Learning Protein-Ligand Binding in Hyperbolic Space

Jianhui Wang[1,2*] **Wenyu Zhu**[1*] **Bowen Gao**[1,3*] **Xin Hong**[1]
Ya-Qin Zhang[1] **Wei-Ying Ma**[1] **Yanyan Lan**[1,4†]

[1]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
[2]University of Electronic Science and Technology of China, Chengdu, China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[4]Beijing Academy of Artificial Intelligence, Beijing, China

Abstract

Protein-ligand binding prediction is central to virtual screening and affinity ranking, two fundamental tasks in drug discovery. While recent retrieval-based methods embed ligands and protein pockets into Euclidean space for similarity-based search, the geometry of Euclidean embeddings often fails to capture the hierarchical structure and fine-grained affinity variations intrinsic to molecular interactions. In this work, we propose HypSeek, a hyperbolic representation learning framework that embeds ligands, protein pockets, and sequences into Lorentz-model hyperbolic space. By leveraging the exponential geometry and negative curvature of hyperbolic space, HypSeek enables expressive, affinity-sensitive embeddings that can effectively model both global activity and subtle functional differences-particularly in challenging cases such as activity cliffs, where structurally similar ligands exhibit large affinity gaps. Our mode unifies virtual screening and affinity ranking in a single framework, introducing a protein-guided three-tower architecture to enhance representational structure. HypSeek improves early enrichment in virtual screening on DUD-E from 42.63 to 51.44 (+20.7%) and affinity ranking correlation on JACS from 0.5774 to 0.7239 (+25.4%), demonstrating the benefits of hyperbolic geometry across both tasks and highlighting its potential as a powerful inductive bias for protein-ligand modeling. Our code is publicly available at `https://github.com/jianhuiwemi/HypSeek`.

1 Introduction

Modeling protein–ligand interactions is critical for drug discovery, where accurate binding affinity prediction underpins both large-scale virtual screening and fine-grained ligand prioritization. Virtual screening seeks to identify molecules likely to bind a given protein target from large compound libraries, often containing millions or even billions of candidates. Approaches such as molecular docking [1, 2] estimate binding compatibility by sampling ligand poses and scoring them with physics-based functions. While effective in small-scale settings, these methods are computationally intensive and scale poorly to modern library sizes. Unlike virtual screening, which emphasizes identifying likely binders from vast libraries, affinity ranking focuses on ordering a smaller set of candidate ligands by predicted binding strength, with physics-based techniques like free energy perturbation (FEP+) [3] offering high accuracy at the cost of extensive molecular dynamics simulations. These limitations restrict the practicality of traditional methods in early-stage drug discovery pipelines.

A notable shift in virtual screening came with DrugCLIP [4], which reframed the task as a dense retrieval problem. Rather than predicting binding affinity or docking poses, DrugCLIP learns

*Equal contribution.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

contrastive embeddings of ligands and protein pockets such that interacting pairs are close in a shared Euclidean space. This design enables efficient similarity-based retrieval and allows for scalable screening across billion-scale compound libraries. Despite its promising performance and efficiency, DrugCLIP struggles to capture fine-grained interaction patterns which are essential for downstream affinity ranking. Recently, LigUnity [5] extends the retrieval-based framework by unifying virtual screening and affinity ranking into a single training objective. It combines contrastive learning for global interaction patterns with listwise ranking to model pocket-specific ligand preferences, aiming to jointly learn both binding likelihood and relative affinity within a unified embedding space.

While retrieval-based methods have shown strong potential, they typically embed ligands and protein pockets into Euclidean space, where distances grow linearly and the geometry does not explicitly encourage separation based on functional or activity-related differences. As a result, standard Euclidean training objectives may fail to emphasize fine-grained distinctions in binding strength, especially when molecular structures are similar.

To enrich the embedding geometry and better capture complex protein–ligand interactions, we propose HypSeek, a retrieval-based model that embeds ligands, pockets, and protein sequences into hyperbolic space. Unlike previous dual-tower designs, HypSeek adopts a protein-guided three-tower architecture during training to promote more structured representations. The curvature of hyperbolic space enables affinity-sensitive encoding through both angular direction and radial depth, providing greater expressivity than linear Euclidean geometry. This design not only enhances fine-grained affinity discrimination, but also offers a natural mechanism to address activity cliffs—cases where structurally similar ligands exhibit large differences in binding strength. While Euclidean embeddings often enforce functional similarity among structurally similar ligands, hyperbolic geometry allows such ligands to diverge meaningfully in the embedding space, reflecting differences in interaction modes or physicochemical properties. During inference, we retain efficient similarity computation via Euclidean inner products over hyperbolically shaped representations, preserving scalability without sacrificing expressiveness.

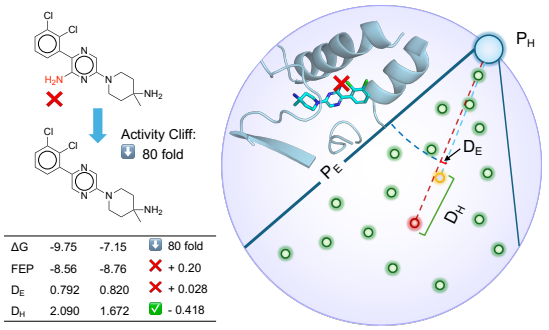

Figure 1: Illustration of how hyperbolic geometry distinguishes activity cliffs (PDB ID: 5EHR). **Left**: Two structurally similar ligands (Ligand ID: 5OD vs. its amino-substituent-removed derivative) show an ∼80-fold affinity difference. **Right**: The yellow and red points denote the two ligands; the blue point is the pocket. Dashed lines show distances in hyperbolic (red/light blue) and Euclidean (dark blue) space. Euclidean embeddings preserve structural similarity but fail to reflect affinity gaps, while hyperbolic embeddings separate such pairs via both radial and angular dimensions (D_H, green), enabling affinity-sensitive representations.

We evaluate HypSeek across both large-scale virtual screening and fine-grained affinity ranking tasks. On the DUD-E [6] benchmark, HypSeek improves $EF_{1\%}$ from **42.63 to 51.44 (+20.7%)**, demonstrating strong retrieval performance across targets. For affinity ranking, it increases Spearman correlation on the JACS [3] dataset from **0.5774 to 0.7239 (+25.4%)**, consistently outperforming Euclidean baselines. These results highlight the benefits of hyperbolic geometry in capturing both global activity and nuanced affinity variation within a unified embedding space.

In summary, our contributions are as follows:

- We propose a hyperbolic embedding framework for protein–ligand modeling, where the geometry naturally captures hierarchical interactions and targets **the critical challenge of activity cliffs** by enabling structured separation of similar ligands with divergent affinities.
- We introduce **HypSeek**, a dense retrieval model with a protein-guided three-tower architecture that integrates structure and sequence information to learn affinity-aware representations in hyperbolic space.

- HypSeek achieves strong performance on both virtual screening and affinity ranking, capturing fine-grained binding differences more effectively than Euclidean baselines while maintaining scalable inference.

2 Method

2.1 Problem Setting

Our goal is to predict the binding affinity between protein pockets and candidate ligands. The training data are organized by assay, where each assay is an experimental setup designed to evaluate ligand binding against a specific protein target. Each assay includes one protein and a subset of ligands from the full compound library that have been experimentally screened, yielding binary activity labels and optionally affinity values. Crucially, affinity values are only comparable within the same assay due to differences in experimental conditions (e.g., pH, temperature, cofactors), assay protocols (e.g., cell-based or target-based), and measurement types (e.g., IC_{50}, K_d, K_i).

Therefore, the task is formulated as learning relative binding strength rankings within each assay rather than predicting absolute affinities across assays. Let $\mathcal{A}$ denote the set of assays. For each assay $A_i \in \mathcal{A}$, let $\mathcal{L}_i$ be the set of tested ligands, and $v_i(\ell)$ be the affinity value of ligand $\ell \in \mathcal{L}_i$. Each assay corresponds to a target protein, represented by both its amino acid sequence and a set of candidate pocket structures $\mathcal{P}_i$. During training, one pocket from $\mathcal{P}_i$ is sampled to represent the structure, and combined with the sequence information to encode the full target. The model is trained to embed both targets and ligands into a shared hyperbolic space, enabling retrieval of active ligands and ranking them by relative binding strengths within each assay.

