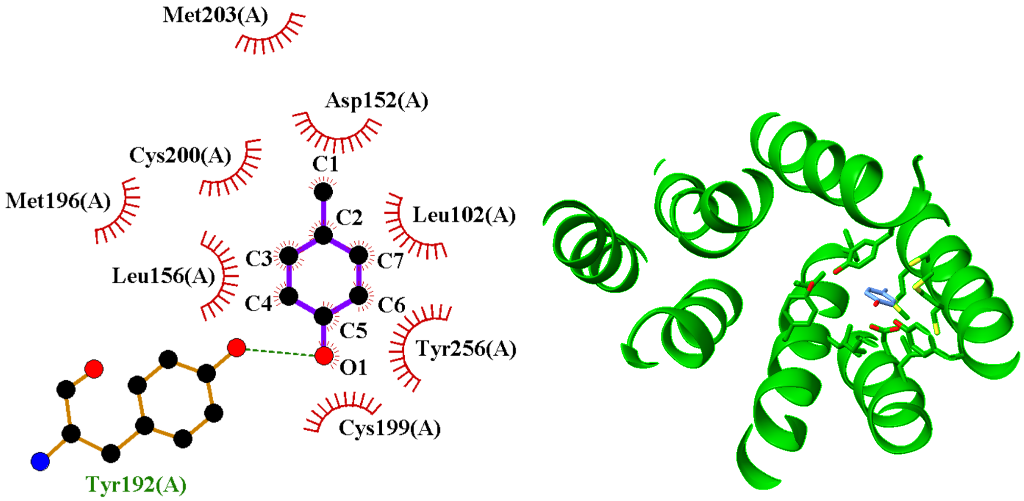
E Negative societal impacts

While our method has the potential to greatly expedite the drug discovery process, which is undoubtedly advantageous, it is important to consider the potential implications it may have on drug auditing and clinic trials. The increased speed and efficiency may inadvertently create additional pressures and challenges for regulatory bodies responsible for ensuring the safety and efficacy of new drugs.



(a) (b)

Figure 6: Visualization of the docking pose of OR2T5 and 2-nonenal complex. The 2D interaction pattern is generated with LigPlot+. Interactions between OR2T5 and 2-nonenal are mainly hydrophobic interactions.

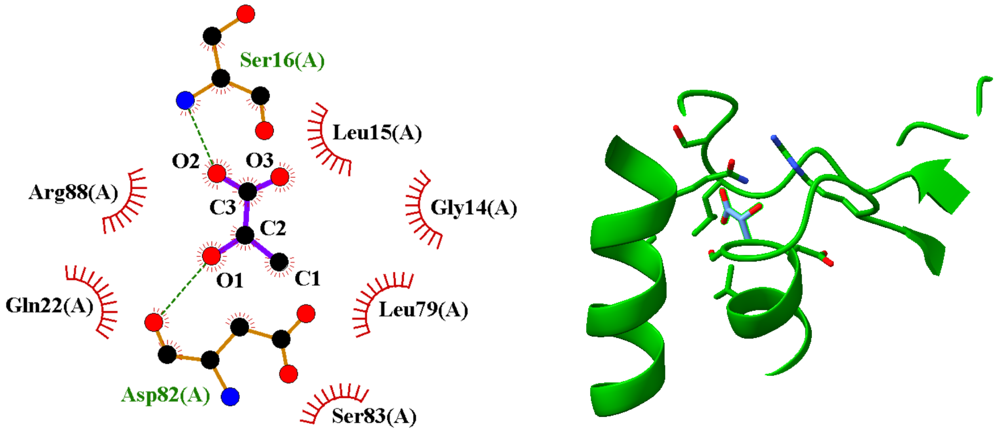


(a) (b)

Figure 7: Visualization of the docking pose of OR2T11 and p-cresol complex. The 2D interaction pattern is generated with LigPlot+. Tyr192 of OR2T11 and O1 of p-cresol form a hydrogen bond. Tyr256 could have potential $\pi - \pi$ interaction with p-cresol.

Table 8: Results of Human Expert Evaluation.

	5kdt	6g2o	1n5x	7ksi	8etr
Glide [11]	2	2	4	**7**	4
DrugCLIP	**8**	**8**	**6**	3	**6**



(a)

(b)

Figure 8: Visualization of the docking pose of OR4C3 and D-lactic acid complex. The 2D interaction pattern is generated with LigPlot+. Ser16 and Asp82 interact with D-lactic acid via hydrogen bonds.