From High Dimensions to Human Comprehension: Exploring Dimensionality Reduction for Chemical Space Visualization

Alexey A. Orlov, Tagir N. Akhmetshin, Dragos Horvath, Gilles Marcou, Alexandre Varnek*

Laboratory of Chemoinformatics, UMR 7140 CNRS, University of Strasbourg, 4, Blaise Pascal Str., 67000 Strasbourg, France

Abstract

Dimensionality reduction is an important exploratory data analysis method that allows high-dimensional data to be represented in a human-interpretable lower-dimensional space. It is extensively applied in the analysis of chemical libraries, where chemical structure data — represented as high-dimensional feature vectors—are transformed into 2D or 3D chemical space maps. In this paper, commonly used dimensionality reduction techniques — Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), and Generative Topographic Mapping (GTM) — are evaluated in terms of neighborhood preservation and visualization capability of sets of small molecules from the ChEMBL database.

keywords: dimensionality reduction, chemical libraries, chemical space, chemography, principal component analysis, t-distributed Stochastic Neighbor Embedding, Uniform Manifold Approximation and Projection, Generative Topographic Mapping

Introduction

Dimensionality reduction (DR) is an important machine learning (ML) technique used to produce a compressed low-dimensional embedding of a given high-dimensional dataset, serving either as a data preprocessing step for further application of other machine learning algorithms or as a tool for visualizing human-interpretable 2 or 3 dimensions (2D and 3D).[1–3] As a visualization tool DR techniques are ubiquitous and are widely used in a variety of fields, including extensive application for omics studies[4], and chemical space analysis[5]. In the latter, they have coined the term

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

"chemography" by analogy to geography[6]. Chemography aims at converting the data on chemical structures that are frequently represented as a feature vector of high dimensionality in the form of a 2D chemical space map. Beyond their illustrativeness and artistic visual appeal[7], chemical space maps can be combined with other tools, such as deep generative models [8,9] to effectively guide chemical space exploration, or accelerate similarity-based virtual screening[10].

Numerous benchmarking studies have been conducted to compare DR methods, both for tasks in specific domains and across numerous fields.[4,11–13]. These studies highlight non-linear DR algorithms t-Distributed Stochastic Neighbor Embedding (t-SNE)[14], Uniform Manifold Approximation and Projection (UMAP)[15] as the best-performing methods. However, a linear DR method, Principal Component Analysis (PCA), is also very popular and is sometimes reported as more efficient[16]. Therefore, there is no single method that is universally superior; the choice of method should be guided by its suitability for a particular set of tasks. While chemical datasets have been benchmarked in some studies[11,17–19], a detailed discussion of the DR methods' relevance in the context of chemical space analysis for medicinal chemistry-relevant small organic molecules is lacking in the literature.

This paper compares dimensionality reduction (DR) techniques for exploring chemical space. Specifically, we evaluate the effectiveness of three non-linear methods — t-SNE, UMAP, and Generative Topographic Mapping (GTM)—and one linear method, PCA, commonly used for visualizing chemical spaces[20–22]. The analysis utilizes subsamples from the ChEMBL database[23], focusing on compounds tested against specific biological targets. Various representations of different dimensionalities were used to describe chemical compounds and a grid-based search was conducted to optimize hyperparameters with neighborhood preservation as the objective. The results confirmed the strong performance of the non-linear methods in neighborhood preservation. Additionally, scatterplot diagnostics (scagnostics)[24] were applied to quantitatively assess the characteristics of the chemical space maps that can be relevant to human perception. The strengths and weaknesses of these methods are discussed, highlighting their effectiveness and potential limitations.

Methods

Workflow for comparing dimensionality reduction approaches for chemical space analysis

The dimensionality reduction techniques were assessed in two ways: the accuracy of neighborhood preservation and visual interpretability. The general scheme for the comparative analysis of DR techniques used in this paper is presented in Figure 1. To optimize hyperparameters, a grid-based search was conducted using the percentage of preserved nearest 20 neighbors from the high-dimensional space as the optimization metric. The optimized models were then evaluated using

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

additional neighborhood preservation metrics. Additionally, to quantitatively assess the visual interpretability of the plots, scatter diagnostics[24] (scagnostics) were calculated providing relevance of the method's visualization for better human perception.

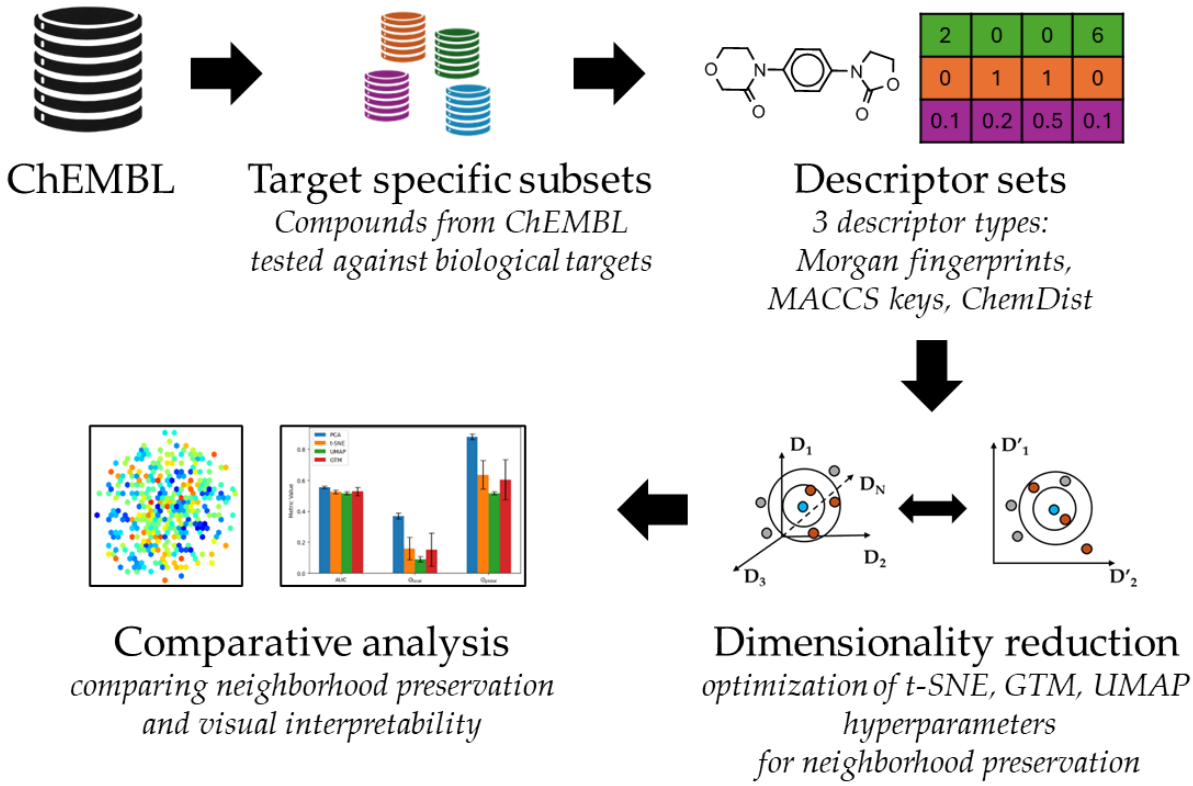


Figure 1. Workflow for comparing dimensionality reduction methods for chemical space visualization. Initially, 13 datasets with low intrinsic dimensionality are retrieved from ChEMBL. These datasets are then processed using three distinct descriptor types: Morgan count fingerprints, MACCS keys, and ChemDist. Subsequently, dimensionality reduction techniques, including t-SNE, GTM, and UMAP, are applied, with hyperparameters optimized for neighborhood preservation. A comparative analysis focused on evaluating the preservation of neighborhood structures and the visual interpretability of the chemical space maps (low-dimensional embeddings) is performed.

In summary, the objective of this study is to evaluate the performance of various dimensionality reduction (DR) methods across different scenarios:

1. Comparison of Neighborhood Preservation for *in-sample* DR: This involves assessing how well the methods maintain neighborhood relationships within a series of target-specific ChEMBL subsets, using the entire dataset for training.
2. Comparison of Neighborhood Preservation for *out-of-sample* DR: The neighborhood preservation was assessed in a Leave-One-Library-Out Scenario (LOLO). This focuses on

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

evaluating the DR methods when applied to new data, where one library is excluded during training.

