S^{2}Drug: Bridging Protein Sequence and 3D Structure in Contrastive Representation Learning for Virtual Screening

Bowei He[1], Bowen Gao[2], Yankai Chen[3], Yanyan Lan[2, 4*], Chen Ma[1*], Philip S. Yu[3], Ya-Qin Zhang[2], Wei-Ying Ma[2]

[1] City University of Hong Kong
[2] Institute for AI Industry Research (AIR), Tsinghua University
[3] University of Illinois Chicago
[4] Beijing Academy of Artificial Intelligence

Abstract

Virtual screening (VS) is an essential task in drug discovery, focusing on the identification of small-molecule ligands that bind to specific protein pockets. Existing deep learning methods, from early regression models to recent contrastive learning approaches, primarily rely on structural data while overlooking protein sequences, which are more accessible and can enhance generalizability. However, directly integrating protein sequences poses challenges due to the redundancy and noise in large-scale protein-ligand datasets. To address these limitations, we propose **S^{2}Drug**, a two-stage framework that explicitly incorporates protein **S**equence information and 3D **S**tructure context in protein-ligand contrastive representation learning. In the first stage, we perform protein sequence pretraining on ChemBL using an ESM2-based backbone, combined with a tailored data sampling strategy to reduce redundancy and noise on both protein and ligand sides. In the second stage, we fine-tune on PDBBind by fusing sequence and structure information through a residue-level gating module, while introducing an auxiliary binding site prediction task. This auxiliary task guides the model to accurately localize binding residues within the protein sequence and capture their 3D spatial arrangement, thereby refining protein-ligand matching. Across multiple benchmarks, S^{2}Drug consistently improves virtual screening performance and achieves strong results on binding site prediction, demonstrating the value of bridging sequence and structure in contrastive learning.

Introduction

Virtual screening (VS) is regarded as a cornerstone of modern drug discovery, as it enables researchers to computationally evaluate large libraries of chemical compounds to identify potential drug candidates that bind to a target protein (Shoichet 2004). By leveraging high-performance computation algorithms and hardwares, VS significantly reduces the cost, time, and labor required in early-stage hit identification, especially when compared to traditional high-throughput empirical screening over trillions of or even more drug-like molecules (Lyu, Irwin, and Shoichet 2023).

*Corresponding Authors: lanyanyan@air.tsinghua.edu.cn and chenma@cityu.edu.hk.

Molecular docking and deep learning methods are two types of the most mainstream approaches for virtual screening. The docking-based methods (Friesner et al. 2004) simulates the physical binding process between a ligand and a protein to estimate binding affinity, while learning-based methods leverage data-driven models to predict interactions without explicit simulation. Notably, the success of deep learning in related fields such as graph learning and natural language processing has fueled growing interest in learning-based approaches. Early works in this category consider the VS as a binding affinity regression or classification task (Wu et al. 2022; Zhang et al. 2023). Recent works like DrugCLIP (Gao et al. 2023) and DrugHash (Han, Hong, and Li 2025) model the VS as a retrieval task, where protein pockets are regarded as queries to retrieve matching ligands. To enable this, they leverage contrastive learning to align protein and ligand representations by pulling binding pairs closer and pushing apart non-binding pairs in the embedding space. However, whether molecular docking or existing learning methods, most approaches still rely exclusively on 3D structures (Maia et al. 2020), neglecting the employment of protein sequence information. In fact, this reliance on single-conformer atom-level structural information can make models overly sensitive to input perturbations, and less robust to pocket conformational flexibility-a common challenge in practical drug discovery. Furthermore, determining protein 3D structures requires techniques such as X-ray crystallography and Cryo-EM, which are time-consuming and costly. These limitations hinder the expansion of large-scale training datasets, ultimately restricting the performance and generalizability of VS methods.

The protein research in recent decades follows a fundamental principle: *"Sequence determines structure, and structure determines function"* (Anfinsen 1973). This implies that amino acid sequences fundamentally encode the information necessary for protein folding and function, which in turn influences how proteins interact with ligands, especially their affinity to ligands in specific sites. Early works (Shen et al. 2023) have explored the introduction of protein sequence information for VS. However, they merely leverage sequence representations encoded by pretrained protein LSTM (Alley et al. 2019)

or Transformer (Rives et al. 2021) models, without explicitly learning the protein sequence-ligand interactions. Although large-scale protein sequence–ligand interaction datasets such as ChemBL (Gaulton et al. 2012) have been available for years, their potential to directly supervise the learning of sequence-level binding preferences remains underexplored. Nevertheless, directly integrating such datasets into VS model training poses significant challenges. Because the data redundancy and noise in both protein and ligand sides-such as homology/functional redundancy, binding affinity variability, and non-specific binders-may significantly hinder the model performance and generalizability. Moreover, relying solely on protein sequence information while discarding structural context severely limits the model's ability to identify binding pockets, model spatial interactions (e.g., shape complementarity), and account for conformational flexibility. These factors are critical for accurate and generalizable virtual screening.

To tackle such challenges, we propose **S^{2}Drug**, a two-stage contrastive learning framework that bridges protein **S**equence and 3D **S**tructure information to enable more accurate virtual screening. (1) In the first stage, we perform protein sequence-ligand contrastive representation pretraining on the large-scale dataset ChemBL. During this stage, we design a bilateral data-sampling strategy: on the protein side, we apply homology-aware downweighting and functional deduplication to reduce protein-side redundancy; on the ligand side, we apply affinity variability filtering and frequent hitter removal to mitigate noise. (2) In the next stage, we introduce a residue-level gating module to fuse protein sequence and 3D structure information. This enables fine-tuning of the learned representations on the small-scale PDBBind dataset, which provides high-resolution pocket structures. In addition, we introduce an auxiliary binding site prediction task that determines whether each residue belongs to the binding site. Given that a pocket is essentially the spatial cavity-shape aggregation of binding site residues scattered across the sequence, this task enhances the model's understanding of protein 3D folding-particularly in pocket regions-thereby further improving VS performance.

