MACHINE LEARNING
Science and Technology

PAPER

Flow-based fragment identification via binding site-specific latent representations

Rebecca M Neeser[1], **Ilia Igashov**[1], **Arne Schneuing**[1], **Michael Bronstein**[1,2,3,4], **Philippe Schwaller**[1,5] **and Bruno Correia**[1,*]

[1] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2] Oxford University, Oxford, United Kingdom
[3] Proxima, NY, United States of America
[4] Aithyra Research Institute for Biomedical Artificial Intelligence, Vienna, Austria
[5] National Centre of Competence in Research (NCCR) Catalysis EPFL, Lausanne, Switzerland
[*] Author to whom any correspondence should be addressed.

E-mail: bruno.correia@epfl.ch

OPEN ACCESS

RECEIVED
19 September 2025

REVISED
30 March 2026

ACCEPTED FOR PUBLICATION
9 April 2026

PUBLISHED
14 May 2026

Keywords: fragment based drug design, machine learning, fragment identification, drug design, representation learning, generative modelling

Supplementary material for this article is available online

Abstract

Fragment-based drug design is a promising strategy leveraging the binding of small chemical moieties that can efficiently guide drug discovery. The initial step of fragment identification remains challenging, as fragments often bind weakly and non-specifically. We developed a protein-fragment encoder that relies on a contrastive learning approach to map both molecular fragments and protein surfaces in a shared latent space. The encoder captures interaction-relevant features and achieves strong discrimination between binding and non-binding regions, reaching ROC–area under the curve values of 0.92 on pocket surfaces and enrichment factors of 22.85 across full protein surfaces. Building on this representation, our generative method LatentFrag produces chemically realistic fragment identities and positions conditioned on the protein surface. LatentFrag improves fragment recovery over docking-based virtual screening, achieving a sampling hit rate more than four times higher at a fraction of its computational cost providing a valuable starting point for fragment hit discovery. We further show the practical utility of LatentFrag and extend the workflow to full ligand design tasks. Together, these approaches contribute to advancing fragment identification and provide valuable tools for fragment-based drug discovery.

1. Introduction

Hit identification remains a critical challenge in drug discovery, despite advances in screening techniques and computational tools (Hasselgren *et al* 2024, Jalencas *et al* 2024). Traditional high-throughput screening (HTS) workflows to identify ligands targeting proteins of interest have limitations, particularly in exploring efficiently large chemical spaces and identifying actionable starting points for drug design (Edfeldt *et al* 2011). Fragment-based drug design (FBDD) offers a promising alternative, leveraging smaller chemical moieties that can be combined to form potent ligands (Congreve *et al* 2008, Hubbard *et al* 2011).

Fragment screening has several advantages over conventional HTS approaches, that use larger, drug-like molecules. Smaller sized fragments typically exhibit higher ligand efficiency, a measure of binding energy per atom, and their smaller size allows exploration of bigger chemical space (Congreve *et al* 2008). Although fragments generally have very low binding affinities than larger ligands, the combination of multiple key interactions with the target protein can yield a more specific ligand (Edfeldt *et al* 2011, Yu *et al* 2020).

Recent machine learning (ML) approaches enable rapid exploration of chemical space, yet many FBDD methods do not account for a protein structure, rely on prior knowledge of fragment

hits (McCorkindale *et al* 2022), or depend on classical molecular docking (Bian *et al* 2018, Marchand *et al* 2018). ML-based structure-based drug design (SBDD) incorporates three-dimensional information but frequently generates unrealistic or synthetically inaccessible molecules when designing full ligands in an all-atom approach (Buttenschoen *et al* 2024, Yang *et al* 2024). This challenge is less pronounced in FBDD as the use of known fragments constrains the chemically accessible space. However, despite *in silico* methods trying to tackle this problem (Imrie *et al* 2020, Neeser *et al* 2023, Igashov *et al* 2024, Ferla *et al* 2025), challenges of merging, linking and growing fragments to a full ligand remain. Fragment docking often is unable to localise binding sites within protein pockets as fragments often lack the necessary structural information content for docking algorithms to identify good binding poses and meaningful docking scores. These challenges further emphasise the need for new, more effective, structure-based FBDD solutions (De Esch *et al* 2021).

To address these limitations, we introduce a novel structure-based fragment screening approach that learns a protein-fragment representation through contrastive training. Our encoder maps protein surfaces and molecular fragments into a shared latent space. These rich fragment representations capture aspects of binding interactions with the target while maintaining chemical relevance through fragment similarity-based penalties. The resulting latent representations can be directly used in virtual screening (VS). Inspired by Igashov *et al* (2022), who performed ligand and fragment screening based on pocket representations but using a model trained in protein–protein interactions, our model is specifically tailored for fragment screening and does not depend on the availability of fragments in crystal structures.

Beyond this, we extend the method to a generative framework for structure-based fragment identification called LatentFrag. Instead of relying on a fixed fragment library, our generative approach, explicitly conditioned on the protein pocket, learns distributions of fragments and their spatial organisation directly in the latent space. This approach does not require a decoder, but rather queries a fragment library for the most similar fragment embedding ensuring chemically realistic fragments. This mechanism allows us to replace or expand the fragment library without retraining the generative model or even without re-sampling once embeddings were generated for one target. By representing ligands as fragment graphs during training, LatentFrag is able to generate more than one fragment per model call increasing efficiency. To train our models, we curated a dataset of pairs of proteins and fragmented drug-like ligands from the protein data bank (PDB) (Berman *et al* 2000). The fragmentation was carried out in a manner that yields synthetically sensible fragments while ensuring sufficient binding specificity by limiting the minimal size (Degen *et al* 2008).

By assessing both ‘hard’ recovery metrics and ‘soft’ pharmacophoric similarity, we provide insights into the model’s ability to identify meaningful fragment hits. Our analysis demonstrates improvements over VS baselines, like docking, while improving computational efficiency. We further demonstrate in a case study how our framework can be complemented with existing methods (Igashov *et al* 2024) to extend our fragment hits to full drug-like ligands. We showcase this pipeline by targeting the therapeutically relevant protein c-Met finding many fragments recovering known interactions but also many of which have new interactions to the target. The most promising fragments are subsequently connected resulting in a structure with improved *in silico* properties over the reference ligand BMS-777607.

2. Previous work

Molecular representation learning (MRL) is a well-established field, with numerous studies refining and adapting methods for specific tasks such as property and reaction prediction (Guo *et al* 2022). Various MRL approaches leverage language models as encoders (Shin *et al* 2019, Chithrananda *et al* 2020, Li *et al* 2021) or introduce specialised representations, such as UniMol, which incorporates 3D conformers (Zhou *et al* 2023), and MolR, which is tailored for reaction-based learning (Wang *et al* 2021). However, these methods typically do not explicitly account for protein targets and are not primarily designed for hit screening or drug discovery. Gao *et al* (2023) proposed DrugCLIP, which contrastively learns pocket and ligand representations for the task of VS based on the UniMol encoder architecture (Zhou *et al* 2023). This method conceptually shares many aspects with our work but encodes the pocket globally and full ligands making the task of fragment placement impracticable. While none of the aforementioned approaches focus specifically on molecular fragments, Chakravarti *et al* (2018) propose a fragment-based method, though it remains centred on chemical properties relevant to tasks like property prediction rather than interaction-driven applications. A concurrent work by Lohmann *et al* (2024) proposes a conceptually similar encoder of both protein and full ligand in 3D, which allows to analyse

protein pockets in latent space. However, the representation is obtained by training for the task of affinity prediction and based on the protein graph instead of the surface.