3. Quantitative Evaluation of Visualizations: The visualizations generated by the DR methods are quantitatively assessed using scagnostics metrics.

This contribution does not account Neighborhood Behavior[25] (NB, the hypothesis that structurally similar compounds exhibit similar biological properties). It only focuses on the question of the impact of DR on the neighborhood of items, and does not further explore the question whether resulting maps are effectively regrouping activity-related molecules. Map NB-compliance is expectedly worse than original descriptor space compliance, but this issue has been extensively explored at least for one method, GTM, known to support highly NB-compliant property "landscapes"[26,27]. Therefore, the biological properties of the compounds were not included into the analysis.

Data collection and preprocessing

Subsets of chemical compounds were retrieved from a pool of preprocessed target specific subsets from ChEMBL version 33 database,[23] prepared according to an in-house protocol as previously described [28,29]. The selection of datasets was based on two criteria: each subset contains more than 400 compounds, and the intrinsic dimensionality, calculated using Fisher's separability algorithm[30] on the data represented as Morgan count fingerprints (see below) shall cover a wide range of values. In addition to these target-specific subsets, three random subsets of sizes 500, 1500, and 9269 were also retrieved from ChEMBL.

Descriptor calculation

Descriptors were calculated using the RDKit (v.2022.09.5) library[31]. Compounds were represented as Morgan count fingerprints with radius 2 and fingerprint size 1024. For each dataset, all zero-variance features were removed, and the remaining features were standardized.

Three types of descriptors with varying number of dimensions were used: Morgan count fingerprints[32], MACCS keys[33], and embeddings from deep neural network[34] (ChemDist).

Morgan count fingerprints and MACCS keys were calculated using the RDKit (v.2022.09.5) library[31]. For Morgan count fingerprints radius 2 and fingerprint size 1024 were used. Default RDKit parameters were used to generate MACCS keys.

ChemDist embeddings were obtained using the pretrained network[34]. The default parameters were used.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

For each dataset, all zero-variance features were removed, and the remaining features were standardized before applying a dimensionality reduction algorithm.

Dimensionality reduction methods

The following implementations of dimensionality reduction algorithms were used: the PCA algorithm implemented in scikit-learn (version 1.4.1.post1); the t-SNE algorithm from the OpenTSNE (version 1.0.1) library[35]; the UMAP algorithm from the umap-learn (version 0.5.5) library[15]; and an in-house algorithm for GTM, which is available upon request[36].

Defining "Neighbors" in descriptor and latent space respectively.

In latent space (on the maps) the first k neighbors of an item i are found by calculating the Euclidean distances between the projection of i and all the projections of the remaining set members and ranking the latter.

In descriptor spaces, "default" neighbor definition used the same approach, based on respective Euclidean distances. However, an alternative definition of distance as the complement of the Tanimoto similarity score (1-T) was also employed. No normalization of descriptors was undertaken for calculating Tanimoto similarity values.

Neighborhood preservation analysis

As a primary metric of the neighborhood preservation for the optimization, an average number of nearest neighbors preserved between the original and the latent spaces was used[11]:

$$P_{NN}(k) = \sum_{i=1}^{N} \frac{S_{ik}}{k \times N}, \quad (1)$$

where $P_{NN}(k)$ is the neighborhood preservation score, k represents the number of considered nearest neighbors, S_{ik} is the number of the shared k-nearest neighbors of the i-th compound from the N populating the latent space and original spaces.

Additionally, the following metrics for evaluating neighborhood preservation suggested in the literature were calculated[37]: co-k-nearest neighbor size (Q_{NN}) (2), an area under Q_{NN} curve AUC(Q_{NN}) (3), local continuity meta criterion (LCMC) (4), the local (Q_{local}) (5) and global (Q_{global}) (6) properties, trustworthiness (7), continuity (8).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The calculation of the metrics is based on building co-ranking matrix (Q) (Figure 2). The Q_{kl} elements of the matrix Q count how many samples of rank k became of rank l[37]. The off-diagonal elements of Q are being used to calculate all the metrics. In the case of ties in the ranking, the ranks were selected randomly.

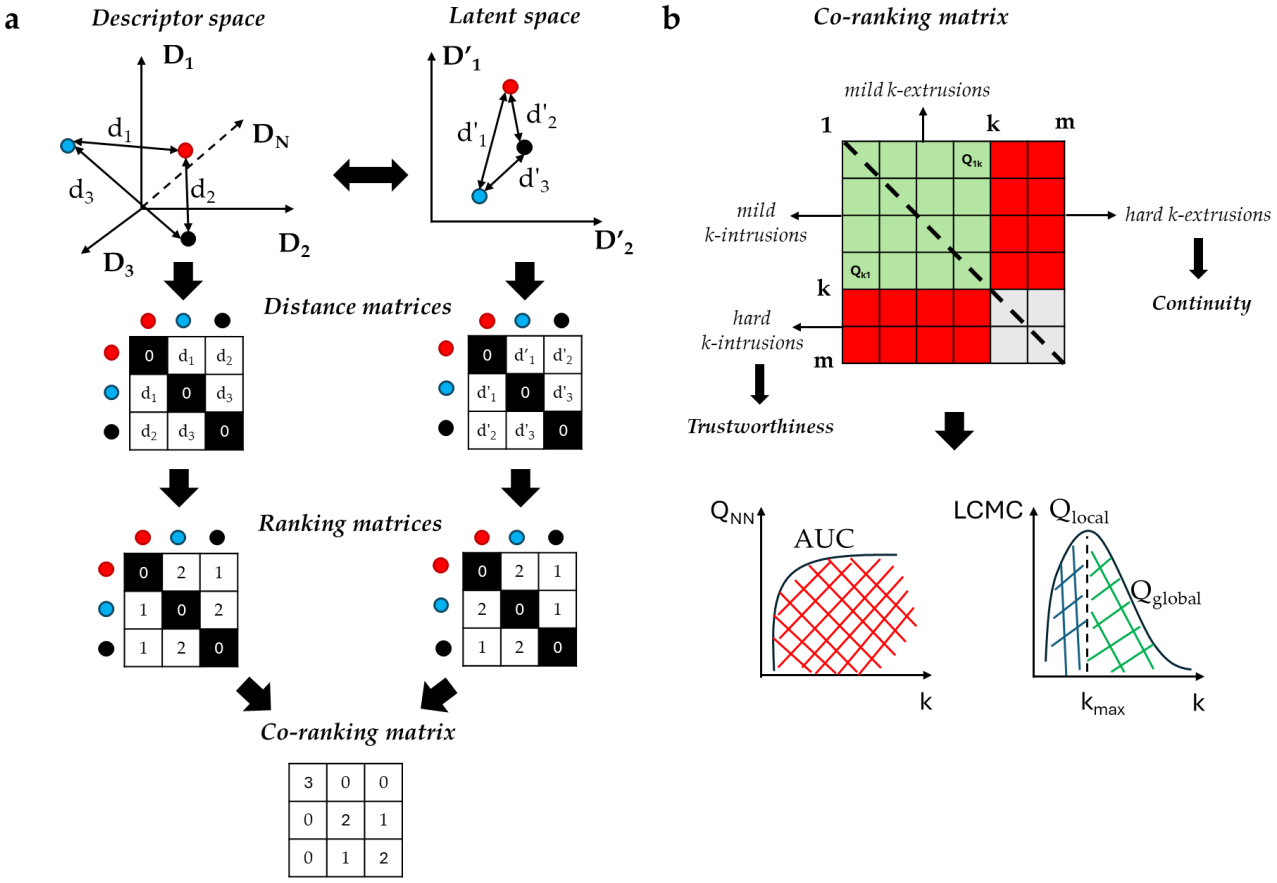


Figure 2. Metrics used for the analysis of neighborhood preservation **(a)** The descriptor space and the corresponding latent space are compared through (i) distances between data points that are then (ii) converted into a ranking of the nearest neighbors for each instance. The co-ranking matrix Q with elements (Q_{kl}) is calculated based on the ranking matrices. **(b)** The illustration of the calculation of metrics based on the co-ranking matrix and equations (2)-(9).

