PL-JEPA: JOINT PREDICTIVE AND GENERATIVE LEARNING FOR STRUCTURE-BASED DRUG DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure-based drug discovery is typically formulated as two separate tasks. Virtual screening ranks existing molecules for a protein pocket, whereas structure-based molecular generation constructs new molecules conditioned on that pocket. Although both tasks are governed by the molecular constraints imposed by the same binding pocket, they are commonly modeled separately. We introduce PL-JEPA, a protein-ligand joint embedding predictive architecture that represents these shared constraints as a binding embedding in ligand representation space. The binding embedding provides a common interface for molecular recognition and realization. It can be directly matched with ligand representations for virtual screening or used to condition an equivariant diffusion decoder for 3D molecular generation. PL-JEPA progressively learns this shared interface through predictive alignment, representation-conditioned generation, and joint predictive and generative training, making the binding embedding both searchable and realizable. A single checkpoint supports virtual screening, 3D molecular generation, and target fishing. On DUD-E and LIT-PCBA, PL-JEPA improves $\text{BEDROC}_{80.5}$ over DrugCLIP by 36.2% and 84.3%, reaching 68.80 and 11.48, respectively. On CrossDocked2020, it achieves a median Vina Dock score of -8.23 kcal/mol and a diversity of 0.78. On CASF-2016, it improves target fishing Acc@5 from 62.59 to 65.96. Ablation studies show that the three-stage training progressively establishes and integrates the searchable and realizable properties of the binding embedding.

1 INTRODUCTION

A central goal of drug discovery is to identify small molecules that can modulate therapeutically relevant protein targets. When the three-dimensional structure of a target is available, structure-based drug discovery uses this structural information to guide the search for such molecules. In the deep-learning era, this search is commonly pursued through two major computational routes. Virtual screening (Gao et al., 2023) identifies promising candidates by ranking molecules from a predefined library against a target pocket, whereas structure-based molecular generation (Peng et al., 2022; Guan et al., 2023a;b; Qu et al., 2024) directly constructs new molecules conditioned on the pocket, enabling exploration beyond existing libraries.

Despite addressing the same drug-discovery objective, virtual screening and structure-based generation are typically modeled as separate learning problems. Contrastive approaches such as DrugCLIP optimize pocket–ligand representations for retrieval (Gao et al., 2023), whereas structure-based generators learn autoregressive or diffusion-based molecular construction (Peng et al., 2022; Guan et al., 2023a;b; Qu et al., 2024). This computational separation, however, does not reflect the physical basis shared by the two tasks. The geometry and physicochemical environment of a binding pocket impose constraints on the molecular characteristics of compatible ligands, while allowing substantial variation in their scaffolds and composition. Virtual screening therefore searches among existing molecules for those that satisfy these pocket-induced constraints, whereas structure-basedgeneration constructs new molecules under the same constraints.

This view naturally connects to joint embedding predictive learning (LeCun, 2022; Assran et al., 2023), which learns by predicting target representations in latent space rather than reconstructing observations in full. In structure based drug discovery, this suggests predicting not a particular ligand

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

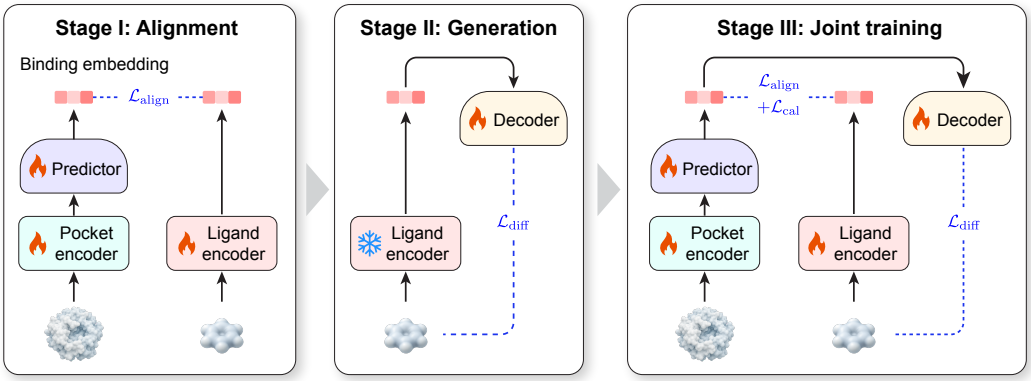


Figure 1: **Overview of PL-JEPA.** PL-JEPA progressively learns a shared *binding embedding* for molecular recognition and realization. Stage I establishes a searchable representation space through predictive alignment. Stage II learns to reconstruct 3D molecules from ligand embeddings. Stage III jointly optimizes the predictive and generative pathways, making the binding embedding both searchable and realizable.

but a latent representation of the molecular characteristics implied by a binding pocket. We refer to this representation as the *binding embedding*.

A useful binding embedding must therefore satisfy two requirements. It must provide a stable fixed dimensional interface across pockets and ligands of varying size and structure, while preserving both the discriminative information needed to identify compatible ligands and the geometric and chemical information needed for molecular construction. Virtual screening identifies existing molecules whose representations match the binding embedding, whereas structure-basedmolecular generation uses it to construct new molecular structures.

To learn a binding embedding with these properties, we introduce PL-JEPA, a unified predictive and generative framework. A predictor maps the pocket representation into the ligand representation space, where the predicted binding embedding can be directly matched with ligand representations for virtual screening or used to condition a molecular decoder for 3D molecular generation. PL-JEPA progressively learns this shared interface through alignment, representation-conditioned generation, and joint predictive and generative training, enabling the binding embedding to support both molecular recognition and realization.

The learned binding embedding supports virtual screening, structure-based molecular generation, and target fishing within a unified representation space. On DUD-E and LIT-PCBA, PL-JEPA improves $\text{BEDROC}_{80.5}$ over DrugCLIP by 36.2% and 84.3%, reaching 68.80 and 11.48, respectively. On CrossDocked2020, it generates molecules with a median Vina Dock score of -8.23 kcal/mol and a diversity of 0.78. On CASF-2016, the same compatibility space enables reverse target retrieval without task-specific retraining and improves Acc@5 from 62.59 for DrugCLIP to 65.96. Ablation studies further show that alignment, generation, joint training, and calibration progressively establish and integrate the searchable and realizable properties of the binding embedding.

Our contributions are:

- We reformulate virtual screening and structure-based molecular generation as recognition and realization of the same pocket induced molecular constraints, represented by a shared binding embedding.
- We develop PL-JEPA to predict the binding embedding in ligand representation space and progressively shape it through predictive alignment, representation-conditioned generation, and joint training, making the same representation both searchable and realizable.
- We demonstrate that the learned representation improves virtual screening over DrugCLIP, supports structure-based 3D molecular generation, and enables reverse target retrieval without task-specific retraining.

