Multi-Granular Contrastive Alignment and Fusion for Fragment-Enhanced Virtual Screening

Haichuan Tan [*1,2] **Bowen Gao** [*1,2] **Jiaxin Li** [3] **Yanwen Huang** [4] **Wenyu Zhu** [1] **Jianhui Wang** [5] **Yinjun Jia** [1] **Yuanhuan Mo** [6] **Ya-Qin Zhang** [1] **Wei-Ying Ma** [1] **Yanyan Lan** [1,7]

Abstract

Virtual screening (VS) accelerates drug discovery by identifying bioactive molecules from large libraries. Recent deep learning methods treat VS as a dense retrieval task, embedding protein pockets and molecules into a shared space. However, these models rely on whole-molecule representations, limiting their ability to capture fine-grained, fragment-level interactions—despite the fragment-centric nature of the growing importance of fragment-based drug discovery (FBDD). We introduce FragCLIP, a fragment-centric, two-stage retrieval framework with multi-granular contrastive learning. Stage one learns to jointly embed protein pockets, molecules, and fragments, guided by non-covalent interaction (NCI) supervision. Stage two fuses molecule- and fragment-level embeddings into a unified representation. This design enables accurate alignment of interaction-relevant fragments with compatible pockets while preserving efficiency. FragCLIP boosts early enrichment (EF1) on DUD-E from 31.89 to 37.23 and outperforms docking and deep learning baselines on a new fragment-level benchmark FragBench. FragCLIP bridges molecular and substructure-level reasoning, offering a practical foundation for structure-based virtual screening in realistic FBDD workflows.

[*]Equal contribution [1]Institute for AI Industry Research (AIR), Tsinghua University [2]Department of Computer Science and Technology, Tsinghua University [3]School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications [4]Department of Pharmaceutical Science, Peking University [5]University of Electronic Science and Technology of China [6]School of Software Engineering, South China University of Technology [7]Beijing Academy of Artificial Intelligence (BAAI). Correspondence to: Yanyan Lan <lanyanyan@air.tsinghua.edu.cn>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Structure-based virtual screening (SBVS)(Maia et al., 2020; Lyu et al., 2019) has become a cornerstone of modern drug discovery, enabling the identification of bioactive molecules from large chemical libraries by modeling their interactions with protein targets. Recent progress in deep learning has reframed this task as a retrieval problem: dual-encoder models independently embed molecules and protein pockets into a shared latent space and compute binding relevance via similarity. This design, exemplified by models such as DrugCLIP (Gao et al., 2023), supports large-scale screening with impressive scalability and speed.

Despite their efficiency, current dual-encoder methods use coarse, whole-molecule representations that overlook fine-grained binding interactions. This is a critical gap, given the fragment-centric nature of real-world chemical libraries like Enamine REAL (Shivanyuk et al., 2007) and the central role of fragment-based drug discovery (FBDD) in early-stage medicinal chemistry (Jinsong et al., 2024).

To address this modeling bottleneck, we propose a fragment-centric and interaction-aware framework, **FragCLIP**, that bridges global molecular representation with localized fragment-level reasoning, as shown in Figure 1. Our approach introduces a multi-encoder architecture that jointly embeds protein pockets, full molecules, and chemically meaningful molecular fragments. Through hierarchical contrastive learning, the model captures structural signals across granularities, aligning both molecules and fragments with protein pockets in a unified representation space.

We train FragCLIP with interaction-aware fragment supervision derived from non-covalent interactions (NCIs) identified by PLIP (Salentin et al., 2015), enabling the model to distinguish binding-relevant fragments from non-interacting decoys. A fusion module then integrates molecule- and fragment-level embeddings into a unified retrieval representation. This two-stage setup preserves dual-encoder efficiency while capturing fine-grained structural detail.

FragCLIP is evaluated on standard molecule-level benchmarks and a new fragment-level virtual screening (F-VS) benchmark built from DUD-E with NCI-based labels. It

outperforms both docking and retrieval-based baselines, and uniquely enables direct retrieval from fragment libraries—bridging predictive screening with fragment-based design and generation.