Co-k-nearest neighbor size

$$Q_{NN}(k) = \frac{1}{km}\sum_{i=1}^{k}\sum_{j=1}^{k} Q_{ij} \qquad (2),$$

where k is a row and a column index, m – number of rows/columns in the coranking matrix, Q_{ij} - elements of the co-ranking matrix Q.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

This measures the number cases where the neighborhood is preserved in a given tolerance, up to rank k. It corresponds to the green region in the Q matrix represented in Figure 2b.

Area under the Q_{NN} curve

$$AUC = \frac{1}{m}\sum_{k=1}^{m} Q_{NN}(k) \quad (3),$$

where k is a row index, m – number of rows/columns in the co-ranking matrix Q, $Q_{NN}(k)$ – co-k-nearest neighbor size as calculated by the equation 2.

AUC characterizes the global neighborhood preservation based on the Q_{NN} curve as shown in Figure 2b.

Local Continuity Meta Criterion (LCMC)

$$LCMC(k) = Q_{NN}(k) - \frac{k}{m-1} \quad (4),$$

where k is a row index, m – number of rows/columns in the co-ranking matrix Q, $Q_{NN}(k)$ – co-k-nearest neighbor size as calculated by the equation 2.

LCMC is a normalized (by the number of neighbors ($\frac{k}{m-1}$) that can be retrieved randomly) version of Q_{NN}. [37]

$$LCMC\ maximum\ point k_{max} = \arg\max_{k} LCMC(k) \quad (5),$$

where k is a row index in the co-ranking matrix Q, LCMC (k) – local continuity meta criterion as calculated by the equation 4.

The inflection point of the LCMC curve corresponding to the number of neighbors (Figure 2b).

Local and global property metric

$$Q_{local} = \frac{1}{k_{max}}\sum_{k=1}^{k_{max}} Q_{NN}(k) \quad (6),$$

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

$$Q_{global} = \frac{1}{m-k_{max}} \sum_{k=k_{max}}^{m-1} Q_{NN}(k) \quad (7)$$

where k_{max} is an inflection point in LCMC curve as calculated by the equation 5, k is a row index, m – number of rows/columns in the co-ranking matrix.

Q_{local} and Q_{global} represent local and global neighborhood preservation, respectively, as calculated from the LCMC curve (Figure 2b). These correspond to the green and red areas in Figure 2b when k equals k_{max}.

Trustworthiness and continuity

$$T(k) = 1 - \frac{2}{mk(2m-3k-1)} \sum_{i=k}^{m} \sum_{j=1}^{k} Q_{ij} \times (i-k) \quad (8),$$

$$C(k) = 1 - \frac{2}{mk(2m-3k-1)} \sum_{i=1}^{k} \sum_{j=k}^{m} Q_{ij} \times (j-k) \quad (9),$$

where k is a row index, m – number of rows/columns in the co-ranking matrix Q, Q_{ij} - elements of the co-ranking matrix.

Trustworthiness and continuity correspond to hard intrusions (bottom left corner in Figure 2b) and hard extrusions (top right corner in Figure 2b), respectively. Hard intrusions occur when data points (compounds) that are distant in the descriptor space appear close in the latent space. Hard extrusions happen when compounds that are close in the descriptor space appear far apart in the latent space.

Quantitative assessment of chemical space visualization using scagnostics

To quantitatively evaluate the interpretability of the visualization we used scagnostics (scatterplot diagnostics)[24,38] – visual representation quality metrics, which map a visual pattern to a real number and are frequently used to found visualizations that contain interesting patterns in an automated manner[39]. Scagnostics were calculated using an R package scagnostics[40] (version 0.2-6).

The scagnostics were calculated according to equations 10-19 as suggested in the original publications by Wilkinson et al.[24,38]. In brief, scagnostics are calculated using three principle geometric concepts: the minimum spanning tree, the convex hull, and the alpha hull (Figure 3).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

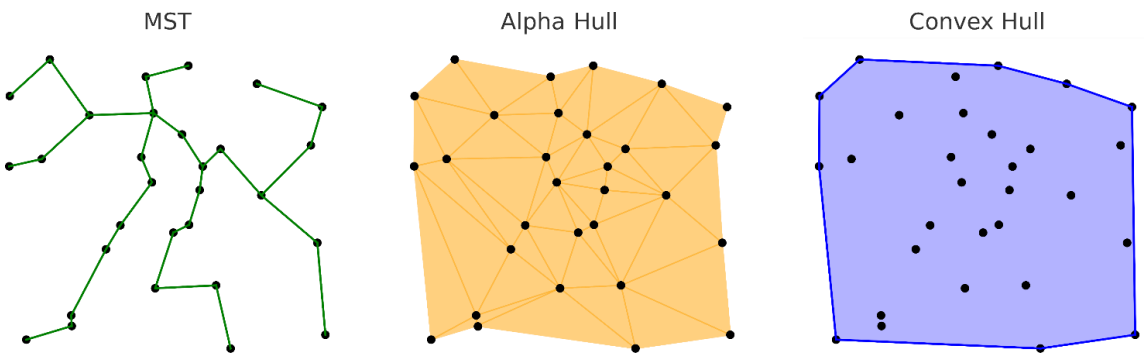


Figure 3. Three fundamental geometric constructs utilized in the calculation of scagnostics: the Minimum Spanning Tree (MST), the Alpha Hull, and the Convex Hull. The MST (left panel) represents the shortest possible tree that connects all points. The Alpha Hull (middle panel) provides a nuanced boundary of the point set by incorporating concavities, thus capturing more detailed structural features and revealing local patterns that the convex hull might overlook. The Convex Hull (right panel) represents the smallest convex polygon encompassing all points, offering a broad overview of the dataset's outer boundary.

In the following equations (10-19), H stands for the convex hull, A stands for the alpha hull, and T stands for the minimum spanning tree.

The *Outlying* scagnostic is calculated based on the minimum spanning tree T and measures the ratio of outlying dots, i.e., dots that are relatively far from others in the plot. It is defined as the proportion of the total edge length of the minimum spanning tree T that is accounted for by the total length of edges adjacent to these outlying points[38]:

$$c_{\text{outlying}} = \frac{\text{length}(T_{\text{outliers}})}{\text{length}(T)} \quad (10),$$

where length(T) is the total edge length for the MST graph, length ($T_{outliers}$) is the total edge length for the outliers in the MST graph.

An outlier is defined to be a vertex with degree 1 and associated edge weight greater than w, where w is calculated using equation (11):

$$w = q_{75} + 1.5(q_{75} - q_{25}) \quad (11),$$

where q_{75} and q_{25} are the 75th and 25th percentiles of the MST edge lengths.

Several following metrics (12-15) characterize the shape of the set of scattered points.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The *Convex* scagnostic characterizes the convexity of the 2D point distribution's shape is calculated as the ratio of the area of the alpha hull and the area of the convex hull

$$c_{convex} = \frac{area(A)}{area(H)} \quad (12),$$

where area(A) and area(H) are the areas of the alpha hull and the area of the convex hull respectively.

Skinny scagnostic evaluates the ratio of perimeter to area of a polygon:

$$c_{skinny} = 1 - \frac{\sqrt{4\pi area(A)}}{perimeter(A)} \quad (13),$$

where area(A) and perimeter (A) are the area and the perimeter of the alpha hull respectively.

The *Stringy* scagnostic defines how "path-like" is MST and is calculated using the equation 14:

$$c_{stringy} = \frac{diameter(T)}{length(T)} \quad (14),$$

where diameter(T) and length(T) are the diameter and length of all MST edges.

Straight scagnostic is defined as the Euclidean distance between the points at the ends of the longest shortest path of the MST divided by the diameter:

$$c_{straight} = \frac{dist(t_j, t_k)}{diameter(T)} \quad (15),$$

where t_j and t_k are the vertices in T on which the diameter is defined.

