Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Molecular Function-guided Unsupervised Similarity-based Enrichment for Low-Data Virtual Screening

Alexander Hagg

alexander.hagg@h-brs.de

Hochschule Bonn-Rhein-Sieg University of Applied Sciences

Dirk Reith

Hochschule Bonn-Rhein-Sieg University of Applied Sciences

Matthias Preller

Hochschule Bonn-Rhein-Sieg University of Applied Sciences

Karl N. Kirschner

Hochschule Bonn-Rhein-Sieg University of Applied Sciences

Research Article

Keywords: virtual screening, unsupervised learning, manifold learning, molecular descriptors, low-data regimes, cheminformatics

Posted Date: June 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9781932/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Molecular Function-guided Unsupervised Similarity-based Enrichment for Low-Data Virtual Screening

Alexander Hagg[1,2*], Dirk Reith[1,2,3], Matthias Preller[2,4,5], Karl N. Kirschner[2,6]

[1*]Department of Engineering and Communication, University of Applied Sciences Bonn-Rhein-Sieg, Grantham-Allee 20, Sankt Augustin, 53757, Germany.
[2*]Institute of Technology, Resource and Energy-efficient Engineering, University of Applied Sciences Bonn-Rhein-Sieg, Grantham-Allee 20, Sankt Augustin, 53757, Germany.
[3]Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, 53754, Germany.
[4]Department of Natural Sciences, University of Applied Sciences Bonn-Rhein-Sieg, Von-Liebig-Straße 20, Rheinbach, 53359, Germany.
[5]Institute for Functional Gene Analytics, University of Applied Sciences Bonn-Rhein-Sieg, Von-Liebig-Straße 20, Rheinbach, 53359, Germany.
[6*]Department of Computer Science, University of Applied Sciences Bonn-Rhein-Sieg, Grantham-Allee 20, Sankt Augustin, 53757, Germany.

*Corresponding author(s). E-mail(s): alexander.hagg@h-brs.de;
Contributing authors: dirk.reith@h-brs.de; matthias.preller@h-brs.de; karl.kirschner@h-brs.de;

Abstract

Identifying bioactive compounds for orphan or understudied protein targets remains a significant hurdle in drug discovery due to the lack of sufficient structure-activity relationship (SAR) data. We present MolFuSE (Molecular Function-guided Similarity Explorer), an unsupervised, projection-based enrichment framework designed for low-data virtual screening. MolFuSE leverages manifold learning - specifically uniform manifold approximation and projection (UMAP) - to create low-dimensional similarity spaces based on a reference set

of ligands known to bind proteins with the same molecular function as the target. Through systematic evaluation across nine tyrosine transferase targets, we demonstrate that UMAP embeddings of Mordred 2D descriptors act as a potent "potency filter", significantly outperforming traditional topological fingerprints in the early enrichment of high-potency inhibitors (<100 nM). While extended-connectivity fingerprints (ECFP4) fingerprints achieve superior global ranking (receiver operating characteristic (ROC)-area under the curve (AUC) 0.961 vs. 0.863), the descriptor-based approach provides superior early recognition (enrichment factor at 1% (EF@1%)) when the reference set is appropriately curated. Crucially, we identify a critical point in data scaling: performance peaks at a reference set size of approximately 10,000 compounds and degrades for larger sets due to descriptor-space dilution from structurally diverse weak binders. MolFuSE provides a computationally efficient, interpretable tool for prioritizing high-affinity candidates when target-specific data is unavailable.

Scientific Contribution

We introduce a data-pooling strategy for low-data virtual screening that uses molecular-function annotations to aggregate "pseudo-active" reference sets, bypassing the need for target-specific ligands or structural data. Evaluating this framework across eight diverse protein families, we show that non-linear UMAP projections of physicochemical descriptors act as a potent potency filter, preferentially prioritizing high-affinity binders (< 100 nM) in early enrichment compared to traditional topological fingerprints. We provide evidence of a critical scaling threshold where reference sets exceeding approximately 10,000 diverse compounds dilute the descriptor space, establishing a practical blueprint for unsupervised screening across distinct target classes.

Keywords: virtual screening, unsupervised learning, manifold learning, molecular descriptors, low-data regimes, cheminformatics

1 Introduction

Identifying novel bioactive compounds for emerging therapeutic targets remains a persistent challenge in early-stage drug discovery. For well-characterized protein targets with extensive SAR data, machine learning approaches have demonstrated remarkable success in virtual screening campaigns Tran-Nguyen et al. (2023); Ghislat et al. (2021); Carlsson and Luttens (2024); Kumar and Acharya (2024); Chowdhury et al. (2025). However, the majority of potential therapeutic targets – particularly those in understudied protein families or emerging disease pathways – lack the hundreds to thousands of known bioactive molecules required for robust supervised learning models Ramsundar et al. (2019).

The Low-Data Virtual Screening Challenge

This low-data regime problem is particularly acute for novel targets like myosins, orphan G-protein-coupled receptors, Tang et al. (2012); Jobe and Vijayan (2024) and

other undruggable proteins, Xie et al. (2023) for which experimental screening campaigns have yet to be conducted at scale. While large pharmaceutical companies may leverage proprietary databases, this data remains inaccessible to the broader research community. Furthermore, for truly novel targets – such as those identified in emerging pathogens or rare diseases – historical screening data may not exist. Traditional ligand-based virtual screening (LBVS) methods rely on the *similar property principle*: structurally similar compounds are likely to exhibit similar biological activities Wilkins and Randić (1980); Johnson et al. (1990); Klabunde (2007); Willett (2011) and proteins with the same function bind the same set of small molecules Shortridge et al. (2011). This principle underpins similarity searching, where known active compounds for a specific target protein serve as queries to identify novel candidates from large chemical libraries. Compounds are typically encoded as fingerprints (e.g., ECFP4 Rogers and Hahn (2010)) or physicochemical descriptors, enabling rapid comparison via metrics such as the Tanimoto coefficient Tanimoto (1958). However, if few or no ligands are known, then the pairwise similarity comparison becomes undefined. Structure-based virtual screening (SBVS) via molecular docking Appelt et al. (1991); Kitchen et al. (2004); Lionta et al. (2014) offers an alternative pathway through the lock-and-key principleFischer (1894), but requires high-quality experimental structures – frequently unavailable for novel targets – and computationally expensive pose optimization. To reduce the SBVS computational cost, an LBVS prefiltering can be done. In practice, early retrieval performance in virtual screening is commonly quantified using the enrichment factor at 1% (EF@1%). Literature analyses indicate that traditional 2D LBVS methods typically yield median EF@1% values of approximately 16.3 (10%/90% percentiles [12.7, 21.7]), whereas 3D shape-based methods can achieve higher median values near 27.8 (10%/90% percentiles [27.6, 28.0]) on custom benchmarking sets Sciabola et al. (2022). Crucially, these metrics are not directly comparable across studies due to differing target compositions and baseline difficulties. For instance, structure-based molecular docking protocols evaluated on distinct, potentially more challenging benchmarks like the DUD-E dataset show mean EF@1% values of approximately 8.09 for standard tools like AutoDock Vina Ni et al. (2024), though multi-reference implementations can yield higher rates. These historical metrics establish a performance context for low-data screening frameworks.

Manifold Learning from Related Proteins

Homology-based similarity search has been used to identify ligands for a reference target when no initial ligands are known, by screening compounds that bind to homologous targets Schuffenhauer et al. (2003). The authors introduce a descriptor based on four molecular features and calculate the molecular similarity between the candidate compound and the reference set by measuring the distance to its nearest neighbor(s). While the method yields high-quality rankings, the linear similarity space is entirely determined by the four features. Consequently, it may not perform well in highly diverse molecular spaces.

Following developments in chemogenomics, which have identified that certain molecular families can bind similar proteins, cross-target supervised approaches have been developed to screen the chemical space against entire protein families when known

ligands for a target protein are absent Jacob and Vert (2008). When comparing binding affinity predictions using family-specific versus generic protein–ligand complexes, the family-specific models consistently yielded better predictions Wang et al. (2015).

Supervised approaches, as described above, require datasets with affinity annotations, whereas unsupervised approaches can benefit from richer datasets. Herein, we exploit unsupervised learning methods, specifically *manifold learning*, from related proteins. For example, if a novel transferase target has no known inhibitor, one might leverage the 100,000+ known transferase inhibitors from associated targets to guide the screening. To our knowledge, this "representation-based transfer" concept – defined as collections of bioactive compounds grouped by broad protein family annotations (e.g., all transferase inhibitors and all G-protein-coupled receptors) – has not been extensively explored as a manifold learning method in virtual screening, particularly for non-linear approaches.

While using functional annotations to group proteins is an established concept Gao and Skolnick (2013), our approach specifically applies this concept to bootstrap virtual screening by enriching an ample, diverse purchasable chemical space (i.e., from ZINC Irwin and Shoichet (2005)) with a targeted set of functionally relevant compounds. We hypothesize that a non-linear similarity space derived from the combined ZINC and molecular function (MF)-specific datasets provides a more informative chemical landscape, enabling the identification of promising candidates, even for targets with few known ligands. In the search for ligands that show affinity to novel or orphan target proteins, we define a reference set (RS) as the set containing ligands with known affinity values for proteins within the same molecular function family. Aggregating known ligands for all proteins associated with a target's function yields a significantly larger and more diverse RS of "pseudo-active" compounds, guiding the search more effectively and covering a larger portion of the potentially active molecular space. However, to better characterize such a diverse set, a richer (i.e., higher-dimensional) descriptor set is needed, which in turn poses its own problem.

Dimensionality Reduction for Chemical Space Navigation

High-dimensional descriptors (i.e., encoding spaces) suffer from the curse of dimensionality, where pairwise distances converge to the same magnitude, making comparisons meaningless Bellman (1966); Beyer et al. (1999). To mitigate this information loss, (non-)linear projections can be used to find an embedding space whose dimensionality is close to the number of dimensions necessary to fully describe the data (a.k.a., the "intrinsic dimensionality") Bennett (1965). Since projection methods are lossy by themselves, and it is usually not possible to determine the intrinsic dimensionality, the embedding space dimensionality can be chosen to allow a certain amount of information loss. For example, in principal component analysis (PCA) the dimensionality reduction is achieved by retaining the number of principal components required to account for a predefined percentage of the total data.