2. Related Works

Structure-based virtual screening (SBVS) identifies bioactive molecules by modeling interactions between protein pockets and ligands. While traditional methods rely on docking (Trott & Olson, 2010), recent deep learning models (Zhou et al., 2023) improve generalization using graph and 3D encoders. Retrieval-based methods like DrugCLIP (Gao et al., 2023) and UniMol scale well via contrastive learning but focus only on whole molecules.

In contrast, fragment-based drug discovery (FBDD) targets small functional substructures (Bon et al., 2022; Hajduk & Greer, 2007), yet fragment-level screening remains largely unexplored. We address this gap by integrating fragment representations into a contrastive retrieval framework and introducing new fragment-level screening tasks with both weak and strong interaction supervision.

3. Method

3.1. Preliminary

In structure-based virtual screening, the goal is to identify molecules that bind to a target pocket. Let $\mathcal{P}$, $\mathcal{M}$, and $\mathcal{F}$ denote the spaces of protein pockets, molecules, and fragments, respectively. Traditional deep learning approaches model this as a binary classification task over pairs $(p, m) \in \mathcal{P} \times \mathcal{M}$, predicting binding likelihood. Recent methods reframe this as a dense retrieval problem using a dual-encoder architecture with embedding functions $f_p(p)$ and $f_m(m)$ for protein pockets and molecules, respectively. These functions map inputs into a shared latent space, where similarity (e.g., cosine similarity) supports efficient large-scale screening.

We extend this framework by introducing fragments as a third modality. Each fragment $f \in \mathcal{F}_m \subset \mathcal{F}$ is a substructure of molecule m, and is encoded by a fragment encoder $f_f(f)$. Our model operates on triplets $(p, m, \mathcal{F}_m)$, embedding all three modalities—pockets, molecules, and fragments—into a unified space. This design enables fragment-level supervision, two-stage retrieval, and supports downstream fragment-based drug design workflows.

3.2. Fragment Segmentation and Importance Labeling

To support fragment-level learning and retrieval, we decompose molecules into chemically meaningful substructures using the BRICS algorithm (Degen et al., 2008), generating synthetically accessible and interpretable fragments. We exhaustively extract all BRICS-based substructures and retain those with 8–24 heavy atoms, consistent with typical fragment library sizes like Enamine REAL. To reduce redundancy and promote diversity, we cluster fragments based on fingerprint similarity. This curated fragment set underpins our multi-resolution alignment of protein pockets, molecules, and their substructures.

We assign fragment-level binding relevance using PLIP (Salentin et al., 2015), a rule-based tool that detects non-covalent interactions (NCIs) in protein–ligand complexes. Applied to all training structures (Wang et al., 2005), PLIP provides atomic-level interaction annotations. A fragment is labeled as *positive* if any of its atoms participate in an interaction; otherwise, it is *negative*. This supervision enables the model to distinguish functionally relevant fragments from inactive ones.

3.3. Multi-Granularity Alignment Learning

We aim to learn a structured embedding space that captures both global molecular semantics and local fragment-level binding signals for accurate protein–ligand alignment. To achieve this, we use a multi-encoder architecture with a protein encoder f_p, a molecule encoder f_m, and a fragment encoder f_f, trained jointly via contrastive objectives.

The protein encoder maps 3D pockets into a latent space, while f_m and f_f operate on molecular graphs. We decouple their optimization: f_m captures scaffold-level features, and f_f focuses on interaction-relevant substructures.