We summarize our contributions as follows:

- We propose a two-stage contrastive learning framework, namely S^{2}Drug, which performs sequence pretraining and sequence-structure fusion finetuning for virtual screening.
- We devise a data sampling strategy to harness large-scale protein sequence-ligand datasets, ensuring high-quality pretraining data by reducing redundancy and noise.
- We incorporate a binding site prediction auxiliary task in the fine-tuning stage to enhance model's understanding of binding pocket locations and their spatial configurations.
- We evaluate our S^{2}Drug framework on two most typical virtual screening datasets and obtain consistent superior performance over baselines. Additionally, S^{2}Drug exhibits compelling results on binding site prediction.

Related Works

Virtual Screening Existing methods for virtual screening can be broadly categorized into two main approaches: docking-based and learning-based methods. Docking-based methods, such as Glide-SP (Friesner et al. 2004) and Vina (Trott and Olson 2010), is a computational technique that predicts how a small molecule (ligand) binds to a target protein's binding site, estimating both the binding pose and interaction strength. However, this type of methods are often time-consuming because they must search numerous ligand conformations and evaluate each with computationally expensive scoring functions. On the other hand, learning-based methods leverage machine learning and deep learning techniques to predict ligand-protein pocket interactions more efficiently. Within learning-based methods, there are three primary subcategories: regression (Öztürk, Özgür, and Ozkirimli 2018; Zheng, Fan, and Mu 2019; Zhang et al. 2023), classification (Wu et al. 2022; Yazdani-Jahromi et al. 2022), and retrieval (Gao et al. 2023; Han, Hong, and Li 2025) approaches. Among them, while structural data has been extensively used, the potential of protein sequence data, which is more readily available and can provide complementary information, has been largely neglected.

Protein Representation Learning Though the importance of molecular representation learning in drug discovery tasks has been widely recognized (Wang et al. 2022), the most advanced protein representation learning methods in the virtual screening still stay in the structure learning stage (Han, Hong, and Li 2025). They utilize pretrained SE(3) Transformers to model 3D atom-level interactions, and adopt contrastive learning to align protein and ligand representations. In contrast, out of virtual screening domain, the protein sequence representation learning has achieved great progress in recent years due to the widely available sequence data, low annotation cost without structure determination, and the model architecture derived from language models (Lin et al. 2023; Abramson et al. 2024). Besides, there is also increasing interest in incorporating both sequence and structure information to conduct more comprehensive protein representation learning (Lee et al. 2023; Hu et al. 2024; Li et al. 2024). However, how to employ protein sequence representation learning in virtual screening, especially sequence-structure collaborative learning, is still less explored, which is the focus of this work.

Methodology

In this section, we first define the problem, and then detail the two stages of our S^{2}Drug framework: (1) sequence model pretraining and (2) sequence-structure fusion finetuning. We illustrate the framework in Figure 1.

Problem Formulation

In this work, we define two interrelated tasks to enhance virtual screening in drug discovery. Note we use the term "ligand" to refer to candidate small molecules, regardless of whether binding has been experimentally confirmed.

Main Task: Virtual Screening The primary objective is to predict the binding likelihood between a specific protein binding pocket and a ligand, enabling the identification of potential drug candidates from a ligand library. Given a protein P, which includes its amino acid sequence $S(P)$ and

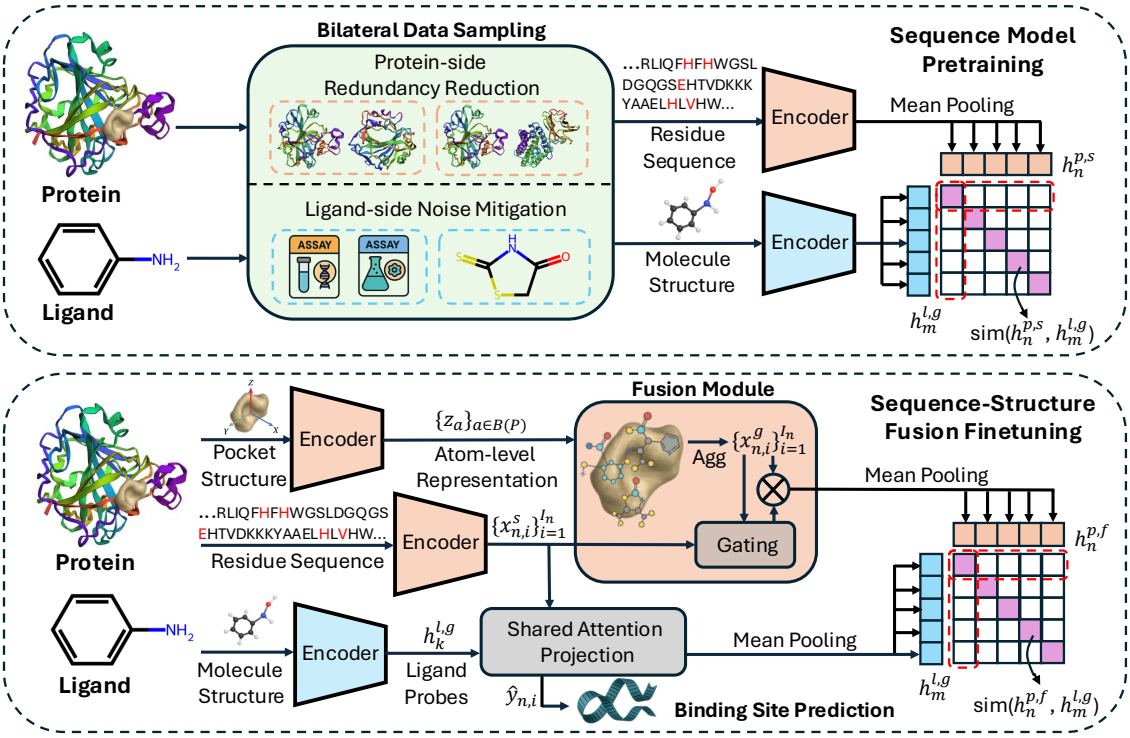

Figure 1: The illustration of our proposed two-stage contrastive representation learning framework S^{2}Drug for bridging protein sequence and 3D structure. The red characters indicate pocket residues, which are however agnostic to the sequence encoder.