2 Related Works

Structure-based drug discovery. Structure-based drug discovery is commonly approached through virtual screening and structure-based molecular generation. In virtual screening, DrugCLIP aligns pocket and ligand representations for efficient similarity-based ranking (Gao et al., 2023), while LigUnity extends shared pocket–ligand representations to virtual screening and hit-to-lead affinity prediction (Feng et al., 2025a). In structure-based molecular generation, Pocket2Mol builds 3D ligands atom by atom (Peng et al., 2022), TargetDiff uses equivariant diffusion for target-conditioned generation (Guan et al., 2023a), DecompDiff incorporates decomposition-based priors (Guan et al., 2023b), and MolCRAFT generates molecules in a continuous parameter space (Qu et al., 2024). Recent work has begun to connect retrieval and generation. BindCLIP jointly trains contrastive screening and diffusion-based binding-pose prediction, but its diffusion model receives an existing ligand and primarily serves representation learning for screening (Qiao et al., 2026). Another work learns pocket–ligand embeddings for screening and then trains a chemical language model to generate molecular strings conditioned on frozen embeddings (Navarro et al., 2026). In contrast, PL-JEPA represents the molecular constraints imposed by a binding pocket as a shared binding embedding that supports both recognition and realization.

Joint-embedding predictive learning. Joint-embedding predictive architectures (JEPAs) learn to predict target representations in latent space instead of reconstructing observations (LeCun, 2022). Within a single modality, I-JEPA predicts features of masked image regions from visible context (Assran et al., 2023), and V-JEPA extends feature prediction to masked spatiotemporal regions of video (Bardes et al., 2024). Multimodal JEPAs predict across distinct input and target modalities. M3-JEPA uses a multi-gate mixture-of-experts predictor for cross-modal embedding alignment, optimized with contrastive and regularization losses (Lei et al., 2025). VL-JEPA predicts text embeddings from visual inputs and can decode them into text when needed, although its decoder does not participate in the main predictive training objective (Chen et al., 2026). These methods demonstrate that latent prediction can provide a shared interface across modalities, but output generation is not jointly used to shape the predicted representation for both recognition and realization. PL-JEPA extends this principle to structure-based drug discovery by predicting a binding embedding in ligand representation space and jointly shaping it through predictive alignment and molecular generation, making the same representation both searchable and realizable.

3 Method

3.1 A Unified Formulation of Structure-based Drug Discovery

Denote a protein pocket $\mathcal{P} = (\boldsymbol{a}^p, \boldsymbol{x}^p)$ and a ligand $\mathcal{M} = (\boldsymbol{a}^m, \boldsymbol{x}^m)$ each consist of atom types and 3D coordinates, with N_p and N_m atoms respectively. Given a dataset of paired complexes $\mathbb{D} := \{(\mathcal{P}_i, \mathcal{M}_i)\}_{i=1}^{M}$, our core abstraction is that structure-based drug discovery reduces to two sequential maps:

$$\boxed{\mathcal{P} \xrightarrow{\text{encode \& predict}} S_y \xrightarrow{\text{retrieve/generate}} \mathcal{M}} \tag{1}$$

where $S_y \in \mathbb{R}^d$ is the *binding embedding*, an abstract, pocket-induced representation of the molecular characteristics that govern ligand compatibility, rather than an encoding of any particular ligand. The pocket-to-embedding map is realized by the pocket encoder followed by the latent predictor, $\hat{S}_y = P(E_p(\mathcal{P}))$. The embedding-to-molecule map admits two instantiations: retrieval measures the similarity between $\hat{S}_y$ and embeddings produced by a ligand encoder, while generation decodes $\hat{S}_y$ into a novel 3D ligand via an equivariant diffusion decoder. The two tasks are thus unified through the shared abstraction S_y, rather than implemented as separate pipelines. These four modules, the pocket encoder, latent predictor, ligand encoder, and equivariant decoder, are trained in three stages described below. Figure 1 summarizes the architecture and the three-stage training procedure.

3.2 Stage I: Binding Embedding Pretraining

The first stage jointly trains the pocket encoder E_p, ligand encoder E_m, and predictor P using $\mathcal{L}_{\text{align}}$. Its goal is to imbue the binding embedding with target–ligand binding information, using positive

and negative pairs to encode binding specificity, and to provide a well-initialized encoder–predictor for subsequent stages.

3.2.1 POCKET AND LIGAND ENCODERS

We use two independently parameterized Uni-Mol encoders (Zhou et al., 2023). Given atom tokens, pairwise distances, and edge types, the pocket and ligand encoders consume sequences of the form $[\texttt{CLS}, \texttt{atom}_1, \ldots, \texttt{atom}_N, \texttt{SEP}]$ and produce

$$H^p = E_p(\mathcal{P}) \in \mathbb{R}^{L_p \times d}, \qquad H^m = E_m(\mathcal{M}) \in \mathbb{R}^{L_m \times d}, \qquad L_p = N_p + 2, \quad L_m = N_m + 2, \tag{2}$$

where $d = 512$. Following the global representation used in DrugCLIP (Gao et al., 2023), we take the `[CLS]` states as the pocket and ligand representations,

$$S_p = H_0^p \in \mathbb{R}^d, \qquad S_y = H_0^m \in \mathbb{R}^d. \tag{3}$$

3.2.2 POCKET-TO-LIGAND LATENT PREDICTION

The encoders map variable-size pocket and ligand structures to a fixed-dimensional space that captures both topological and geometric features, but S_p and S_y are not yet aligned. Rather than imposing a direct cross-modal projection, we introduce a predictor P that maps the pocket representation into the ligand representation space, producing the predicted binding embedding $\hat{S}_y = P(S_p)$. We implement P as a lightweight residual MLP operating on the pocket `[CLS]` representation. The residual formulation preserves the original pocket representation while learning its transformation toward the binding embedding space, and the residual branch is initialized such that the predictor starts close to an identity mapping. The resulting $\hat{S}_y$ has the same dimensionality as S_y. Detailed architectural specifications are provided in the Appendix.

3.2.3 PREDICTIVE ALIGNMENT IN BINDING EMBEDDING SPACE

Following DrugCLIP (Gao et al., 2023), we optimize a symmetric contrastive objective over a mini-batch $\mathcal{B}$ of B paired complexes. For each pair i, we define the normalized representations $\hat{\boldsymbol{u}}_i = \hat{S}_{y,i}/\|\hat{S}_{y,i}\|_2$ and $\boldsymbol{u}_i = S_{y,i}/\|S_{y,i}\|_2$, with pairwise logits $\ell_{ij} = \tau\, \hat{\boldsymbol{u}}_i^\top \boldsymbol{u}_j$ and a fixed logit scale $\tau = 14$.

To avoid treating duplicated complexes as negatives, we define $D_{ij} = 1$ when $i \neq j$ and the two examples share either the pocket identifier or canonical ligand SMILES, and set $D_{ii} = 0$. The diagonal pair (i, i) is the positive pocket–ligand pair, while non-duplicate off-diagonal pairs form the negative pool. The symmetric alignment objective is

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\mathbb{E}_{i\in\mathcal{B}}\left[-\log\frac{e^{\ell_{ii}}}{\sum_{j\in\mathcal{B}:\, D_{ij}=0} e^{\ell_{ij}}} - \log\frac{e^{\ell_{ii}}}{\sum_{j\in\mathcal{B}:\, D_{ji}=0} e^{\ell_{ji}}}\right]. \tag{4}$$

The first term aligns pockets to ligands, while the second enforces the reverse ligand-to-pocket direction. In distributed training, normalized representations are gathered across workers before constructing the negative pool. The resulting $\hat{S}_y$ serves as an aligned surrogate for S_y, providing a stable initialization for decoder pretraining and joint training.