Recent advances in projective dimensionality reduction, particularly UMAP McInnes et al. (2018), enable the construction of low-dimensional molecular similarity spaces that preserve both local neighborhoods and global structure. This technique has been successfully applied to chemical space visualization (i.e., chemography) Oprea

and Gottfries (2001); Medina-Franco et al. (2008) and has very recently been benchmarked against other non-linear projection methods for that purpose Orlov et al. (2025).

Unlike linear techniques such as PCA, which often struggle to preserve non-linear local similarity relationships Lovrić et al. (2021); Orlov et al. (2025), or alternative non-linear methods like t-SNE, which do not preserve global spatial relationships and are computationally difficult to scale to large datasets Kobak and Berens (2019), UMAP balances local and global structures while facilitating out-of-sample mapping without model retraining Becht et al. (2019). While modern deep learning representation methods – such as Graph Neural Networks and pre-trained Transformer models like ChemBERTa or MolBERT – capture highly complex chemical features Chithrananda et al. (2020); Fabian et al. (2020); Liu et al. (2025), they typically require substantial computational resources and many training examples for fine-tuning, alongside offering limited direct interpretability compared to explicit distance-based metric spaces. This makes unsupervised manifold learning an attractive alternative to navigate these limitations. However, whether embeddings trained on a generic RS can effectively prioritize high-potency compounds for specific low-data targets remains an open question. In this work, we investigate this problem using MolFuSE.

MolFuSE Framework and Contributions

MolFuSE (Molecular Function-guided Similarity Explorer) is a projection-based enrichment framework designed explicitly for low-data virtual screening scenarios. MolFuSE operates under the following three core design principles:

1. **Projection-only strategy**: Dimensionality reduction models are fit exclusively on the large RS plus decoy libraries, ensuring that held-out target actives do not influence the learned similarity space – mimicking realistic prospective screening conditions.
2. **Exact nearest-neighbor scoring**: Compounds are ranked according to their minimum Euclidean distance to the RS within the manifold (i.e., the embedded space). This provides interpretable retrieval paths and facilitates post-hoc analysis of enrichment drivers.
3. **Potency-stratified evaluation**: Performance is assessed separately for high-potency (affinity < 100 nM), medium-potency (100–1000 nM), and weak (1000–10000 nM) actives, revealing method-specific biases toward different potency regimes.

Through four systematic experimental phases, we explore MolFuSE's performance using several criteria: (1) hyperparameter optimization of dimensionality reduction methods (i.e., PCA vs. UMAP) and molecular representations (i.e., 1,611 Mordred 2D descriptors vs. 2048-bit ECFP4 fingerprints), (2) sensitivity to RS affinity cutoffs before scoring (i.e., filtering RS by affinity thresholds), (3) robustness to RS size (i.e., random subsampling from 10 to 100,000+ compounds with full model retraining), and (4) cross-target generalization (i.e., across eight target proteins from various molecular function families with varying natural RS sizes).

Our findings reveal that UMAP embeddings of Mordred descriptors achieve state-of-the-art enrichment for high-potency transferase inhibitors, substantially outperforming both traditional fingerprint methods and literature-reported 3D docking benchmarks. Critically, we demonstrate that adequate screening is achievable with an RS as small as 1,000 compounds, and identify a previously unreported critical point where an extensive RS degrades performance–likely due to descriptor space dilution from structurally diverse low-affinity binders. These results establish MolFuSE as a practical, computationally efficient tool for hit discovery in low-data regimes, with implications for orphan-target prioritization and early-stage drug-discovery workflows.

2 Methods

Our experimental framework relies on the interplay between three distinct compound sets: (1) the RS, comprising known binders to the transferase target protein family; (2) the Target Actives Set, a held-out collection of ligands with known affinity for specific targets (e.g., tyrosine-protein kinase ABL1) that are used for evaluation; and (3) the Decoy Set, a large library of presumed non-binders used to challenge the retrieval capability.

2.1 Datasets

Reference Set: Transferase Inhibitors

The dataset was constructed from ChEMBL (version 34) Gaulton et al. (2012) by querying all compounds with reported bioactivity with human transferase targets. Compounds were required to have standard binding affinity measurements (EC_{50}, IC_{50}, K_i, or K_d) in nanomolar units, and valid SMILES strings with successful RDKit parsing.

For compounds with multiple measurements against the same or different transferase targets, we applied median affinity deduplication: the median affinity across all assays was computed and assigned as the compound's representative value. This approach provides robustness against multi-assay variability and batch effects common in large-scale screening databases Fourches et al. (2010). The final RS contained approximately 191,000 unique transferase inhibitors with median affinity values ranging from sub-nanomolar to >10 μM.

Target Actives Set: Tyrosine Transferase Inhibitors

For Phase 1–3 experiments, we selected Tyrosine-Protein Kinase/Transferase ABL1 (CHEMBL1862) as the primary evaluation target since it has many known actives, making it a convenient target for this work. Known actives were retrieved from ChEMBL using the P00519 identifier, requiring affinity measurements between 1 nM and 10,000 nM and the removal of active compounds from the RS dataset, via canonical SMILES matching, to prevent data leakage between training and test sets.

We also divided the active set into three potency tiers for evaluation:

- **High**: affinity $<$ 100 nM (N = 1,562 compounds, 46.9%)
- **Medium**: 100 nM $\leq$ affinity $<$ 1,000 nM (N = 729, 21.9%)

- **Weak**: 1,000 nM $\leq$ affinity $\leq$ 10,000 nM (N = 975, 29.3%)

For Phase 4's cross-target validation, we selected eight additional targets spanning diverse structural families and RS sizes (43,000–425,000 compounds): ABL2, EGFR, ERBB2, SRC, LCK, YES1, FYN and LYN.

Decoy Set: ZINC Database

To simulate realistic screening conditions, we incorporated a decoy library from the ZINC database Irwin and Shoichet (2005), representing commercially available "screening-like" compounds with no known transferase activity. After removing overlap with the RS or target actives (by canonical SMILES), the final decoy set contained 1,293,566 compounds. The decoy-to-active ratio is 58:1 with strict SMILES-based deduplication.

2.2 Molecular Representations

To capture diverse aspects of chemical similarity, we utilized two distinct encoding strategies – explicit physicochemical descriptors (Mordred) to capture molecular properties, and topological fingerprints (ECFP4) to capture substructural features.

Mordred 2D Descriptors

A total of 1,611 two-dimensional (2D) molecular descriptors were utilized in this study, calculated via the Mordred library Moriwaki et al. (2018). These descriptors are summarized by category, descriptor family, and weighting properties in the supporting information. For a full listing, we refer to the code repository (see 5.1).

Preprocessing the resulting data consisted of the following steps:

1. **Zero-variance filtering**: Descriptors with variance $< 10^{-12}$ (effectively constant across all compounds), or those where no descriptor values were calculable, were removed, yielding approximately 1,611 informative descriptors.
2. **Median imputation**: Missing descriptor values, around 5.95% on average per descriptor (e.g., from SMILES parsing failures or undefined mathematical operations), were replaced with the median descriptor value computed across the RS and decoy set.
3. **Standardization**: The remaining descriptors were z-score normalized, fit exclusively on the combined RS and decoy set. Held-out target actives were transformed using the same scaler parameters, ensuring no data leakage.

StandardScaler was applied *only* to Mordred descriptors. Fingerprint representations (see below) were used directly without scaling.

ECFP4 Fingerprints

For comparison, we generated 2048-bit Extended Connectivity Fingerprints with radius 2 (ECFP4) Rogers and Hahn (2010); Landrum (2013). Fingerprints were stored as binary vectors (0/1 for absence/presence) and used without preprocessing.

2.3 Dimensionality Reduction Algorithms

PCA and UMAP models were fit on the combined RS and decoy set (N = 1,484,566 compounds: 191,000 RS + 1,293,566 decoys). Held-out target actives were projected into the learned embedding space using the respective `transform` method, ensuring strict separation between training and test sets. Both methods were used to project the dataset into 2D, 5D, 10D and 20D embedding spaces. Both methods used either the binary fingerprints directly or z-score-normalized Mordred descriptors.

The following key UMAP hyperparameters were optimized in Phase 1:

- **n_neighbors** $\in \{10, 20, 50, 100, 500\}$: Controls local vs. global structure preservation. Lower values emphasize tight local clusters; higher values capture broader manifold geometry.
- **min_dist** $\in \{0.0, 0.01, 0.1\}$: Minimum distance between embedded points. Lower values allow tighter clusters; higher values produce more dispersed embeddings.
- **n_components** $\in \{2, 5, 10, 20\}$: Dimensionality of the target embedding space.
- **metric**: Distance metric for manifold construction. Employed were the `euclidean` metric for Mordred descriptors, and `jaccard` for ECFP4 fingerprints due to their binary bit-vector nature.
- **random_state**: Set to `None` to enable parallel processing across independent replicates and for variance assessment.

To ensure reproducibility, all UMAP runs used five independent replicates, while the deterministic PCA runs are reported as single values. The RS subsampling (Phase 3) used five independent random samples per size, and the cross-target validation (Phase 4) used five replicates per target. Performance metrics are reported as mean ± standard deviation across replicates.

2.4 Scoring and Ranking

After embedding, all compounds (i.e., RS, decoys, and held-out actives) were represented as points in the low-dimensional space. For each held-out active compound, we computed its ranking score as:

$$\text{score}(x_{\text{active}}) = -\min_{x_{\text{RS}} \in \text{RS}} d(x_{\text{active}}, x_{\text{RS}}) \tag{1}$$

where $d(x_{\text{active}}, x_{\text{RS}})$ is the Euclidean distance in the embedded space. The negative sign ensures that compounds *closer* to the RS receive *higher* scores. This exact 1-nearest-neighbor (1-NN) approach was implemented using a kd-tree for computational efficiency with Euclidean metrics. The evaluation set compounds (i.e., target actives and decoys) were ranked by descending score, and enrichment metrics were subsequently computed.