Monotonic scagnostic is defined as the square of Spearman correlation coefficient ($r_{spearman}$) between coordinates and characterizes the presence of a clear trend on the plot:

$$c_{monotonic} = r^2_{spearman} \quad (16),$$

Skewed scagnostic evaluates the relative density of points in a scattered configuration:

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

$$c_{skew} = \frac{(q_{90} - q_{50})}{(q_{90} - q_{10})} \quad (17),$$

where q_{90}, q_{50}, q_{10} are quantiles of the MST edge lengths.

Clumpy scagnostic characterizes the presence of clusters and is calculated as:

$$c_{clumpy}(T) = \max_j \left[1 - \max_k \frac{length(e_k)}{length(e_j)}\right], \quad (18)$$

where j indexes edges in the MST and k indexes edges in each runt set derived from an edge indexed by j. The runt set corresponds to an edge that is the smaller of the two subsets of edges that are still connected to each of the two vertices in e j after deleting edges in the MST with lengths less than length($e_{j)}$.

Striate scagnostic assesses the presence of multiple parallel lines and defined as:

$$c_{striate}(T) = \frac{1}{|V^{(2)}|} \sum_{v \in V^{(2)}} |\cos \theta_{e(v,a)e(v,b)}|, \quad (19)$$

$V^{(2)} \subseteq V$ be the set of all vertices of degree 2 in V.

The metrics used in this study are summarized in Table 1.

Table 1. List of measures used to assess the neighborhood and scagnostics of dimensionality reduction techniques.

Name	Range	Comment	Equation
Neighborhood preservation metrics			
Neighborhood preservation score (P_{NN})	0-1 (1 – all neighbors preserved at given k, 0 – no neighbors preserved at given k)	Real-valued metric: Characterizes the preservation of neighbors without considering their ranks in the descriptor and latent spaces.	(1)

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Co-k-nearest neighbor size (Q_{NN} (k))	0-1 (1 – ideal neighborhood preservation at given k)	Real valued metric: Evaluates the preservation of neighbors at a given nearest neighborhood size k, considering their ranks in both descriptor and latent spaces.	(2)
Area under the Q_{NN} curve (AUC)	0.5-1 (1 – ideal neighborhood preservation)	Real-valued metric: Summarizes the global preservation of neighbors based on Q_{NN}.	(3)
Local Continuity Meta Criterion (LCMC (k))	0-1 (1 – all neighbors preserved at given k, 0 – no neighbors preserved at given k)	Real valued metric: Q_{NN} (k) value is normalized by the number of neighbors that can be drawn randomly for a given k	(4)
k_{max}	1-N, where N – the number of data points in the dataset (larger k_{max} values signify larger preserved local neighborhoods)	Integer: The maximum value point of the LCMC curve.	(5)
Q_{local}	0-1 (1 – high local neighborhood preservation, 0 – low local neighborhood preservation)	Real number: Metric characterizing local neighborhood preservation based on LCMC for ($k<k_{max}$)	(6)
Q_{global}	0-1 (1 – high global	Real number: Metric characterizing global	(7)

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

	neighborhood preservation, 0 – low global neighborhood preservation)	neighborhood preservation based on LCMC for k>kmax	
Trustworthiness	0-1 (1 – no hard intrusions, 0 – large number of hard extrusions)	Real number: Indicates the presence of hard intrusions.	(8)
Continuity	0-1 (1 – no hard extrusions, 0 – large number of hard extrusions)	Real number: Indicates the presence of hard extrusions.	(9)
Scagnostics			
Outlying	0-1	Characterizes the presence of outlying data points on a scatter plot.	(10)
Convex	0-1	Characterizes various aspects of the shape distribution of data points on a scatter plot.	(12)
Stringy	0-1		(13)
Straight	0-1		(14)
Monotonic	0-1	Determines if a clear trend can be identified on a scatter plot.	(15)
Skewed	0-1		(16)

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Clumpy	0-1	Provides characteristics of the density of the data point distribution on a scatter plot.	(17)
Striated	0-1	Characterizes the coherence of the plots.	(18)

Optimization of hyperparameters

The following hyperparameters were optimized for non-linear methods, with a total of 72 parameters tested for each method.

The PCA calculations have been performed using scikit-learn v. 1.5.0 software. The implementation is based on the singular value decomposition. The default ('auto') parameter of 'svd_solver' was used. Only the 2 first principal components are used to project the datasets.

For t-SNE, the hyperparameters were chosen according to the suggestions from Gove et al.[41] Perplexity values were chosen to be [1, 2, 4, 8, 16, 32, 64, 128]. Exaggeration values were chosen to be [1, 2, 3, 4, 5, 6, 8, 16, 32]. The learning rate was kept as the default of OpenTSNE, since t-SNE is more robust to changes in learning rate around our empirical hyperparameter guideline than to changes in perplexity or exaggeration.[35] Fast Fourier Transform accelerated interpolation method was used to calculate gradients.

For UMAP, the parameter grid included 9 values for nearest neighbors (n_neighbors): [2, 4, 6, 8, 16, 32, 64, 128, 256] and 8 values for minimal distance (min_dist): [0.0, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8, 0.99].

The GTM parameter grid encompassed configurations such as the number of nodes set to 225, 625, and 1600; the number of basis functions set to 100, 400, and 1225; regularization coefficients (reg_coeff) of 1, 10, and 100; and basis widths of 0.1, 0.4, 0.8, and 1.2.

Intrinsic dimension analysis

Intrinsic dimension analysis with Fisher separability algorithm was performed using scikit-dimension library (v. 0.3.3).[42]

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Results

Neighborhood preservation analysis

Neighborhood preservation analysis for in-sample dimensionality reduction

Although numerous chemical datasets are available for benchmarking supervised machine learning methods[43–45], to our knowledge there are no datasets designed to evaluate the quality of DR neighborhood preservation and visualization. In this study, we focused on small organic molecules tested against specific ChEMBL targets. Following observations on the importance of low intrinsic dimension for achieving meaningful visualization[11,16,42], these datasets were chosen to cover a wide range of intrinsic dimension values as assessed by the Fisher separation method using Morgan count fingerprints as features[30]. In total, 103 datasets with the number of compounds from 406 to 4376 were selected (Supplementary Figure SF1) and intrinsic dimensionality ranged between 3 and 26.

One of the most common methods to evaluate the usability of a visualization obtained using a DR technique is to assess how well close neighbors in the original space are preserved in the latent space[2]. While numerous metrics have been suggested for this type of evaluation, in this work, we focused on one of the simplest: the number of k-neighbors preserved in the latent space, also known as the neighborhood hit[11]. To optimize hyperparameters, a grid-based search was conducted using the percentage of preserved nearest 20 neighbors from the high-dimensional space as the optimization metric. All non-linear techniques were able to retrieve, on average, 40% to 75% of the 20 closest neighbors depending on the descriptor set, outperforming PCA by 20% or more (Figure 4a, Supplementary Table ST1). On average, a lower dimensionality of the ambient space data corresponds to a higher preservation score (Figure 4), while the relative performance of the methods remains consistent across different descriptor sets.