3.3 STAGE II: DECODER PRETRAINING

The second stage freezes E_m and conditions the decoder on the binding embedding provided by ligand encoder, $\boldsymbol{g} = S_y = E_m(\mathcal{M})_0$. This stage trains the decoder under oracle conditions so that the embedding-to-molecule map is well-posed before the pocket pathway is introduced. Only the decoder parameters θ_{dec} are optimized using $\mathcal{L}_{\text{diff}}(S_y)$. Because the condition and denoising target are derived from the same ligand, the decoder learns the representation-to-geometry mapping without prediction error from the pocket pathway.

3.3.1 EQUIVARIANT DECODER

The decoder D_θ is a conditional diffusion generative model for 3D molecules. Its backbone uses E(3)-equivariant graph neural networks (EGNNs) (Satorras et al., 2021): each message-passing block

constructs invariant messages from node features and squared pairwise distances, and the network outputs an equivariant coordinate head $D_\theta^{(\epsilon)}$ and an invariant atom-type head $D_\theta^{(a)}$. To perform conditional generation, the decoder receives a condition $\boldsymbol{g} \in \mathbb{R}^d$ derived from the pocket embedding; at each backbone layer, a GeoRCG-style conditioning module injects $\boldsymbol{g}$ into the ligand-node features through gated residual updates (Li et al., 2025). Following UniGEM (Feng et al., 2025b), coordinates are generated by equivariant diffusion, while atom types are predicted from the denoising trajectory.

3.3.2 Coordinate Diffusion

We perform atom-wise diffusion in the ligand-centered zero-center-of-mass subspace following (Ho et al., 2020; Feng et al., 2025b). At timestep t, the forward process is

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \tag{5}$$

with the noise schedule following Feng et al. (2025b); Hoogeboom et al. (2022). The equivariant output of the backbone predicts the centered noise, and the training objective is

$$\mathcal{L}_{\text{coord}}(\boldsymbol{g}) = \mathbb{E}_{q(\boldsymbol{x}_0,t,\boldsymbol{\epsilon})}\left[\left\| D_\theta^{(\epsilon)}(\boldsymbol{x}_t, t, \boldsymbol{g}) - \boldsymbol{\epsilon} \right\|_2^2\right], \qquad t \in [1, T], \tag{6}$$

where $D_\theta^{(\epsilon)}(\boldsymbol{x}_t, t, \boldsymbol{g})$ is the equivariant coordinate prediction conditioned on the pocket embedding $\boldsymbol{g}$, and both $\boldsymbol{x}_t$ and $\boldsymbol{\epsilon}$ lie in the zero-center-of-mass subspace.

3.3.3 Atom-Type Prediction

UniGEM (Feng et al., 2025b) separates molecular generation into a high-noise nucleation phase and a low-noise growth phase, where atom-level semantics become identifiable. Following this design, PL-JEPA applies atom-type supervision only in the low-noise regime $t \le t_n$, with $t_n = 10, T = 1000$. A three-block low-noise branch with an additional two-layer MLP head predicts $C = 23$ atom channels, and the invariant objective is

$$\mathcal{L}_{\text{atom}}(\boldsymbol{g}) = \mathbb{E}_{q(\boldsymbol{x}_0,t,\boldsymbol{\epsilon})}\left[\left\| D_\theta^{(a)}(\boldsymbol{x}_t, t, \boldsymbol{g}) - \boldsymbol{a} \right\|_1\right], \qquad t \in [1, t_n], \tag{7}$$

where $\boldsymbol{a}$ is the one-hot atom-type vector corresponding to $\boldsymbol{x}_0$ and $D_\theta^{(a)}(\boldsymbol{x}_t, t, \boldsymbol{g})$ is the invariant output of the backbone with the low-noise prediction head. The complete decoder objective is

$$\mathcal{L}_{\text{diff}}(\boldsymbol{g}) = \mathcal{L}_{\text{coord}}(\boldsymbol{g}) + \mathcal{L}_{\text{atom}}(\boldsymbol{g}). \tag{8}$$

3.4 Stage III: Joint Predictive and Generative Learning

The final stage jointly optimizes all components, requiring the binding embedding to support both retrieval and generation and to retain both discriminative and geometrically reconstructive information.

3.4.1 Latent Calibration Objective

The contrastive alignment objective operates on normalized representations and therefore mainly constrains their angular geometry. Molecular generation, however, conditions the decoder on the unnormalized predicted representation $\hat{S}_y$. To reduce the mismatch between the pocket-predicted representation $\hat{S}_y$ and the binding embedding S_y used as the decoder condition, we further introduce a latent calibration objective.

We define the normalized residual as

$$\boldsymbol{r} = \frac{\hat{S}_y - \text{sg}(S_y)}{\sigma}, \tag{9}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and σ is a scaling factor. The smooth-ℓ_1 penalty is defined as

$$\rho_\beta(r) = \begin{cases} \dfrac{r^2}{2\beta}, & |r| < \beta, \\ |r| - \dfrac{\beta}{2}, & \text{otherwise}, \end{cases} \tag{10}$$

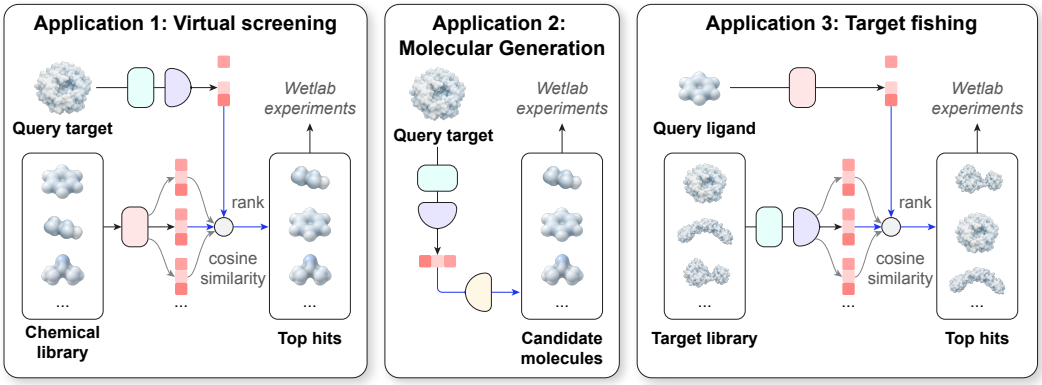


Figure 2: **Applications of PL-JEPA.** The shared binding embedding supports virtual screening, molecular generation, and target fishing within a unified representation space. For virtual screening, a query pocket is matched against a chemical library to rank candidate ligands. For molecular generation, the pocket-predicted binding embedding conditions the decoder to generate 3D molecules. For target fishing, a query ligand is matched against a pocket library to rank potential targets.

where β controls the transition from quadratic to linear regions. The calibration objective is then

$$\mathcal{L}_{\text{cal}} = \mathbb{E}_{(\mathcal{P},\mathcal{M})\in\mathbb{D}} \left[\frac{1}{d} \sum_{k=1}^{d} \rho_\beta(r_k) \right]. \tag{11}$$

While $\mathcal{L}_{\text{align}}$ organizes the representation space for similarity-based retrieval, $\mathcal{L}_{\text{cal}}$ directly reduces the discrepancy between the unnormalized pocket prediction and its ligand-space target, making $\hat{S}_y$ better suited as the conditioning representation for molecular generation.