2.5 Computational Implementation

All preprocessing, dimensionality reduction, and scoring were implemented in Python 3.12. SMILES parsing, ECFP4 fingerprint generation Rogers and Hahn (2010), EF@1%, ROC-AUC, and Boltzmann-enhanced discrimination of ROC (BEDROC)

were computed using RDKit 2023.03.1 Landrum et al. (2023). Mordred community edition 2.0.6 Moriwaki et al. (2018) was used for 2D descriptor computation. Data manipulation and numerical computation were done using Pandas Wes McKinney (2010); pandas development team (2020) and NumPy Harris et al. (2020). The scikit-learn 1.3.0 package Pedregosa et al. (2011) was used to compute PCA, StandardScaler, KDTree, imputation and zero-variance filtering. The fast umap-learn 0.5.3 library McInnes et al. (2018); Healy and McInnes (2024) was used for computing UMAP.

2.6 Experimental Design

Evaluation Metrics

While global metrics like ROC-AUC and precision-recall (PR)-AUC assess overall ranking quality, they can fail to capture "early enrichment" – the ability to find true actives in the top $x\%$ of predictions Truchon and Bayly (2007); Kirchmair et al. (2008); Liu et al. (2019); Maia et al. (2020); Sciabola et al. (2022). A high global AUC does not guarantee that actives are positioned early enough for experimental validation Truchon and Bayly (2007). Consequently, we utilize EF@1% and BEDROC to specifically evaluate early enrichment. The EF@1% quantifies how many true actives appear in the top 1% of ranked compounds relative to random selection Truchon and Bayly (2007).

$$EF_{1\%} = \frac{\left(n_{\text{active in top 1\%}}/n_{\text{top 1\%}}\right)}{\left(n_{\text{active}}/N_{\text{total}}\right)} \tag{2}$$

where N_{total} represents the total number of compounds, and n represents the number of compounds in the respective subsets.

It was noted that EF@1% is not sensitive to ranking performance before and after the cutoff threshold Truchon and Bayly (2007). The BEDROC metric was introduced to address this limitation by giving more weight to top-ranked compounds. Consequently, it provides a more robust measure of early enrichment that is less sensitive to a specific, arbitrary cutoff value. It is defined as:

$$\text{BEDROC}(\alpha) = \frac{RIE(\alpha) - RIE_{\min}(\alpha)}{RIE_{\max}(\alpha) - RIE_{\min}(\alpha)} \tag{3}$$

where

$$RIE(\alpha) = \frac{\sum_{i=1}^{n} e^{\frac{-\alpha r_i}{N_{\text{total}}}}}{\left\langle \sum_{i=1}^{n} e^{\frac{-\alpha r_i}{N_{\text{total}}}} \right\rangle_{\text{random}}} = \frac{N_{\text{total}}}{n} \frac{\sum_{i=1}^{n} e^{\frac{-\alpha r_i}{N_{\text{total}}}}}{\frac{1-e^{-\alpha}}{e^{\frac{\alpha}{N_{\text{total}}}}-1}} \tag{4}$$

$$RIE_{\min}(\alpha) = \frac{N_{\text{total}}}{n} \frac{1 - e^{\frac{\alpha n}{N_{\text{total}}}}}{1 - e^{\alpha}} \tag{5}$$

$$RIE_{\max}(\alpha) = \frac{N_{\text{total}}}{n} \frac{1 - e^{\frac{-\alpha n}{N_{\text{total}}}}}{1 - e^{-\alpha}} \tag{6}$$

The α is a tunable exponential weighting factor, r_i is the rank of the i^{th} active compound (from 1 to N_{total}), n is the number of active compounds, and N_{total} is the total number of compounds within the dataset Truchon and Bayly (2007); Riniker

and Landrum (2013). The "early enrichment" factor α controls the weight given to the top-ranked compounds, tuning the metric towards quantifying early enrichment. A higher α places greater importance on compounds at the top of the ranked list. An $\alpha = 20$ is a common assignment where approximately 80% of the BEDROC score is determined by the top 8% of the ranked compounds. An $\alpha \approx 160$ is a highly stringent setting where 80% of the score is derived from just the top 1% of the ranked list.

Stratified Potency Tiers

To assess method-specific biases, we computed separate EF@1% values for high-potency (affinity $<$ 100 nM), medium-potency (100–1000 nM), and weak (1000–10,000 nM) actives. This stratification reveals whether methods preferentially retrieve tight binders vs. weaker inhibitors – a critical distinction for hit-to-lead optimization.

Phase 1 – Hyperparameter Optimization

We performed a systematic grid search across the following methods, representations, and hyperparameters:

- **Dimensionality reduction methods**: PCA and UMAP
- **Molecular representations**: Mordred (1,611 descriptors) and ECFP4 (2048-bit)
- **UMAP hyperparameters**: n_neighbors $\in \{10, 20, 50, 100, 500\}$, min_dist $\in \{0.0, 0.01, 0.1\}$, and n_components $\in \{2, 5, 10, 20\}$
- **PCA hyperparameters**: n_components $\in \{2, 5, 10, 20\}$

This yielded 515 total experimental runs. For each configuration, we computed overall EF@1% and potency-stratified EF@1% values. The optimal configuration was selected as the highest-performing method assessed using the overall EF@1%, with secondary analysis of high-potency-stratified performance.

Phase 2 – Affinity Cutoff Sensitivity (Re-scoring Only)

Using the optimal hyperparameters identified in Phase 1, we investigated if filtering the RS by affinity thresholds improves the enrichment. We re-scored the *same* Phase 1 embeddings (i.e., no model retraining) using RS filtered to include affinities a) $\leq$ 100 nM (i.e., high-potency compounds), b) $\leq$ 1,000 nM, c) $\leq$ 10,000 nM, and d) $\leq$ 100,000 nM. For each cutoff, exact 1-NN distances were recomputed to the filtered RS, and enrichment metrics were re-evaluated. This design isolates the effect of RS potency composition from embedding geometry.

Phase 3 – Reference Set Size Ablation (Full Retraining)

To assess robustness to RS size, we performed the following systematic ablation study:

1. **Random subsampling**: The full RS was randomly subsampled to target sizes: 10, 100, 1,000, 10,000, and 100,000. Subsampling was performed *without replacement* from RS to test generalization capacity across diverse RS scales.
2. **Model retraining**: For each RS subsample, we restandardized the values and retrained the dimensionality reduction model (UMAP or PCA) from scratch using the optimal hyperparameters found in Phase 1 (i.e., UMAP 2D, n_neighbors=10, min_dist=0.0 for Mordred descriptors).

3. **Evaluation**: Held-out actives were projected into each newly trained embedding space and scored via exact 1-NN against the subsampled RS. The stratifying potency enrichment was then evaluated based on the optimal affinity cutoff (i.e., 100 nM).
4. **Replicates**: 5 independent random subsampling and retraining runs were performed per RS size to assess the variance.

This design tests the following hypothesis that larger RS provides better coverage of chemical space, improving retrieval, and, conversely, smaller set sizes degrade performance. The random subsampling approach (rather than curation or active learning) ensures that observed effects reflect *scale alone*, not content quality.

Phase 4 – Cross-Target Generalization

To validate generalizability beyond a single target, we applied the optimal hyperparameters from Phases 1–3 (i.e., see above plus 100 nM affinity cutoff) to eight additional tyrosine transferase targets with varying RS sizes: ABL2 ($\sim$ 43,000), EGFR ($\sim$ 86,000), ERBB2 ($\sim$ 95,000), SRC ($\sim$ 180,000), LCK ($\sim$ 225,000), YES1 ($\sim$ 280,000), FYN ($\sim$ 340,000), and LYN ($\sim$ 425,000 compounds). For each target, we:

1. constructed target-specific RS by excluding compounds with known activity against the held-out target (i.e., leakage prevention),
2. retrained the UMAP model on target-specific RS + ZINC decoys, and
3. evaluated potency-stratified enrichment on held-out actives (5 replicates per target).

We computed Spearman correlation between natural RS size and high-potency EF@1% to quantify the relationship between the RS scaling and the screening performance. This design tests whether the Phase 3 findings (optimal size 10,000, degradation at 100,000+) generalize across targets with naturally occurring RS sizes spanning 1–2 orders of magnitude.

Phase 5 – Ablation and Decoy Bias Study

In this phase, we validated MolFuSE's approach by conducting three control experiments that tested the necessity of dimensionality reduction and the magnitude of decoy bias.

Database Bias Negative Control

To test if MolFuSE exhibits systematic bias toward ChEMBL-like compounds regardless of target specificity, we scored non-transferase actives against transferase-trained models. We selected seven diverse non-transferase target families from Phase 4 and scored their known actives using the respective transferase RS. Under the null hypothesis of no database bias, we expect random performance of EF@1% $\approx$ 1.0. Elevated enrichment would indicate the model preferentially ranks ChEMBL compounds over ZINC compounds, independent of biological relevance. However, critically, if this enrichment remains substantially lower than the target-specific performance, the

databases are only *partially* separable with target family specific models increasing enrichment.

Raw Descriptor Ablation

To determine if UMAP dimensionality reduction is necessary, we scored candidates using the raw 1,611-dimensional Mordred descriptors with scaled Euclidean distance (1-NN). This control (35 replicates across seven Phase 4 target families) directly tests whether high-dimensional descriptor spaces provide adequate discrimination without the need for manifold learning. We compare raw descriptor performance against Phase 4 (descriptors + UMAP, averaged across all targets) using identical descriptors and scaling, isolating UMAP 's contribution. The Phase 4 results represent an average behavior over diverse protein families, and thus provides a generalizable baseline rather than target-specific optimization.

Fingerprint Ablation

To benchmark MolFuSE against established standards, we implemented raw ECFP4 fingerprint similarity using Tanimoto/Jaccard distance (1-NN, no UMAP). We compare against Phase 1 (ECFP4 + UMAP) across five replicates to quantify if UMAP's learned metric provides meaningful improvement over established methods. This control also verifies that MolFuSE's performance gains are not simply due to the use of ECFP4 representations.

The raw descriptor baseline and database bias control experiments used all eight Phase 4 target families (i.e., Antimicrobial, Antioxidant, Cytokine, Heparin-binding, Lyase, Motor Protein, Oxidoreductase, and Transferase; 35 total replicates) to assess generalizability across diverse protein families. The fingerprint baseline experiment (raw ECFP4 vs. Phase 1 ECFP4 + UMAP) used ABL1 transferase (UniProt: P00519) exclusively, with 3,325 ABL1 actives and 1,293,529 ZINC decoys (5 replicates).