Training is guided by three contrastive losses:

$$\mathcal{L}_{\text{p-m}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(f_p(p_i), f_m(m_i))/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(f_p(p_i), f_m(m_j))/\tau)} \tag{1}$$

$$\mathcal{L}_{\text{p-f}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(f_p(p_i), f_f(f_i))/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(f_p(p_i), f_f(f_j))/\tau)} \tag{2}$$

$$\mathcal{L}_{\text{m-f}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(f_m(m_i), f_f(f_i))/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(f_m(m_i), f_f(f_j))/\tau)} \tag{3}$$

These losses jointly align full molecules and key fragments with the same pocket while preserving intra-ligand coherence. The total objective is:

$$\mathcal{L}\text{align} = \mathcal{L}\text{p-m} + \lambda_1\mathcal{L}\text{p-f} + \lambda_2\mathcal{L}\text{m-f}. \tag{4}$$

3.4. Multi-Granular Fusion for Fragment-Aware Retrieval

To unify molecule- and fragment-level signals for retrieval, we introduce a learnable fusion module f_{fusion} that generates

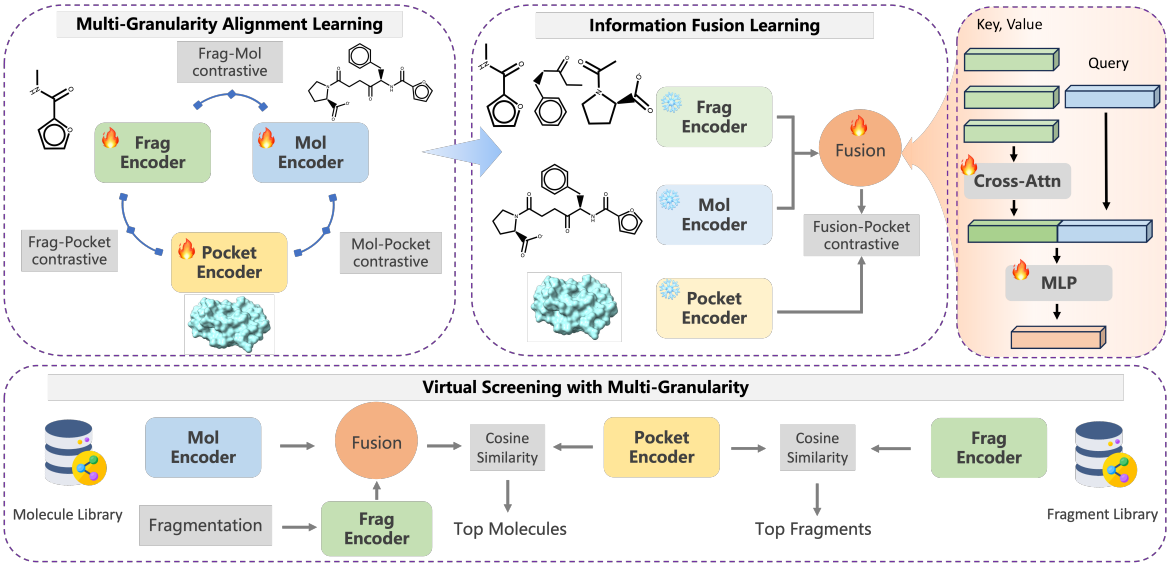

Figure 1. The framework of FragCLIP consists of two main stages: Multi-Granularity Alignment Learning and Information Fusion Learning. During virtual screening, FragCLIP supports both molecule-level and fragment-level screening.

a pocket-aware embedding by aggregating global and local chemical information. In this second training stage, the encoders are frozen, and only the fusion module is trained.

Given a full-molecule embedding $f_m(m)$ and a set of its associated fragment embeddings $\{f_f(f_i)\}_{i=1}^k$, the fusion module performs a cross-attention operation where the molecule embedding acts as the query and the fragment embeddings serve as keys and values. The attention output is concatenated with the original molecule embedding and passed through a multilayer perceptron (MLP) to yield the final fused representation:

$$z_{\text{fusion}} = \text{MLP}\left(f_m(m) \,\|\, \text{Attn}(f_m(m), \{f_f(f_i)\}_{i=1}^k)\right), \tag{5}$$

This design allows the model to adaptively attend to fragment-level substructures conditioned on the molecule context, highlighting those most relevant to potential binding interactions.