2.2 Multimodal Encoding and Lorentz Mapping

Let x^p and x^m denote the atom-based inputs (coordinates and types) for a protein pocket and ligand, respectively, and let $S = (s_1, \ldots, s_L)$ denote the amino acid sequence of a target protein. We define three encoder functions: g_ϕ and f_θ as SE(3)-equivariant 3D graph transformers for pockets and ligands (following DrugCLIP [4]), and h_ψ as a protein sequence encoder based on ESM-2 [7]. As illustrated in Figure 2, each encoder maps its input to a vector in $\mathbb{R}^{d_{\mathrm{euc}}}$:

$$E_{\mathrm{poc}} = g_\phi(x^p), E_{\mathrm{mol}} = f_\theta(x^m), E_{\mathrm{seq}} = h_\psi(S). \tag{1}$$

We then lift these Euclidean embeddings to hyperbolic space via the exponential map defined in Eq. (22):

$$\mathbf{h}_{\mathrm{poc}} = \exp_{\mathbf{0}}^{\kappa}\big(E_{\mathrm{poc}}\big), \mathbf{h}_{\mathrm{mol}} = \exp_{\mathbf{0}}^{\kappa}\big(E_{\mathrm{mol}}\big), \mathbf{h}_{\mathrm{seq}} = \exp_{\mathbf{0}}^{\kappa}\big(E_{\mathrm{seq}}\big). \tag{2}$$

The resulting hyperbolic embeddings $\mathbf{h}_{\mathrm{mol}}, \mathbf{h}_{\mathrm{poc}}, \mathbf{h}_{\mathrm{seq}} \in \mathbb{L}^n$ are subsequently employed in both the training and the inference stage.

2.3 Contrastive and Ranking as the Foundation.

We retain the in-batch contrastive retrieval losses of DrugCLIP [4] and LigUnity's listwise ranking term [5], applied to the hyperbolic embeddings $\tilde{\mathbf{h}}_u$. For each assay A_i with query modality $u \in \{\mathrm{poc}, \mathrm{seq}\}$ and its B candidate ligands $\{v_j\}$, we compute similarity logits $s_{i,j} = \frac{1}{\tau}\langle \tilde{\mathbf{h}}_{u_i}, \tilde{\mathbf{h}}_{v_j} \rangle$.

We adopt a symmetric InfoNCE objective over each assay A_i. Let $L_i \subseteq \{1, \ldots, B\}$ denote the indices of true binders for u_i. We compute:

$$\mathcal{L}^{(i)}_{\mathrm{p}\to\mathrm{l}} = -\frac{1}{|L_i|}\sum_{k\in L_i} \log \frac{\exp(s_{i,k})}{\sum_{j=1}^{B}\exp(s_{i,j})}, \tag{3}$$

$$\mathcal{L}^{(i)}_{\mathrm{l}\to\mathrm{p}} = -\frac{1}{|L_i|}\sum_{k\in L_i} \log \frac{\exp(s_{i,k})}{\sum_{n=1}^{B}\exp(s_{i,n})}, \tag{4}$$

The total contrastive loss is then

$$\mathcal{L}_{\mathrm{contrast}} = \frac{1}{2}\sum_i \left(\mathcal{L}^{(i)}_{\mathrm{p}\to\mathrm{l}} + \mathcal{L}^{(i)}_{\mathrm{l}\to\mathrm{p}}\right). \tag{5}$$

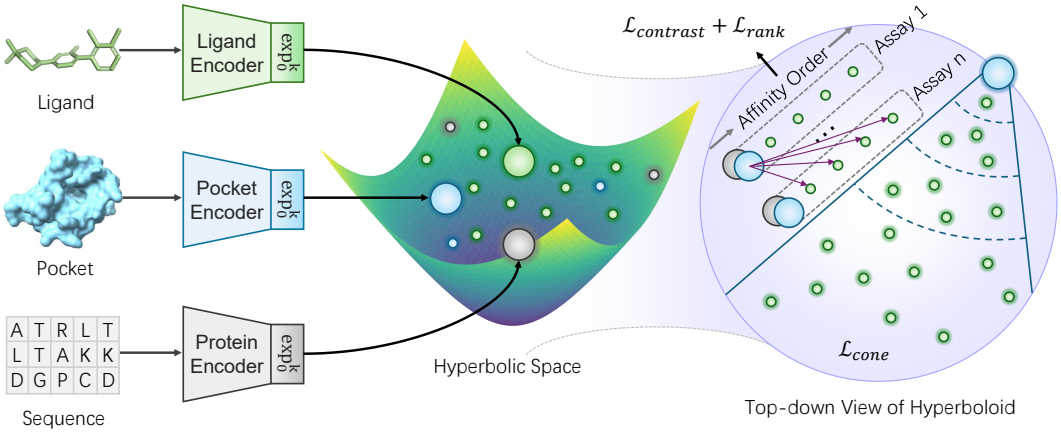

Figure 2: Overall architecture of HypSeek: three encoders lift ligands, pockets and protein sequences to a shared hyperbolic space (left); contrastive and list-wise ranking losses align pocket/sequence with ligands while the cone–hierarchy loss imposes radial–angular tiers around each pocket (right).

For each assay A_i the screened ligands are sorted by measured affinity, yielding an ordered list $(v_{i,1}, \ldots, v_{i,B})$. Following the Plackett–Luce model [8], the probability of selecting ligand $v_{i,k}$ at step k (from the remaining set $\mathcal{R}_{i,k} = \{k, k+1, \ldots, B\}$) is

$$p_{i,k}(v_{i,k}) \;=\; \frac{\exp\big(s_{i,k}\big)}{\sum_{j \in \mathcal{R}_{i,k}} \exp\big(s_{i,j}\big)}, \tag{6}$$

where $s_{i,k} = \langle \tilde{\mathbf{h}}_{u_i}, \tilde{\mathbf{h}}_{v_{i,k}} \rangle / \tau$. We use the decay $\mu_k = \frac{1}{\sqrt{B}\,\log(k+1)}$. The listwise loss for assay A_i is therefore

$$\mathcal{L}_{\text{rank}}^{(i)} \;=\; -\sum_{k=1}^{B} \mu_k \log p_{i,k}\big(v_{i,k}\big). \tag{7}$$

2.4 Hyperbolic Geometry as a Structural Prior

Beyond simply embedding pockets and ligands into a shared hyperbolic space, we aim to further leverage the geometric structure of $\mathbb{L}^n$ to encode fine-grained inductive biases about binding affinity. The exponential capacity of hyperbolic space allows for natural modeling of hierarchical relationships, while the Lorentz model enables cone-based entailment mechanisms. We therefore introduce a cone–hierarchy learning process that exploits both the radial and angular dimensions of hyperbolic space to reflect the graded nature of ligand binding strength.

Within an assay A_i, the protein pocket is represented by a Lorentz-model vector $\mathbf{h}_{\text{poc},i} \in \mathbb{L}^n$, and every screened ligand $j \in \mathcal{L}_i$ has its own embedding $\mathbf{h}_{\text{mol},ij} \in \mathbb{L}^n$. Each hyperbolic vector splits into a time-like coordinate and an n-dimensional spatial part: $\mathbf{h}_{\text{poc},i} = \big(p_{0,i},\ \tilde{\mathbf{p}}_i\big), \mathbf{h}_{\text{mol},ij} = \big(m_{0,ij},\ \tilde{\mathbf{m}}_{ij}\big)$, with $p_{0,i}, m_{0,ij} \in \mathbb{R}$ and $\tilde{\mathbf{p}}_i, \tilde{\mathbf{m}}_{ij} \in \mathbb{R}^n$. These components satisfy the hyperboloid constraint $p_{0,i}^2 - \|\tilde{\mathbf{p}}_i\|^2 = m_{0,ij}^2 - \|\tilde{\mathbf{m}}_{ij}\|^2 = 1/\kappa$.