Consistently, all non-linear methods demonstrated similar trends in other neighborhood preservation metrics, avoiding significant intrusions and exclusions—cases where compounds positioned far apart in the original space appear close on the map, and *vice versa*, where compounds far apart on the map are actually close in the original space (Figure 4). For non-linear methods, co-k-nearest neighbor size (Q_{NN}) and Local Continuity Meta Criterion (LCMC) exhibited a sharp increase for low k-values, indicating their strong performance in preserving the closest neighbors (Figure 4). In contrast, PCA demonstrated a more uniform performance across various values of k.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

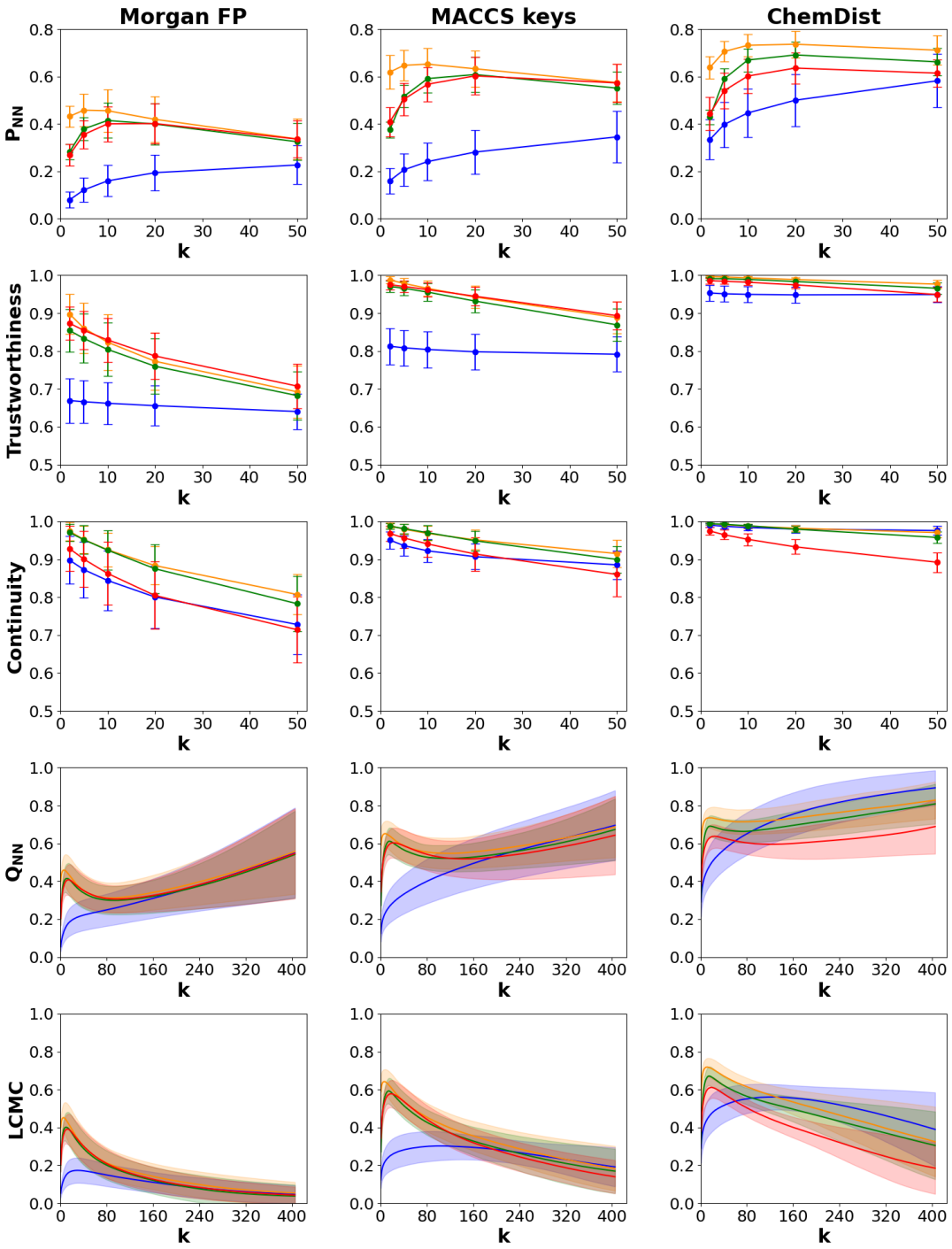


Figure 4. Average neighborhood preservation metrics for optimized models across 59 ChEMBL subsets for various feature sets (Morgan fingerprints, MACCS keys, ChemDist embeddings). The models' hyperparameters were selected to maximize the preservation of neighbors among the 20 nearest ones (Euclidean distance in the original space was used). Color scheme: PCA – blue, t-SNE – orange, UMAP – green, GTM – red. The ratio of nearest neighbors (P_{NN}) preserved at different k-values, trustworthiness, continuity, co-k-nearest neighbor size (Q_{NN}), and Local Continuity Meta Criterion (LCMC) as functions of the k-nearest neighbors are shown. Standard deviation values

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

calculated across datasets are shown as bars or filled areas. Corresponding AUC, Q_{local}, Q_{global}, k-max values can be found in Supplementary Table ST1.

While all the methods were able to preserve a significant number of neighboring compounds for the aforementioned datasets, the percentage of the preserved neighbors for nine randomly selected ChEMBL datasets was significantly lower (Supplementary Table ST1, Supplementary Figure SF2) for all considered descriptor spaces. As was shown before, this discrepancy suggests that the effectiveness of these DR can vary considerably depending on the dataset's characteristics[16,42]. If the data inherently resides in a high-dimensional space, traditional dimensionality reduction methods might struggle to preserve the neighborhood structure accurately[16,42]. Therefore, assessing the intrinsic dimension of the datasets is important for evaluating the applicability of the DR for the particular dataset. The correlation between intrinsic dimension and ID was observed in datasets using Morgan fingerprints and MACCS keys as features (Figure 5, Supplementary Figure SF3, SF4). To put this into a chemical perspective, congeneric organic compound series are much less dimensional than random collections. Random sets of a few thousand ChEMBL compounds consist almost entirely of singletons, compounds for which the nearest neighbor being very distant. Therefore, neighborhood preservation scores are meaningful only when there are items within a relevant neighborhood. However, further investigation is required to draw more solid conclusions across various dataset sizes and feature sets.

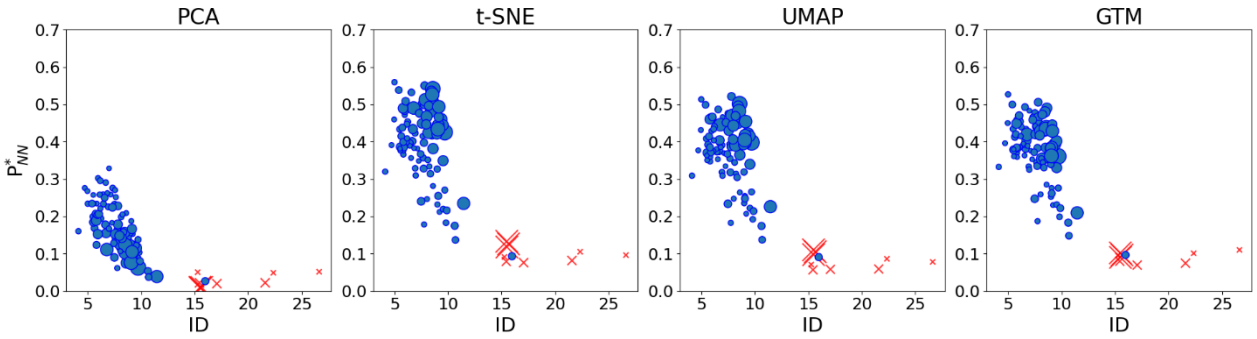


Figure 5. The figure illustrates the negative correlation between the adjusted neighborhood preservation (P^*_{NN}), normalized by the number of neighbors that can be selected randomly, and the intrinsic dimension (ID) calculated using the Fisher algorithm. This correlation is observed across different dimensionality reduction techniques (PCA, t-SNE, UMAP, GTM) utilizing Morgan fingerprints as features. The size of the data points reflects the number of compounds in the dataset. The random ChEMBL subsets are shown as red crosses.

There were not many datasets with large sizes and low intrinsic dimensions (Figure 6, Supplementary Figure SF1). To assess the possibility of having a relatively large dataset with low intrinsic dimension, we selected 18 partially overlapping datasets (low-ID datasets, Supplementary Table ST1, Supplementary Figure SF5). Each dataset contained between 411 and 1756 compounds and had an intrinsic dimension of less than 6 when represented as Morgan count fingerprints. When

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

combined into a single dataset, there were 16287 unique compounds, and the intrinsic dimension was equal to 7.5. The results for the fused dataset were similar to those for the individual datasets: in both cases, all non-linear methods significantly outperformed PCA in terms of preserving neighborhood behavior, exhibiting performance similar to that observed in the case of the individual libraries (Supplementary Table ST1, Supplementary Figure SF5). Among non-linear methods, t-SNE and UMAP outperformed GTM in preserving the closest nearest neighbors in all descriptor spaces.