The task of computational fragment identification has been mostly dependent on fragment docking (Bian *et al* 2018, Marchand *et al* 2018), which is less accurate than ligand docking. FRESCO (McCorkindale *et al* 2022) is a ML-based method that implicitly considers target structure through pharmacophore distributions. The extraction of those, however, requires known hits from fragment screens and respective crystal structures, which is often not available in a drug discovery campaign. The closest approach to ours by Igashov *et al* (2022) matches protein pocket embeddings in order to find related fragment hits, making it also limited by the availability of crystal structures.

3. Methods

3.1. Protein-fragment contrastive learning

The protein-fragment encoder is trained contrastively and is designed to produce expressive latent embeddings for both fragments and protein surfaces (for details see appendix A). Thus, the latent vectors capture critical features for binding interactions and are uniquely suited to the task of fragment identification.

Training involves maximising the cosine similarity between embeddings of fragments and protein surface points within 3 Å of any fragment atom (positive pairs $\mathcal{P}^+$), while minimising the similarity for other surface points elsewhere on the protein (negative pairs $\mathcal{P}^-$), ensuring a robust distinction. Negative examples are selected from the pocket to include both convex and concave protein surface geometries as well as points sampled uniformly. This prevents overfitting to concave regions, abundant among positive examples. The surface curvature is predicted on the fly by a concurrently trained classifier. Protein surface embeddings h_p are parametrised by a geodesic convolutional neural network similar to dMaSIF (Sverrisson *et al* 2021) while fragments are processed by a graph transformer (Dwivedi *et al* 2020, Vignac *et al* 2022) yielding embedding h_f. The contrastive training loss is the mean of the following positive and negative contributions:

$$\mathcal{L}_{\text{pos}} = -\mathbb{E}_{(p,f)\in\mathcal{P}^+}\left[w(p,f)\log\sigma\left(\cos\left(h_p,h_f\right)\right)\right] \tag{1}$$

$$\mathcal{L}_{\text{neg}} = -\mathbb{E}_{(p,f)\in\mathcal{P}^-}\left[\log\sigma\left(-\cos\left(h_p,h_f\right)\right)\right] \tag{2}$$

with the sigmoid function σ, the cosine similarity $\cos(h_p,h_f)$ (cf. equation (A.1)), and $w(p,f)$, which scales the contribution of positive pairs by their distances d (cf. equation (A.4)). This up-weights positive points that are further away, encouraging the model to focus on points at the edge of the binding regions. Further details are provided in appendix A.1.

To ensure chemical relevance, a FSP is incorporated via a hinge loss. The FSP is computed between the true fragment embedding h_f^+ and a fragment h_f^- from the library with Tanimoto similarity below 0.1:

$$\mathcal{L}_{\text{FSP}} = \text{ReLU}\left(\cos\left(h_f^+,h_f^-\right) - c\right) \tag{3}$$

with margin c. This loss discourages molecules with low chemical similarity from having similar embeddings and in turn encourages diversity of embeddings, as shown in figure 1(A). In order to integrate interaction-specific information we employ an additional classification loss. A multi-label classifier is incorporated that predicts the types of NCI, if present, that each protein surface point can engage in. This encourages the differentiation of the pocket surface and aims at increasing sensitivity to different fragments. The final training loss is the weighted sum of the contrastive loss and the auxiliary FSP and NCI losses including L2 regularisation (cf. equation (A.9)).

Notably, the relative positions of the molecules are only used for assigning positive and negative examples during training, and do not directly influence the embeddings as fragments are represented as 2D graphs. This design promotes flexible, position and conformation-agnostic representations of fragment and surface features, enabling their broad applicability in different FBDD scenarios.

3.2. LatentFrag: fragment identification via flow matching

To identify and place relevant fragments in a given protein pocket, we introduce LatentFrag. LatentFrag is a generative modelling approach using flow matching (Lipman *et al* 2022), representing proteins as surface point clouds and ligands as coarse fragment graphs. Protein surface points are featurised by latent vectors and ligand nodes representing fragments are defined by a latent embedding (fragment type)

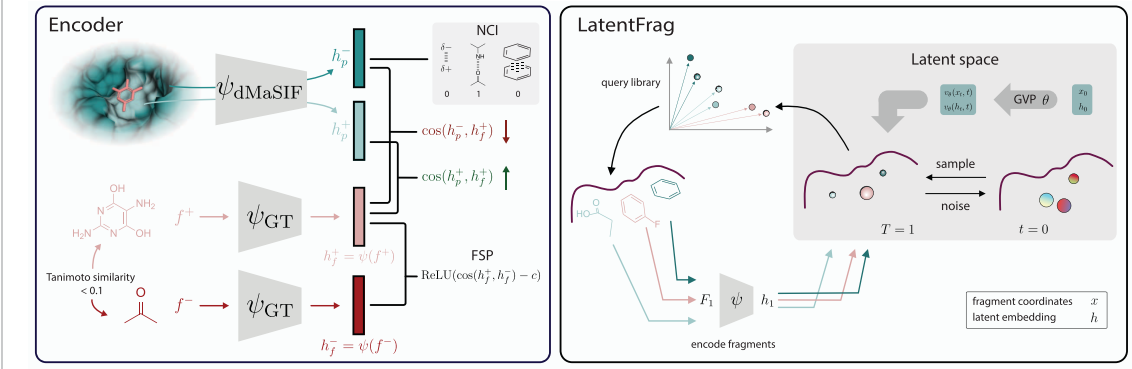

Figure 1. *Left:* Encoder pipeline, maximising the cosine similarity $\cos(\cdot)$ between fragment h_f and contacting pocket surface embeddings h_p while minimising the similarity between negative pairs. Additionally, a fragment similarity penalty (FSP) and loss on classifying potential non-covalent interactions (NCIs) from protein surfaces are incorporated. *Right:* LatentFrag: Generative pipeline starting with encoding and the flow matching model operating in latent space..

and the arithmetic mean of their coordinates (centroid). The latent vectors are learned embeddings from our protein-fragment encoder. A schematic overview of the fragment identification process is shown in figure 1(B) and the neural network is detailed in figure B.1.