three-dimensional structural information $G(P)$, we denote its binding pocket as $B(P) \subseteq P$, defined by a subset of residues and their spatial configuration. For a ligand L, represented by its molecular structure $G(L)$, we aim to learn a scoring function $s(B(P), L) : \mathcal{B} \times \mathcal{L} \to \mathbb{R}$, where $\mathcal{B}$ is the space of protein binding pockets, and $\mathcal{L}$ is the space of ligands. The function $s(B(P), L)$ outputs a real-valued score indicating the likelihood of ligand L binding to the pocket $B(P)$, which is used to rank ligands for a given protein target based on their affinity to the specified binding pocket.

Auxiliary Task: Binding Site Prediction The auxiliary task focuses on identifying the amino acid residues in the protein sequence that constitute the binding site, thereby enhancing the performance of the main task. For a given protein P, its sequence is represented as $S(P) = (r_1, r_2, \ldots, r_I)$, where r_i is the i-th amino acid residue, and I is the sequence length. We predict a binary label $y_i \in \{0, 1\}$ for each residue s_i, where $y_i = 1$ indicates that the residue is part of the binding site, and $y_i = 0$ indicates otherwise. This task is formulated as a sequence labeling problem, where the model outputs a prediction vector $\hat{Y}(P) = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_I)$, with $\hat{y}_i$ representing the predicted probability that residue i belongs to the binding site.

Sequence Model Pretraining

Although large-scale protein language models have demonstrated strong performance across a wide range of tasks, they are primarily trained on general sequence corpora without explicit supervision from protein–ligand interactions. To effectively model residue-level representations that are sensitive to binding specificity, it is necessary to further adapt these sequence models using protein–ligand interaction data.

Bilateral Data Sampling To ensure high-quality and generalizable supervision from large-scale protein–ligand interaction datasets such as ChemBL, we introduce a holistic strategy that jointly filters and subsamples the training data from both the protein and ligand sides. Let $\mathcal{D}_0 = \{(P_n, L_n, a_n)\}_1^{|\mathcal{D}_0|}$ denote the initial dataset with 745K protein–ligand–affinity triplets, where P_n is the protein with $S(P_n)$ as corresponding residue sequence, L_n is the ligand, and $a_n \in \mathbb{R}$ is the reported binding affinity. We construct a cleaned subset $\mathcal{D} \subset \mathcal{D}_0$ via the following two-sided strategy:

Step 1: Protein-side Redundancy Reduction Protein data in ChemBL suffers from severe homology and functional redundancy, which may encourage the memorization of family-specific patterns rather than the learning generalizable interaction rules.

(1) Homology-aware Downweighting. We compute pairwise residue sequence similarity $\mathrm{SeqSim}(S(P_n), S(P_m)) \in [0, 1]$, based on normalized sequence alignment scores. Then we cluster the sequences using MMseqs2 at a 40% identity threshold (Kallenborn et al. 2024), yielding clusters $\mathcal{C}^{hom} = \{C_1^{hom}, \ldots, C_M^{hom}\}$. The sampling probability of each protein $P_n \in C_m^{hom}$ is defined as:

$$\Pr(P_n) = \frac{1}{|C_m^{hom}|^{\alpha}}, \quad \alpha \in (0, 1]. \tag{1}$$

We set $\alpha = 0.5$ to penalize large protein families while retaining sufficient diversity.

(2) Functional Deduplication. Functional isoforms or mutants may be overrepresented, contributing negligible additional information. Based on UniProt or Gene Ontology function annotations $\phi(P_i)$, we define groups:

$$C_k^{fun} = \{P_n \mid \phi(P_n) = \phi_k\}. \tag{2}$$

We retain a single representative per C_k^{fun}, prioritizing proteins with greater ligand diversity.

Step 2: Ligand-side Noise Mitigation The noise in the ligand side arises from binding affinity variability across assays and from non-specific promiscuous compounds.

(1) Affinity Variability Filtering. For duplicate protein-ligand pairs with different affinity values obtained under different assay conditions $\{(P_n, L_n, a_n^j\}_{j=1}^J$, we calculate:

$$\sigma_n = \mathrm{StdDev}(\log a_n^1, \ldots, \log a_n^J). \tag{3}$$

Only instances with $\sigma_n < \delta$ (we use $\delta = 1.0$) are retained, and the canonical affinity is set as the mean log-value.

(2) Frequent Hitter Removal. Some compounds exhibit promiscuous binding activity across multiple targets, not due to specific molecular recognition but rather because of reactive substructures. These artifacts can mislead representation learning. Let $f(L_n)$ be the number of proteins a ligand binds. We discard ligands where $f(L_n) > T$ (with $T = 20$) unless a majority of affinities indicate strong binding, i.e., high affinity values. Besides, ligand molecules with known such reactive substructures (e.g., PAINS) are filtered out.