For each experiment, we computed enrichment metrics (EF@1% and ROC-AUC) across multiple replicates and performed independent t-tests comparing Phase 5 (no UMAP) against Phase 1/4 baselines (with UMAP). Significance thresholds are set to $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

3 Results

Results are reported in the same order as described in the previous section. It is important to understand what values the metrics would assume in the worst case, where the screening model would produce a uniform-randomly distributed ranking (see Table 1). In that case, the expected value of the first term in Eq. 3 is exactly 1.0. For Phases 1, 2, 3 and 5, the same expected BEDROC metric will result since the same number of actives ($n = 5,505$) and total molecules ($N = 1,718,819$) were used.

3.1 Phase 1 – Hyperparameter Optimization

Performance variation was determined through systematic evaluation of 515 hyperparameter configurations. This evaluation comprehensively covered two dimensionality

Table 1 Expected metric values for a random ranking based on set sizes used in Phases 1–3, 5.

Metric	Expected Value
EF@1%	1.0000
BEDROC ($\alpha = 20$)	0.0516
BEDROC ($\alpha = 160$)	0.0080
ROC-AUC	0.5000
PR-AUC $= R_a$	0.0032

reduction methods (PCA, UMAP) and two molecular representations (Mordred descriptors, ECFP4 fingerprints). The extensive hyperparameter grids revealed substantial performance variation.

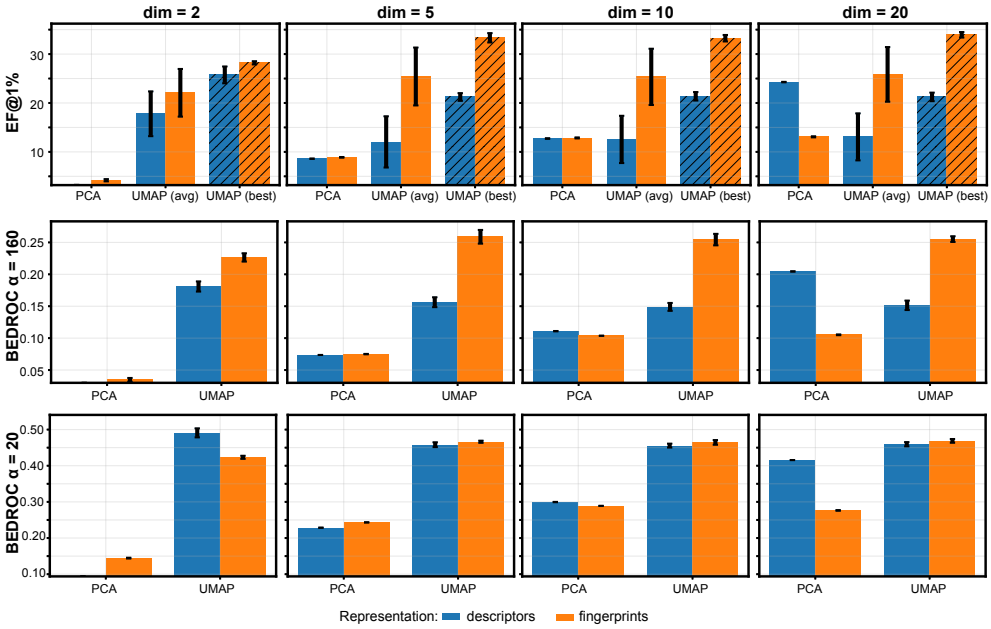


Fig. 1 Phase 1 - Dimensionality-Stratified EF@1% Comparison. EF@1%, BEDROC ($\alpha = 20/160$) across four embedding dimensionalities (2D, 5D, 10D, 20D) for PCA, UMAP-average (mean across all hyperparameter combinations), and UMAP-best (optimal hyperparameters). Blue bars: Mordred descriptors. Orange bars: ECFP4 fingerprints. Error bars represent standard deviation across replicates.

3.1.1 Overall Enrichment Performance Across Dimensionalities

Examination of enrichment performance across embedding dimensionalities shows that ECFP4 fingerprints consistently outperform Mordred descriptors when evaluated on enrichment (Figure 1). The optimal configuration for EF@1% was UMAP with ECFP4 fingerprints (20 dimensions, n_neighbors=10, min_dist=0.0, Jaccard metric), achieving a value of 33.94 ± 0.56. PCA baselines performed substantially worse. The best performance for descriptors was an EF@1% of 24.27 at 2D and for fingerprints an EF@1% of 13.08 at 20D), confirming that non-linear manifold learning is critical for enrichment.

Fingerprints' structural encoding provides a consistent advantage across all dimensionalities for general retrieval. However, as detailed in the subsequent section, this trend starts to invert when specifically targeting high-potency compounds.

3.1.2 Potency-Stratified Analysis

An important shift in performance emerges when evaluating enrichment stratified by compound potency. While fingerprints dominate global metrics, Mordred descriptors coupled with UMAP achieve better enrichment for high-potency actives (affinity $<$ 100 nM), at least relative to the mean enrichment over the entire set.

As shown in Figure 2, the BEDROC ($\alpha = 20$) scores for Mordred descriptors on high-potency targets significantly exceed those of medium and weak potency tiers. Conversely, ECFP4 fingerprints show better performance for the medium potency tier. This suggests that while fingerprints effectively retrieve topologically similar scaffolds, the physicochemical properties encoded by Mordred descriptors are better suited for clustering the tightest binders in the embedding space. However, fingerprints still seem to outperform descriptor-based UMAP similarity spaces.

3.1.3 Optimal Hyperparameter Landscape

Analysis of the 515-run grid (Figure 3) revealed clear topological preferences for the embedding. Low values for the number of neighbors (`n_neighbors`=10–20) consistently performed best, indicating that preserving local neighborhood structures is more critical for 1-NN retrieval than preserving the global manifold geometry. Furthermore, tight clustering parameters (`min_dist`=0.0–0.01) outperformed dispersed embeddings (`min_dist`=0.1), suggesting that dense packing of the RS aids in creating discriminative decision boundaries for the 1-NN scorer.

3.2 Phase 2 – Affinity Cutoff Sensitivity Reveals Optimal Scoring Against High-Potency Reference Set

Using the optimal embeddings from Phase 1, we investigated the impact of RS quality on retrieval performance. We rescored held-out actives against subsets of the RS, filtered by varying affinity thresholds (100 nM to 100,000 nM).

Figure 4 shows a strong negative correlation between the inclusion of weak binders (affinity $>$ 1 μM) into RS and enrichment performance. Including weak binders creates "noise" in the embedding space, allowing decoys to score well by appearing similar

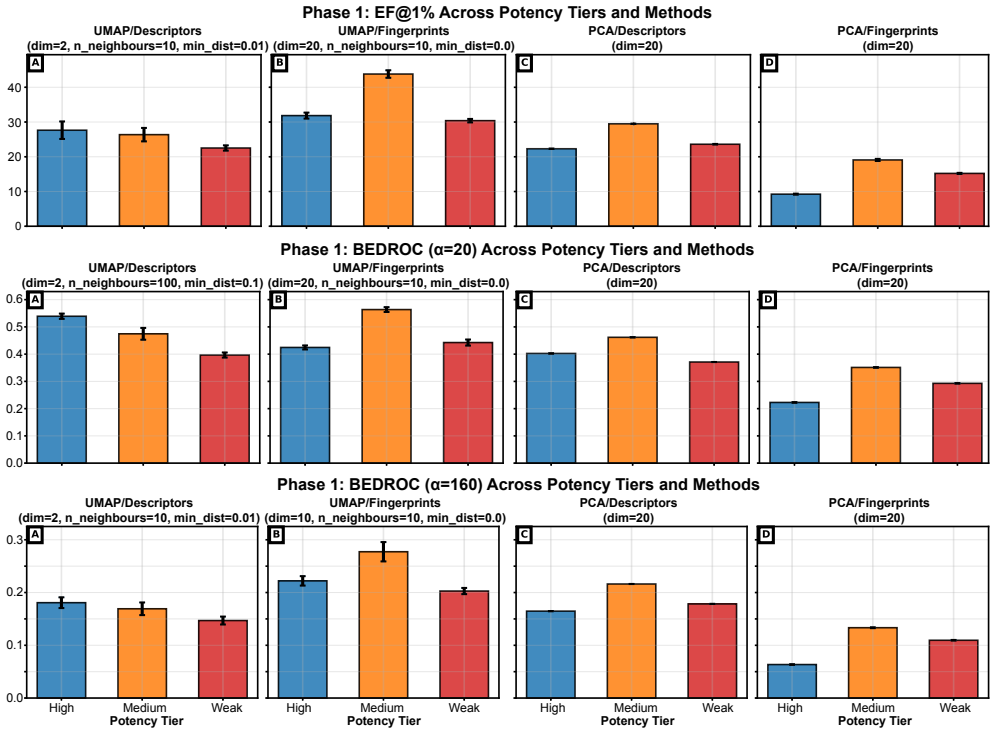


Fig. 2 Phase 1 - Potency-Stratified Enrichment. EF@1%, BEDROC ($\alpha = 20/160$) for best-performing configurations across four method combinations, stratified by active potency tier (High: affinity < 100 nM, Medium: 100–1000 nM, Weak: 1000–10,000 nM). Error bars represent standard deviation across replicates.

to weak, non-specific binders. The same effect is also clearly seen in Figure 5, where the performance for the different potency tiers of the target compounds is plotted. In virtual screening, we are more interested in retrieving strong binders (blue line) than weak binders (red line). Descriptor-based UMAP similarity spaces outperform fingerprint-based similarity spaces in terms of EF@1% and BEDROC ($\alpha = 20$) scores.

One can also note that the choice of metric leads to different choices for UMAP, either for descriptors or fingerprints. Filtering the RS to include only high-potency compounds (< 100 nM) resulted in the highest enrichment across all metrics for the UMAP methods. Consequently, we adopted the 100 nM affinity cutoff as the default for subsequent experiments to maximize the signal-to-noise ratio in the reference manifold. PCA on descriptors is influenced the least by changing the filtering of RS and PCA on fingerprints is generally the worst performer. We therefore choose to ignore PCA in the rest of the experiments.