The fused representation is trained to align with the corresponding pocket embedding via a contrastive loss:

$$\mathcal{L}_{\text{fusion}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(f_p(p_i), z_{\text{fusion},i})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(f_p(p_i), z_{\text{fusion},j})/\tau)}. \tag{6}$$

This objective encourages the fused embedding to reflect both molecule-level semantics and fragment-level interaction cues.

Ensemble for fragment-aware retrieval To integrate information across molecular granularities, we adopt a score-level ensemble strategy. The ligand score is computed as the cosine similarity between molecule and pocket embeddings. For the fragment score, each molecule is decomposed into BRICS fragments; we compute pocket similarity for each, and average the top-3 maxima to emphasize key substructures.

The fusion score is obtained by passing the molecule and its fragments through a cross-attention-based fusion module, producing a fused embedding compared to the pocket via cosine similarity.

The final retrieval score combines all three:

$$\text{score} = \text{ligand\_score} + \alpha \cdot \text{frag\_score} + \beta \cdot \text{fusion\_score}, \tag{7}$$

with $\alpha = \beta = 0.8$ tuned on validation. This ensemble boosts accuracy by combining scaffold-, fragment-, and fusion-level signals.

3.5. FragBench: Fragment Retrieval Benchmark Creation

To evaluate fragment-level retrieval under realistic conditions, we build **FragBench** from the DUD-E (Mysinger et al., 2012) dataset. Since DUD-E lacks protein–ligand complex structures, we generate docking poses using Schrödinger GLIDE. Ligands are fragmented via BRICS, and non-covalent interactions (NCIs) are extracted to label fragments. A fragment is labeled as positive if it comes from an active molecule and includes at least one atom involved in direct interaction; negatives are sampled from decoys.

To reduce redundancy, we cluster fragments by fingerprint similarity and keep up to two per cluster. The dataset is balanced to match the DUD-E active-to-decoy ratio (1:12.7),

resulting in 19,825 positives and 1,301,610 negatives. This benchmark offers a diverse, interaction-aware testbed for fragment-level retrieval evaluation.

4. Experiments

4.1. Experiment Settings

4.1.1. Tasks

We evaluate our method on two tasks: **molecule-level virtual screening (M-VS)** and **fragment-level virtual screening (F-VS)**. For M-VS, we follow the standard active/decoy splits from established benchmarks. For F-VS, we use the FragBench dataset introduced in Section 3.5.

4.1.2. Baseline Models

We compare FragCLIP with several representative virtual screening methods. For both molecule-level and fragment-level virtual screening, we benchmark against classical docking algorithms such as AutoDock (Trott & Olson, 2010) Vina and Schrödinger's Glide (Halgren et al., 2004) , as well as representative deep learning-based approaches, including (Durrant & McCammon, 2011; Ballester & Mitchell, 2010; Stepniewska-Dziubinska et al., 2018; Liangzhen Zheng & Mu, 2019; Zhang et al., 2023; Gao et al., 2023).

4.1.3. Metrics

We evaluate model performance using standard virtual screening metrics (Gao et al., 2023): AUC, enrichment factor (EF), and BEDROC, which emphasizes early recognition. For F-VS, metrics are computed by ranking fragments; for M-VS, by ranking full molecules.

4.2. Molecule Level Retrieval

On the molecule-level DUD-E (Mysinger et al., 2012) benchmark, FragCLIP consistently eclipses both traditional docking and leading learning-based methods, as shown in Table 1. Compared with the baseline DrugCLIP, FragCLIP boosts BEDROC (early recognition) from 50.52 $\rightarrow$ 59.32, meaning more true actives are recovered at the top of the ranked list. Notably, the fragment-only variant ("FragCLIP w/o fusion") already outperforms every baseline, and the multi-granular fusion module contributes an additional 6 BEDROC points. These gains demonstrate that combining global molecular context with fragment-level cues yields a markedly richer representation while preserving dual-encoder efficiency.

4.3. Fragment Level Retrieval

On FragBench, the fragment-level virtual-screening benchmark, FragCLIP markedly surpasses both classical docking protocols and DrugCLIP baselines. It yields higher overall discrimination and substantially stronger early-enrichment. These improvements empirically validate the benefit of explicitly encoding fragment-level interaction cues and establish FragCLIP as a robust screening framework for fragment-based drug-discovery workflows.