3.4.2 Joint Objective

The final stage replaces the oracle ligand condition with the deployable binding embedding: $S_p = E_p(\mathcal{P})_0$ and $\boldsymbol{g} = \hat{S}_y = P(S_p)$. The pocket encoder, predictor, ligand encoder and decoder are optimized jointly. The objective is

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}(\hat{S}_y) + \lambda_{\text{cal}}\mathcal{L}_{\text{cal}}. \tag{12}$$

Here, $\mathcal{L}_{\text{diff}}(\hat{S}_y)$ adapts the decoder to the predicted pocket representation rather than the oracle binding embedding, while $\mathcal{L}_{\text{cal}}$ in Equation 11 reduces the representation mismatch introduced by replacing S_y with $\hat{S}_y$. By jointly optimizing retrieval and generation objectives, the model learns a more informative binding embedding $\hat{S}_y$ that serves as a shared representation for both bidirectional retrieval and molecular generation.

3.5 Inference

The learned binding embedding provides a common interface for molecular recognition and realization. The same pocket-predicted representation can be used for similarity-based retrieval or as the condition for molecular generation, allowing a single checkpoint to support virtual screening, structure-based molecular generation, and target fishing without task-specific representation learning. For retrieval, we define the normalized pocket prediction $\hat{\boldsymbol{u}}(\mathcal{P}) = P(E_p(\mathcal{P})_0)/\|P(E_p(\mathcal{P})_0)\|_2$ and ligand embedding $\boldsymbol{u}(\mathcal{M}) = E_m(\mathcal{M})_0/\|E_m(\mathcal{M})_0\|_2$, with their cosine similarity $s(\mathcal{P},\mathcal{M}) = \hat{\boldsymbol{u}}(\mathcal{P})^\top \boldsymbol{u}(\mathcal{M})$ serving as the common compatibility score.

Virtual screening. Given a query pocket $\mathcal{P}$, we compute $\hat{\boldsymbol{u}}(\mathcal{P})$ once and rank a precomputed library of ligand embeddings by $s(\mathcal{P},\mathcal{M}_j)$.

Table 1: Virtual screening performance on DUD-E and LIT-PCBA.

Benchmark	Method	AUROC ↑	BEDROC$_{80.5}$ ↑	EF1% ↑
DUD-E	Glide-SP	76.70	40.70	16.18
	AutoDock Vina	71.60	—	7.32
	DrugCLIP	80.93	50.52	31.89
	LigUnity	89.22	65.26	42.63
	PL-JEPA (Ours)	**91.21**	**68.80**	**44.13**
LIT-PCBA	Glide-SP	53.15	4.00	3.41
	GNINA	60.93	5.40	4.63
	DrugCLIP	57.17	6.23	5.51
	LigUnity	59.85	11.33	6.47
	PL-JEPA (Ours)	**61.01**	**11.48**	**6.64**

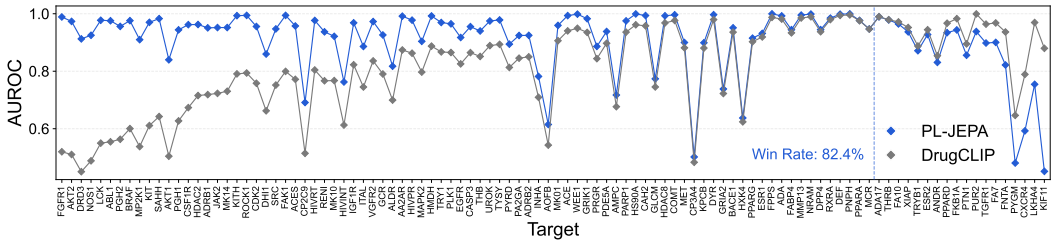


Figure 3: Target-wise AUROC comparison between PL-JEPA and DrugCLIP on DUD-E.

Structure-based molecular generation. Given a query pocket $\mathcal{P}$, we compute the predicted binding embedding $\hat{S}_y = P(E_p(\mathcal{P})_0)$ and use it directly to condition the molecular decoder for 3D ligand generation.

Target fishing. Given a query ligand $\mathcal{M}$, we compute its embedding $\boldsymbol{u}(\mathcal{M})$ once and rank candidate pockets by the same compatibility score $s(\mathcal{P}_i, \mathcal{M})$. Target fishing therefore performs reverse retrieval in the same learned compatibility space without retraining or an additional prediction head.

4 Experiments

4.1 Experimental Setup

We train PL-JEPA following the three-stage training procedure described in Section 3 on the baseline training data. The pocket and ligand encoders are initialized from pretrained Uni-Mol weights. Unless otherwise specified, all evaluations use the jointly trained model without task-specific representation learning or prediction heads. Additional training and implementation details are provided in Appendix C.

4.2 Virtual Screening

Protocol. We evaluate virtual screening on DUD-E (Mysinger et al., 2012) and LIT-PCBA (Tran-Nguyen et al., 2020), following the benchmark setup of DrugCLIP (Gao et al., 2023). DUD-E contains 102 targets with property-matched decoys, whereas LIT-PCBA contains 15 targets with experimentally measured inactive compounds and is designed to reduce ligand-based biases. Following prior work, we rank candidate ligands independently for each target and macro-average the results across targets. We report AUROC for overall ranking performance, and BEDROC ($\alpha = 80.5$) and EF1% for early enrichment (Truchon & Bayly, 2007).

Results. As shown in Table 1, PL-JEPA achieves the best performance across all three metrics on both DUD-E and LIT-PCBA. Compared with the closely related DrugCLIP baseline, PL-JEPA

Table 2: Structure-based molecular generation performance on CrossDocked2020.

Method	Vina Dock ↓		QED ↑	SA ↑	Diversity ↑
	Mean	Median			
Reference ligand	-7.45	-7.26	0.48	0.73	—
Pocket2Mol	-7.15	-6.79	**0.56**	**0.74**	0.69
TargetDiff	-7.80	-7.91	0.48	0.58	0.72
DecompDiff-R	-7.03	-7.16	0.51	0.66	0.73
MolCRAFT	-7.92	-8.01	0.50	0.69	0.72
PL-JEPA (Ours)	**-8.12**	**-8.23**	0.46	0.61	**0.78**

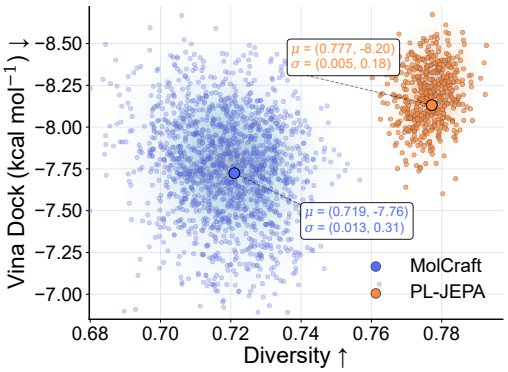


Figure 4: Distribution of binding affinity and molecular diversity for molecules generated by PL-JEPA and MolCRAFT on CrossDocked2020.