A key observation is that as the affinity cutoff becomes stricter, EF@1% and BEDROC scores increase dramatically for UMAP, and descriptor-based UMAP similarity spaces outperform fingerprint-based similarity spaces. This proves that the

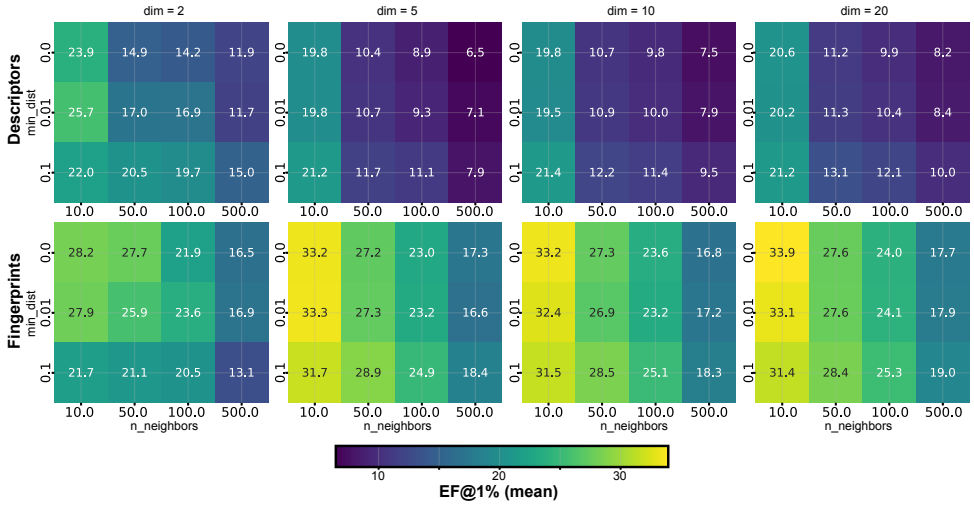


Fig. 3 Phase 1 - Optimal Hyperparameters and Sensitivity. EF@1% values for all UMAP hyperparameter configurations. Low values for the number of neighbors (n neighbors=10–20) consistently performed best.

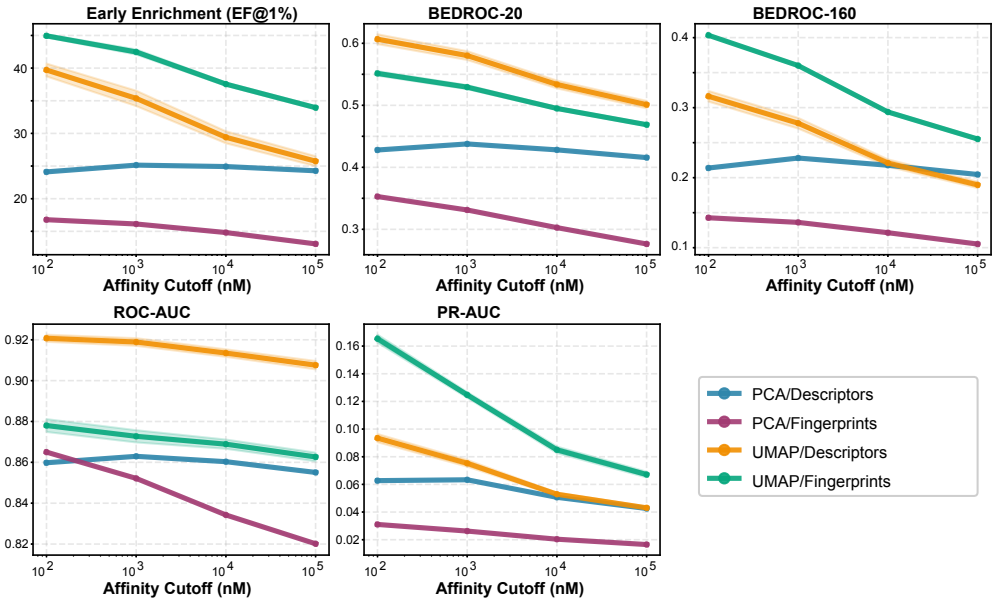


Fig. 4 Phase 2 - Affinity cutoff sensitivity for all metrics and methods, shaded area shows 95% confidence interval.

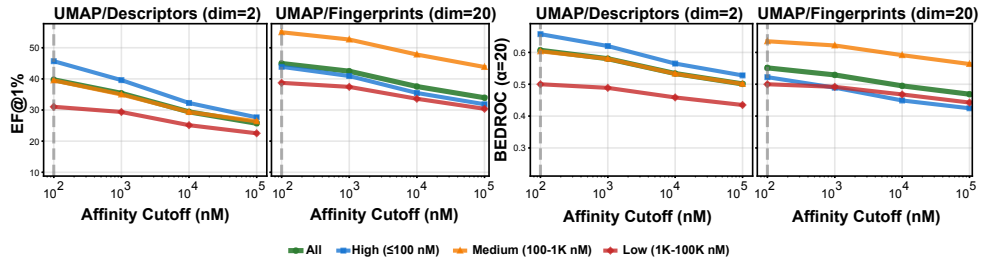


Fig. 5 Phase 2 - Affinity cutoff sensitivity for stratified potency sets for all metrics and methods.

molecular-function-based RS is not just a random background, but is a functional reference.

3.3 Phase 3 – Reference Set Size Ablation Reveals Robustness and Critical Point

To test the limits of the "low-data" assumption, we systematically subsampled the RS from 10 to 100,000 (about 50,000 fewer than the original data set) compounds and retrained the models. Figure 6 illustrates the relationship between RS size and enrichment metrics, revealing a striking non-monotonic trajectory. At all RS sizes, Mordred descriptors outperformed ECFP4 fingerprints for high-potency retrieval, reinforcing the conclusion that rich physicochemical descriptor spaces better capture affinity-relevant features than structural fingerprints. These trends are consistent across all metrics employed.

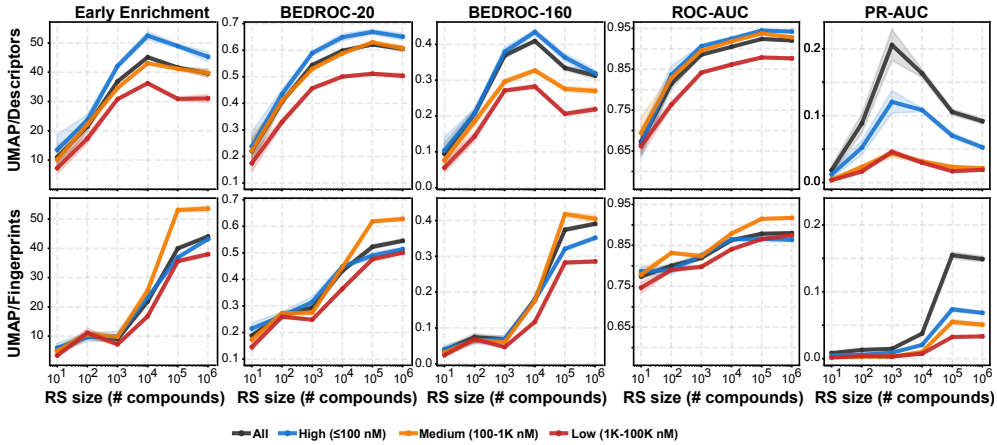


Fig. 6 Phase 3 – Sensitivity to RS size for all potency tiers, all metrics and methods. Metric values are plotted on y-axes.

Robustness in Low-Data Regimes: MolFuSE demonstrates remarkable efficiency with small datasets. Using UMAP on the descriptors of a small RS of just 1,000 compounds achieves an EF@1% of 41.77 ± 1.38. This strengthens the method's applicability to orphan targets where limited bioactivity data is available.

However, contrary to the assumption that "more data is better", when applying UMAP on descriptors, performance peaks at approximately 10,000 compounds (EF@1% 53.62 ± 1.70) and significantly degrades as the set size increases to 100,000+ compounds. This degradation (Figure 6, right columns) suggests a descriptor space dilution effect. As the RS grows to encompass the full range of transferase inhibitor space, the UMAP embedding must accommodate increasingly diverse molecular structures. This forces the manifold to change in ways that disperse the high-potency clusters, making the 1-NN distance less discriminative. This finding highlights a critical trade-off between chemical space coverage and embedding specificity. Consequently, curating RS to exclude very weak binders (e.g., affinity > 10 μM) may improve performance, even if it reduces the total compound count.

3.4 Phase 4 – Cross-Target Generalization

To ensure that the results are not artifacts of the ABL1 target, we applied the optimal configuration (i.e., UMAP 2D, Mordred, n_neighbors=10, min_dist=0.01, 100 nM cutoff) to eight additional tyrosine transferases with varying natural RS sizes (43,000 to 425,000 compounds). For each target protein's data set, we trained a UMAP model. Figure 7 summarizes the cross-target performance. Phase 3 overestimated the generalization capability on smaller RS; however, the linear trend of the correlation between RS size and enrichment is clearly visible. Furthermore, the models were retrained without rehyperparameterizing UMAP for the other targets since this is a very computationally expensive operation.

Two additional notable observations include:

1. **Consistency:** The trend line (dashed red) indicates a positive correlation between available data and performance, although the gains plateau, mirroring the saturation effects seen in Phase 3.
2. **Metric Stability:** The method maintains high early enrichment (EF@1%) across diverse transferase sub-families, validating the generalizability of the manifold learning approach.

These results confirm that the parameters optimized on ABL1 transfer effectively to related protein targets, supporting the utility of MolFuSE as a algorithmic tool for protein family-wide screening.