The similarity between chemical compounds is typically analyzed using the Tanimoto score rather than Euclidean distance[46]. These metrics do not necessarily produce the same neighborhoods, and numerical transformation of the descriptors can significantly alter the results. We assessed whether the neighborhood preservation metrics would differ if the methods were optimized while keeping track of nearest neighbors in the descriptor space with Tanimoto similarity for the low-ID ChEMBL datasets. All methods preserved more nearest neighbors when using Tanimoto similarity to evaluate neighborhoods in the descriptor space (Supplementary Table ST1). Since the Tanimoto kernel can be used in combination with all methods as a distance metric in the original space (for example, as in kernel PCA[2] and GTM[47]), its usage presents a promising avenue for further enhancing neighborhood preservation.

Neighborhood preservation for out-of-sample dimensionality reduction

While standard dimensionality reduction techniques can be straightforwardly applied to small and medium-sized libraries, their application to large (millions to tens of millions) and ultra-large (over 1 billion) datasets remains challenging because it is time-consuming, and resource-intensive and often necessitates the use of certain approximations[48,49]. To handle such large volumes of data, a common approach is to select a subsample of the entire dataset, often referred to as a frameset or reference set, and then project the remaining data points onto the map built using this subset of the original data[22,50]. In this case, the DR algorithm should be able to project new (out-of-sample) data onto already built embedding. We assessed the algorithms for the effectiveness of out-of-sample projection using a leave-one-library-out (LOLO) scenario. In this scenario, a library was removed from the pool of 18 low-ID ChEMBL libraries, the method was fitted to the remaining data (the frameset), its parameters optimized towards neighborhood preservation, the removed (out-of-sample) library was projected onto the built embedding and neighborhood preservation metrics were calculated. On average, GTM demonstrated more robust out-of-sample neighborhood preservation compared to other non-linear methods (Supplementary Table ST1, Figure 6), preserving more neighbors in 2 out of 3 descriptor spaces.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

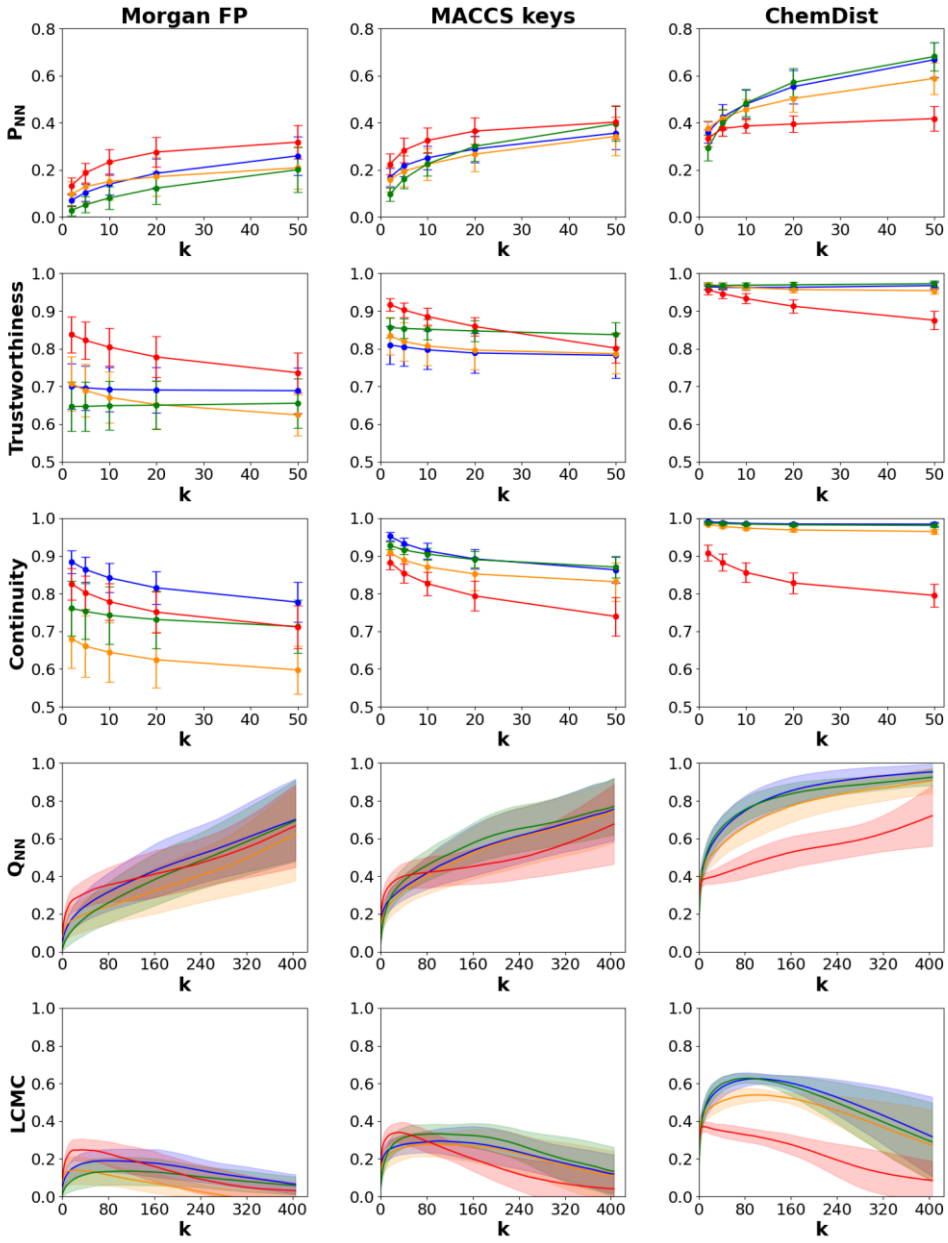


Figure 6. Average Nearest Neighbor (P_{NN}) preservation metric for leave-one library-out (LOLO) setup. As with the in-sample case (Figure 4), non-linear methods outperform PCA in neighborhood preservation for Morgan FP and MACCS keys, albeit with lower P_{NN} values. Among them, GTM demonstrated the most robust performance. The models' hyperparameters were selected to maximize the preservation of neighbors among the 20 nearest ones using Euclidean distance in the original descriptor space with Morgan fingerprints, MACCS keys and ChemDist embeddings used as feature sets. Color scheme: PCA – blue, t-SNE – orange, UMAP – green, GTM – red. The ratio of nearest neighbors (P_{NN}) preserved at different k-values, trustworthiness, continuity, co-k-nearest neighbor size (Q_{NN}), and Local Continuity Meta Criterion (LCMC) as functions of the k-nearest neighbors are shown. Standard deviation values calculated across datasets are shown as bars or filled areas. Corresponding AUC, Q_{local}, Q_{global}, k-max values can be found in Supplementary Table ST1.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Quantitative analysis of chemical space maps visualization using scagnostics

While neighborhood preservation is an important parameter for assessing the quality of a low-dimensional embedding, the primary goal of DR-based visualization is to present the data in a form that can be easily understood by humans. Such visualizations should reveal data patterns within the dataset. For instance, chemical space maps built in this work show that neighborhood preservation is not evenly distributed across them: in some areas, nearly all of the 20 closest neighbors are preserved, while in other areas, the percentage of preserved neighbors is much lower (Figure 7).