The geodesic distance $d_{i,j} = d_{\mathbb{L}}(\mathbf{h}_{\text{poc},i}, \mathbf{h}_{\text{mol},i,j})$ is computed via Eq. (20). The exterior angle at the pocket,

$$\phi_{i,j} = \arccos\!\Big(\frac{m_{0,i,j} + \kappa\big(\langle \tilde{\mathbf{p}}_i, \tilde{\mathbf{m}}_{i,j} \rangle - p_{0,i} m_{0,i,j}\big) p_{0,i}}{\|\tilde{\mathbf{p}}_i\| \sqrt{\big[\kappa(\langle \tilde{\mathbf{p}}_i, \tilde{\mathbf{m}}_{i,j} \rangle - p_{0,i} m_{0,i,j})\big]^2 - 1}}\Big), \tag{8}$$

follows from the hyperbolic law of cosines and measures how far the ligand "leans" away from the pocket direction.

Each pocket defines a surface of admissible directions. Its half-aperture angle is formulated as [9, 10]

$$\omega_i = \arcsin\Big(\tfrac{2r_0}{\sqrt{\kappa}\,\|\tilde{\mathbf{p}}_i\|}\Big), \tag{9}$$

with a small constant $r_0 > 0$ to keep the expression bounded near the origin; larger $\|\tilde{\mathbf{p}}_i\|$ (a pocket already pushed towards the boundary) therefore yields a narrower cone.

Given the assay–specific affinity values $\{v_{i,j}\}_{j=1}^{|\mathcal{L}_i|}$, we draw K thresholds $t_0 < t_1 < \cdots < t_K$ and assign each ligand a bucket index

$$b_{i,j} \;=\; \big\{\, k \in \{0,\dots,K\} \;:\; v_{i,j} \in [t_k, t_{k+1}) \,\big\}. \tag{10}$$

Bucket 0 therefore collects the weakest binders and bucket K the strongest. For every ligand we derive a bucket–specific radial limit $r_{i,j}$ and angular–scaling factor $\eta_{i,j}$

$$r_{i,j} = r_0 + b_{i,j}\,\Delta r, \qquad \eta_{i,j} = \eta_0 - b_{i,j}\,\Delta\eta, \tag{11}$$

where r_0 and η_0 are the base radius/angle for the weakest tier, and $\Delta r, \Delta\eta > 0$ are the per-tier increments. Smaller $b_{i,j}$ thus yields a smaller radius cap and a larger cone. We penalise violations in radius and angle:

$$L_{\text{rad}} = \frac{1}{\sqrt{N}} \sum_{i,j} \max\big(d_{i,j} - r_{i,j},\, 0\big), \tag{12}$$

$$L_{\text{ang}} = \frac{1}{\sqrt{N}} \sum_{i,j} \max\big(\phi_{i,j} - \eta_{i,j}\,\omega_i,\, 0\big), \tag{13}$$

and combine them as

$$\mathcal{L}_{\text{cone}} = \lambda_{\text{rad}}\, L_{\text{rad}} + \lambda_{\text{ang}}\, L_{\text{ang}}. \tag{14}$$

We furthur introduce two regularization terms that operate on angular structure and intra-assay heterogeneity, respectively. To prevent trivial angular collapse, we introduce a fixed angular margin $m > 0$ beyond the cone boundary:

$$R_{\text{ang}} = \frac{1}{\sqrt{N}} \sum_{i,j} \max\big(\phi_{i,j} - \eta_{i,j}\omega_i + m,\, 0\big), \tag{15}$$

We also re-weight active ligands within each assay using rank-based weights $w_{i,j}$ and intra-assay softmax scores $p_{i,j}$:

$$R_{\text{het}} = \frac{1}{\max(C,1)} \sum_{i} \sum_{\substack{j \\ v_{i,j} < v_{\text{th}}}} -w_{i,j} \log p_{i,j}, \tag{16}$$

where C is the number of assays with at least one active ligand, and v_{th} is a predefined affinity threshold.

2.5 Addressing Activity Cliffs with Hyperbolic Geometry

While structurally similar ligands often cluster in Euclidean space, such geometry can underrepresent functional differences—especially in activity cliffs, where minor structural changes lead to large affinity shifts. As formalized in Proposition 1, hyperbolic space provides exponentially greater separation via angular variation, offering a principled mechanism for distinguishing such cases. The theoretical derivation is provided in Appendix.

Proposition 1. ***(Hyperbolic Separation of Activity Cliffs)*** *Let ℓ_1, ℓ_2 be structurally similar ligands with large affinity differences. Under constant radial norm and small angular deviation, hyperbolic embeddings yield significantly larger geodesic distance than their Euclidean counterparts:*

$$d_{\mathbb{H}}(h_H(\ell_1), h_H(\ell_2)) \gg d_E(h_E(\ell_1), h_E(\ell_2)).$$

This highlights the capacity of hyperbolic geometry to distinguish functionally divergent ligands without distorting local structural similarity.

2.6 Training and Inference

The core learning signal is driven by the pocket–ligand relationship. Accordingly, we apply hyperbolic regularisation only to the structure-based (pocket) branch, where geometric alignment in Lorentz space is both meaningful and effective. The sequence pathway provides complementary information to enhance generalisation, but does not participate in hyperbolic supervision.

Our full training objective is given by:

$$\mathcal{L}_{\text{total}} = \underbrace{\alpha_{\text{poc}}\Big(\mathcal{L}_{\text{cont}}^{\text{poc}\leftrightarrow\text{lig}} + \lambda_{\text{rank}}\,\mathcal{L}_{\text{rank}}^{\text{poc}}\Big)}_{\text{pocket}\,\leftrightarrow\,\text{ligand}} + \underbrace{\alpha_{\text{seq}}\Big(\mathcal{L}_{\text{cont}}^{\text{seq}\leftrightarrow\text{lig}} + \lambda_{\text{rank}}\,\mathcal{L}_{\text{rank}}^{\text{seq}}\Big)}_{\text{sequence}\,\leftrightarrow\,\text{ligand}} + \underbrace{\gamma_{\text{cone}}\,\mathcal{L}_{\text{cone}} + \lambda_{\text{ang}}\,R_{\text{ang}} + \lambda_{\text{het}}\,R_{\text{het}}}_{\text{pocket}\,\leftrightarrow\,\text{ligand}}. \tag{17}$$

At inference time, we simply embed a query pocket and each candidate ligand into hyperbolic space, extract their spatial components $\tilde{\mathbf{h}}_{\text{poc}}$ and $\tilde{\mathbf{h}}_{\text{mol},j}$, and compute similarity scores by their inner product $s_j = \tilde{\mathbf{h}}_{\text{poc}}^{\top}\,\tilde{\mathbf{h}}_{\text{mol},j}$. We then rank all ligands in descending order of s_j.

3 Experiments

3.1 Quantitative Results

Virtual Screening. As shown in Table 1, HypSeek substantially outperforms all baselines across both DUD-E and LIT-PCBA. On DUD-E, HypSeek achieves an AUROC of 0.9435, improving over the next best method (LigUnity) by more than 5 points, and delivers a $\text{BEDROC}_{80.5}$ of 0.7892, nearly 0.14 higher than LigUnity. Its $\text{EF}_{1\%}$ of 51.44 is more than 20 points above the highest competing model, demonstrating exceptional early retrieval of actives. Similarly, on the more challenging LIT-PCBA benchmark, HypSeek attains the top AUROC (0.6210), the highest $\text{BEDROC}_{80.5}$ (0.1196), and an $\text{EF}_{1\%}$ of 6.81, consistently surpassing both docking-based and deep learning approaches. These results highlight HypSeek's superior ability to rank true binders early in the list, making it particularly well suited for high-throughput virtual screening applications.

Table 1: Virtual-screening results on the DUD-E and LIT-PCBA benchmarks.