Table 1. Molecule-level virtual screening performance on the DUD-E dataset. Metrics are averaged across targets.

Method	AUC ↑	BEDROC ↑	EF ↑		
			0.5%	1%	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
RFscore	65.21	12.41	4.90	4.52	2.98
Pafnucy	63.11	16.50	4.24	3.86	3.76
OnionNet	59.71	8.62	2.84	2.84	2.20
Planet	71.60	–	10.23	8.83	5.40
DrugCLIP	80.93	50.52	38.07	31.89	10.66
FragCLIP (w/o Fusion)	84.76	53.61	40.64	33.56	11.39
FragCLIP	**85.44**	**59.32**	**42.93**	**37.23**	**12.45**

Table 2. Fragment-level virtual screening performance on the FragBench dataset. Metrics are averaged across targets.

Method	AUC ↑	BEDROC ↑	EF ↑		
			0.5%	1%	5%
Vina	0.54	0.06	4.39	3.55	2.16
Glide	0.59	0.10	10.22	7.16	3.11
DrugCLIP	0.61	0.18	16.50	11.23	4.04
FragCLIP	**0.74**	**0.30**	**26.36**	**19.07**	**7.05**

5. Limitations and Future Works

While FragCLIP is evaluated standalone, real-world FBDD demands integration with fragment growing, generative design, and iterative feedback—embedding FragCLIP into such closed-loop pipelines is a key direction ahead.

6. Conclusion

FragCLIP introduces a fragment-centric framework for structure-based virtual screening, jointly embedding protein pockets, full molecules, and chemically meaningful fragments. By incorporating a learnable fusion module, it achieves strong early-stage enrichment at both molecule and fragment levels while retaining the efficiency of dual

encoders. Crucially, FragCLIP enables direct fragment retrieval, bridging predictive screening with fragment-based drug design and paving the way for more integrated, fine-grained discovery workflows.

Acknowledgements

This work is supported by Beijing Academy of Artificial Intelligence (BAAI).

References

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

Bon, M., Bilsland, A., Bower, J., and McAulay, K. Fragment-based drug discovery—the importance of high-quality molecule libraries. *Molecular Oncology*, 16(21): 3761–3777, 2022.

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., and Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, October 2008. doi: 10.1002/cmdc.200800178.

Durrant, J. D. and McCammon, J. A. Nnscore 2.0: a neural-network receptor–ligand scoring function. *Journal of Chemical Information and Modeling*, 51(11):2897–2903, 2011.

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

Hajduk, P. J. and Greer, J. A decade of fragment-based drug design: Strategic advances and lessons learned. *Nature Reviews Drug Discovery*, 6:211–219, 2007. doi: 10.1038/nrd2220.

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., and Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of Medicinal Chemistry*, 47(7):1750–1759, 2004.

Jinsong, S., Qifeng, J., Xing, C., Hao, Y., and Wang, L. Molecular fragmentation as a crucial step in the ai-based drug development pathway. *Communications Chemistry*, 7(1):20, 2024.

Liangzhen Zheng, J. F. and Mu, Y. Onionnet: A multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS Omega*, 4(14):15956–15965, 2019.

Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K., et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

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

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

Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., and Schroeder, M. Plip: fully automated protein–ligand interaction profiler. *Nucleic acids research*, 43(W1):W443–W447, 2015.

Shivanyuk, A. N., Ryabukhin, S. V., Tolmachev, A., Bogolyubsky, A., Mykytenko, D., Chupryna, A., Heilman, W., and Kostyuk, A. Enamine real database: Making chemical diversity real. *Chemistry today*, 25(6):58–59, 2007.

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018.

Trott, O. and Olson, A. J. 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, 2010.

Wang, R., Fang, X., Lu, Y., Yang, C.-Y., and Wang, S. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005.

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

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-mol: A universal 3d molecular representation learning framework. 2023.