While individual data points can be mostly recognized in smaller datasets (Figure 7a), larger datasets (Figure 7) present a challenge as most zones are too dense to explore effectively on a static image. To alleviate this problem we use a hexagonal grid to render the density of data points covered by the grid. Alternatively, one can apply interpolation techniques such as kernel density estimation or Voronoi diagrams[51,52]. In contrast to other methods, grid-based visualization is a built-in feature of the GTM, allowing data to be visualized not only as scatter plots but also as grid-based landscapes without the need for auxiliary binarization tools, which can be especially attractive for the visualization of large-scale datasets.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

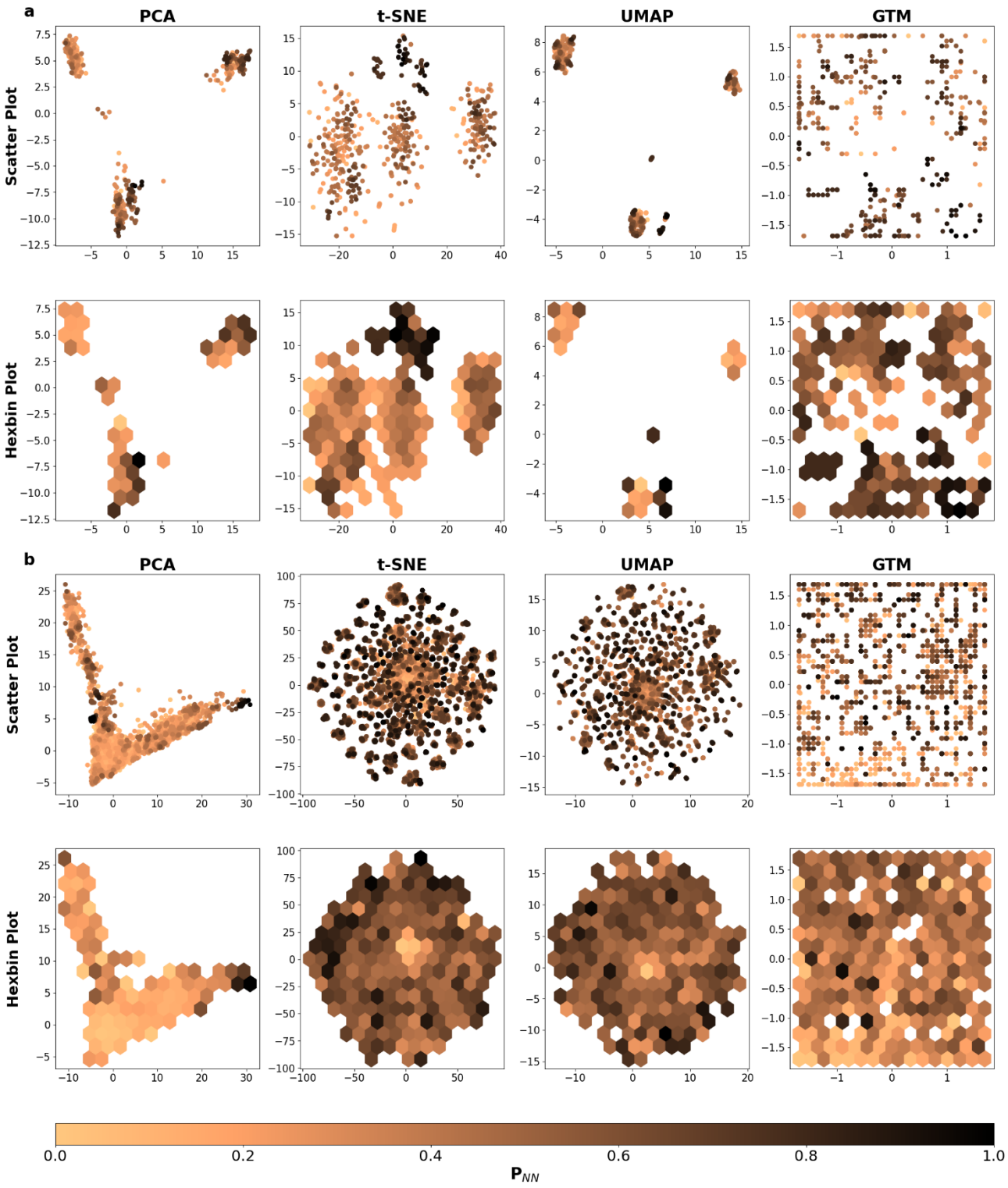


Figure 7. Scatter and hexbin plot visualizations of chemical space using PCA, t-SNE, UMAP, and GTM using Morgan count fingerprints as descriptors. These visualizations are shown for one out of 18 low-ID datasets (CHEMBL3638344) (**a**) and the combined dataset of 18 low-ID ChEMBL subsets (**b**). The color scheme corresponds to P_{NN} (k=20): black indicates all neighbors are preserved, while pale brown indicates none are preserved.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

An orthogonal approach to neighborhood preservation for comparing DR techniques is to assess the interpretability of the visualization, specifically how effectively a human can comprehend the patterns shown on the map.[53] The analysis of factors influencing human perception of statistical pattern visualization is an active and evolving area of research[39,54,55]. One of the most frequently used metrics to assess the ease of visualization for scatter plots are scatterplot diagnostics, commonly known as scagnostics[39,54–56]. Scagnostics provide a quantitative way to evaluate the visual characteristics of scatterplots, such as shape, density, skewness, and the presence of outliers and can help to determine which plots are more likely to be easily understood by human observers[56]. For example, they were found to be aligned with human perception of correlations, clusters, and trends[39].

Scagnostics were calculated for all low-dimensional embeddings built in this work (Figure 8, Supplementary Figure SF6). They show high variance in the obtained values, indicating that even with similar neighborhood scores, one method may be preferred over another in terms of visualization quality. For instance, scagnostics calculated for embeddings of the dataset CHEMBL3638344 (Figure 8a) highlight different characteristics of GTM, t-SNE, and UMAP-based generalizations. UMAP shows clear clustering of compounds, resulting in high Clumpy values, while the GTM plot offers complementary more striated representation. Additionally, different descriptors show varying scagnostic values across different methods (Supplementary Figure SF6), providing further options for choosing the most relevant representation for visualization.

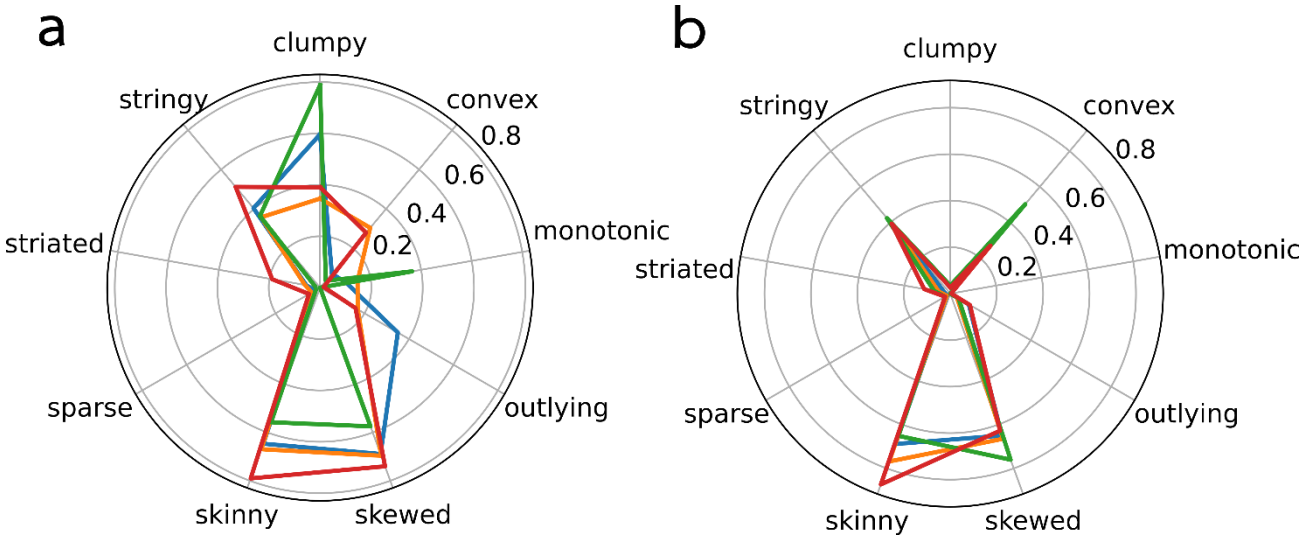


Figure 8. Radar chart representation of scagnostics calculated for scatter plot visualizations of chemical space using PCA, t-SNE, UMAP, and GTM with parameters optimized for preserving the 20 nearest neighbors using Morgan count fingerprints as descriptors (as shown in Figure 7). These visualizations are shown for CHEMBL3638344 dataset (**a**) and the combined dataset of 18 low-ID ChEMBL subsets (**b**). Color scheme: PCA – blue, t-SNE – orange, UMAP – green, GTM – red.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Discussion

Benchmarking dimensionality reduction methods for chemical space exploration

Representing chemical structure data embedded in chemical libraries in a manner suitable for chemist comprehension poses a significant challenge. Among the various approaches, one can use similarity heatmaps[20], network-based approaches[57,58], scaffolds[59], and DR techniques. The latter represents one of the main strategies, especially in the context of big data[36,49]. When used properly, DR methods can provide valuable insights into the inner structure of chemical spaces, as demonstrated in numerous studies[20,21]. While reducing complex data to two dimensions may result in information loss, this information can potentially be recovered by chemists analyzing the maps, as can be seen when combining "human intuition" with machine learning.[60,61] However, to be maximally useful for a chemist, a good dimensionality reduction method should have the following features[2,62,63] :

- The produced projections should be a sufficiently accurate lower-dimensional (2D or 3D) representation of the input data;
- The method should provide options for projection of new data points to facilitate library comparison and large chemical space data analysis.
- The method should be big-data compatible, ensuring fast training and new data projection with minimal resources.