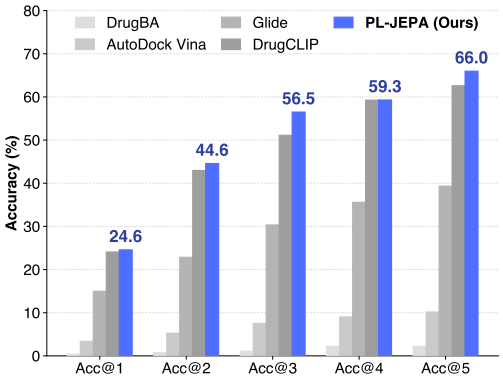


Figure 5: Target fishing on CASF-2016.

achieves an average relative improvement of 29.1% on DUD-E and 37.2% on LIT-PCBA across the three metrics. Notably, PL-JEPA performs virtual screening directly by matching the pocket-predicted binding embedding against ligand representations, without a task-specific scoring head, demonstrating that the learned binding embedding provides a directly searchable representation of pocket–ligand compatibility. Figure 3 further shows that this improvement is consistent across targets: PL-JEPA achieves higher AUROC than DrugCLIP on 84 of the 102 DUD-E targets (82.4%), indicating that the aggregate gain is broadly distributed rather than driven by a small subset of targets.

4.3 Structure-based Molecular Generation

Protocol. We evaluate structure-based molecular generation on the standard 100-pocket CrossDocked2020 test set (Francoeur et al., 2020), following the benchmark protocol used by Pocket2Mol, TargetDiff, DecompDiff, and MolCRAFT (Peng et al., 2022; Guan et al., 2023a;b; Qu et al., 2024). We report Vina Dock score, quantitative estimate of drug-likeness (QED) (Bickerton et al., 2012), normalized synthetic accessibility (SA) (Ertl & Schuffenhauer, 2009), and molecular diversity, measured as the mean pairwise Tanimoto distance among molecules generated for the same pocket. Lower Vina Dock scores are better, while higher QED, SA, and diversity are preferred.For Figure 4, we generated molecular samples using our checkpoint and the released MolCRAFT checkpoint under the same CrossDocked2020 evaluation protocol, as the original MolCRAFT-generated samples were unavailable to us.

Results. As shown in Table 2, PL-JEPA achieves the best Vina Dock scores and molecular diversity among all methods, with a mean/median Vina Dock score of $-8.12/-8.23$ and a diversity of 0.78. Notably, its generated molecules achieve more favorable docking scores than the reference ligands, while maintaining reasonable QED and synthetic accessibility. Figure 4 further illustrates the joint behavior of binding quality and diversity. Compared with MolCRAFT, the PL-JEPA distribution shifts toward both higher molecular diversity and more favorable Vina Dock scores, rather than improving one at the expense of the other. Together with the virtual screening results, these findings

Table 3: Ablation of the training components in PL-JEPA.

Training components				Virtual Screening (DUD-E)			Molecular Generation (CrossDocked2020)				
Align.	Gen.	Joint	Cal.	AUROC ↑	$\text{BEDROC}_{80.5}$ ↑	EF@1% ↑	Vina Dock ↓		QED ↑	SA ↑	Div. ↑
							Mean	Median			
✓	✓	–	–	89.18	58.81	36.94	−5.94	−6.03	0.37	0.52	0.73
✓	–	✓	–	89.16	66.75	41.16	−4.73	−4.69	0.34	0.51	0.52
✓	✓	✓	–	90.32	65.34	40.29	−7.92	−7.97	0.45	0.58	0.77
✓	✓	✓	✓	**91.21**	**68.80**	**44.13**	**−8.12**	**−8.23**	**0.46**	**0.61**	**0.78**

show that the pocket-predicted binding embedding is not only directly searchable for molecular recognition, but also contains sufficient structural and chemical information for molecular realization.

4.4 Target Fishing

Protocol. We evaluate target fishing on CASF-2016 (Su et al., 2019), following the benchmark protocol of DrugCLIP (Gao et al., 2023). Each ligand is queried against the candidate pocket pool to recover its cognate target. We report Acc@k, defined as the fraction of query ligands for which the cognate target is ranked among the top-k candidate pockets.

Results. As shown in Figure 5, PL-JEPA achieves the best target-fishing performance across all retrieval depths. More importantly, the same learned compatibility space supports both pocket-to-ligand virtual screening and ligand-to-pocket target fishing without task-specific retraining. This reverse retrieval capability provides further evidence that PL-JEPA captures a shared notion of pocket–ligand compatibility rather than a representation specialized for a single retrieval direction.

4.5 Ablation Study

We ablate the three training stages and calibration objective while keeping the model architecture and evaluation protocols unchanged. As shown in Table 3, when alignment and generation are trained independently, the aligned representation already supports strong virtual screening, but the two capabilities remain disconnected. Directly introducing joint training without generation-stage pretraining retains competitive retrieval performance but substantially degrades molecular realization, yielding considerably worse Vina Dock scores and reducing diversity from 0.73 to 0.52. In contrast, performing generation-stage training before joint optimization substantially improves molecular generation while further increasing AUROC. This suggests that the generation stage first establishes the mapping from ligand-space representations to molecular structures, after which joint training can couple this realization capability with the pocket-predicted binding embedding. Finally, the calibration objective consistently improves both virtual screening and molecular generation, indicating that reducing the discrepancy between pocket-predicted and ligand embeddings helps maintain a latent interface that is simultaneously searchable and realizable. Overall, the ablation shows that searchability and realizability are progressively established and integrated through training. Alignment establishes a searchable compatibility space, generation learns to realize ligand representations into molecular structures, joint training couples the two capabilities through the binding embedding, and calibration further improves the consistency of this shared latent interface.

5 Conclusion

We introduced PL-JEPA, a unified predictive and generative framework for structure-based drug discovery. PL-JEPA represents the molecular constraints imposed by a binding pocket as a binding embedding in ligand representation space, providing a common interface for molecular recognition and realization. Through predictive alignment, representation-conditioned generation, and joint training, the binding embedding becomes both searchable for virtual screening and realizable for 3D molecular generation. The same learned compatibility space further supports target fishing without task-specific retraining. Across virtual screening, molecular generation, and target fishing benchmarks, PL-JEPA demonstrates that a shared binding embedding can effectively connect molecular retrieval and generation within a single framework.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing and to assist with the retrieval and discovery of related work. All AI-assisted content was carefully reviewed, verified, and, where necessary, revised by the authors. In particular, retrieved literature and associated claims were checked against the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training objectives, training procedure, evaluation protocols and dataset we used in the main text. Implementation details and hyperparameters are documented in the appendix. Additional experimental settings and ablation details are also provided to facilitate reproduction of the reported results.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01499. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01499.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2. Featured Certification.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012. doi: 10.1038/nchem.1243.

Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: Joint embedding predictive architecture for vision-language. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tjimrqc2BU.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1:8, 2009. doi: 10.1186/1758-2946-1-8.