LatentFrag learns to map noise to structured data by matching modelled probability flows. This enables efficient sampling from complex distributions. Two flows are employed: a spherical flow for the latent fragment embeddings, which assumes a unit sphere prior, and a Euclidean flow for the centroids, with a Gaussian prior. Once the latent fragment representations have been generated with LatentFrag, we query the precomputed library by cosine similarity and place the most similar fragments. This approach ensures fragment consistency with respect to chemical plausibility and geometry. We use a library with 41 224 unique fragments extracted from the PDB (Berman *et al* 2000) (appendix C).

The predicted coordinate represents the centre of the fragment, and obtaining its orientation requires downstream docking processes. This framework offers flexibility while aligning with FBDD workflows where a successful campaign is dependent on robust fragment identification. Detailed information on the generative frameworks is described in the appendix B.

3.3. Evaluation of fragment identification task

We evaluated fragment identification through two approaches: VS using our latent embeddings (Latent VS) and the generative framework for sampling fragment embeddings and their centroids. These are compared against VS based on docking (Docking VS) and a random baseline, evaluated on 100 protein targets with $\leqslant 30\%$ sequence similarity to the training set.

For Latent VS, fragments are ranked by the sum of cosine similarities between the fragment embeddings and the protein pocket surface points. The top 100 fragments per target are then selected for evaluation. Using LatentFrag, for each surface 100 samples are generated with number of fragments corresponding to the reference number of fragments. Generated latent representations are used to query a library for the closest fragments based on cosine similarity. Sampled fragments are subsequently docked using Gnina (McNutt *et al* 2021) within a restricted volume around predicted centroids. The random baseline is established by slightly noising the centroids of the ground truth fragments and randomly assigning fragments from the library to them. We evaluate using three key metrics:

- **Hard Recovery—Sampling Hits:** total number of generated fragments exactly matching references in 2D
- **Hard Recovery—Unique Fragments:** number of unique reference fragments recovered
- **Soft Recovery:** shape and pharmacophoric similarity of the docked fragment to the full reference ligand via SuCOS score (Leung *et al* 2019)

The reference fragment count for the calculation of recovery rates often exceeds generated fragments per target and sample due to data augmentation combining BRICS fragmentation (Degen *et al* 2008) with graph partitioning, allowing reference fragments to be substructures of others. Correspondingly, recovery rates are calculated by dividing the number of hits by number of samples (hit rate) or number of reference fragments also found in the library (recovery rate). The VS baselines are divided by the total

Table 1. Metrics evaluating the encoder on the test set and comparing to the theoretical random baseline. All scores are computed on the cosine similarity between fragment and protein surface representations with points labelled as true being 3 Å away from the fragment.

Method	Surface set	ROC AUC ↑	EF_1 ↑	AUPR ↑
Encoder	Whole	0.92	22.85	0.31
	Pocket	0.76	2.28	0.39
Random	Whole	0.5	1.00	0.01
	Pocket	0.5	1.00	0.16

number of selected fragments for evaluation (100 fragments × 100 targets). Detailed information can be found in appendix D.2.

3.4. Data

To build the fragment library, we extracted protein-ligand structures from the PDB (Berman *et al* 2000) and remove ligands irrelevant to the task such as solvents and buffers. Subsequently ligands are fragmented using BRICS rules (Degen *et al* 2008) while not allowing double bonds to be broken and reassemble fragments to reach a minimum fragment size of 8 heavy atoms in a combinatorial manner. This data processing pipeline makes sure that all fragments are sensible from a medicinal chemistry standpoint and carry enough information with respect to interactions to proteins. The dataset is split into training, validation and test following the approach used for HoloProt (Somnath *et al* 2021), which is based on precomputed 30% sequence similarity. For more detailed information we refer to appendix C.

The protein-fragment encoder is trained on individual fragments paired with the respective contacting protein surface ($\leqslant$5 Å) and the protein ligand interaction profiler (Adasme *et al* 2021) assigns interactions. The generative model is trained on the same dataset but with all fragments originating from a ligand as one data point. We further make use of the data augmentation outlined above, in which fragments below the defined size threshold are recombined, leading to multiple possible fragment combinations per ligand. We further restrict the protein surface to the pocket by discarding points further than 7 Å from the full ligand.

4. Results

4.1. Latent representation

We evaluated the quality of the learned embeddings by studying whether the similarity between the protein and fragment embeddings allows to recover the binding region of the fragment. True binding regions are defined as protein surfaces within 3 Å of the fragment. To quantify this, we computed cosine similarities between a fragment and its target surface embeddings and applied several metrics: ROC area under the curve (AUC) (area under the receiver operating characteristic curve), AUPR (area under the precision recall curve) and EF_1 (enrichment factor in the 1st percentile). We performed this study for both the whole protein surface and the pocket surface only (table 1). All metrics are compared against a baseline of randomly selecting points on the protein/pocket surface as positive (binding) regions.

The encoder achieved ROC AUC values of 0.76 and 0.92 for the whole protein and pocket surfaces, respectively. The corresponding enrichment factors were 22.85 for the whole surface and 2.28 for the pocket surface. The AUPR values were 0.31 for the whole surface and 0.39 for the pocket surface, compared to random baselines of 0.01 and 0.16, respectively. Furthermore, the visualisation in figure 2(A) shows localisation of high similarity regions corresponding to binding surfaces.

We ablated the effect of the additional FSP and NCI losses and the complete results can be found in table E.4. Removing the FSP loss resulted in similar sensitivity but increased the average cosine similarity among all library fragments from 0.25 to 0.45. The effect of the NCI loss is less pronounced but does increase the metrics discussed above.

We further investigated different similarity metrics on a random subset of 1000 fragments. Cosine similarity between latent representations correlate with Tanimoto similarity (Spearman $\rho = 0.45$) based on `RDKit` fingerprints (2048 bits). Correlations with pharmacophoric and shape similarity (after alignment) as determined by ROSHAMBO (Atwi *et al* 2024) (for details see appendix D) are $\rho = 0.38$ and $\rho = 0.25$, respectively.

Lastly, we assessed the learned latent space using t-SNE dimensionality reduction, which shows clustering of chemically related fragments that is not observed with standard `RDKit` fingerprints (figure 2(B)