Method	**DUD-E** (n = 102)			**LIT-PCBA** (n = 15)		
	AUROC	$\text{BEDROC}_{80.5}$	$\text{EF}_{1\%}$	AUROC	$\text{BEDROC}_{80.5}$	$\text{EF}_{1\%}$
Glide-SP [1]	0.7670	0.4070	16.18	0.5315	0.4000	3.41
Surflex [11]	0.7426	0.2387	13.35	0.5147	—	2.50
DeepDTA [12]	0.5836	0.0513	2.28	0.5627	0.0253	1.47
Gnina [13]	0.7817	0.2994	17.73	0.6093	0.0540	4.63
BigBind [14]	0.5014	0.0240	1.18	0.6278	0.0502	3.79
RTMScore [15]	0.7529	0.4341	27.10	0.5247	0.0388	2.94
Tankbind [16]	0.7509	0.3300	13.00	0.5970	0.0389	2.90
DrugCLIP [4]	0.8093	0.5052	31.89	0.5717	0.0623	5.51
GenScore [17]	0.8160	0.4726	28.53	0.5957	0.0654	5.14
Planet [18]	0.7160	—	8.83	0.5731	—	3.87
EquiScore [19]	0.7760	0.4320	17.68	0.5678	0.0490	3.51
DrugHash [20]	0.8373	0.5716	37.18	0.5458	0.0714	6.14
$\text{LigUnity}_{\text{poc}}$ [5]	0.8922	0.6526	42.63	0.5985	0.1133	6.47
HypSeek	**0.9435**	**0.7892**	**51.44**	**0.6210**	**0.1196**	**6.81**

Table 2 reports ROC Enrichment (RE) metrics on the DUD–E benchmark under the fine-tuning setting. Notably, HypSeek achieves $\text{RE}_{0.5\%} = 137.15$, surpassing even the few-shot $\text{DrugCLIP}_{\text{FT}}$ result of 118.10, highlighting its exceptional ability to enrich actives early in the ranking.

Table 2: ROC–enrichment (RE) on the DUD–E benchmark.

Method	AUROC	$RE_{0.5\%}$	$RE_{1\%}$	$RE_{2\%}$	$RE_{5\%}$
Graph CNN [21]	0.8860	44.41	29.75	19.41	10.74
DrugVQA [22]	0.9720	88.17	58.71	35.06	17.39
AttentionSiteDTI [23]	0.9710	101.74	59.92	35.07	16.74
COSP [24]	0.9010	51.05	35.98	23.68	12.21
$\text{DrugCLIP}_{\text{ZS}}$ [4]	0.8093	73.97	41.79	23.68	11.16
$\text{DrugCLIP}_{\text{FT}}$ [4]	**0.9659**	118.10	67.17	37.17	16.59
$\text{LigUnity}_{\text{poc}}$ [5]	0.8922	104.69	57.47	33.76	13.88
HypSeek	0.9435	**137.15**	**73.16**	**38.80**	**16.60**

Affinity Ranking. We evaluate HypSeek on the JACS and Merck datasets using five independent random seeds to assess both accuracy and robustness. We report two sets of our results: "ensemble," which averages the five models' predictions before computing metrics, and "mean_{std}," which gives the mean and standard deviation of Pearson's r and Spearman's ρ across the five runs. As shown in Table 3, on JACS HypSeek (ensemble) achieves Pearson $r = 0.7742$ and Spearman $\rho = 0.7819$, closely matching the physics-based FEP+ (Pearson $r = 0.7811$, Spearman $\rho = 0.7595$) and significantly outperforming all deep-learning baselines. On Merck, HypSeek (ensemble) attains Pearson $r = 0.6120$ and Spearman $\rho = 0.5447$, leading the non-physics methods. Moreover, HypSeek's standard deviations are lower than those reported for LigUnity's mean_{std} results, indicating more consistent performance across random seeds.

Table 3: Affinity ranking results on the JACS and MERCK benchmark datasets.

Type	**Method**	**JACS**		**Merck**	
		Pearson r	Spearman ρ	Pearson r	Spearman ρ
Physics	FEP+ [3]	0.7811	0.7595	0.6960	0.6798
	MM-GB/SA [25]	0.1489	0.2011	0.1299	0.1299
DL	PBCNet [26]	0.3939	0.3799	0.4058	0.4075
	EHIGN [27]	0.5787	0.5814	0.4246	0.3830
	GET [28]	0.4034	0.3753	0.4203	0.4214
	BindNet [29]	0.5481	0.5368	0.4037	0.3477
	Boltz-2 [30]	0.5231	0.5285	0.4298	0.4013
	$\text{LigUnity}_{\text{poc}}$ (ensemble) [5]	0.6454	0.6460	0.5997	0.5554
	$\text{LigUnity}_{\text{poc}}$ (mean_{std}) [5]	$0.5705_{0.1955}$	$0.5774_{0.2097}$	$0.5323_{0.1865}$	$0.4994_{0.1773}$
Ours	**HypSeek (ensemble)**	0.7742	0.7819	0.6120	0.5447
	HypSeek (mean_{std})	$\mathbf{0.7186_{0.1157}}$	$\mathbf{0.7239_{0.1321}}$	$\mathbf{0.5606_{0.1738}}$	$\mathbf{0.5034_{0.1739}}$

3.2 Ablation and Analysis of HypSeek

Impact of Key Components. As summarised in Table 4, switching off hyperbolic–specific terms (no hyp) already degrades virtual–screening performance on DUD-E ($\text{BEDROC}_{80.5}$ drops from 0.7892 to 0.7671; $\text{EF}_{1\%}$ from 51.44 to 49.14), while the Euclidean baseline is markedly worse. The advantage becomes even more pronounced for affinity ranking on JACS, where Pearson r falls from 0.7518 to 0.6839 without hyperbolic supervision and to 0.5978 in purely Euclidean space. In the affinity ranking task, due to limited computational resources, we conducted each ablation with a single random seed. Ablating either the angular or heterogeneity regulariser alone (no R_{ang}, no R_{het}) yields intermediate losses, confirming that both angle control and intra-assay weighting contribute complementary signals beyond the core cone loss. Finally, removing the protein sequence pathway (no Seq) also degrades performance, indicating that protein-sequence features serve mainly as an auxiliary signal that further shapes the embeddings.

Table 4: Ablation results on the DUD-E and JACS benchmarks.

Setting	Module				DUD-E (n = 102)		JACS	
	$\mathcal{L}_{\text{cone}}$	R_{ang}	R_{het}	Seq	$\text{BEDROC}_{80.5}$	$\text{EF}_{1\%}$	Pearson r	Spearman ρ
Hyperbolic space								
Full model	✓	✓	✓	✓	0.7892	51.44	0.7518	0.7580
– no hyp	×	×	×	✓	0.7671	49.14	0.6839	0.6906
– no R_{ang}	✓	×	✓	✓	0.7856	50.52	0.7340	0.7529
– no R_{het}	✓	✓	×	✓	0.7773	50.42	0.7047	0.7074
– no Seq	✓	✓	✓	×	0.7351	47.70	0.7194	0.7050
Euclidean space								
Contrastive + rank	×	×	×	×	0.6565	42.87	0.5978	0.6060

Pairwise Affinity Prediction. Figure 3 (A)-(B) demonstrate the behavior of Euclidean and hyperbolic models across varying ECFP4 [31] similarity. Both models perform similarly on dissimilar ligand pairs, but as the ligands become more structurally similar, Euclidean accuracy and correlation decrease significantly. In contrast, the hyperbolic model maintains strong performance, even in these highly similar pairs. This suggests that the richer geometry information in hyperbolic space, which better accommodates relationships between molecules, is more effective at capturing subtle affinity shifts typical of situations where structurally similar molecules exhibit significantly different biological activity. These differences are often compressed in Euclidean space, where the geometry may fail to distinguish between such subtle shifts.

Embedding Visualization. Ligand embeddings are first reduced via HoroPCA [32] and visualized using CO-SNE [33]. Without hyperbolic constraints (Figure 3C), embeddings collapse near the origin with overlapping targets. With the full HypSeek objective (Figure 3D), clear target-wise clusters and radial affinity gradients emerge. This contrast illustrates how the cone–hierarchy constraints introduced by HypSeek structure the hyperbolic manifold, enabling a more effective representation of the complex relationships between ligands in hyperbolic space.

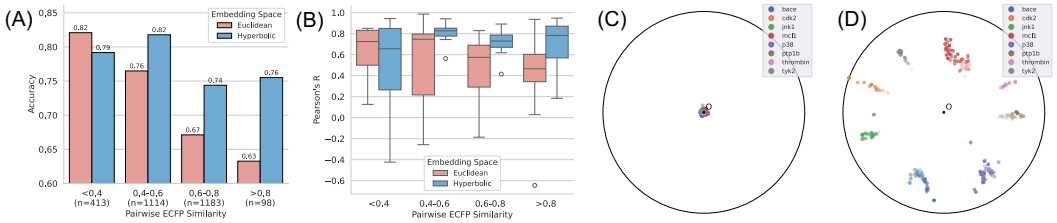

Figure 3: Pairwise analysis and CO-SNE visualization on the JACS benchmark. (A) Accuracy of affinity change prediction on ligand pairs with different ECFP4 similarity, comparing Euclidean and hyperbolic spaces; (B) Pearson's R between predicted score difference and ground truth affinity gap; (C) CO-SNE visualization of ligand embeddings in hyperbolic space without the hyperbolic constraint loss; (D) CO-SNE visualization of our HypSeek ligand embeddings.