Bin Feng, Zijing Liu, Hao Li, Mingjun Yang, Junjie Zou, He Cao, Yu Li, Lei Zhang, and Sheng Wang. Hierarchical affinity landscape navigation through learning a shared pocket-ligand space. *Patterns*, 6(10):101371, 2025a. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2025.101371. URL https://www.sciencedirect.com/science/article/pii/S2666389925002193.

Shikun Feng, Yuyan Ni, Lu yan, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. UniGEM: A unified approach to generation and property prediction for molecules. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=Lb91pXwZMR.

Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of Chemical Information and Modeling*, 60 (9):4200–4215, 09 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00411. URL https://doi.org/10.1021/acs.jcim.0c00411.

Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. Drugclip: Contrastive protein-molecule representation learning for virtual screening. In A. Oh, T. Naumann, A. Globerson, K. Saenko,

M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 44595–44614. Curran Associates, Inc., 2023. doi: 10.52202/075280-1932. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8bd31288ad8e9a31d519fdeede7ee47d-Paper-Conference.pdf.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=kJqXEPXMsE0.

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. DecompDiff: Diffusion models with decomposed priors for structure-based drug design. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11827–11846. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/guan23a.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022.

Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Huazhen Huang, Qingqing Gu, Yetao Wu, and Luo Ji. M3-JEPA: Multimodal alignment via multi-gate moe based on the joint-embedding predictive architecture. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=tYwKQMMjJA.

Zian Li, Cai Zhou, Xiyuan Wang, Xingang Peng, and Muhan Zhang. Geometric representation condition improves equivariant molecule generation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36921–36953. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25dz.html.

Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 07 2012. ISSN 0022-2623. doi: 10.1021/jm300687e. URL https://doi.org/10.1021/jm300687e.

Carles Navarro, Philipp Tholke, and Gianni de Fabritiis. Structure-guided molecular design with contrastive 3d protein-ligand learning, 2026. URL https://arxiv.org/abs/2604.19562.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2Mol: Efficient molecular sampling based on 3D protein pockets. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17644–17655. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/peng22b.html.

Anjie Qiao, Zhen Wang, Yaliang Li, Jiahua Rao, and Yuedong Yang. Bindclip: A unified contrastive-generative representation learning framework for virtual screening. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 11854–11865,

New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3819040. URL https://doi.org/10.1145/3770855.3819040.

Yanru Qu, Keyue Qiu, Yuxuan Song, Jingjing Gong, Jiawei Han, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. MolCRAFT: Structure-based drug design in continuous parameter space. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41749–41768. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/qu24a.html.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: The casf-2016 update. *Journal of Chemical Information and Modeling*, 59(2):895–913, 02 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00545. URL https://doi.org/10.1021/acs.jcim.8b00545.

Viet-Khoa Tran-Nguyen, Celien Jacquemard, and Didier Rognan. Lit-pcba: An unbiased data set for machine learning and virtual screening. *Journal of Chemical Information and Modeling*, 60(9):4263–4273, 09 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00155. URL https://doi.org/10.1021/acs.jcim.0c00155.

Jean-François Truchon and Christopher I. Bayly. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of Chemical Information and Modeling*, 47(2):488–508, 03 2007. ISSN 1549-9596. doi: 10.1021/ci600426e. URL https://doi.org/10.1021/ci600426e.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A Supplementary Experiments

A.1 Visualization of Binding Embeddings

For each of the 102 DUD-E targets, we extract the ℓ_2-normalized 512-dimensional pocket representation used for retrieval, yielding the same target set for DrugCLIP and PL-JEPA. We project each model's representations to two dimensions independently with t-SNE (cosine distance, perplexity 15, 2,000 iterations, fixed seed), using a shared eight-family color map.

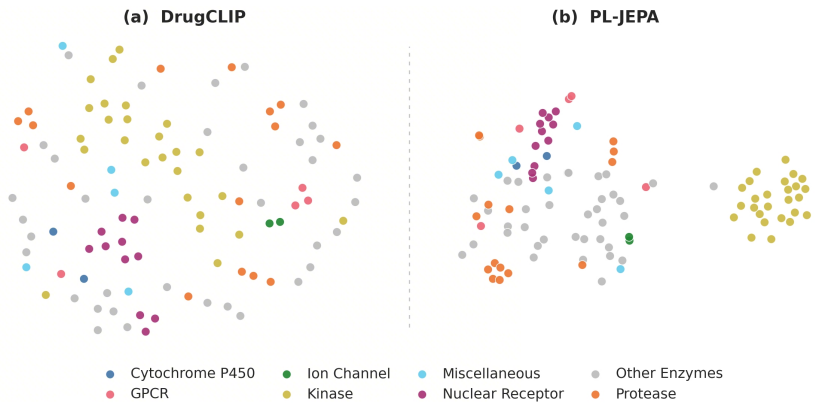


Figure 6: t-SNE of binding embeddings colored by DUD-E target protein family.

As shown in Figure 6, DrugCLIP representations overlap across protein families, whereas PL-JEPA representations exhibit clearer family-level organization. Kinases form a compact, well-separated cluster under PL-JEPA, and nuclear receptors show stronger local concentration. This is consistent with measurements in the original 512-dimensional space: PL-JEPA raises same-vs-different family pair AUC from 0.560 to 0.726 and five-nearest-neighbor family purity from 0.506 to 0.643. The separation is not expected to be uniform: *Miscellaneous* and *Other Enzymes* are heterogeneous DUD-E categories spanning different folds and functions, while Cytochrome P450 and Ion Channel each contain only two targets. The visualization thus supports a coarse-grained family-level organization in PL-JEPA, but should not be read as evidence that every family forms a homogeneous pocket class or that two-dimensional distances quantify biological similarity.

A.2 Visualization of Hard Screening Cases

Figure 7 provides a molecule-level view of representative hard cases on which PL-JEPA corrects a DrugCLIP ranking failure. Specifically, morgan, USR, USRCAT, and aligned-shape scores measure similarity (larger is more similar), while the NPR distance measures global shape proportion (smaller is more similar). The displayed conformers are deterministic ETKDGv3 geometries optimized with MMFF/UFF and shape-aligned, not docking poses or experimental bioactive conformations. The selected pairs are intrinsically hard: the active and decoy often exhibit similar generated 3D shape, high USR and small NPR distance—scaffold-hopping pairs that preserve a similar global envelope. DrugCLIP ranks the decoy above the active with a small margin, while PL-JEPA reverses the order with a clear gap. AutoDock Vina assigns favorable scores to both partners ($E < -7$ kcal/mol). These examples suggest that the joint predictive alignment and decoder supervision can yield a representation that discriminates active–decoy pairs on which both contrastive retrieval and fixed docking provide little signal.

A.3 High-Affinity and Diverse Generation for the Same Pocket

Figure 8 illustrates molecules generated for the same protein pocket. The generated candidates achieve high predicted binding affinity while spanning diverse molecular structures, showing that PL-JEPA can preserve structural diversity without sacrificing pocket-specific affinity.