Step 3: Joint Subsampling with Rebalancing After cleaning, we subsample remaining data with rebalancing:

$$\mathcal{D} = \mathrm{Sample}_{(P,L,a)\sim\mathcal{D}_0}\left[\Pr(P) \cdot \mathbb{I}_{\mathrm{clean}(P,L,a)} \cdot w_{\mathrm{lig}}(L)\right], \tag{4}$$

where $\mathbb{I}_{\mathrm{clean}(P,L,a)} \in \{0, 1\}$ indicates that a triplet passes all filtering rules, and $w_{\mathrm{lig}}(L) \propto 1/f(L)$ downweights high-frequency ligands.

Representation Learning To learn modality-specific representations that simultaneously capture protein sequence semantics and ligand structural properties, we first employ modality-aligned encoders for both proteins and ligands. Then, a symmetric contrastive training objective is applied to align their embeddings in a shared latent space.

Sequence Encoder for Proteins We adopt the pretrained ESM2 model (Lin et al. 2023) to encode the amino acid residue sequences. Given a protein P_n with its residue sequence $S(P_n) = (r_1, \ldots, r_I)$, the sequence encoder $\mathrm{Seq}^p(\cdot)$ maps it to a fixed-length embedding $h_n^{p,s} \in \mathbb{R}^{d_s}$:

$$h_n^{p,s} = \mathrm{MeanPool}(\mathrm{Seq}^p(S(P_n))), \tag{5}$$

where $\mathrm{Seq}^p(\cdot)$ is initialized from the 650M-parameter ESM2 model. The output $h_n^{p,s}$ is the mean pooling of tokens' representation from the final Transformer layer.

Structure Encoder for Ligands For ligand molecules, we adopt the Uni-Mol molecular encoder (Zhou et al. 2023), which encodes 3D conformers using atom types and pairwise Euclidean distances. Given a ligand L_i represented by structual information $G(L_n)$ including 3D atomic coordinates and associated atom types, the encoder $\mathrm{Stru}^l(\cdot)$ outputs an embedding $h_n^{l,g} \in \mathbb{R}^{d_g}$ after mean pooling over atoms:

$$h_n^{l,g} = \mathrm{MeanPool}(\mathrm{Stru}^l(G(L_n))). \tag{6}$$

To align its representation space with the protein sequence encoder, we project both $h_n^{p,s}$ and $h_n^{l,g}$ into a shared embedding space via two-layer MLPs with layer normalization.

Contrastive Training Objective We train the protein and ligand encoders using a symmetric contrastive objective that brings matching pairs closer while pushing apart non-matching ones. For a batch of N protein–ligand pairs $\{(P_n, L_n)\}_{n=1}^N$, we compute the pairwise cosine similarity:

$$\mathrm{sim}(h_n^{p,s}, h_m^{l,g}) = \frac{\langle h_n^{p,s}, h_m^{l,g}\rangle}{\|h_n^{p,s}\| \cdot \|h_m^{l,g}\|}. \tag{7}$$

The contrastive loss is defined via InfoNCE (Oord 2018):

$$\begin{aligned}\mathcal{L}_{\mathrm{pc}} = -\frac{1}{N}\sum_{n=1}^{N} \log \frac{\exp(\mathrm{sim}(h_n^{p,s}, h_n^{l,g})/\tau)}{\sum_{m=1}^N \exp(\mathrm{sim}(h_n^{p,s}, h_m^{l,g})/\tau)} + \\ \log \frac{\exp(\mathrm{sim}(h_n^{p,s}, h_n^{l,g})/\tau)}{\sum_{m=1}^N \exp(\mathrm{sim}(h_m^{p,s}, h_n^{l,g})/\tau)},\end{aligned} \tag{8}$$

where τ is a temperature hyperparameter controlling the sharpness of the similarity distribution.

Sequence-Structure Fusion Finetuning

To enhance virtual screening accuracy, it is crucial for the model to incorporate both the evolutionary information embedded at protein sequences and the fine-grained geometric context provided by 3D pocket structures. While structural encoders are effective in modeling atomic interactions, they often suffer from sensitivity to conformational perturbations and may not generalize well to novel proteins. To address this, we propose a fusion-based finetuning strategy that combines above pretrained protein sequence embeddings with structure-aware representations. They are jointly trained using a contrastive interaction objective and an auxiliary binding site prediction task on the PDBBind dataset.

Sequence-Structure Fusion Module Given a protein P_n, we obtain residue-level sequence representation $\{x_{n,i}^s\}_{i=1}^{I_n}$ from the encoder Seq^p and the atom-level structural representation $\{z_a\}_{a\in B(P)}$ from pocket structure encoder Stru^p initialized from Uni-Mol. For each pocket residue $r_i \subset B(P)$, we perform masked mean pooling over its constituent atoms to aggregate the structure-based embedding:

$$x_{n,i}^g = \frac{1}{|r_i|}\sum_{a\in r_i} \mathbb{I}_{a\in r_i} \cdot z_a, \tag{9}$$

where $\mathbb{I}_{a\in r_i}$ indicates the atom a is in the pocket residue r_i. We then apply a *gating* mechanism to adaptively fuse the sequence and structure information, thus obtaining $x_{n,i}^f$:

$$\begin{aligned}\beta_{n,i} &= \sigma\left(W_\beta^\top [W_s x_{n,i}^s; W_g x_{n,i}^g] + b_\beta\right), \\ x_{n,i}^f &= \beta_{n,i} \cdot W_s x_{n,i}^s + (1 - \beta_{n,i}) \cdot W_g x_{n,i}^g,\end{aligned} \tag{10}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and W_β, b_β are learnable parameters. Also, W_s, W_g are learnable matrices to project $x^s_{n,i}$ and $x^g_{n,i}$ into shared representation space. This fusion mechanism allows the model to dynamically prioritize the most informative modality for each residue, depending on the local sequence-structure alignment. Finally, after two Transformer layers, we conduct mean pooling over pocket residues to obtain fused pocket representation $h^{p,f}_n$.