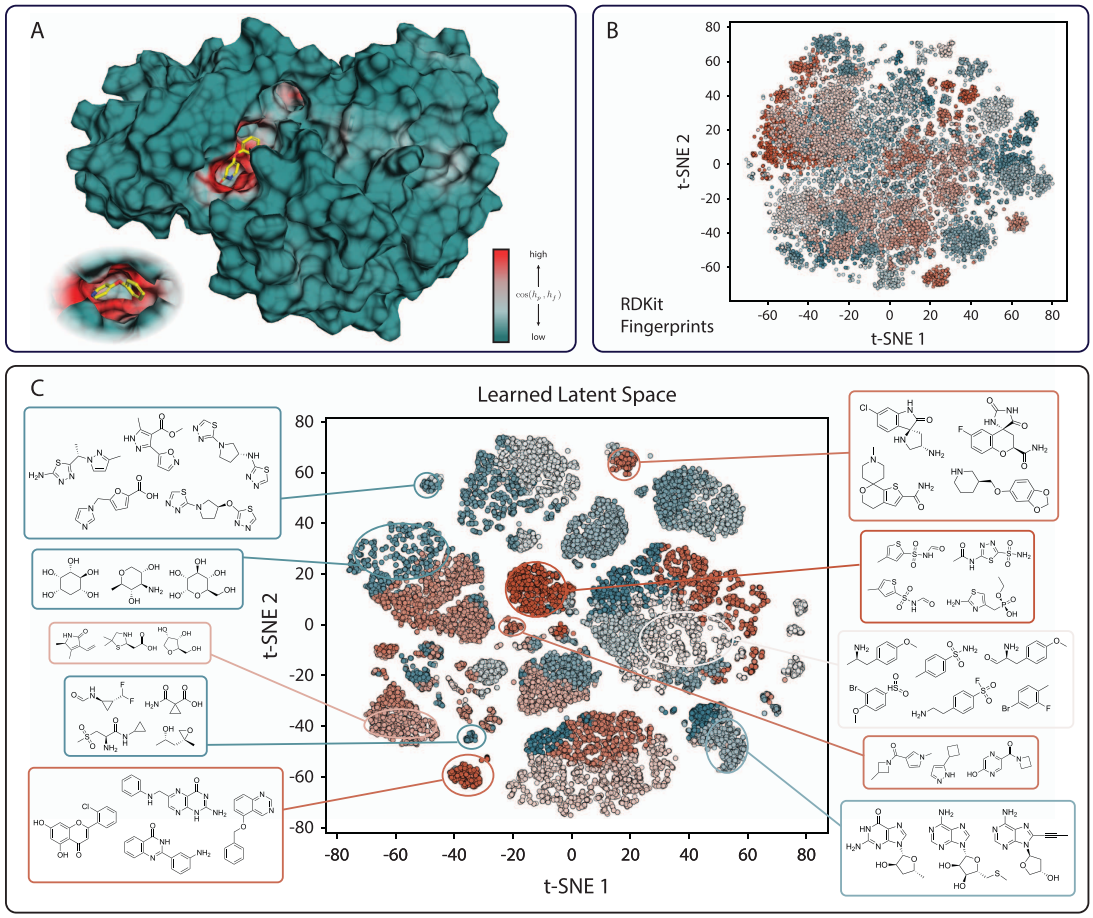

Figure 2. *A:* Test set protein surface coloured by cosine similarity of every point to the fragment visually showing the sensitivity of this representation (PDB ID: 6Q4I). More examples are shown in figure E.6. *B:* t-SNE dimensionality reduction of the fragment library using `RDKit` fingerprints (2048 bits). Each point corresponds to a fragment, coloured according to K-means cluster assignment ($k = 30$) performed directly on the fingerprint vectors. *C:* t-SNE dimensionality reduction of the fragment library in latent space with representative molecules. Fragments are embedded using the trained model and clustered using K-means ($k = 30$) applied directly to the learned embeddings.

and (C). This is reflected in clustering metrics: for K-means with $k = 30$, the silhouette score (Rousseeuw *et al* 1987) is 0.23 for the learned embeddings compared to 0.01 for the `RDKit` fingerprints. More broadly, silhouette scores for the learned embeddings remain between 0.18 and 0.23 across $k = 10$–50, whereas those for fingerprints fluctuate around zero (-0.005 to 0.015), indicating no meaningful cluster structure (figure E.5). We also report the Calinski–Harabasz index (Caliński *et al* 1974), defined as the ratio of between-cluster to within-cluster sum of squares, where higher values indicate more compact and well-separated clusters. At $k = 30$, this score is substantially higher for the learned latent space (7019 vs. 220), consistent with the silhouette analysis.

4.2. Fragment identification

A straightforward application of our learned latent space is VS, focusing on its ability to identify relevant fragments by similarity to the target pocket surface. We call this sampling approach Latent VS. LatentFrag is thus a natural extension of this approach making use of generative modelling, which enables the sampling of novel fragment embeddings directly in the learned space. We compare our generative pipeline LatentFrag to Latent VS, to VS based on docking scores obtained with Gnina (McNutt *et al* 2021), dubbed Docking VS, and a random baseline. We compare these approaches on 100 pockets from our test. For detailed information on evaluation and metrics we refer to appendix D.2.

To evaluate the performance of LatentFrag we used hard recovery (figure 3(A)) which quantifies exact matches to reference fragments. We chose recovery as our main metric over frequently used docking scores because it directly reflects the model's ability to retrieve chemically relevant fragments that are known to bind. LatentFrag achieved higher recovery rates than Latent VS and Docking VS, while random sampling did not recover reference fragments. LatentFrag achieved a sampling hit rate of 0.554% for hard recovery (table E.7). It is worth noting, that the different approaches result in large differences

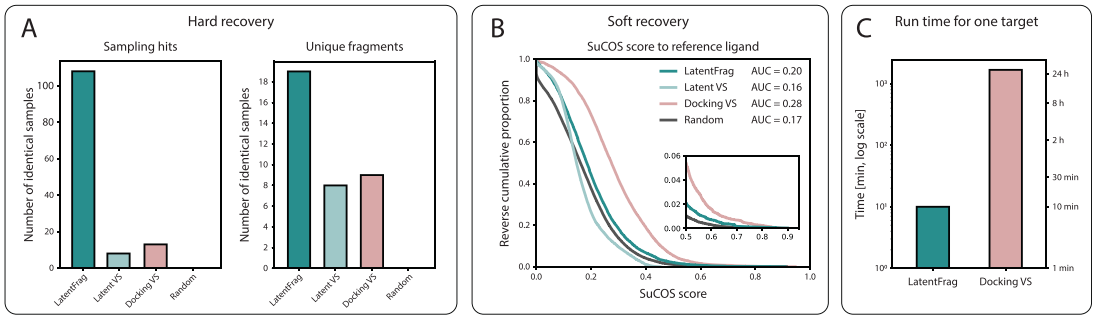

Figure 3. *A:* 'Hard' recovery metrics for different sampling strategies. Both number of sampling hits (*left*), which captures how many of the sampled fragments recover the ground truth and unique fragment recovery (*right*), describing the number of unique recovered fragments, are shown. The RMSD refers to the distance between centroids. *B:* Soft recovery metrics based on the SuCOS score relative to the full reference ligand are shown as reverse cumulative proportion with the corresponding AUC. *C:* Comparison of averaged rounded run times for one single target between LatentFrag and Docking VS. Bars indicate runtime in minutes in logarithmic scale (bottom axis) while the top axis shows equivalent times in linear time. *Sampling strategies:* **LatentFrag** makes use of our generative framework, **Latent VS** screens the whole fragment library based on embedding similarity to the pocket, **Docking VS** makes use of docking scores to screen the fragment library and finally **random** samples fragments randomly.