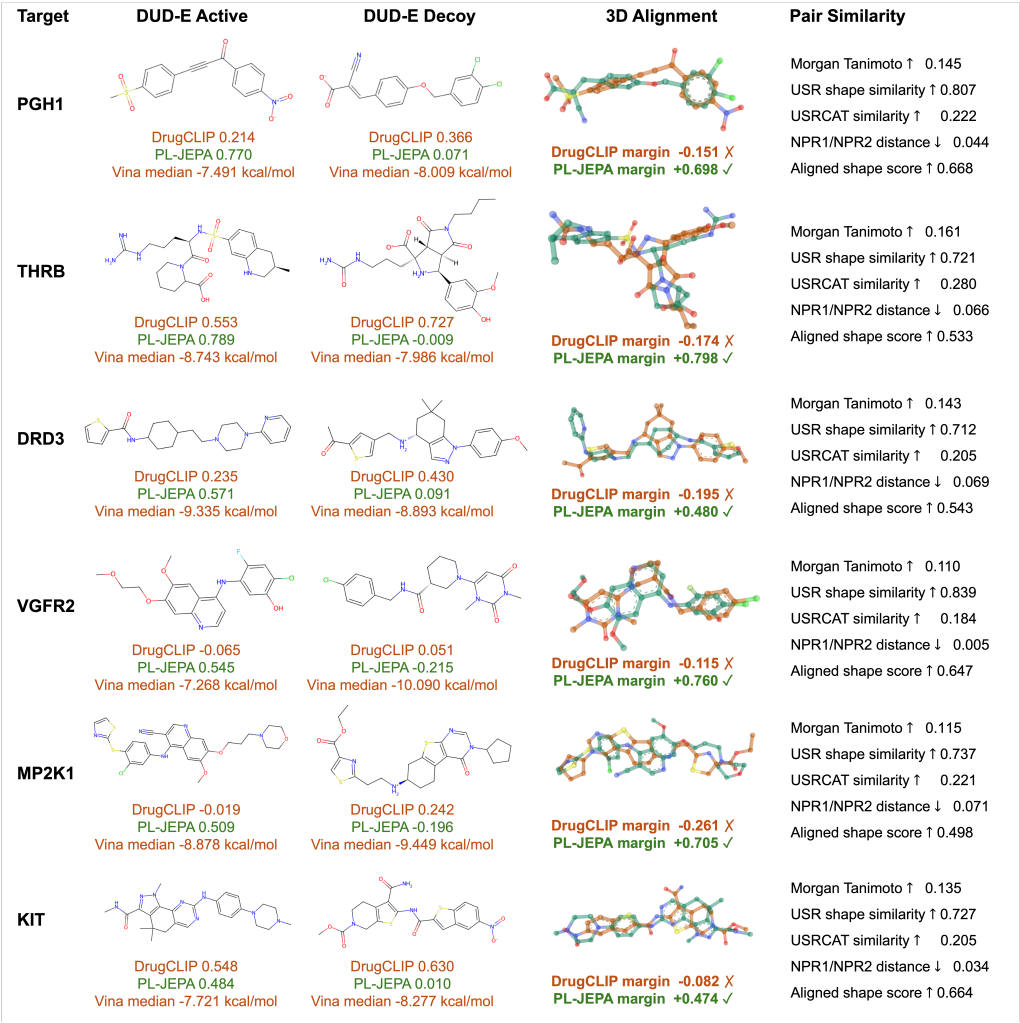


Figure 7: Active-decoy pairs where DrugCLIP misranks the DUD-E active but PL-JEPA recovers the correct ordering. For each target: left, 2D structures and screening scores of the active and decoy; middle, shape-aligned 3D conformers; right, ranking margins and similarity metrics. The margin is defined as the active score minus the decoy score, thus a negative margin indicates a misordering.

B Detailed Architecture

B.1 Input representation and encoders

A pocket and a ligand are represented as atom-type sequences together with pairwise Euclidean distances and edge types. Both branches use independently parameterized Uni-Mol encoders with hidden width $d = 512$. Each sequence contains `[CLS]` and `[SEP]` tokens. Pocket inputs are truncated to at most 256 atoms in the training and retrieval pipelines; the decoder supports at most 163 ligand nodes and predicts 23 atom channels. Coordinates passed to the generative model are expressed in the ligand zero-center-of-mass frame.

The predictor consumes only the two global encoder states:

$$S_p = H_0^p \in \mathbb{R}^{512}, \qquad S_y = H_0^m \in \mathbb{R}^{512}. \tag{13}$$

The pocket and ligand token sequences are still used internally by their Uni-Mol encoders, but no atom-level token memory is passed to the predictor.

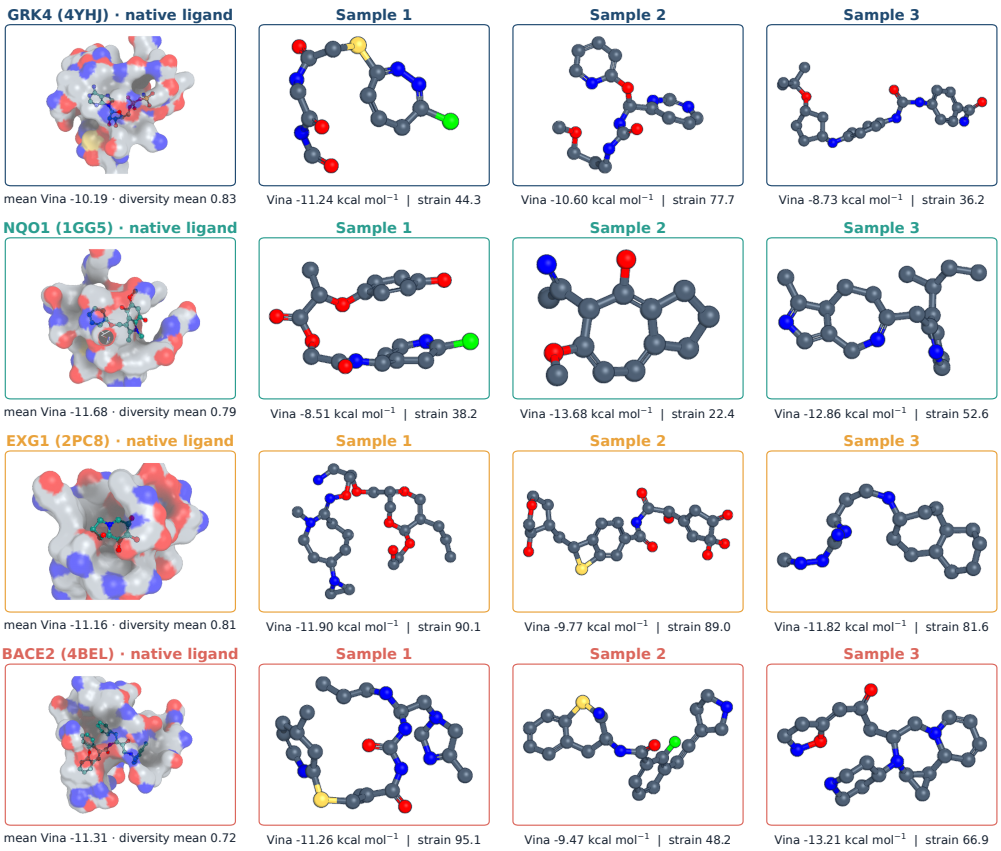


Figure 8: High-affinity and structurally diverse molecules generated for the same protein pocket. The examples demonstrate that PL-JEPA can produce multiple distinct molecular candidates with strong predicted affinity to a shared pocket, rather than collapsing to a narrow structural mode.