Auxiliary Binding Site Prediction As we know, a pocket is a spatial aggregation of residues in the three-dimensional conformation, rather than a continuous segment in the primary sequence. These residues come together to form a binding cavity as a result of protein folding that brings them into close proximity in 3D space. Therefore, conducting binding site prediction on the entire protein sequence helps the model infer its spatial arrangement-particularly the conformation and further biochemical properties of the pocket region-ultimately facilitating the accurate virtual screening process. To obtain the residue distribution of the binding pocket on the whole residue sequence, for each protein, we sample K ligand probes $\{L_k\}_{k=1}^K$ and compute their interaction relevance with residues via a *shared attention projection*:

$$\alpha^k_{n,i} = \frac{\exp(W_r x^s_{n,i} \cdot W_l h^{l,g}_k)}{\sum_{i=1}^{I_n} \exp(W_r x^s_{n,i} \cdot W_l h^{l,g}_k)}, \tag{11}$$

where $h^{l,g}_k$ is the structural representation of the k-th probe ligand from Stru^l, and W_r, W_l are trainable projection matrices. To avoid information leakage, we only utilize the sequence representation $x^s_{n,i}$ here. We then average the attention scores to obtain the per-residue binding probability:

$$\hat{y}_{n,i} = \frac{1}{K} \sum_{k=1}^{K} \alpha^k_{n,i}. \tag{12}$$

The prediction is trained with summed binary cross-entropy:

$$\begin{aligned} \mathcal{L}_{\text{bsp}} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{I_n} [y_{n,i} \log \hat{y}_{n,i} + \\ (1 - y_{n,i}) \log(1 - \hat{y}_{n,i})]. \end{aligned} \tag{13}$$

Training Objective To align protein–ligand interactions while emphasizing spatially grounded residue position, we jointly optimize a dual loss objective. For contrastive training, we encode a protein pocket into $h^{p,f}_n$ via pooling over fused residue embeddings, and a ligand into $h^{l,g}_n$ using the Stru^l encoder. The contrastive loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{fc}} = -\frac{1}{N} \sum_{n=1}^{N} \log \frac{\exp(\text{sim}(h^{p,f}_n, h^{l,g}_n)/\tau)}{\sum_{m=1}^{N} \exp(\text{sim}(h^{p,f}_n, h^{l,g}_m)/\tau)} + \\ \log \frac{\exp(\text{sim}(h^{p,f}_n, h^{l,g}_n)/\tau)}{\sum_{m=1}^{N} \exp(\text{sim}(h^{p,f}_m, h^{l,g}_n)/\tau)}, \end{aligned} \tag{14}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, N is the batch size, and τ is the temperature parameter, similar to pretraining phase. The final training objective combines two losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fc}} + \lambda \cdot \mathcal{L}_{\text{bsp}}, \tag{15}$$

where λ is a balancing coefficient. This learning scheme enables the model to simultaneously learn global interaction alignment and local spatial specificity, improving both virtual screening accuracy and binding site interpretability.

Experiment

Evaluation Settings

Datasets We evaluate the virtual screening performance of various methods on the DUD-E (Mysinger et al. 2012) and LIT-PCBA (Tran-Nguyen, Jacquemard, and Rognan 2020). As for the binding site prediction task, HOLO4K, COACH420 (Krivák and Hoksza 2018), ASD (Liu et al. 2020) are utilized as the benchmarks. To ensure fairness, all the evaluations in this paper follow the *zero-shot* setting.

Baselines For virtual screening evaluation, we take several shared baselines over above two datasets: *docking*: Glide-SP (Friesner et al. 2004), *learning*: Planet (Zhang et al. 2023), Banana (Brocidiacono et al. 2023), SVSBI (Shen et al. 2023), DrugCLIP (Gao et al. 2023), DrugHash (Han, Hong, and Li 2025). Besides, we also take specialized baselines on DUD-E: *docking*: Vina (Trott and Olson 2010), *learning*: NN-score (Durrant and McCammon 2011), RF-score (Ballester and Mitchell 2010), Pafnucy (Stepniewska-Dziubinska, Zielenkiewicz, and Siedlecki 2017). On LIT-PCBA, we compare with following exclusive baselines: *docking*: Surflex (Spitzer and Jain 2012), *learning*: DeepDTA (Öztürk, Özgür, and Ozkirimli 2018), Gnina (McNutt et al. 2021). As for the evaluation of binding site prediction, we adopt P2Rank (Krivák and Hoksza 2018), VN-EGNN (Sestak et al. 2023), DiffDock (Corso et al. 2023), and VN-EGNNrank (Schneckenreiter et al. 2025) as the baselines. Given these methods directly predict the pocket center coordinates, residues whose heavy atoms are within a fixed distance threshold (8 Å) from the predicted center are considered as binding site residues.

Metrics Following previous works (Gao et al. 2023; Han, Hong, and Li 2025), we utilize the following metrics to evaluate virtual screening accuracy: the area under the receiver operating characteristic curve (AUROC), the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC), enrichment factor (EF). For binding site prediction, we take F1 Score and the area under the precision-recall curve (PR-AUC), which is specially suitable for this task due to the extremely unbalanced labels. All reported metrics are scaled to percentage values for presentation clarity.

Implementation Details We conduct our experiments on a local server with 1.6 TB RAM, 8 × NVIDIA RTX A6000 GPUs, and dual Intel Xeon Silver 4309Y processors (16 cores, 32 threads @ 2.80 GHz). The epoch, learning rate, and batch size for pretraining stage are set as 10, 1e-3, and 128, respectively. During the finetuning stage, the epoch and batch size values are set to 50 and 48, while the learning rate keeps unchanged. As for the parameter initialization of sequence encoder and structure encoders, we employ the 650M version of ESM2 pretrained model and the 181MB version of Uni-Mol pretrained models, respectively. Statistical significance is assessed using paired t-tests, with per-

formance improvements over best baselines considered significant at $p < 0.01$ (marked by * in tables). We run each experiment for 5 times and report the average results. The code has been provided in supplementary materials.