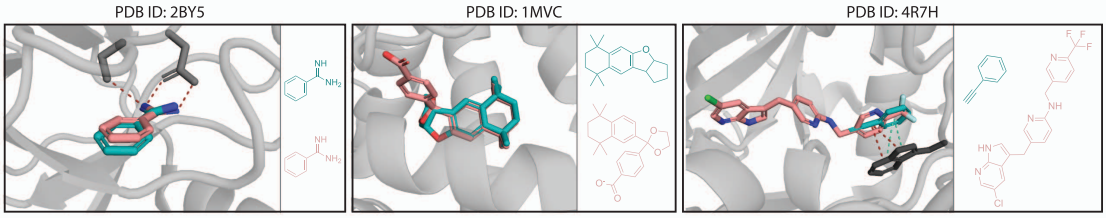

Figure 4. Selected LatentFrag samples (*teal*) to test set pockets compared to the respective reference ligand (*light red*) with corresponding reference interactions (*red lines*). *Left:* Sample that recovered the ground truth, which is also the full reference ligand. *Middle:* Sample with high SuCOS score. *Right:* Sample with low docking efficiency including predicted π–π stacking similar to the reference.

in the number of sampled/top fragments. For Latent VS we select the top 100 fragments based on the sum of cosine similarities to the pocket and similarly for Docking VS the top 100 based on docking scores while in both cases the whole fragment library was screened. However, for the random baseline and generative modelling we sample 100 times as many fragments as there are in the reference. This is accounted for in the calculation of the recovery rate.

Additionally, we assessed soft recovery (figure 3(B)), which tries to assess the overlap of the sampled fragments to the complete reference ligands in terms of shape and pharmacophoric patterns using the SuCOS score (Leung *et al* 2019), which computes the overlap of pharmacophore features and molecular shape between two molecules, yielding a value between 0 (no overlap) and 1 (identical). This evaluation shows partial pharmacophore overlap for generative samples. However, Docking VS achieved higher soft recovery compared to LatentFrag.

Importantly, docking the full fragment library as for Docking VS required approximately 28 h per target, while LatentFrag required approximately 10 min (figure 3(C)).

Lastly, figure 4 shows three selected samples to test set pockets from our generative pipeline. The first example recovers the full reference ligand and given the close overlap will likely recover also the ground truth hydrogen bonds. The second sample has a high SuCOS score to the reference and visually exhibits close overlap with the reference. When selecting by docking efficiency (docking score divided by atom count), we observe recovering of π–π interactions similar to the reference ligand. For additional results we refer to appendix E.

4.3. Case study: proto-oncogene c-Met

We next showcase the potential of LatentFrag in a case study on the disease-relevant target c-Met (hepatocyte growth factor receptor). The proto-oncogene c-Met is known to promote the growth of several solid tumours. There are drugs approved for non-small cell lung cancer that inhibit c-Met but many of these inhibitors are rendered insufficient due to emergence of resistance. This resistance is attributed to mutations near the active site (Zhang *et al* 2014, Collie *et al* 2019). In this case study we investigate a

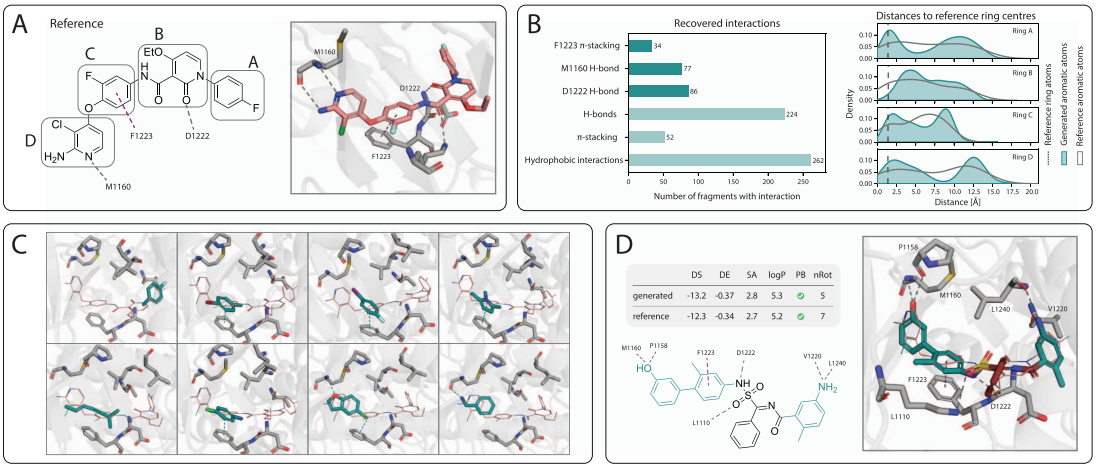

Figure 5. *A*: Structure of reference BMS-777607 with hydrogen bonds (grey) and $\pi - \pi$ stacking (magenta) to its target c-Met (PDB ID: 6SDD). *B*: Recovered interactions and spatial sampling of generated fragments. *Left:* Number of fragments recovering the reference interactions to M1160, D1222 and F1223 as well as specific interactions (hydrogen bonds, $\pi - \pi$ stacking and hydrophobic interactions) to any residue in the pocket. *Right:* Kernel density estimates of distances between generated aromatic atoms (teal) and the centres of reference rings of moieties A–D compred to the aromatic atoms of the reference (grey). The dashed black line shows the distance of the ring atoms of the respective reference rings to their centre. More detailed distributions are shown in figure F.11. *C*: Selected LatentFrag sampled fragments (teal) with top (low) docking efficiencies superimposed on the reference (magenta). *D*: Sampled fragments (teal) connected by a DiffLinker generated linker (black) compared to the reference ligand (magenta). The table shows the docking score (DS), docking efficiency (DE), synthetic accessibility score (SA), octanol-water partition coefficient (logP), PoseBusters benchmark success and number of rotatable bonds.

D1228V mutant of c-Met, whose crystal structure bound to the investigational compound BMS-777607 (PDB ID: 6SDD) is in our test set. No similar ligand is present in the training set (highest Tanimoto similarity is 0.75; see figure F.12). BMS-777607, as a type-II inhibitor, extends deep into the binding pocket beyond the active site and thus is assumed to be more active against mutated c-Met (Zhang *et al* 2014). Figure 5(A) displays the structure of BMS-777607 and highlights four motifs with distinct functions often shared among type II inhibitors. These motifs were identified based on their role in interacting with the target and we will henceforth call them *moieties* in order to distinguish from the *fragments* obtained by BRICS-fragmentation. Moiety A occupies a deep hydrophobic pocket and is often *para*-substituted. Moiety B usually has at least one amide group forming a hydrogen bond with D1222 and does not always have a ring. Moiety C is relatively conserved and forms a $\pi - \pi$ interaction with F1223. Moreover, moiety D forms a hydrogen bond between the pyridine Nitrogen and the backbone of M1160 and in doing so anchors the hinge region of c-Met (Damghani *et al* 2022).