B.2 Binding Embedding Predictor

The predictor maps the pocket `[CLS]` representation $S_p \in \mathbb{R}^{512}$ to a predicted binding embedding $\hat{S}_y \in \mathbb{R}^{512}$ in the ligand representation space. We implement it as a residual MLP:

$$\hat{S}_y = S_p + W_2\ \mathrm{SiLU}(W_1\,\mathrm{LN}(S_p) + b_1) + b_2. \tag{14}$$

The output projection W_2 and bias b_2 are initialized to zero, so the predictor starts as an identity map and learns a residual transformation toward the ligand binding space. Inference requires only the pocket `[CLS]` state.

Retrieval uses FP32 ℓ_2 normalization,

$$\hat{u}_i = \frac{\hat{S}_{y,i}}{\|\hat{S}_{y,i}\|_2}, \qquad u_i = \frac{S_{y,i}}{\|S_{y,i}\|_2}, \tag{15}$$

whereas generation and raw-latent calibration use the corresponding unnormalized states. Keeping the raw magnitude prevents retrieval normalization from discarding information needed by the decoder-facing pathway.

B.3 Duplicate-masked symmetric retrieval objective

For a distributed batch, representations and string identifiers are gathered across all workers before constructing the negative pool. The score matrix is

$$\ell_{ij} = 14\,\hat{u}_i^\top u_j. \tag{16}$$

An off-diagonal entry is excluded when two samples share either a pocket identifier or a canonical ligand SMILES. In implementation, excluded logits receive a large negative offset before the two directional cross-entropies are evaluated. The loss is the average of pocket-to-ligand and ligand-to-pocket cross-entropy. The final configuration uses global negatives and no hard-negative reweighting or pairwise late-interaction residual.

B.4 Conditional equivariant decoder

The decoder has latent width 512 and node hidden width 256. Its main denoising backbone contains four E(3)-equivariant message-passing blocks. Each block contains two condition-injection blocks that attend to the 512-dimensional condition with four heads; the condition feed-forward expansion factor is four. Condition and attention dropout are both 0.1. The decoder uses sum aggregation, one invariant sublayer per message-passing block, and a three-block low-noise atom branch. The coordinate head is initialized with gain 10^{-3} and uses the stabilized legacy $\tanh$ coordinate update with range 15. Bond prediction and direct pocket-geometry conditioning are disabled in the final model.

The coordinate process is a 1,000-step DDPM with a degree-two polynomial noise schedule and precision 10^{-5}. Timesteps are sampled with the implementation's stratified legacy sampler. Coordinate loss is weighted uniformly. Atom supervision is a masked ℓ_1 objective over 23 atom channels and is active only for $t \leq 10$. Coordinate and atom losses both have unit weight. At inference, sampling maintains the ligand zero-center-of-mass constraint; the final velocity is projected to the same subspace.

C Training Curriculum

C.1 Stage I: Binding Embedding Pretraining

The pocket encoder, ligand encoder, and CLS-only predictor are first trained for 20 epochs with the duplicate-masked symmetric contrastive objective. This produces the Native-512 retrieval warm start used by Stage III. The fixed score scale is 14.

C.2 Stage II: Decoder Pretraining

The decoder is independently pretrained for 20 epochs using the frozen true ligand latent S_y as its condition. Stage II configuration uses batch size 32, AdamW with learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 10^{-12}, AMSGrad, and gradient clipping at norm 1.0. A linear warmup is followed by cosine decay to 10^{-5}. Decoder parameters are tracked with EMA decay 0.9999; the EMA weights initialize Stage III.

C.3 Stage III: Joint Predictive and Generative Learning

Joint training starts from the Stage I retrieval model and the Stage II decoder. It uses live pocket predictions rather than cached latents, and the decoder is conditioned only on $\hat{S}_y$. From the beginning of this stage, the predictor, pocket encoder, ligand encoder, and decoder are optimized jointly for 20 epochs. The objective is

$$\mathcal{L}_{\text{joint}} = 1.00\,\mathcal{L}_{\text{align}} + 1.00\,\mathcal{L}_{\text{diff}}(\hat{S}_y) + 0.25\,\mathcal{L}_{\text{cal}}, \tag{17}$$

where the calibration residual is

$$r = \frac{\hat{S}_y - \text{sg}(S_y)}{0.95}. \tag{18}$$

The diffusion gradient is passed to the predicted-latent pathway, allowing the generative objective to update the predictor and pocket encoder together with the decoder.

The run uses AdamW with AMSGrad, base learning rate 10^{-3}, decoder learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 10^{-12}, and global-norm clipping at 1.0. A 300-step warmup is followed by cosine decay to 0.10 of each parameter group's initial rate. The ligand encoder uses a 0.20 learning-rate multiplier. The configured batch size is 128 per distributed worker, training uses

eight workers, automatic mixed precision is enabled with initial loss scale 4, and the decoder EMA decay is 0.9999.

D CONSOLIDATED HYPERPARAMETERS

Table 4 summarizes the training curriculum, including the conditioning input, trainable modules, number of epochs, batch size, and loss weights in each stage. Table 5 reports the architecture and optimization hyperparameters used to reproduce the final reported results.

Table 4: Training curriculum. "Trainable modules" lists modules updated in each phase.

Phase	Condition	Trainable modules	Epochs	Batch	Main loss weights
Stage I	N/A	pocket, ligand, predictor	20	32	$\lambda_{\text{align}} = 1.00$
Stage II	true ligand latent	decoder	20	32	$\lambda_{\text{coord}} = 1.00$, $\lambda_{\text{atom}} = 1.00$
Stage III	predicted latent	pocket, ligand, predictor, decoder	20	128	$\lambda_{\text{align}} = 1.00$, $\lambda_{\text{diff}} = 1.00$, $\lambda_{\text{cal}} = 0.25$

Table 5: Architecture and optimization settings used to produce the final reported results.

Component	Hyperparameter	Value
Encoders	hidden dimension	512
Encoders	maximum pocket atoms	256
Predictor	input	pocket / ligand `[CLS]` states
Predictor	retrieval map	identity-initialized residual MLP
Retrieval	logit scale / negatives	14 / global
Decoder	latent / hidden width	512 / 256
Decoder	EGNN blocks / atom blocks	4 / 3
Decoder	conditioning blocks per layer	2
Decoder	conditioning heads / dropout	4 / 0.1
Decoder	maximum nodes / atom channels	163 / 23
Diffusion	steps / schedule	1,000 / polynomial, power 2
Diffusion	atom-supervision threshold	$t \leq 10$
Optimization	AdamW betas / epsilon	$(0.9, 0.999)$ / 10^{-8}
Optimization	weight decay / clip norm	10^{-12} / 1.0
Optimization	base LR / decoder LR	10^{-3} / 10^{-4}
Optimization	ligand LR multiplier	0.20
Optimization	EMA decay	0.9999
Joint loss	align / diffusion / calibration	1.00 / 1.00 / 0.25
Calibration	scale / smooth-ℓ_1 beta	0.95 / 1.00