Results and Analysis

	AUROC	BEDROC	$EF^{0.5\%}$	$EF^{1\%}$	$EF^{5\%}$
Glide-SP	76.70	40.70	19.39	16.18	7.23
Vina	71.60	-	9.13	7.32	4.44
NN-score	68.30	12.20	4.16	4.02	3.12
RF-Score	65.21	12.41	4.90	4.52	2.98
Pafnucy	63.11	16.50	4.24	3.86	3.76
Planet	71.60	-	10.23	8.83	5.40
Banana	50.14	2.40	1.19	1.18	1.01
SVSBI	76.28	42.35	25.64	12.82	7.03
DrugCLIP	79.45	47.82	37.86	30.76	10.10
DrugHash	83.73	57.16	43.03	37.18	12.07
S^2Drug	**92.46***	**79.25***	**58.37***	**43.06***	**18.82***

Table 1: Virtual screening performance comparison of different methods on DUD-E. The bold indicates best result.

	AUROC	BEDROC	$EF^{0.5\%}$	$EF^{1\%}$	$EF^{5\%}$
Surflex	51.47	-	-	2.50	-
Glide-SP	53.15	4.00	3.17	3.41	2.01
DeepDTA	56.27	2.53	-	1.47	-
Planet	57.31	-	4.64	3.87	2.43
Gnina	60.93	5.40	-	4.63	-
Banana	**62.78**	5.02	3.98	3.79	2.83
SVSBI	53.77	3.84	3.05	2.71	1.92
DrugCLIP	56.36	6.78	7.77	5.66	2.32
DrugHash	54.58	7.14	9.65	6.14	2.42
S^2Drug	58.23	**8.69***	**11.44***	**7.38***	**2.97***

Table 2: Virtual screening performance comparison of different methods on LIT-PCBA. The bold indicates best result.

Overall Screening Performance To evaluate whether our proposed S^2Drug can effectively identify small-molecule ligands binding to the given protein pockets, we conduct experiments on DUD-E and LIT-PCBA datasets, with results shown in Tables 1 and 2. From these results, we can observe that S^2Drug significantly outperforms all docking-based and learning-based baselines on all metrics. In particular, S^2Drug outperforms recent retrieval-based methods including DrugCLIP and DrugHash by a large margin (e.g., 13.01 and 8.73 points on AUROC, respectively), even though these methods also employ the contrastive representation learning. This demonstrates that our two-stage representation learning framework-designed to bridge protein sequence and 3D structure-can indeed improve VS accuracy.

Evaluation on Homology Exclusion Scenarios To assess the S^2Drug's generalization ability, we conduct the experiments by varying the identity cutoffs between testing targets and training data in DUD-E. We set the identify cutoff as 90%, 60%, 30% and HMM, respectively, where the sequence identity decreases progressively. According to the results shown in Figure 2, we may notice that no matter under any homology exclusion threshold, S^2Drug can always achieve much higher virtual screening accuracy over DrugCLIP. This validates that S^2Drug indeed exhibits strong generalizability even under HMM scenario which enforces strict remote homology exclusion. Moreover, we observe that under 90% and 60% identity cutoff settings, S^2 still outperforms the performance of DrugHash and DrugCLIP obtained under the no exclusion setting. This evidence suggests our method significantly reduces the dependency on high train-test distribution similarity. This advantage can be attributed to the bilateral data sampling strategy used during the pretraining, which effectively mitigates bias and overfitting by filtering out redundant proteins and noisy ligands.

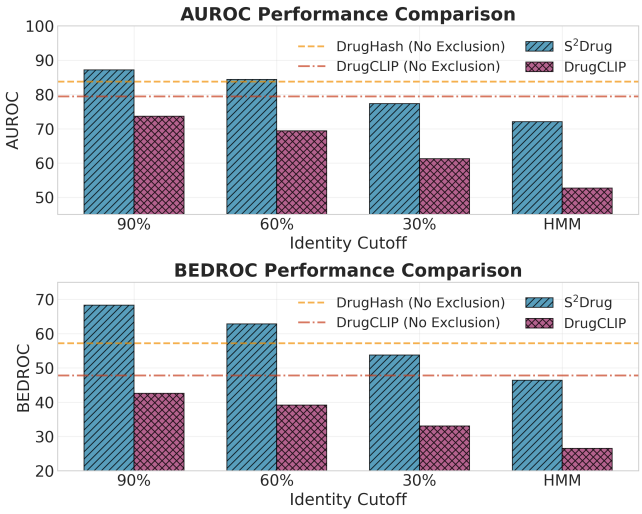

Figure 2: Virtual screening experiments on homology exclusion scenarios to evaluate method generalizability.

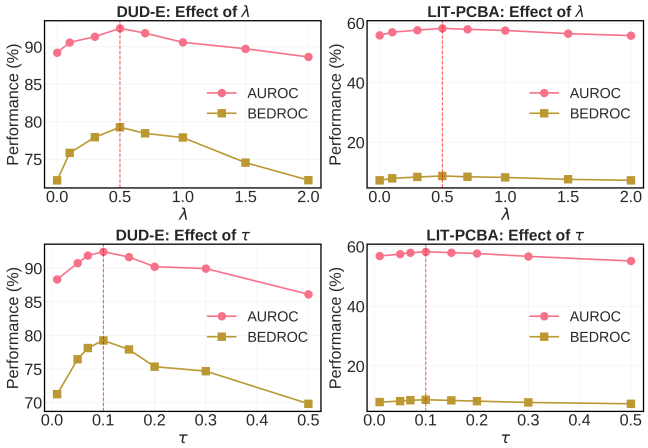

Figure 3: The hyperparameter robustness analysis for balancing coefficient λ and temperature parameter τ.