3.5 Phase 5 – Ablation and Decoy Bias Study

Database Bias Negative Control

Table 2 shows that non-transferase actives scored against transferase-trained models achieved EF@1% = 13.5 ± 11.2 and ROC-AUC = 0.655 ± 0.107 across seven target families. This substantially exceeds random performance (EF@1% = 1.0, Table 1),

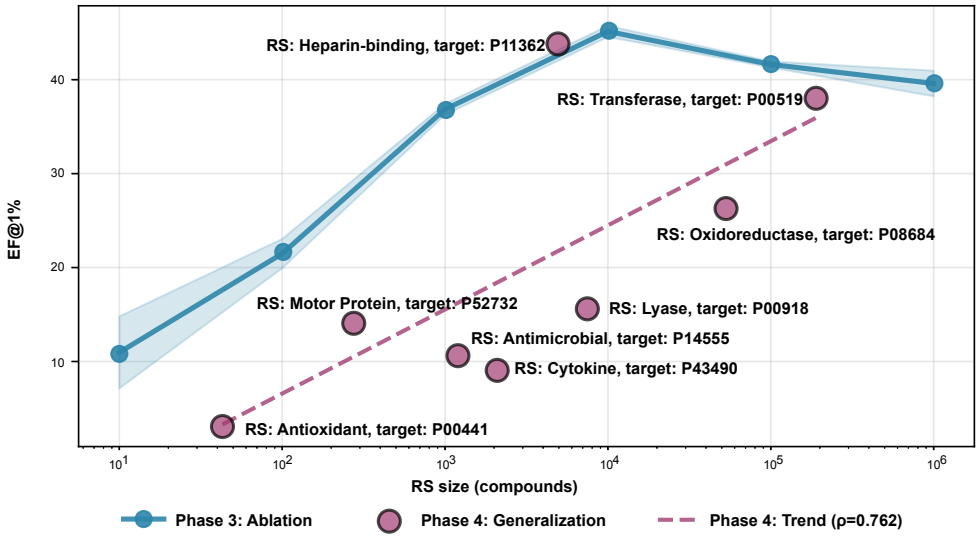


Fig. 7 Phase 4 - Generalization performance (EF@1%) of hyperparameters used in Phase 3 on other RS and target proteins.

Table 2 Validation & Baseline Experiments vs Phase 1/4 Baselines.

Method	EF@1%	ROC-AUC	p-value
Database Bias Control			
Non-kinase actives vs kinase model	13.5 ± 11.2	0.655 ± 0.107	N/A
Descriptor Comparison (Raw Descriptors)			
Phase 5 (raw descriptors, NO UMAP)	7.8 ± 9.1	0.504 ± 0.121	–
Phase 4 (descriptors+UMAP)	20.7 ± 14.7	0.729 ± 0.130	$< 0.001^{***}$
Fingerprint Comparison (Tanimoto)			
Phase 5 (raw ECFP4, NO UMAP)	30.1 ± 0.0	0.961 ± 0.000	–
Phase 1 (ECFP4+UMAP)	33.9 ± 0.5	0.863 ± 0.004	$< 0.001^{***}$

confirming the presence of database bias: the model preferentially ranks ChEMBL compounds over ZINC decoys independent of target-specific biological relevance.

This bias likely arises from two sources: (1) ChEMBL compounds are drug-like compounds with measured bioactivity, exhibiting restricted physicochemical properties (e.g., Lipinski's rule of five), while ZINC contains broader chemical diversity that includes non-drug-like structures; (2) RS itself is derived from ChEMBL, creating implicit training set homogeneity. Future work should explore alternative decoy sets (e.g., property-matched DUD-E decoys) to isolate target-specific signals more cleanly, though such matched sets introduce their own biases.

However, it is important to note that this bias effect (EF@1% = 13.5) is approximately 2.5-fold *lower* than target-specific performance (EF@1% = 30-34 for matched

transferase targets). This quantitative difference demonstrates that the ChEMBL and ZINC databases are only *partially* separable, not fully trivially distinguishable. Target-specific biological signals clearly dominate MolFuSE's predictions. The model is not merely learning to distinguish "ChEMBL-like" from "ZINC-like" compounds, but instead capturing genuine protein-ligand interaction patterns superimposed on a non-negligible background of database bias.

Raw Descriptor Ablation

Using the raw 1,611-dimensional descriptors without UMAP achieved an EF@1% of 7.8 ± 9.1 and ROC-AUC of 0.504 ± 0.121 across the 7 target families (Table 2). In contrast, the descriptors + UMAP achieved an average EF@1% = 20.7 ± 14.7 and ROC-AUC = 0.729 ± 0.130 over the same targets and descriptors ($p < 0.001$, 166% relative improvement). These results reflect average behavior across all target families, capturing UMAP 's generalizable benefit rather than target-specific tuning. This improved performance demonstrates that UMAP is not merely a computational convenience but is necessary for effective descriptor-based screening.

The poor performance of using only high-dimensional descriptors can be attributed to the curse of dimensionality: in 1,611-dimensional space, the Euclidean distance metrics become uninformative as all pairwise distances converge Beyer et al. (1999). The nearest-neighbor search in such spaces approaches random selection because the ratio of nearest-to-farthest neighbor distances tends toward 1.0. UMAP addresses this by learning a low-dimensional manifold that preserves local neighborhood structures, effectively denoising the descriptor space and recovering biologically meaningful similarity relationships. Without this manifold learning step, the high-dimensional descriptor space is notably less useful in ligand-based screening.

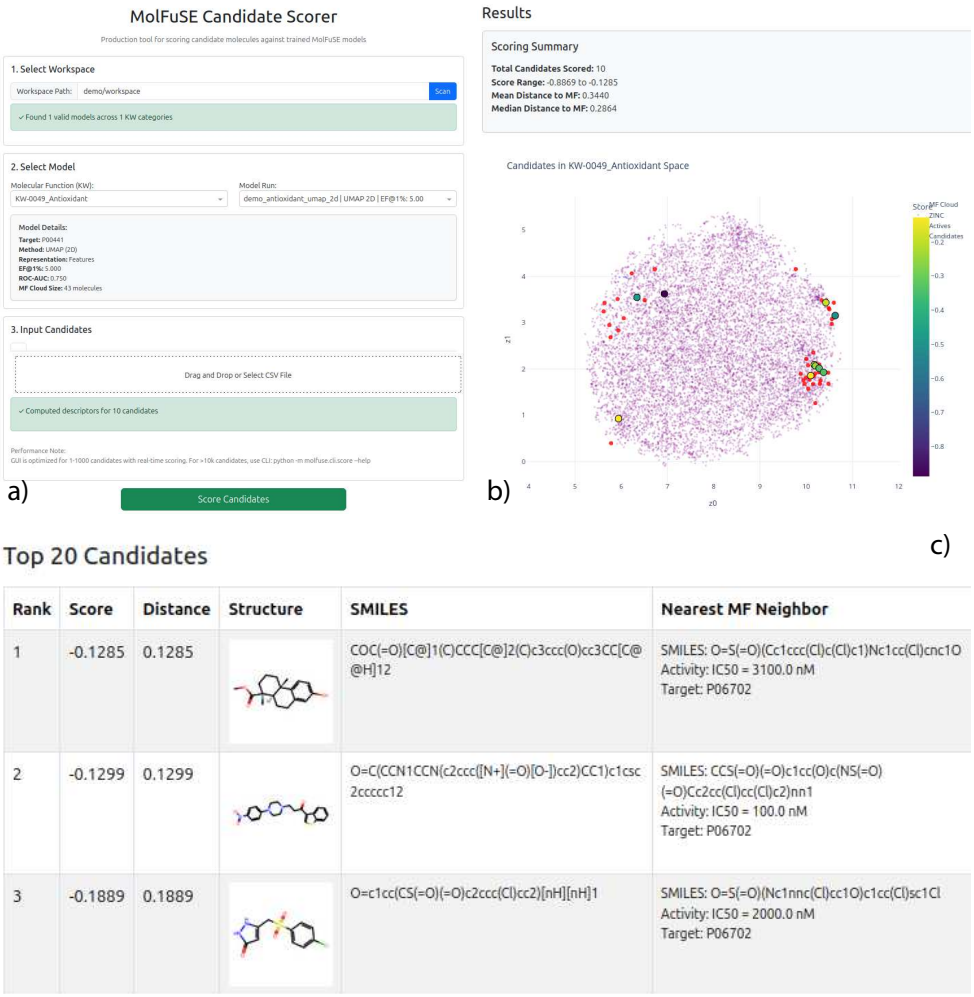
Fingerprint Ablation

Raw ECFP4 fingerprints with Tanimoto similarity (i.e., without UMAP) achieved EF@1% = 30.1 ± 0.0 and ROC-AUC = 0.961 ± 0.000 (Table 2). This represents strong performance and establishes a high baseline for comparison. Completion of Phase 1 (ECFP4 + UMAP) achieved an EF@1% = 33.9 ± 0.5 ($p < 0.001$), an 11% improvement, and a ROC-AUC = 0.863 ± 0.004. Comparison of these metric values shows an unusual pattern: raw ECFP4 has higher global ranking quality (ΔROC-AUC = +0.1; +10%). In contrast, ECFP4 + UMAP has superior early enrichment ($\Delta EF@1\%$ = +3.8; +11%). This indicates UMAP prioritizes top-ranked predictions – the most critical region for practical virtual screening where experimental validation costs dominate. The trade-off is acceptable: a 10% reduction in global AUC is offset by an 11% gain in the actionable top 1% of predictions.

3.6 MolFuSE Explorer

Akin to tools like the SwissSimilarity web tool Zoete et al. (2016), which predicts the targets of bioactive compounds based on a combination of 2D and 3D similarity measures and linear projections with known ligands, we introduce a web application, shown in Figure 8, designed for interactive MolFuSE navigation and analysis of the

RS-guided similarity spaces. This goes beyond applications such as VSFlow Jung et al. (2023) that do not involve dimensionality reduction techniques.



Top 20 Candidates

Rank	Score	Distance	Structure	SMILES	Nearest MF Neighbor
1	-0.1285	0.1285		COC(=O)[C@]1(C)CCC[C@]2(C)c3ccc(O)cc3CC[C@@H]12	SMILES: O=S(=O)(Cc1ccc(Cl)c(Cl)c1)Nc1cc(Cl)cnc1O Activity: IC50 = 3100.0 nM Target: P06702
2	-0.1299	0.1299		O=C(CCN1CCN(c2ccc([N+](=O)[O-])cc2)CC1)c1csc2ccccc12	SMILES: CCS(=O)(=O)c1cc(O)c(NS(=O)(=O)Cc2cc(Cl)cc(Cl)c2)nn1 Activity: IC50 = 100.0 nM Target: P06702
3	-0.1889	0.1889		O=c1cc(CS(=O)(=O)c2ccc(Cl)cc2)[nH][nH]1	SMILES: O=S(=O)(Nc1nnc(Cl)cc1O)c1cc(Cl)sc1Cl Activity: IC50 = 2000.0 nM Target: P06702

Fig. 8 The MolFuSE Explorer. a) User selects workspace, model and input candidates. b) The resulting similarity space is calculated by the backend. c) Candidates are scored, and top candidates are revealed.