We sample new fragments 100 times using LatentFrag as described above. Every sample contains three fragments matching the number of fragments in the reference. Therefore, the reference fragments obtained by BRICS-fragmentation are not identical to the interaction-defining moieties A, B, C and D described above. The comparison of all generated fragments to the reference reveals that the model consistently recovered modes for moieties A, C and D preferentially sampling aromatic atoms at distances similar to the reference rings of the respective moieties (figure 5(B) right). Solely the ring of moiety B shows less overlap as less aromatic rings are placed in the same position. However, moiety B does not primarily interact with the target through the aromatic ring and the hydrogen bond to D1222 is frequently recovered by sampled fragments (figure 5(B) left). Interactions to F1223 ($\pi - \pi$ stacking) and M1160 (hydrogen bond) are also recovered several times apart from many new interactions to residues the reference did not target. Figure 5(C) presents the top 10 fragments with respect to docking efficiency.

Next, we extend our approach beyond fragment sampling to the full workflow of FBDD. By combining our method with established tools for fragment linking, we illustrate how it can be used to generate complete, drug-like ligands tailored to a target pocket. For this, we select two non-overlapping fragments from the aforementioned top fragments and make use of DiffLinker to generate 50 different linkers (for details see appendix F.2). DiffLinker is a diffusion model able to generate linkers conditioned on the input fragments and the protein pocket (Igashov *et al* 2024). All designed linkers with the exception of three recover the missing hydrogen bond to D1222 from the reference. The three samples not recovering the interaction form a hydrogen bond to L1110 instead. Figure 5(D) showcases one selected ligand with favourable profile. The novel ligand results in an improved docking score and docking efficiency compared to the reference while recovering all of the reference interactions and additionally forming

Table 2. Benchmark results on the reduced CrossDocked test set comparing the combination of LatentFrag with DiffLinker (LatentFrag+) to SBDD methods. *valid* is the fraction of valid and connected molecules for which all following metrics are computed. *uniq.* is short for uniqueness and calculated based on canonical SMILES. *PB* describes the fraction passing all PoseBusters (Buttenschoen *et al* 2024) filters. *DS* stands for the Vina docking score obtained with Gnina (McNutt *et al* 2021). HB and hphobic abbreviate hydrogen bonds and hydrophobic interactions, respectively. All metrics labelled with an asterisk * indicate metrics normalised by the size of the molecule. For DS, this is equivalent to the docking efficiency..

Model	valid	uniq.	PB	QED	SA	size	DS*	HB*	$\pi-\pi$*	hphobic*
Reference	1.00	0.94	0.96	0.49	3.52	21.1	–0.35	0.201	0.011	0.147
DrugFlow	0.85	0.95	0.76	0.52	3.59	19.9	–0.36	0.213	0.012	0.162
TargetDiff	0.88	0.99	0.54	0.49	4.66	22.8	–0.35	0.209	0.009	0.179
FLOWR	0.88	0.85	0.88	0.53	3.75	20.7	–0.34	0.205	0.015	0.122
LatentFrag+	0.88	0.79	0.81	0.53	3.37	20.3	–0.31	0.207	0.022	0.166

new ones to L1110, V1220, L1240, and P1158. The molecule fulfils all PoseBusters (Buttenschoen *et al* 2024) criteria suggesting a physically valid pose and does not clash with the protein. The low synthetic accessibility score (Ertl *et al* 2009) of 2.8 indicates low chemical complexity allowing synthesis of the compound. Lastly, our novel ligand contains less rotatable bonds compared to the reference, which is advantageous due to lowered conformational entropy.

To further contextualise this workflow, we performed a small-scale benchmark on the CrossDocked (Francoeur *et al* 2020) test set comparing the full pipeline (LatentFrag followed by DiffLinker (Igashov *et al* 2024) to pocket-conditioned full ligand design approaches (table 2), namely DrugFlow (Schneuing *et al* 2025), TargetDiff (Guan *et al* 2023) and FLOWR (Cremer *et al* 2025). We observe comparable structure validity, docking, and interaction metrics. Full details are provided in appendix F.1.

5. Discussion

The encoder demonstrates strong ability to distinguish interacting from non-interacting protein surface regions, as reflected by high ROC AUC values. A value of 0.5 would correspond to random guessing. The enrichment factor EF_1 measures how many more true points are identified in the 1st percentile compared to random selection ($EF_1 = 1$) and thus assesses early retrieval performance. The high enrichment for the whole surface demonstrates that the model is capable of clearly distinguishing the pocket from the rest of the surface. Identifying the specific binding region within a pocket is a much harder task, which is reflected in the much lower EF_1 for the pocket. However, a 2.28-fold enrichment over random still highlights that even in the pocket itself the model chooses binding surface points twice as likely as non-interacting areas. Similarly, the AUPR of 0.39 substantially exceeds the random baseline of 0.16, highlighting the model's ability to identify true interaction sites in a highly imbalanced setting. Together, these results demonstrate that the learned latent space captures interaction-relevant features of protein surfaces.

Beyond classification performance, the FSP contributes to increased diversity in latent representations. Without this loss, embeddings exhibited higher average similarity, suggesting a form of mode collapse. This can be detrimental for downstream fragment screening tasks. The FSP encourages higher entropy and chemical specificity, ensuring that the model learns meaningful distinctions between fragments rather than merely separating fragments from protein surfaces. The moderate correlation with pharmacophore similarity further suggests that the latent representation captures relevant interaction features beyond structural similarity as pharmacophores can represent interactions with a protein such as hydrogen bond donors or aromatic rings.

Qualitative analysis supports these findings. As shown in figure 2(A), the encoder localises interaction hotspots on the protein surface, with highest similarity observed between fragments and their corresponding binding regions. Notably, this localisation emerges without explicit information about fragment pose or relative orientation, indicating that the latent representations encode complementary interaction features. This ability to identify binding-relevant regions directly from surface geometry and fragment structure provides a strong foundation for downstream applications such as VS and generative modelling.

Building on this representation, LatentFrag demonstrates improved fragment recovery compared to similarity-based screening and docking-based VS. Importantly, LatentFrag not only recovered more unique fragments but also did so repeatedly, achieving a sampling hit rate more than four times higher than docking-based VS. The observed hard recovery rate for LatentFrag approaches experimentally

observed fragment screening hit rates (1%–2%) at a fraction of the time and resource cost (Jalencas *et al* 2024). While recovery of chemically exact matches remains challenging, these results indicate that the model effectively enriches for relevant chemical matter and identifies promising fragment hits to narrow the experimental search.

Interestingly, docking-based screening performs better on the soft shape and pharmacophore similarity metrics. This may be explained by the larger average fragment size and explicit pose optimisation during docking, which can improve geometric alignment. In contrast, LatentFrag samples smaller fragments with fewer atoms on average (13 vs 19; figure E.8), which may reduce performance on geometry-based similarity metrics despite capturing relevant interaction features.

A key practical advantage of LatentFrag is computational efficiency. Fragment generation is approximately 180 × faster than docking-based VS for a single target. Moreover, once sampled, different fragment libraries or library expansions can be tested retrospectively by simply querying the new library based on the previously sampled embeddings. However, for docking, this further increases computational cost.