4 Conclusion

We introduced HypSeek, a hyperbolic protein–ligand binding prediction model that embeds ligands, protein pockets, and sequences into a shared hyperbolic space using a three-tower architecture. By leveraging the negative curvature and exponential geometry of hyperbolic space, HypSeek captures both global interaction patterns and fine-grained affinity differences—especially in challenging cases like activity cliffs, where Euclidean embeddings often fail. Meanwhile, it retains efficient retrieval through inner product similarity, enabling large-scale virtual screening. Extensive experiments show that HypSeek consistently outperforms existing baselines across both screening and ranking tasks. HypSeek provides a geometry-aware solution for binding prediction.

Acknowledgments

This work was supported by the National Key R&D Program of China No. 2025ZD1802501, the National Natural Science Foundation of China (NSFC) No. 62506193, and the Beijing Academy of Artificial Intelligence.

References

[1] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, March 2004.

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

[3] Lingle Wang, Yujie Wu, Yuqing Deng, Byungchan Kim, Levi Pierce, Goran Krilov, Dmitry Lupyan, Shaughnessy Robinson, Markus K Dahlgren, Jeremy Greenwood, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7):2695–2703, 2015.

[4] 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. *Advances in Neural Information Processing Systems*, 36:44595–44614, 2023.

[5] Bin Feng, Zijing Liu, Mingjun Yang, Junjie Zou, He Cao, Yu Li, Lei Zhang, and Sheng Wang. A foundation model for protein-ligand affinity prediction through jointly optimizing virtual screening and hit-to-lead optimization. *bioRxiv*, pages 2025–02, 2025.

[6] 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, 2012.

[7] Zeming Lin, Hakan Akin, Roshan M. Rao, Rajan Das, Phineus Doshi, Tristan Bepler, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[8] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 129–136, New York, NY, USA, 2007. Association for Computing Machinery.

[9] Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3231–3241. Association for Computational Linguistics, 2019.

[10] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 7694–7731. PMLR, 2023.

[11] Russell Spitzer and Ajay N Jain. Surflex-dock: Docking benchmarks and real-world application. *Journal of computer-aided molecular design*, 26:687–699, 2012.

[12] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[13] Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):1–20, 2021.

[14] Michael Brocidiacono, Paul Francoeur, Rishal Aggarwal, Konstantin Popov, David Koes, and Alexander Tropsha. Bigbind: Learning from nonstructural data for structure-based virtual screening. 2022.

[15] Chao Shen, Xujun Zhang, Yafeng Deng, Junbo Gao, Dong Wang, Lei Xu, Peichen Pan, Tingjun Hou, and Yu Kang. Boosting Protein-Ligand Binding Pose Prediction and Virtual Screening Based on Residue-Atom Distance Likelihood Potential and Graph Transformer. *Journal of Medicinal Chemistry*, 65(15):10691–10706, August 2022.

[16] Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction. *Advances in Neural Information Processing Systems*, 35:7236–7249, December 2022.

[17] C. Shen, X. Zhang, C.-Y. Hsieh, Y. Deng, D. Wang, L. Xu, J. Wu, D. Li, Y. Kang, T. Hou, and P. Pan. A generalized protein-ligand scoring framework with balanced scoring, docking, ranking and screening powers. *Chemical Science*, 14(30):8129–8146, July 2023.

[18] Xiangying Zhang, Haotian Gao, Haojie Wang, Zhihang Chen, Zhe Zhang, Xinchong Chen, Yan Li, Yifei Qi, and Renxiao Wang. PLANET: A Multi-objective Graph Neural Network Model for Protein–Ligand Binding Affinity Prediction. *Journal of Chemical Information and Modeling*, 64(7):2205–2220, April 2024. Publisher: American Chemical Society.

[19] Duanhua Cao, Geng Chen, Jiaxin Jiang, Jie Yu, Runze Zhang, Mingan Chen, Wei Zhang, Lifan Chen, Feisheng Zhong, Yingying Zhang, Chenghao Lu, Xutong Li, Xiaomin Luo, Sulin Zhang, and Mingyue Zheng. Generic protein–ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. *Nature Machine Intelligence*, 6(6):688–700, June 2024. Publisher: Nature Publishing Group.

[20] Jin Han, Yun Hong, and Wu-Jun Li. Drughash: Hashing based contrastive learning for virtual screening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):17041–17049, Apr. 2025.

[21] Wen Torng and Russ B. Altman. Graph convolutional neural networks for predicting drug-target interactions. *Journal of Chemical Information and Modeling*, 59(10):4131–4149, 2019. PMID: 31580672.

[22] Shuangjia Zheng, Yongjian Li, Sheng Chen, Jun Xu, and Yuedong Yang. Predicting drug–protein interaction using quasi-visual question-answering system. *Nature Machine Intelligence*, 2(2):134–140, 2020.

[23] Mehdi Yazdani-Jahromi, Niloofar Yousefi, Aida Tayebi, Elayaraja Kolanthai, Craig J Neal, Sudipta Seal, and Ozlem Ozmen Garibay. Attentionsitedti: an interpretable graph-based model for drug-target interaction prediction using nlp sentence-level relation classification. *Briefings in Bioinformatics*, 23(4):bbac272, 07 2022.

[24] Zhangyang Gao, Cheng Tan, Jun Xia, and Stan Z. Li. Co-supervised pre-training of pocket and ligand. In *Machine Learning and Knowledge Discovery in Databases: Research Track, ECML PKDD 2023*, volume 13956 of *Lecture Notes in Computer Science*, pages 405–421, Turin, Italy, 2023. Springer.

[25] Samuel Genheden and Ulf Ryde. The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities. *Expert opinion on drug discovery*, 10(5):449–461, 2015.

[26] Jie Yu, Zhaojun Li, Geng Chen, Xiangtai Kong, Jie Hu, Dingyan Wang, Duanhua Cao, Yanbei Li, Ruifeng Huo, Gang Wang, et al. Computing the relative binding affinity of ligands based on a pairwise binding comparison network. *Nature Computational Science*, 3(10):860–872, 2023.

[27] Ziduo Yang, Weihe Zhong, Qiujie Lv, Tiejun Dong, Guanxing Chen, and Calvin Yu-Chian Chen. Interaction-based inductive bias in graph neural networks: enhancing protein-ligand binding affinity predictions from 3d structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[28] Xiangzhe Kong, Wenbing Huang, and Yang Liu. Generalist equivariant transformer towards 3D molecular interaction learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25149–25175. PMLR, 21–27 Jul 2024.

[29] Shikun Feng, Minghao Li, Yinjun Jia, Wei-Ying Ma, and Yanyan Lan. Protein-ligand binding representation learning from fine-grained interactions. In *The Twelfth International Conference on Learning Representations*, 2024.

[30] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025.

[31] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. PMID: 20426451.

[32] Ines Chami, Albert Gu, Dat P Nguyen, and Christopher Re. Horopca: Hyperbolic dimensionality reduction via horospherical projections. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1419–1429. PMLR, 18–24 Jul 2021.

[33] Yunhui Guo, Haoran Guo, and Stella X. Yu. Co-sne: Dimensionality reduction and visualization for hyperbolic data. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, 2022.

[34] Andrew T. McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. GNINA 1.0: molecular docking with deep learning. *Journal of Cheminformatics*, 13(1), June 2021.

[35] Xujun Zhang, Odin Zhang, Chao Shen, Wanglin Qu, Shicheng Chen, Hanqun Cao, Yu Kang, Zhe Wang, Ercheng Wang, Jintu Zhang, Yafeng Deng, Furui Liu, Tianyue Wang, Hongyan Du, Langcheng Wang, Peichen Pan, Guangyong Chen, Chang-Yu Hsieh, and Tingjun Hou. Efficient and accurate large library ligand docking with KarmaDock. *Nature Computational Science*, 3(9):789–804, September 2023. Publisher: Nature Publishing Group.