The primary interface is an interactive 2D scatter plot that visualizes the selected space. Points are color-coded to distinguish between MF-actives, ZINC decoys, and user-uploaded compounds. Users can dynamically filter the displayed data based on activity thresholds and other properties. A key capability is the projection of user-uploaded custom compounds into the current RS-guided space using a trained UMAP model. This allows users to assess their novel compounds in the context of known

RS-relevant actives and the broader purchasable chemical space. To aid candidate identification, a 1-Nearest Neighbors (1-NN) search can be performed within the 2D space to find the most similar ZINC or custom compounds to the aggregated set of ChEMBL actives.

4 Discussion

This study demonstrates that in low-data virtual screening regimes, high-performing prediction models can be developed using manifolds learned from homologous protein data. Our findings regarding potency stratification, data saturation, and embedding dimensionality suggest that MolFuSE operates best through a similarity-based scoring mechanism governed by physicochemical compatibility (using Mordred descriptors) rather than pure substructural analogy ECFP4 fingerprints.

Physicochemical Descriptors Drive High-Potency Early Enrichment

A central finding in the hyperparameter and representation analysis (Phase 1) is the performance divergence between topological fingerprints (ECFP4) and physicochemical descriptors (Mordred) when evaluated against high-potency targets. While fingerprints generally offer robust global retrieval – evidenced by high ROC-AUC scores in baseline comparisons – descriptor-based UMAP embeddings demonstrated a specific advantage in concentrating sub-100 nM binders at the very top of the active rankings. This advantage, reflected in the higher high-potency discrimination of descriptor-based UMAP embeddings at optimized RS sizes, suggests that tight binding is driven by specific physicochemical property profiles that are explicitly encoded by Mordred descriptors, whereas fingerprints excel at broader structural retrieval. While fingerprints excel at identifying structural analogs (scaffold hopping), descriptor-based manifold learning appears to filter the potency better, resulting in clustering compounds with the requisite physical properties for tight binding regardless of their scaffold similarity. This distinction supports the use of descriptor-based embeddings specifically for the early recognition problem, where limited experimental data hinders the prediction of top-ranked candidates.

Descriptor Space Dilution and the Limits of Data Scaling

Contrary to the prevailing big data paradigm in deep learning, where performance typically scales with dataset size, our reference set (RS) ablation studies (Phase 3) identified a critical performance ceiling. The observation that enrichment peaks at an RS size of approximately 10,000 compounds and significantly degrades as the set expands to 100,000+ provides a nuanced view of the "Similar Property Principle" Shortridge et al. (2011) in low-dimensional embeddings. We attribute this degradation to descriptor space dilution. As the RS grows to encompass the historical diversity of a protein family's active compounds, it inevitably incorporates a wider array of distinct molecular structures and weaker binders. In the context of a fixed-dimensional UMAP embedding, preserving the global structure of this heterogeneous data forces the manifold to expand, dispersing the dense local clusters of high-potency compounds required for effective 1-NN retrieval. This is further supported by Phase

2 results, where excluding weak binders ($\geq 100nm$) from the scoring set improved performance. Consequently, for embedding-based similarity search, the quality and physicochemical specificity of the reference data appear more critical than volume. This finding is particularly relevant for "low-data" targets, as it implies that massive historical datasets are not required if a smaller, curated set of highly specific ligands is available.

Dimensionality Reduction: Necessity and Trade-offs

The ablation study clarifies the role of UMAP not merely as a visualization tool, but as a necessary pre-processing step for physicochemical screening. The failure of raw high-dimensional descriptors (EF@1% ≈ 7.8) compared to their UMAP-projected counterparts (EF@1% ≈ 20.7) confirms that the "curse of dimensionality" renders Euclidean metrics ineffective in the raw 1,611-dimensional space. UMAP mitigates this by learning a manifold that preserves local neighborhood structures, effectively denoising the descriptor space. However, this gain in early enrichment comes at the cost of global ranking quality. While the ECFP4-based MolFuSE workflow improves early enrichment EF@1% by 11% over the raw Tanimoto baseline, the descriptor-based approach provides more actionable prioritization of high-potency hits at the cost of a lower global ROC-AUC. This trade-off is consistent with UMAP's algorithmic focus on local versus global structure preservation. For practical virtual screening, where only the top 0.1% to 1% of a library is typically screened, optimizing for EF@1% and BEDROC is operationally more valuable than maintaining a correct ranking order among non-active decoys.

Database Bias and Generalization

Finally, the detection of a moderate baseline preference for ChEMBL compounds over ZINC decoys underscores the challenge of database bias in retrospective validation. While the target-specific biological signal was found to be approximately 2.5-fold stronger than this background bias, absolute performance metrics must be interpreted with this inflation in mind. Despite this, the consistent performance across eight diverse tyrosine transferase targets (Phase 4) demonstrates that the method generalizes beyond the optimization target (ABL1), with performance dropping with the size of RS. It should be noted that the cross-target validation was performed without rehyperparameterizing UMAP for each specific protein family; thus, the reported results likely represent a conservative lower bound of the method's potential performance when tuned for specific target families.

The Phase 5 controls establish three key findings:

1. **Partial database bias exists but does not dominate.** MolFuSE exhibits moderate ChEMBL-over-ZINC enrichment for irrelevant targets (EF@1% = 13.5), confirming database provenance introduces bias. However, this effect is 2.5-fold weaker than target-specific performance (EF@1% = 30-34), demonstrating that biological signals dominate over database artifacts. Absolute performance metrics should be interpreted with this bias acknowledged.
2. **UMAP is essential for descriptor-based screening.** Compared to the 1,611-dimensional raw descriptors (EF@1% = 7.8), UMAP embeddings increase

performance by 166% (EF@1% = 20.7, $p < 0.001$). The curse of dimensionality renders high-dimensional Euclidean spaces ineffective for nearest-neighbor search; manifold learning is necessary, not optional. Notably, Phase 4 UMAP results represent average cross-target performance, suggesting this benefit generalizes across protein families.

3. **MolFuSE modestly improves upon industry-standard Tanimoto.**
   Raw ECFP4 fingerprints with Tanimoto similarity provide a strong baseline (EF@1% = 30.1), which UMAP integration improves by 12.6% (EF@1% = 33.9, $p < 0.001$). The gain is concentrated in early enrichment (top 1%), the most critical region for experimental validation. While UMAP introduces a trade-off by reducing global ROC-AUC (0.863 vs. 0.961), it notably improves actionable, top-ranked predictions, representing a good compromise for practical screening applications.

These results validate MolFuSE's core hypothesis, that manifold learning enhances ligand-based virtual screening, while honestly characterizing limitations and establishing that simpler baselines (Tanimoto) remain competitive.

5 Conclusions

Herein, we have presented MolFuSE, a projection-based virtual screening framework that leverages a reference set (RS) of ligands that bind proteins with the same molecular function as the target protein. By leveraging UMAP dimensionality reduction, this method constructs a lower-dimensional embedding space that significantly enhances early recognition performance, achieving state-of-the-art enrichment of high-potency compounds in low-data regimes. Through systematic evaluation across four experimental phases, we demonstrate that this model has:

1. **High-potency discrimination**: UMAP embeddings of Mordred 2D molecular descriptors' explicit encoding of binding-relevant properties (LogP, hydrogen bonding, electronic distribution) substantially outperforms topology-focused ECFP4 fingerprint-based methods in the early recognition (EF@1%) of tight binders, although topological fingerprints maintain better global ranking quality (ROC-AUC).
2. **RS size vs. performance**: Effective screening is achievable with RS as small as 1,000 compounds, enabling application to orphan targets and understudied protein families. Performance peaks at 10,000 compounds and degrades at 100,000+, suggesting that feature space dilution from structurally diverse weak binders impairs retrieval – challenging the assumption that larger RS are universally better. Instead, MolFuSE refines the 'similar property principle' Shortridge et al. (2011) by demonstrating that for high-potency retrieval, the quality and physicochemical specificity of RS are more critical than its size.

MolFuSE addresses an unmet need in early-stage drug discovery: effective virtual screening for targets with limited SAR data. By transferring knowledge from related protein families via RS, MolFuSE achieves performance competitive with or exceeding resource-intensive docking methods, while maintaining computational efficiency

suitable for ultra-large library screening (10^7–10^9 compounds). The method's interpretability – via exact 1-NN scoring and traceable similarity paths – further enhances its utility for medicinal chemistry hit triaging.

5.1 Limitations and Future Directions

While MolFuSE demonstrates robust performance, several avenues for refinement remain. First, UMAP hyperparameters were optimized to a single protein target and then held constant across all targets to ensure a standardized benchmark. Consequently, target-specific rehyperparameterization could yield improved results. While we also identified a bias in the decoy dataset, its impact on MolFuSE's performance was minimal.

A primary limitation is the reliance on Mordred's 2D descriptors:

- **Shape and Stereochemistry**: The current model does not explicitly encode 3D volume/curvature, shape moments, or chiral centers (i.e., enantiomers may receive identical embeddings).
- **Conformational Flexibility**: Conformer preferences and flexibility are currently unrepresented.

All results in this study are retrospective. While high enrichment of ChEMBL-sourced actives is valuable, the ultimate utility of MolFuSE must be confirmed through **prospective experimental validation**. Future efforts will focus on the synthesis and testing of novel scaffolds to confirm that top-tier enrichment translates into viable experimental hit rates ($> 10\%$) and successful lead optimization campaigns. In conclusion, by leveraging non-linear manifold learning, MolFuSE offers a beneficial path to accelerate hit discovery in data-constrained virtual screening workflows, particularly for under-explored therapeutic targets.

Supplementary information. Please find an extended version of the related work and a full listing of the molecular descriptors in the SI.

Declarations

Availability of data and materials

The complete computational pipeline (configuration generators, HPC batch scripts, post-processing analysis tools) is available at: https://github.com/m3-center/molfuse.git. The dataset and an example of a good UMAP model can be found at https://zenodo.org/records/20325360

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partially funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia, grant FKZ 005-2302-0023, and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 528562393 - FIP 26. All calculations were performed on the high-performance computing cluster of the Bonn-Rhein-Sieg University of Applied Sciences (BRSU), Sankt Augustin, Germany. BRSU's computer hardware was partially supported by the Federal Ministry for Education and Research, and by the Ministry for Innovation, Science, Research, and Technology of the state North Rhine-Westphalia (research grant 13FH156IN6).