The case study demonstrates the potential of LatentFrag for the use in drug discovery. The strong chemical and interaction profile similarity to the reference highlights the model's ability to sample relevant chemistry and also demonstrates that the generative model is not only recovering fragments hits but is also biased toward chemical features that are characteristic to true binders. The benchmark on full-ligand design further indicates that fragment placement by LatentFrag provides a strong foundation for ligand construction. As LatentFrag explicitly models fragment-level interactions, it enables a more controlled exploration and subsequent elaboration of binding motifs, complementing direct ligand generation approaches.

Overall, these results suggest that latent-space generative modelling provides an efficient approach for identifying and designing fragments with interaction-relevant chemical features.

6. Conclusions

We introduce a novel protein-fragment encoder trained in a contrastive fashion that jointly learns rich representations of protein surfaces and molecular fragments in a shared latent space. Our approach captures key aspects of protein-fragment interactions while maintaining chemical relevance through fragment similarity-based penalties. We highlight the quality of our learned embeddings by demonstrating the recovery of binding sites of fragments on the whole protein with high sensitivity. Complementary, the encoder further enable both VS directly using the latent embeddings and generative fragment identification.

LatentFrag, our flow matching framework to perform generative fragment identification, correspondingly operates in this latent space sampling both fragment embeddings and their centroids. Importantly, our framework does so in the presence of the target pocket, which is not the case for many fragment based drug design approaches. LatentFrag demonstrated successful fragment recovery both when evaluating exact matches to the reference fragments and when assessing shape and pharmacophore similarity to the full reference ligand. We notably outperform VS by docking, a popular *in silico* approach for hit identification, in recovering known fragments. These results demonstrate its potential for providing initial hit hypotheses for experimental validation.

Importantly, LatentFrag is significantly faster than VS using a popular docking tool on the full fragment library. Furthermore, our method offers the added advantage of flexible choice and expansion of the fragment library even after sampling. This flexibility is important, as building blocks can be interchanged with commercial sources such as Enamine (Enamine 2025), and fragments can be derived not only from crystal structures but also from generated conformers starting from SMILES. By continuing to rely on library-based fragments, synthesizability is maintained, which is a key practical consideration. This makes our approach a cost-effective alternative for early-stage screening in FBDD.

In a case study on the pharmaceutically relevant target c-Met, we recovered most modes of a reference molecule known to be essential for interacting with the target multiple times. Additionally, we successfully captured established interactions but also revealed additional plausible contacts not found in the reference. We then applied an established fragment-linking method to assemble these recovered fragments into an initial full-ligand hit that retained the reference interactions while adding new ones with potential to improve potency.

Beyond fragment identification, our surface embeddings could also be explored for binding site prediction. The results indicate that the protein surface embeddings capture features that are highly relevant for this task, providing an additional critical application.

7. Limitations

There remain limitations and clear directions for improvement. A key limitation is the need to know the binding pocket *a priori*, which may restrict applicability in some settings. The placement of fragments through flow matching, while sufficient for fragment-based drug discovery where recovered fragments can be tested experimentally, may not yet be optimal. Furthermore, reliance on docking tools increases complexity and interpretability of the approach. By only allowing slight refinement of the placed fragment rather the full-pocket docking we aim at decreasing negative impact stemming from unspecific placement by the docking tool. Future improvements should refine the generative pipeline by implementing rigid-body transformations of fragments, improving spatial alignment, and potentially reducing reliance on docking tools. Moreover, while our evaluation demonstrates strong fragment recovery, the recovery ratio itself leaves room for improvement, and metrics based solely on recovery do not fully capture the potential of the sampled fragments. A two-stage approach involving fragment recovery followed by ligand design will likely remain necessary.

Acknowledgments

R M N thanks Proxima (USA) for their support and helpful feedback. I I has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No 945363. M B is partially supported by the EPSRC Turing AI World-Leading Research Fellowship No. EP/X040062/1 and EPSRC AI Hub No. EP/Y028872/1. P S acknowledges support from the NCCR Catalysis (Grant Number 225147), a National Centre of Competence in Research funded by the Swiss National Science Foundation. This work was supported by the Swiss National Science Foundation Grant 310030_197724.

Data availability statement

The data that support the finding owing URL/DOI: https://github.com/rneeser/LatentFrag (Neeser *et al* 2025).

Supplementary Information available at https://doi.org/10.1088/2632-2153/ae5d85/data1.

Author contributions

Rebecca M Neeser 0000-0002-3289-2766
Conceptualization (lead), Data curation (lead), Formal analysis (lead), Methodology (lead), Software (lead), Visualization (lead), Writing – original draft (lead)

Ilia Igashov 0000-0002-6214-2827
Data curation (supporting), Formal analysis (supporting), Methodology (supporting), Software (supporting), Writing – review & editing (equal)

Arne Schneuing 0009-0000-9924-6921
Formal analysis (supporting), Methodology (supporting), Software (supporting), Writing – review & editing (equal)

Michael Bronstein 0000-0002-1262-7252
Conceptualization (supporting), Supervision (supporting), Writing – review & editing (supporting)

Philippe Schwaller 0000-0003-3046-6576
Conceptualization (supporting), Supervision (equal), Writing – review & editing (equal)

Bruno Correia 0000-0002-7377-8636
Conceptualization (supporting), Supervision (equal), Writing – review & editing (equal)

References

Adasme M F, Linnemann K L, Bolz S N, Kaiser F, Sebastian Salentin V J H and Schroeder M 2021 Plip 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA *Nucl. Acids Res.* **49** W530–4

Atwi R, Wang Y, Sciabola S and Antoszewski A 2024 Roshambo: open-source molecular alignment and 3d similarity scoring *J. Chem. Inf. Model.* **64** 8098–104

Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 The protein data bank *Nucl. Acids Res.* **28** 235–42

Bian Y and Xie X-Q 2018 Computational fragment-based drug design: current trends, strategies and applications *AAPS J.* **20** 1–11

Buttenschoen M, Morris G M and Deane C M 2024 Posebusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences *Chem. Sci.* **15** 3130–9

Caliński T and Harabasz J 1974 A dendrite method for cluster analysis *Commun. Stat. Theory Methods* **3** 1–27

Chakravarti S K 2018 Distributed representation of chemical fragments *ACS Omega* **3** 2825–36

Chithrananda S, Grand G, and Ramsundar B 2020 Chemberta: large-scale self-supervised pretraining for molecular property prediction (arXiv:2010.09885)

Collie G W *et al* 2019 Structural and molecular insight into resistance mechanisms of first generation cmet inhibitors *ACS Med. Chem. Lett.* **10** 1322–7

Congreve M, Chessari G, Tisi D and Woodhead A J 2008 Recent developments in fragment-based drug discovery *J. Med. Chem.* **51** 3661–80

Cremer J, Irwin R, Tibot A, Janet J P, Olsson S, and Clevert D-A 2025 Flowr: flow matching for structure-aware de novo, interaction- and fragment-based ligand generation