	DUD-E					LIT-PCBA				
	AUROC	BEDROC	$EF^{0.5\%}$	$EF^{1\%}$	$EF^{5\%}$	AUROC	BEDROC	$EF^{0.5\%}$	$EF^{1\%}$	$EF^{5\%}$
- BDS	88.73	73.91	50.83	38.42	16.54	56.12	7.21	10.06	6.41	2.66
- SSF	87.92	69.85	47.14	35.39	15.12	55.03	6.88	9.73	6.15	2.51
- BSP	89.58	75.60	52.93	39.27	17.36	56.47	7.56	10.49	6.85	2.73
S^2Drug	**92.46***	**79.25***	**58.37***	**43.06***	**18.82***	**58.23***	**8.69***	**11.44***	**7.38***	**2.97***

Table 3: Ablation studies on both DUD-E and LIT-PCBA datasets. The BDS, SSF, and BSP indicate bilateral data sampling, sequence-structure fusion module, and auxiliary binding site prediction task, respectively. The bold indicates the best result.

		P2Rank	VN-EGNN	DiffDock	VN-EGNNrank	S^2Drug
HOLO4K	F1 Score	42.63	47.25	**52.91**	48.84	51.66
	PR-AUC	36.20	41.17	**47.52**	44.51	46.97
COACH420	F1 Score	39.14	40.80	42.25	43.91	**47.32***
	PR-AUC	34.76	35.95	38.58	39.81	**42.62***
ASD	F1 Score	33.67	38.93	40.60	42.18	**48.79***
	PR-AUC	29.58	32.81	34.41	37.44	**43.96***

Table 4: Binding site prediction performance comparison of different methods across three datasets. Bold denotes the best.

Hyperparameter Robustness Analysis To investigate whether the performance of S^2Drug is robust to hyperparameter settings, we vary the values of the balancing coefficient λ in the combined training objective during finetuning stage (Eq. 15) and the temperature parameter τ in the contrastive learning objectives (Eq. 8 and 14). The experiment results have provided in the curves of Figure 3. From the table, we can first find the virtual screening accuracy is relatively robust, especially within the area around the default hyperparameter setting ($\lambda = 0.50$, $\tau = 0.1$). For λ, when its value is set higher than 1.0, this auxiliary binding site prediction task will introduce the optimization conflict and distract model learning away from the primary virtual screening task, leading to underfitting. As for extreme settings of τ, overly low or high values may lead to over-sharpening and over-smoothing issues for learning, respectively.

Ablation Studies To explore the contribution of each proposed technique to the final virtual screening performance, we conduct experiments by individually removing the bilateral data sampling, the sequence-structure fusion module (only finetuning protein sequence encoder on PDBBind data), and the auxiliary binding site prediction task from the overall training framework, respectively. As shown in Table 3, the removal of any of these modules leads to a significant performance drop across all evaluation metrics. This means each of such three parts brings the positive contribution to the overall framework effectiveness. Notably, the lack of sequence-structural fusion module will result in 9.40 and 1.81 points BEDROC decrease on DUDE-E and LIT-PCBA, respectively, highlighting the importance of bridging sequence and structural information for virtual screening.

Auxiliary Binding Site Prediction Performance To validate that our S^2Drug framework benefits both virtual screening and binding site prediction tasks simultaneously, we empirically compare it against several representative binding site prediction baselines. According to the results in Table 4, we can observe that S^2Drug achieves the highest accuracy on 2 out of 3 benchmarks-COACH240 and ASD-showing clear superiority over the strongest baseline. This also validates the central hypothesis of our framework: the binding site prediction task and virtual screening task enhance each other. Binding site information provides useful structural priors for virtual screening, and vice versa. On the evaluation dataset HOLO4K, although S^2Drug ranks the second, it still achieves a competitive performance with a 51.66% F1 Score and 46.97% PR-AUC, which are closed enough to the baseline DiffDock. This narrow gap can be attributed to DiffDock's diffusion-based generative nature, which excels at handling symmetric complexes (common in HOLO4K) by sampling multi-modal pose distributions and mitigating geometric redundancy from symmetric units; these properties ultimately contribute to its stronger results.

Conclusions and Future Works

In this paper, we propose S^2Drug, a two-stage contrastive representation learning framework that explicitly bridges protein sequence and 3D structural information to enhance virtual screening. Our approach pretrains a sequence encoder on large-scale protein–ligand data with bilateral sampling to reduce noise, and fine-tunes by fusing sequence and 3D pocket representations with an auxiliary binding site prediction task for improved spatial specificity. Extensive experiments across diverse benchmarks demonstrate that S^2Drug consistently outperforms baselines on VS accuracy. Additionally, our model achieves accurate binding site localization, validating the complementarity between the two tasks. Future work will explore the integration of protein surface and solvent features to enhance VS accuracy.

Acknowledgments

This work is supported by National Key R&D Program of China No.2025ZD1803103, Beijing Academy of Artificial Intelligence. Besides, this work is supported by the Early Career Scheme (No. CityU 21219323) and the General Research Fund (No. CityU 11220324) of the University Grants Committee (UGC), the NSFC Young Scientists Fund (No. 9240127), and the Donation for Research Projects (No. 9229164 and No. 9220187).

References

Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016): 493–500.

Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; and Church, G. M. 2019. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12): 1315–1322.

Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science*, 181(4096): 223–230.

Ballester, P. J.; and Mitchell, J. B. 2010. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9): 1169–1175.