[36] J. Han, Y. Hong, and W.-J. Li. Drughash: Hashing based contrastive learning for virtual screening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17041–17049, 2025.

[37] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. 30, 2017.

[38] Benjamin Paul Chamberlain, James Clough, and Marc Peter Deisenroth. Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*, 2017.

[39] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *NeurIPS*, 2018.

[40] Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*, 2018.

[41] Silvère Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Trans. Autom. Control.*, 58(9):2217–2229, 2013.

[42] Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[43] Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *ICLR*, 2021.

[44] Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, and Nando de Freitas. Hyperbolic attention networks. In *ICLR*, 2019.

[45] Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *ICLR*, 2024.

[46] Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In *NeurIPS*, 2019.

[47] Liping Wang, Fenyu Hu, Shu Wu, and Liang Wang. Fully hyperbolic graph convolution network for recommendation. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3483–3487, 2021.

[48] P. Mettes, M. Ghadimi Atigh, M. Keller-Ressel, et al. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, 132:3484–3508, 2024.

[49] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. pages 6418–6428, 2020.

[50] Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. Hyperbolic self-paced learning for self-supervised skeleton-based action representations. In *ICLR*, 2023.

[51] Yuanpei Liu, Zhenqi He, and Kai Han. Hyperbolic category discovery, 2025.

[52] Bhuwan Dhingra, Christopher J Shallue, Mohammad Norouzi, Andrew M Dai, and George E Dahl. Embedding text in hyperbolic spaces. *arXiv preprint arXiv:1806.04313*, 2018.

[53] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincare glove: Hyperbolic word embeddings. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[54] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. pages 7694–7731. PMLR, 2023.

[55] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. *arXiv preprint arXiv:2410.06912*, 2024.

[56] Tobia Poppi, Tejaswi Kasarla, Pascal Mettes, Lorenzo Baraldi, and Rita Cucchiara. Hyperbolic safety-aware vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[57] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. pages 3779–3788. PMLR, 2018.

[58] Ya-Wei Eileen Lin, Ronald R Coifman, Gal Mishne, and Ronen Talmon. Hyperbolic diffusion embedding and distance for hierarchical representation learning. pages 21003–21025. PMLR, 2023.

[59] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6418–6428, 2020.

[60] David Mendez, Anna Gaulton, A. Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F. Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J. Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R. Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2018.

[61] Michael K. Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1):D1045–D1053, 10 2015.

[62] Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50(2):302–309, 2017.

[63] Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *Journal of Chemical Information and Modeling*, April 2020. Publisher: American Chemical Society.

[64] Christina EM Schindler, Hannah Baumann, Andreas Blum, Dietrich Bose, Hans-Peter Buchstaller, Lars Burgdorf, Daniel Cappel, Eugene Chekler, Paul Czodrowski, Dieter Dorsch, et al. Large-scale assessment of binding free energy calculations in active drug discovery projects. *Journal of Chemical Information and Modeling*, 60(11):5457–5474, 2020.

A Related Work

Virtual Screening. Structure-based virtual screening traditionally relies on molecular docking methods such as Glide [1] and AutoDock [2], which predict ligand binding poses and evaluate affinities using physics-based scoring functions. Some predict binding affinity directly from protein–ligand complex structures by learning scoring functions [34, 15, 19], while others infer interactions from raw structural inputs [16, 35]. A major shift occurred with DrugCLIP [4], which introduced contrastive retrieval by aligning ligand and pocket embeddings in a shared Euclidean space for billion-scale similarity search. This paradigm has since inspired a range of efficient retrieval methods. For example, DrugHash [36] employs binary hash codes for efficient retrieval with reduced memory cost, and LigUnity [5] integrates listwise ranking with contrastive screening.

Affinity Ranking. Accurate ranking of ligand binding affinities is essential for lead optimization but remains computationally challenging. Physics-based methods such as FEP+ [3] and MM-GB/SA [25] deliver high accuracy via alchemical free-energy calculations and implicit solvent models, respectively, yet they require extensive molecular dynamics sampling. Recent deep learning approaches seek to reduce this cost: PBCNet [26] models pairwise ligand differences with graph neural networks, EHIGN [27] encodes heterogeneous protein–ligand interaction graphs, and LigUnity [5] combines contrastive screening with listwise ranking to jointly address global retrieval and local prioritization.

Hyperbolic Representation Learning. Hyperbolic space has emerged as a powerful embedding manifold for data with latent hierarchical or tree-like structure, owing to its exponential volume growth that preserves hierarchy with low distortion [37, 38]. Early works demonstrated that embedding taxonomies or graphs in Poincaré or Lorentz models captures hierarchical relations more faithfully than Euclidean counterparts [39, 40]. This theoretical appeal led to specialized optimization methods and the design of hyperbolic neural layers, including Riemannian gradient algorithms [41, 42] and Hyperbolic Neural Networks [43], as well as adaptations of convolutional, attention, and graph architectures [44, 45]. Hyperbolic embeddings have demonstrated strong performance across diverse modalities—knowledge graphs and recommender systems [46, 47], vision tasks [48] such as classification and few-shot learning [49, 50, 51], and language modeling [52, 53]. Recent studies further explore multimodal training in hyperbolic space for vision–language models to capture hierarchical semantics [54, 55, 56]. Our work is the first to bring hyperbolic space to protein–ligand retrieval, leveraging its inductive bias to separate fine-grained affinity differences.

B Background

We perform all representation learning in an n-dimensional hyperbolic space of constant negative curvature, using the Lorentz model [57, 58, 54]. This choice affords numerical stability and readily supports geodesic and exponential-map operations.

Let $\mathbb{L}^n$ denote the Lorentz (hyperboloid) model, realized as the upper sheet of a two-sheeted hyperboloid in $\mathbb{R}^{n+1}$. We first equip $\mathbb{R}^{n+1}$ with the Lorentzian inner product

$$\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}} = -\, p_0\, q_0 + \langle \tilde{\mathbf{p}}, \tilde{\mathbf{q}} \rangle_{\mathbb{E}}, \tag{18}$$

where we write $\mathbf{p} = (p_0, \tilde{\mathbf{p}}), p_0 \in \mathbb{R}, \tilde{\mathbf{p}} \in \mathbb{R}^n$ with p_0 the *time*-coordinate and $\tilde{\mathbf{p}}$ the *spatial*-coordinates, and $\langle \cdot, \cdot \rangle_{\mathbb{E}}$ denotes the standard Euclidean inner product.

The Lorentz model is then defined by

$$\mathbb{L}^n = \Big\{ \mathbf{p} \in \mathbb{R}^{n+1} : \langle \mathbf{p}, \mathbf{p} \rangle_{\mathbb{L}} = -\frac{1}{\kappa}, p_0 = \sqrt{\tfrac{1}{\kappa} + \|\tilde{\mathbf{p}}\|^2}, \kappa > 0 \Big\}, \tag{19}$$

where $-\kappa \in \mathbb{R}$ is the curvature of the space.

We can measure distances by integrating the metric along geodesics. The Riemannian metric induced by the Lorentzian inner product gives the length of geodesics on $\mathbb{L}^n$, which in turn defines the hyperbolic distance.

$$d_{\mathbb{L}}(\mathbf{p}, \mathbf{q}) = \frac{1}{\sqrt{\kappa}} \cosh^{-1}\big(-\kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}}\big), \quad \mathbf{p}, \mathbf{q} \in \mathbb{L}^n. \tag{20}$$

At each point $\mathbf{p} \in \mathbb{L}^n$, the tangent space $T_{\mathbf{p}}\mathbb{L}^n$ provides a linear approximation of the manifold. Concretely, any tangent vector $\mathbf{v} \in T_{\mathbf{p}}\mathbb{L}^n \subset \mathbb{R}^{n+1}$ satisfy $\langle \mathbf{p}, \mathbf{v} \rangle_{\mathbb{L}} = 0$, so that

$$T_{\mathbf{p}}\mathbb{L}^n = \big\{ \mathbf{v} \in \mathbb{R}^{n+1} : \langle \mathbf{p}, \mathbf{v} \rangle_{\mathbb{L}} = 0 \big\}. \tag{21}$$

To transfer Euclidean encoder outputs into hyperbolic space, we apply the exponential map at a base point. For any $\mathbf{p} \in \mathbb{L}^n$ and $\mathbf{v} \in T_{\mathbf{p}}\mathbb{L}^n$, the exponential map is

$$\exp_{\mathbf{p}}^{\kappa}(\mathbf{v}) = \cosh\big(\sqrt{\kappa}\,\|\mathbf{v}\|_{\mathbb{L}}\big)\,\mathbf{p} + \frac{\sinh\big(\sqrt{\kappa}\,\|\mathbf{v}\|_{\mathbb{L}}\big)}{\sqrt{\kappa}\,\|\mathbf{v}\|_{\mathbb{L}}}\,\mathbf{v}, \tag{22}$$

where $\|\mathbf{v}\|_{\mathbb{L}} = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle_{\mathbb{L}}}$. In practice, we interpret the output of a Euclidean encoder as a vector in the tangent space at the point $\mathbf{0} = (\frac{1}{\sqrt{\kappa}}, 0, \ldots, 0)^\top$ on the hyperboloid, and then apply the exponential map $\exp_{\mathbf{0}}^{\kappa}$ to lift it onto $\mathbb{L}^n$ [59].