Authors' contributions

A.H.: data curation, conceptualization, investigation, methodology, formal analysis, visualization, writing-original draft preparation, writing-review and editing D.R.: funding acquisition, supervision, writing-review and editing M.P.: funding acquisition, supervision, writing-review and editing K.K.: data curation, validation, formal analysis, supervision, writing-original draft preparation, writing-review and editing

Acknowledgements

The authors thank Rudolf Berrendorf and Javed Razzaq for their continuous development and support of our high performance cluster.

List of abbreviations

1-NN 1-nearest-neighbor. 8, 10–12, 14, 18, 22, 25

AUC area under the curve. 2, 8, 9, 12, 13, 18–20, 22–24

BEDROC Boltzmann-enhanced discrimination of ROC. 8–10, 12–15, 23

ECFP4 extended-connectivity fingerprints. 2, 3, 5, 7, 8, 10, 12–14, 17, 19, 20, 22–24
EF@1% enrichment factor at 1%. 2, 3, 8–16, 18–20, 23, 24

LBVS ligand-based virtual screening. 3

MF molecular function. 4

PCA principal component analysis. 4, 5, 8–10, 13–15
PR precision-recall. 9, 13

ROC receiver operating characteristic. 2, 8, 9, 12, 13, 18–20, 22–24
RS reference set. 4–8, 10, 11, 14, 15, 17–19, 21–24

SAR structure-activity relationship. 1, 2, 24
SBVS structure-based virtual screening. 3

UMAP uniform manifold approximation and projection. 1, 2, 4–6, 8–16, 18–25

References

Appelt K, Bacquet RJ, Bartlett CA, et al (1991) Design of enzyme inhibitors using iterative protein crystallographic analysis. Journal of medicinal chemistry 34(7):1925–1934

Becht E, McInnes L, Healy J, et al (2019) Dimensionality reduction for visualizing single-cell data using umap. Nature biotechnology 37(1):38–44

Bellman R (1966) Dynamic programming. science 153(3731):34–37

Bennett RS (1965) Representation and analysis of signals part xxi. the intrinsic dimensionality of signal collections. Tech. rep.

Beyer K, Goldstein J, Ramakrishnan R, et al (1999) When is "nearest neighbor" meaningful? In: International conference on database theory, Springer, pp 217–235

Carlsson J, Luttens A (2024) Structure-based virtual screening of vast chemical space as a starting point for drug discovery. Current opinion in structural biology 87:102829

Chithrananda S, Grand G, Ramsundar B (2020) Chemberta: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:201009885

Chowdhury IJ, Hossan MI, Khan MTR, et al (2025) Revolutionizing drug discovery: A systematic review of ai and machine learning application. In: 2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), IEEE, pp 1814–1821

Fabian B, Edlich T, Gaspar H, et al (2020) Molecular representation learning with language models and domain-relevant auxiliary tasks. arXiv preprint arXiv:201113230

Fischer E (1894) Einfluss der configuration auf die wirkung der enzyme. Berichte der deutschen chemischen Gesellschaft 27(3):2985–2993

Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and qsar modeling research. Journal of chemical information and modeling 50(7):1189

Gao M, Skolnick J (2013) A comprehensive survey of small-molecule binding pockets in proteins. PLOS Computational Biology 9(10):e1003302. https://doi.org/10.1371/journal.pcbi.1003302, URL https://doi.org/10.1371/journal.pcbi.1003302

Gaulton A, Bellis LJ, Bento AP, et al (2012) Chembl: a large-scale bioactivity database for drug discovery. Nucleic acids research 40(D1):D1100–D1107

Ghislat G, Rahman T, Ballester PJ (2021) Recent progress on the prospective application of machine learning to structure-based virtual screening. Current opinion in

chemical biology 65:28–34

Harris CR, Millman KJ, van der Walt SJ, et al (2020) Array programming with numpy. Nature 585(7825):357–362. URL https://doi.org/10.1038/s41586-020-2649-2

Healy J, McInnes L (2024) Uniform manifold approximation and projection. Nature Reviews Methods Primers 4(1):82. URL https://doi.org/10.1038/s43586-024-00363-x

Irwin JJ, Shoichet BK (2005) Zinc- a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling 45(1):177–182

Jacob L, Vert JP (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. bioinformatics 24(19):2149–2156

Jobe A, Vijayan R (2024) Orphan g protein-coupled receptors: the ongoing search for a home. Frontiers in Pharmacology Volume 15 - 2024. URL https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2024.1349097

Johnson MA, Maggiora GM, et al (1990) Concepts and applications of molecular similarity. Wiley

Jung S, Vatheuer H, Czodrowski P (2023) Vsflow: an open-source ligand-based virtual screening tool. Journal of Cheminformatics 15(1):40

Kirchmair J, Markt P, Distinto S, et al (2008) Evaluation of the performance of 3d virtual screening protocols: Rmsd comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes? Journal of computer-aided molecular design 22(3):213–228

Kitchen DB, Decornez H, Furr JR, et al (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nature reviews Drug discovery 3(11):935–949

Klabunde T (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. British journal of pharmacology 152(1):5–7

Kobak D, Berens P (2019) The art of using t-sne for single-cell transcriptomics. Nature communications 10(1):5416

Kumar N, Acharya V (2024) Advances in machine intelligence-driven virtual screening approaches for big-data. Medicinal Research Reviews 44(3):939–974

Landrum G (2013) Rdkit documentation. Release 1(1-79):4

Landrum G, Tosco P, Kelley B, et al (2023) rdkit/rdkit: 2023_03_1 (q1 2023) release. https://doi.org/10.5281/zenodo.7880616, URL https://doi.org/10.5281/

zenodo.7880616

Lionta E, Spyrou G, K Vassilatis D, et al (2014) Structure-based virtual screening for drug discovery: principles, applications and recent advances. Current topics in medicinal chemistry 14(16):1923–1938

Liu R, AbdulHameed MDM, Wallqvist A (2019) Teaching an old dog new tricks: Strategies that improve early recognition in similarity-based virtual screening. Frontiers in Chemistry Volume 7 - 2019. URL https://www.frontiersin.org/journals/chemistry/articles/10.3389/fchem.2019.00701

Liu Y, Moretti R, Wang Y, et al (2025) Advancements in ligand-based virtual screening through the synergistic integration of graph neural networks and expert-crafted descriptors. Journal of Chemical Information and Modeling 65(10):4898–4905

Lovrić M, DJuričić T, Tran HT, et al (2021) Should we embed in chemistry? a comparison of unsupervised transfer learning with pca, umap, and vae on molecular fingerprints. Pharmaceuticals 14(8):758

Maia EHB, Assis LC, De Oliveira TA, et al (2020) Structure-based virtual screening: from classical to artificial intelligence. Frontiers in chemistry 8:343

McInnes L, Healy J, Saul N, et al (2018) Umap: Uniform manifold approximation and projection. Journal of Open Source Software 3(29)

Wes McKinney (2010) Data Structures for Statistical Computing in Python. In: St'efan van der Walt, Jarrod Millman (eds) Proceedings of the 9th Python in Science Conference, pp 56 – 61, https://doi.org/10.25080/Majora-92bf1922-00a

Medina-Franco JL, Martínez-Mayorga K, Giulianotti MA, et al (2008) Visualization of the chemical space in drug discovery. Current Computer-Aided Drug Design 4(4):322–333

Moriwaki H, Tian YS, Kawashita N, et al (2018) Mordred: a molecular descriptor calculator. Journal of Cheminformatics 10(1):4. URL https://doi.org/10.1186/s13321-018-0258-y

Ni B, Wang H, Khalaf HKS, et al (2024) Autodock-ss: autodock for multiconformational ligand-based virtual screening. Journal of Chemical Information and Modeling 64(9):3779–3789

Oprea TI, Gottfries J (2001) Chemography: the art of navigating in chemical space. Journal of combinatorial chemistry 3(2):157–166

Orlov AA, Akhmetshin TN, Horvath D, et al (2025) From high dimensions to human insight: Exploring dimensionality reduction for chemical space visualization. Molecular Informatics 44(1):e202400265

Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830

Ramsundar B, Eastman P, Walters P, et al (2019) Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more. O'Reilly Media

Riniker S, Landrum GA (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. Journal of cheminformatics 5(1):26

Rogers D, Hahn M (2010) Extended-connectivity fingerprints. Journal of chemical information and modeling 50(5):742–754

Schuffenhauer A, Floersheim P, Acklin P, et al (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. Journal of chemical information and computer sciences 43(2):391–405

Sciabola S, Torella R, Nagata A, et al (2022) Critical assessment of state-of-the-art ligand-based virtual screening methods. Mol Inf 41(11):2200103. URL https://doi.org/10.1002/minf.202200103

Shortridge MD, Bokemper M, Copeland JC, et al (2011) Correlation between protein function and ligand binding profiles. Journal of proteome research 10(5):2538–2545

Tang Xl, Wang Y, Li Dl, et al (2012) Orphan g protein-coupled receptors (gpcrs): biological functions and potential drug targets. Acta Pharmacologica Sinica 33(3):363–371. URL https://doi.org/10.1038/aps.2011.210

Tanimoto TT (1958) Elementary mathematical theory of classification and prediction. Tech. rep.

pandas development team T (2020) pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.3509134, URL https://doi.org/10.5281/zenodo.3509134

Tran-Nguyen VK, Junaid M, Simeon S, et al (2023) A practical guide to machine-learning scoring for structure-based virtual screening. Nature Protocols 18(11):3460–3511

Truchon JF, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. Journal of chemical information and modeling 47(2):488–508

Wang Y, Guo Y, Kuang Q, et al (2015) A comparative study of family-specific protein–ligand complex affinity prediction based on random forest approach. Journal of computer-aided molecular design 29(4):349–360

Wilkins CL, Randić M (1980) A graph theoretical approach to structure-property and structure-activity correlations. Theoretica chimica acta 58(1):45–68. URL https:

//doi.org/10.1007/BF00635723

Willett P (2011) Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance. WIREs Data Mining Knowl Discov 1(3):241–251. URL https://doi.org/10.1002/widm.26

Xie X, Yu T, Li X, et al (2023) Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. Signal Transduction and Targeted Therapy 8(1):335. URL https://doi.org/10.1038/s41392-023-01589-z

Zoete V, Daina A, Bovigny C, et al (2016) Swisssimilarity: a web tool for low to ultra high throughput ligand-based virtual screening

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarymaterial.tex