Damghani T, Elyasi M, Pirhadi S, Haghighijoo Z, and Ghazi S 2022 Type II c-met inhibitors: molecular insight into crucial interactions for effective inhibition *Mol. Diversity* **26** 1–13

De Esch I J P, Erlanson D A, Jahnke W, Johnson C N and Walsh L 2021 Fragment-to-lead medicinal chemistry publications in 2020 *J. Med. Chem.* **65** 84–99

Degen J, Wegscheid-Gerlach C, Zaliani A and Rarey M 2008 On the art of compiling and using'drug-like'chemical fragment spaces *ChemMedChem* **3** 1503

Dwivedi V P and Bresson X 2020 A generalization of transformer networks to graphs (arXiv:2012.09699)

Edfeldt F N B, Folmer R H A and Breeze A L 2011 Fragment screening to predict druggability (ligandability) and lead discovery success *Drug Disc. Today* **16** 284–7

Enamine 2025 Building blocks catalog (available at: https://enamine.net/building-blocks/building-blocks-catalog) (Accessed 15 September 2025)

Ertl P and Schuffenhauer A 2009 Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions *J. Cheminform.* **1** 1–11

Ferla M P, Sánchez-García R, Skyner R E, Gahbauer S, Taylor J C, von Delft F, Marsden B D and Deane C M 2025 Fragmenstein: predicting protein–ligand structures of compounds derived from known crystallographic fragment hits using a strict conserved-binding–based methodology *J. Cheminform.* **17** 4

Francoeur P G, Masuda T, Sunseri J, Jia A, Iovanisci R B, Snyder I and Koes D R 2020 Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design *J. Chem. Inf. Model.* **60** 4200–15

Gao B, Qiang B, Tan H, Jia Y, Ren M, Minsi L, Liu J, Wei-Ying M and Lan Y 2023 Drugclip: contrastive protein-molecule representation learning for virtual screening *Advances in Neural Information Processing Systems* vol 36 pp 44595–614

Guan J, Qian W W, Peng X, Yufeng S, Peng J and Jianzhu M 2023 3d equivariant diffusion for target-aware molecule generation and affinity prediction *The 11th Int. Conf. on Learning Representations*

Guo Z et al 2022 Graph-based molecular representation learning (arXiv:2207.04869)

Hasselgren C and Oprea T I 2024 Artificial intelligence for drug discovery: are we there yet? *Ann. Rev. Pharmacol. Toxicol.* **64** 527–50

Hubbard R E and Murray J B 2011 Experiences in fragment-based lead discovery *Methods in Enzymology* vol 493 (Elsevier) pp 509–31

Igashov I, Jamasb A R, Sadek A, Sverrisson F, Schneuing A, Lio P, Blundell T L, Bronstein M and Correia B 2022 Decoding surface fingerprints for protein-ligand interactions

Igashov I, Stärk H, Vignac C, Schneuing A, Satorras V G, Frossard P, Welling M, Bronstein M and Correia B 2024 Equivariant 3d-conditional diffusion model for molecular linker design *Nat. Mach. Intell.* **6** 417–27

Imrie F, Bradley A R, van der Schaar M and Deane C M 2020 Deep generative models for 3d linker design *J. Chem. Inf. Model.* **60** 1983–95

Jalencas X *et al* 2024 Design, quality and validation of the eu-openscreen fragment library poised to a high-throughput screening collection *RSC Med. Chem.* **15** 1176–88

Leung S, Bodkin M, von Delft F, Brennan P and Morris G 2019 Sucos is better than RMSD for evaluating fragment elaboration and docking poses *ChemRviv Preprint* (available at: https://doi.org/10.26434/chemrxiv.8100203)

Li J and Jiang X 2021 Mol-bert: an effective molecular representation with bert for molecular property prediction *Wireless Commun. Mob. Comput.* **1** 7181815

Lipman Y, Chen R T Q, Ben-Hamu H, Nickel M, and Matt L 2022 Flow matching for generative modeling (arXiv:2210.02747)

Lohmann F, Allenspach S, Atz K, Schiebroek C C G, Hiss J A, and Schneider G 2024 Protein binding site representation in latent space *Mol. Inf.* **44** e202400205

Marchand J-R and Caflisch A 2018 In silico fragment-based drug design with seed *Eur. J. Med. Chem.* **156** 907–17

McCorkindale W, Ahel I, Barr H, Correy G J, Fraser J S, London N, Schuller M, Shurrush K and Lee A A 2022 Fragment-based hit discovery via unsupervised learning of fragment-protein complexes

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 *J. Cheminform.* **13** 43

Neeser R M *et al* 2025 GitHub *LatentFrag* (available at: https://github.com/rneeser/LatentFrag)

Neeser R M, Akdel M, Kovtun D, and Naef L 2023 Reinforcement learning-driven linker design via fast attention-based point cloud alignment (arXiv:2306.08166)

Rousseeuw P J 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis *J. Comput. Appl. Math.* **20** 53–65

Schneuing A, Igashov I, Dobbelstein A W, Castiglione T, Bronstein M M and Correia B 2025 Multi-domain distribution learning for de novo drug design *The 13th Int. Conf. on Learning Representations* (available at: https://openreview.net/forum?id = g3VCIM94ke)

Shin B, Park S, Kang K and Joyce C H 2019 Self-attention based molecule representation for predicting drug-target interaction *Machine Learning for Healthcare Conf.* (PMLR) pp 230–48

Somnath V R, Bunne C and Krause A 2021 Multi-scale representation learning on proteins *Advances in Neural Information Processing Systems* vol 34 pp 25244–55

Sverrisson F, Feydy J, Correia B E and Bronstein M M 2021 Fast end-to-end learning on protein surfaces *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 15272–81

Vignac C, Krawczuk I, Siraudin A, Wang B, Cevher V and Frossard P 2022 Digress: discrete denoising diffusion for graph generation (arXiv:2209.14734)

Wang H, Weijiang Li, Jin X, Cho K, Heng J, Han J and Burke M D 2021 Chemical-reaction-aware molecule representation learning (arXiv:2109.09888)

Yang B, Xiang C and Jianing Li 2024 3d structure-based generative small molecule drug design: are we there yet?

Yu H S, Modugula K, Ichihara O, Kramschuster K, Keng S, Abel R and Wang L 2020 General theory of fragment linking in molecular design: why fragment linking rarely succeeds and how to improve outcomes *J. Chem. Theory Comput.* **17** 450–62

Zhang W, Jing A, Shi D, Peng X, Yinchun J, Liu J, Geng M and Yingxia Li 2014 Discovery of novel type II c-met inhibitors based on BMS-777607 *Eur. J. Med. Chem.* **80** 254–66

Zhou G, Gao Z, Ding Q, Zheng H, Hongteng X, Wei Z, Zhang L and Guolin K 2023 Uni-mol: a universal 3d molecular representation learning framework *The 11th Int. Conf. on Learning Representations*