C Experimental Settings

Implemention Details. We adopt the same curated assay–level training dataset as LigUnity [5], which is constructed from ChEMBL [60], BindingDB [61], and PDBBind [62]. For virtual screening, we strictly exclude any target UniProt IDs present in the DUD-E [6], LIT-PCBA [63] test sets. For affinity ranking tasks, we perform ligand-level deduplication by removing redundant small molecules and non-redundant assay IDs. Training is run on four NVIDIA A100 GPUs for 50 epochs,, using the Adam optimizer with an initial learning rate of 1×10^{-4} and the curvature parameter κ (absolute value of negative curvature) fixed to 1.

Benchmark. In virtual screening, evaluations are performed on DUD-E [6] and LIT-PCBA [63]. DUD-E includes 102 protein targets, each associated with experimentally verified actives and 50 property-matched decoys, designed to test enrichment capability under artificially constructed decoy scenarios. LIT-PCBA, in contrast, contains 15 targets with over 400K experimentally confirmed inactives, offering a more realistic and challenging setting without synthetic decoy bias. For affinity ranking, the evaluation is conducted on JACS [3] and Merck [64]. JACS consists of eight high-quality congeneric series extracted from real lead optimization projects, emphasizing precise ranking within narrow chemical series, while Merck serves as a large-scale benchmark for FEP-based lead optimization with diverse chemical scaffolds and higher experimental noise.

Evaluation Metrics. For virtual screening, we use AUROC, $\text{BEDROC}_{80.5}$, Enrichment Factor (EF), and ROC-enrichment (RE) to assess model performance. For fine-grained affinity ranking, we evaluate using Pearson's and Spearman's rank correlation coefficients. More details are provided in Appendix E.2.

Baselines. We compare our method against a broad spectrum of existing approaches, including classical physics-based docking tools, empirical scoring functions, and modern deep learning models. These baselines reflect diverse modeling paradigms, ranging from structure-based simulations to neural networks trained on large protein–ligand datasets. For affinity ranking benchmarks, we additionally include methods based on free energy perturbation, energy decomposition, and recent representation learning techniques. All baselines are evaluated using their reported protocols or open-source implementations, ensuring consistency with prior work.

D Theoretical Motivation for Hyperbolic Separation of Activity Cliffs

A key challenge in protein–ligand modeling is the presence of *activity cliffs*—cases where structurally similar ligands exhibit large differences in binding affinity. We aim to show, from a geometric perspective, why hyperbolic space is better suited than Euclidean space for separating such ligand pairs.

D.1 Problem Setup

Let $\ell_1, \ell_2 \in \mathbb{R}^n$ be two ligands with high structural similarity, such that their Euclidean distance is small:

$$\|\ell_1 - \ell_2\|_E = \varepsilon, \quad \varepsilon \ll 1 \tag{23}$$

but their binding affinities differ significantly:

$$|f(\ell_1) - f(\ell_2)| \gg 0 \tag{24}$$

Our goal is to learn an embedding $h(\cdot)$ such that:

$$\|h(\ell_1) - h(\ell_2)\| \gg \varepsilon \tag{25}$$

i.e., the embedding space should amplify functional differences despite structural similarity.

D.2 Limitations of Euclidean Geometry

In Euclidean space $\mathbb{R}^d$, distance grows linearly:

$$d_E(x, y) = \|x - y\|_2 \tag{26}$$

Thus, structurally similar ligands must be mapped to nearby locations unless we distort the local geometry, which harms generalization and smoothness.

D.3 Hyperbolic Geometry and Exponential Separation

We consider the Lorentz model of hyperbolic space $\mathbb{H}^n$ with curvature $-\kappa$. The manifold is defined as:

$$\mathbb{H}^n = \{x \in \mathbb{R}^{n+1} \mid \langle x, x\rangle_L = -\frac{1}{\kappa},\ x_0 > 0\} \tag{27}$$

where the Lorentzian inner product is:

$$\langle x, y\rangle_L = -x_0 y_0 + \sum_{i=1}^{n} x_i y_i \tag{28}$$

The geodesic distance between $x, y \in \mathbb{H}^n$ is given by:

$$d_{\mathbb{H}}(x, y) = \frac{1}{\sqrt{\kappa}} \cosh^{-1}\left(-\kappa \langle x, y\rangle_L\right) \tag{29}$$

D.4 Angular Separation and Activity Cliffs

Let $v_1, v_2 \in T_o\mathbb{H}^n$ be tangent vectors at the origin o, representing two structurally similar ligands. Their exponential map into $\mathbb{H}^n$ is:

$$\exp_o(v) = \cosh(\|v\|) \cdot o + \sinh(\|v\|) \cdot \frac{v}{\|v\|} \tag{30}$$

Assume both vectors have the same norm $\|v_1\| \approx \|v_2\| = r$ (i.e. equal radial depth) and a small angular deviation $\theta = \angle(v_1, v_2) \ll 1$. By applying the hyperbolic law of cosines and the expansion $\operatorname{arccosh}(1 + \varepsilon) = \sqrt{2\varepsilon} + \mathcal{O}(\varepsilon^{3/2})$, their geodesic distance satisfies

$$d_{\mathbb{H}}\big(\exp_o(v_1), \exp_o(v_2)\big) \ \approx\ \frac{\sinh r}{\sqrt{\kappa}}\ \theta \ +\ \mathcal{O}(\theta^3). \tag{31}$$

This implies that even small angular differences (e.g., from subtle functional changes) lead to large separations if radial depth (i.e., binding strength) differs.

D.5 Conclusion

Proposition. Let $\ell_1, \ell_2 \in \mathbb{R}^n$ be structurally similar ligands with different affinity labels. Let $h_E : \mathbb{R}^n \to \mathbb{R}^d$ be a Euclidean embedding and $h_H : \mathbb{R}^n \to \mathbb{H}^d$ a hyperbolic embedding. Then under constant radial norm r and small angular separation θ, we have:

$$d_{\mathbb{H}}(h_H(\ell_1), h_H(\ell_2)) \gg d_E(h_E(\ell_1), h_E(\ell_2)) \tag{32}$$

This shows that hyperbolic geometry provides stronger capacity to distinguish *activity cliff pairs*, even under tight structural similarity, without requiring large Euclidean displacement or model distortion.

E Supplementary Analysis and Details

E.1 Analysis of Cross-Target Activity–Cliff Pairs

Table 5 lists 21 ligand pairs whose ECFP [31] similarity is greater than 0.60 yet display large differences in experimental binding free energy Exp ΔG making them representative *activity–cliff* cases for evaluating our embedding space. For comparison, the Euclidean scores in the table are produced by the current state-of-the-art pocket–ligand model LigUnity$_{\mathrm{poc}}$ [5], whereas the hyperbolic scores come from our method.