Brocidiacono, M.; Francoeur, P.; Aggarwal, R.; Popov, K. I.; Koes, D. R.; and Tropsha, A. 2023. BigBind: learning from nonstructural data for structure-based virtual screening. *Journal of Chemical Information and Modeling*, 64(7): 2488–2495.

Corso, G.; StÃ, H.; Jing, B.; Barzilay, R.; Jaakkola, T.; et al. 2023. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. In *International Conference on Learning Representations (ICLR 2023)*.

Durrant, J. D.; and McCammon, J. A. 2011. NNScore 2.0: a neural-network receptor–ligand scoring function. *Journal of chemical information and modeling*, 51(11): 2897–2903.

Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. 2004. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7): 1739–1749.

Gao, B.; Qiang, B.; Tan, H.; Jia, Y.; Ren, M.; Lu, M.; Liu, J.; Ma, W.-Y.; and Lan, Y. 2023. Drugclip: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36: 44595–44614.

Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1): D1100–D1107.

Han, J.; Hong, Y.; and Li, W.-J. 2025. DrugHash: Hashing Based Contrastive Learning for Virtual Screening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17041–17049.

Hu, B.; Tan, C.; Xia, J.; Liu, Y.; Wu, L.; Zheng, J.; Xu, Y.; Huang, Y.; and Li, S. Z. 2024. Learning complete protein representation by dynamically coupling of sequence and structure. *Advances in Neural Information Processing Systems*, 37: 137673–137697.

Kallenborn, F.; Chacon, A.; Hundt, C.; Sirelkhatim, H.; Didi, K.; Cha, S.; Dallago, C.; Mirdita, M.; Schmidt, B.; and Steinegger, M. 2024. GPU-accelerated homology search with MMseqs2. *bioRxiv*, 2024–11.

Krivák, R.; and Hoksza, D. 2018. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10: 1–12.

Lee, Y.; Yu, H.; Lee, J.; and Kim, J. 2023. Pre-training sequence, structure, and surface features for comprehensive protein representation learning. In *The Twelfth International Conference on Learning Representations*.

Li, Z.; Li, M.; Zhu, L.; and Zhang, W. 2024. Improving PTM site prediction by coupling of multi-granularity structure and multi-scale sequence representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 188–196.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Liu, X.; Lu, S.; Song, K.; Shen, Q.; Ni, D.; Li, Q.; He, X.; Zhang, H.; Wang, Q.; Chen, Y.; et al. 2020. Unraveling allosteric landscapes of allosterome with ASD. *Nucleic acids research*, 48(D1): D394–D401.

Lyu, J.; Irwin, J. J.; and Shoichet, B. K. 2023. Modeling the expansion of virtual screening libraries. *Nature chemical biology*, 19(6): 712–718.

Maia, E. H. B.; Assis, L. C.; De Oliveira, T. A.; Da Silva, A. M.; and Taranto, A. G. 2020. Structure-based virtual screening: from classical to artificial intelligence. *Frontiers in chemistry*, 8: 343.

McNutt, A. T.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; and Koes, D. R. 2021. GNINA 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1): 43.

Mysinger, M. M.; Carchia, M.; Irwin, J. J.; and Shoichet, B. K. 2012. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14): 6582–6594.

Oord, A. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Öztürk, H.; Özgür, A.; and Ozkirimli, E. 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17): i821–i829.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118.

Schneckenreiter, L.; Luukkonen, S.; Friedrich, L.; Kuhn, D.; and Klambauer, G. 2025. Ligand-Conditioned Binding Site Prediction Using Contrastive Geometric Learning. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*.

Sestak, F.; Schneckenreiter, L.; Hochreiter, S.; Mayr, A.; and Klambauer, G. 2023. VN-EGNN: Equivariant Graph Neural Networks with Virtual Nodes Enhance Protein Binding Site Identification. In *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*.

Shen, L.; Feng, H.; Qiu, Y.; and Wei, G.-W. 2023. SVSBI: sequence-based virtual screening of biomolecular interactions. *Communications biology*, 6(1): 536.

Shoichet, B. K. 2004. Virtual screening of chemical libraries. *Nature*, 432(7019): 862–865.

Spitzer, R.; and Jain, A. N. 2012. Surflex-Dock: Docking benchmarks and real-world application. *Journal of computer-aided molecular design*, 26: 687–699.

Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; and Siedlecki, P. 2017. Pafnucy-a deep neural network for structure-based drug discovery. *Bioinformatics*, 1050: 19.

Tran-Nguyen, V.-K.; Jacquemard, C.; and Rognan, D. 2020. LIT-PCBA: an unbiased data set for machine learning and virtual screening. *Journal of chemical information and modeling*, 60(9): 4263–4273.

Trott, O.; and Olson, A. J. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2): 455–461.

Wang, Y.; Wang, J.; Cao, Z.; and Barati Farimani, A. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3): 279–287.

Wu, Y.; Gao, M.; Zeng, M.; Zhang, J.; and Li, M. 2022. BridgeDPI: a novel graph neural network for predicting drug–protein interactions. *Bioinformatics*, 38(9): 2571–2578.

Yazdani-Jahromi, M.; Yousefi, N.; Tayebi, A.; Kolanthai, E.; Neal, C. J.; Seal, S.; and Garibay, O. O. 2022. AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification. *Briefings in Bioinformatics*, 23(4): bbac272.

Zhang, X.; Gao, H.; Wang, H.; Chen, Z.; Zhang, Z.; Chen, X.; Li, Y.; Qi, Y.; and Wang, R. 2023. Planet: a multi-objective graph neural network model for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling*, 64(7): 2205–2220.

Zheng, L.; Fan, J.; and Mu, Y. 2019. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega*, 4(14): 15956–15965.

Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; and Ke, G. 2023. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In *The Eleventh International Conference on Learning Representations*.