Directional Agreement with Experimental Affinity. Recall that a smaller (more negative) experimental ΔG indicates a stronger binder, whereas a larger model score indicates stronger binding. Hence, for every pair in Table 5 we expect the sign of $\Delta(\text{score})$ to be opposite to the sign of $\Delta(\text{Exp}\,\Delta G)$. This correspondence is clearly visible: whenever the experimental gap favours molecule B, the hyperbolic score is higher for B (positive ΔHyp), and vice-versa. Euclidean scores occasionally match the sign but the margin is often negligible. Several pairs show that even free-energy perturbation (FEP) [3] predicts the wrong direction of the affinity change, yet the hyperbolic score still aligns with the experimental ordering.

Separation Magnitude. The Euclidean score differences are typically tiny (many are < 0.05), making it hard to tell the two ligands apart. In contrast, the hyperbolic score differences are an order of magnitude larger, providing an immediate visual cue of which ligand the model prefers. This numerical gap illustrates how the hyperbolic embedding stretches *activity-cliff* pairs, whereas the Euclidean embedding leaves them almost collapsed.

Table 5: Cross-Target Activity-Cliff Cases.

Molecule A	Molecule B	PDB ID	ECFP	Exp ΔG (A/B)	FEP ΔG (A/B)	Euc (A/B)	Hyp (A/B)	Δ(Exp ΔG)	ΔFEP	ΔEuc	ΔHyp
		4HW3	0.6731	-6.66/ -8.67	-3.5197/ -8.0306	+0.5356/ +0.5490	+1.5055/ +1.7663	-2.01	-4.5109	+0.0132	+0.2608
		4GIH	0.7674	-7.42/ -9.54	-6.1068/ -9.8942	+0.7420/ +0.7476	+1.1878/ +1.7734	-2.12	-3.7874	+0.0054	+0.5855
		6HVI	0.7097	-10.24/ -7.19	-11.3100/ -7.1480	+0.6760/ +0.6830	+1.9901/ +1.5151	+3.05	+4.16	+0.0073	-0.4750
		2GMX	0.6500	-8.11/ -9.99	-7.8979/ -9.9855	+0.8066/ +0.7866	+1.8717/ +2.2653	-1.88	-2.0876	-0.0200	+0.3936
		1H1Q	0.7273	-11.25/ -8.18	-9.8937/ -8.1376	+0.5684/ +0.5700	+2.0705/ +1.7103	+3.07	+1.7561	+0.0015	-0.3601
		4DJW	0.7719	-9.47/ -11.35	-9.9357/ -11.1010	+0.9110/ +0.9175	+2.1507/ +2.3914	-1.88	-1.1653	+0.0063	+0.2407
		4GIH	0.9048	-11.31/ -9.70	-10.5581/ -9.4767	+0.7390/ +0.7446	+1.8883/ +1.6953	+1.61	+1.0814	+0.0059	-0.1930
		6HVI	0.7213	-9.77/ -7.19	-11.1920/ -7.1480	+0.6953/ +0.6830	+2.0241/ +1.5151	+2.58	+4.0440	-0.0122	-0.5090
		1H1Q	0.7018	-11.11/ -8.18	-9.8570/ -8.1376	+0.5645/ +0.5700	+2.0925/ +1.7103	+2.93	+1.7194	+0.0054	-0.3822
		6HVI	0.6515	-7.69/ -10.71	-8.6460/ -10.5600	+0.7173/ +0.7380	+1.5800/ +2.2676	-3.02	-1.9140	+0.0205	+0.6876
		2GMX	0.6897	-7.51/ -9.68	-8.8421/ -10.7494	+0.7437/ +0.7650	+1.9024/ +2.3753	-2.17	-1.9073	+0.0215	+0.4729
		5EHR	0.6667	-7.15/ -9.75	-8.7560/ -8.5590	+0.8203/ +0.7920	+1.6715/ +2.0898	-2.60	+0.1970	-0.0283	+0.4183
		4HW3	0.7872	-6.66/ -8.90	-3.5197/ -7.2962	+0.5356/ +0.5435	+1.5055/ +1.7195	-2.24	-3.7765	+0.0078	+0.2141
		4DJW	0.8103	-11.35/ -9.42	-11.1010/ -9.4110	+0.9175/ +0.9263	+2.3914/ +2.1834	+1.93	+1.6900	+0.0088	-0.2080
		6HVI	0.7458	-7.93/ -10.24	-6.5430/ -11.3100	+0.6113/ +0.6760	+1.5235/ +1.9901	-2.31	-4.7670	+0.0644	+0.4666
		2QBS	0.7385	-11.42/ -8.72	-9.6890/ -8.8168	+0.9224/ +0.8660	+2.4139/ +2.0398	+2.70	+0.8722	-0.0561	-0.3741
		3FLY	0.6351	-10.23/ -12.26	-9.8951/ -12.1479	+0.2379/ +0.2490	+1.7866/ +2.1387	-2.03	-2.2528	+0.0111	+0.3522
		4UI5	0.7234	-10.05/ -12.08	-9.7050/ -11.9640	+0.9090/ +0.9380	+1.9309/ +2.2185	-2.03	-2.2590	+0.0288	+0.2876
		4PV0	0.8000	-6.82/ -11.83	-10.7470/ -11.1020	+0.6660/ +0.7170	+1.9251/ +2.1345	-5.01	-0.3550	+0.0508	+0.2094
		4GIH	0.7347	-11.70/ -9.00	-10.9067/ -8.8033	+0.7450/ +0.7090	+1.9382/ +1.4873	+2.70	+2.1034	-0.0361	-0.4510
		3FLY	0.6571	-11.85/ -10.23	-12.8311/ -9.8951	+0.2678/ +0.2379	+2.1754/ +1.7866	+1.62	+2.9360	-0.0299	-0.3889

E.2 Evaluation Metrics

Virtual screening asks whether a model can place a handful of true binders at the very top of a ranked list that may contain millions of inactives; affinity ranking asks whether it can preserve the fine-grained order of binding strengths within a chemically related series. Accordingly we employ different metrics.

(1) Virtual Screening Metrics.

AUROC. The area under the ROC curve is the probability that a randomly chosen active (a) scores higher than a randomly chosen inactive (d): $\Pr[s(a) > s(d)]$. Values range from 0.5 (random) to 1.0 (perfect) but treat the whole ranked list uniformly.

BEDROC$_{80.5}$. To emphasise the earliest part of the ranked list, we adopt the Boltzmann-enhanced discrimination of ROC (BEDROC) with focus parameter $\alpha = 80.5$, for which roughly the top 2 % of ranks account for 80 % of the score. Let N be the library size, N_t the number of actives, and $r_i \in [1, N]$ the rank of active i. The normalised form is

$$\begin{aligned} \text{BEDROC}_\alpha = & \frac{\sum_{i=1}^{N_t} e^{-\alpha r_i/N}}{R_\alpha \left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1} \right)} \\ & \times \frac{R_\alpha \sinh(\alpha/2)}{\cosh(\alpha/2) - \cosh\big(\alpha/2 - \alpha R_\alpha\big)} \\ & + \frac{1}{1 - e^{\alpha(1-R_\alpha)}}, \end{aligned} \tag{33}$$

where $R_\alpha = N_t/N$ is the active fraction. Equation (33) is bounded in $[0, 1]$; higher values indicate stronger early enrichment.

Enrichment Factor. The factor at a cut-off $\alpha\%$ quantifies how many actives the model retrieves relative to random ranking:

$$\text{EF}_\alpha = \frac{\text{NTB}_\alpha}{\text{NTB}_t \, \alpha/100}, \tag{34}$$

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

ROC Enrichment (RE). At a false-positive-rate threshold $x\%$ we report

$$\text{RE}(x\%) = \frac{\text{TP}/P}{\text{FP}_{x\%}/N} = \frac{\text{TP}\, N}{P\, \text{FP}_{x\%}}, \tag{35}$$

where N is the library size, P the number of actives, TP the true positives among the top-ranked compounds, and $\text{FP}_{x\%}$ the false positives observed before the FPR reaches $x\%$. A larger RE means stronger early discrimination.

(2)Affinity-ranking metrics.

Within a congeneric series we measure linear and rank agreement between predicted ($\hat{y}$) and experimental (y) affinities.

$$\text{Pearson } r \;=\; \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2} \sqrt{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}}, \tag{36}$$

$$\text{Spearman } \rho \;=\; 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}, \tag{37}$$

where d_i is the rank difference for compound i and n the series size. Both metrics lie in $[-1, 1]$; higher values indicate better agreement (1 is perfect correlation).