The findings of our study in this context are summarized in Figure 8. Among the dimensionality reduction algorithms benchmarked, all non-linear methods proved effective in neighborhood preservation, outperforming PCA. t-SNE demonstrated the strongest performance in preserving the closest neighbors, which is not surprising since it is designed to maximize this criterion. On the other hand, GTM offers more robust out-of-sample performance and out-of-the-box big-data compatible visualization due to its grid-based nature. Therefore, the choice of method should be based on the specific task at hand. For example, if one wants to analyze a small chemical library, t-SNE might be preferred for its performance. On the other hand, for visualizing large libraries and potentially projecting new compounds onto them, GTM might be the better choice.

	PCA	t-SNE	UMAP	GTM
No hyperparameter tuning required	☑	☒	☒	☒
High neighborhood preservation	☒	☑	☑	☑
Robust out-of-sample neighborhood preservation	☑	☒	☒	☑
Out-of-the-box big data-compatible visualization	☒	☒	☒	☑

Figure 8. Visual representations of data reduced by PCA, t-SNE, UMAP, and GTM are displayed. The performance of each technique is assessed based on several criteria: the necessity for

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

hyperparameter tuning, neighborhood preservation quality, interpretability of hyperparameters, robustness of out-of-sample neighborhood preservation, and suitability for out-of-the-box big data visualization.

Conclusions

In this work, the effectiveness of commonly used dimensionality reduction techniques for the visualization of chemical space was assessed across three case studies commonly encountered in practice. It was found that non-linear methods significantly outperformed a linear method (PCA) in neighborhood preservation tasks for several subsets of congeneric organic small molecule compounds. Among non-linear methods t-SNE was shown to excel at preserving the very closest neighbors. For out-of-sample visualization, commonly used methods like t-SNE and UMAP were found to demonstrate less robust behavior as compared to the GTM. Additionally, GTM was recognized for its out-of-the-box, big-data compatible visualization capabilities. However, this work has some limitations, and further improvements can be made to provide a more comprehensive assessment of the performance of DR methods in various scenarios.

Limitations of the current work and future outlook

Data

The datasets used in this study comprise small organic molecule compounds featuring partially overlapping congeneric series of organic molecules, thus covering a very small part of a chemical space. Further research is necessary to design datasets with diverse distributions of chemical similarity for thorough benchmarking results.

A significant aspect of this paper is the focus on compounds from a specific region of chemical space—namely, small organic molecules derived from published medicinal chemistry data. This choice was made deliberately to maintain a clear scope for the study. Consequently, combinatorial libraries (e.g., DNA-encoded libraries, Enamine REAL), which typically exhibit a narrower distribution of chemical similarities and a more densely populated chemical space, were not explored. Additionally, datasets related to materials, polymers, and other chemical entities were not investigated. These areas, where DR techniques are increasingly being applied for visualization[19,64] , present additional layers of complexity and variability. Future studies could expand into these broader chemical spaces to further validate and benchmark DR techniques in diverse contexts.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Algorithms

Only close-to-the original versions of the non-linear DR algorithms (t-SNE, UMAP, GTM) were tested in this paper, while numerous enhancements have been suggested in the literature[1,65]. These enhanced versions may be more suitable than the "vanilla" algorithms in certain scenarios. For example, the question of algorithm run-time was not addressed in our comparison, as multiple optimized versions exist that can significantly reduce computation time, including those capable of running on graphical processing units[66,67] (GPUs). Therefore, benchmarking the original versions in terms of time efficiency would not yield solid conclusions on the applicability of the methods in general.

The hyperparameter grid was fixed in our studies for both in-sample and out-of-sample scenarios, and we did not specifically attempt to alter the hyperparameter grid for the latter, that can potentially be required in such cases[41]. For example, one can choose to lower learning rate while projecting new data onto existing t-SNE embeddings[35]. Alternatively, one may opt to several parametric versions of both UMAP and t-SNE have been proposed for out-of-sample visualizations, which can be more efficient than the setups used in this paper. For instance, the parametric t-SNE was successfully applied for the analysis of chemical space[68]. A thorough analysis of the applicability domains[69] of the DR techniques and, more generally, their out-of-distribution performance[70,71] is left for future studies.

Methodology

While the metrics used in this study are widely applied in the analysis of DR results, further improvements can be made. For instance, some compounds may have identical or very similar distances in the descriptor space, yet they could be ranked differently, impacting the final metrics. Although this paper investigates the influence of various distance thresholds on the neighborhood preservation score, future research could explore the influence on the other neighborhood preservation metrics, as well as the behavior of alternative types of similarity metrics (e.g., graph edit distance).

A significant challenge posed to dimensionality reduction (DR) techniques for chemical space analysis is the rapid expansion of chemical libraries, now encountering up to 10^{26} virtual compounds[72]. In this case, the maps become "too crowded" and lose specific resolution details, as seen in Figure 3c, with too many data points projected onto the same zones. One way to deal with this issue is to organize maps in a hierarchical way[73–75]. Hierarchical versions of all considered methods were developed.[28–31] For example, this approach was applied to build a chemical space atlas[77] – a set of hierarchically organized GTMs. These approaches are to be benchmarked against more recently suggested DR algorithms, such as TMAP[18], which were specifically designed to address the challenges of big data.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

In this work, DR techniques were assessed in terms of neighborhood preservation. The visualization of properties on the maps and the use of labeled data, such as biological activity, as an optimization parameter was not evaluated in this study. This is left for future studies with the evaluation of approaches that incorporate labeled information when building visualizations in a supervised or semi-supervised fashion.[28,78,79] Additionally, future work could include a thorough analysis of the chemical relevance of the obtained maps, such as the distribution and preservation of chemical structure patterns like scaffolds[80]. While metrics like scagnostics can reflect the ease of human perception of scatter plots, future studies could explore other chemistry-relevant aspects of low-dimensional visualizations.

Overall, while an ongoing discussion exists about how effective the DR are for revealing patterns within datasets[81], we believe that these methods, when used properly, represent a promising tool for chemical space analysis. The unification of methods under a common theoretical framework[82,83], along with the development of benchmarking datasets with controlled data complexity as well as protocols for evaluating DR algorithms in various scenarios, will enable a more thorough understanding of which techniques are best suited for specific scenarios.

Data and source code availability

The data and code related to the optimization of the hyperparameters, data analysis, and visualization are available under the GitHub repository: https://github.com/AxelRolov/cdr_bench

Acknowledgments

The authors thank Erik Egyan for preparing the datasets for the analysis using the in-house workflow.

References

1. Ghojogh, B., Crowley, M., Ghodsi, A. & Karray, F. *Elements of Dimensionality Reduction and Manifold Learning*. (Springer International Publishing AG, Cham, 2023). doi:10.1007/978-3-031-10602-6.

2. Lee, J. A. & Verleysen, M. *Nonlinear Dimensionality Reduction*. (Springer, New York, 2007).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

3. Lovrić, M. *et al.* Should We Embed in Chemistry? A Comparison of Unsupervised Transfer Learning with PCA, UMAP, and VAE on Molecular Fingerprints. *Pharmaceuticals* **14**, 758 (2021).

4. Xiang, R. *et al.* A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Front. Genet.* **12**, (2021).

5. Zabolotna, Y. *et al.* Chemography: Searching for Hidden Treasures. *J. Chem. Inf. Model.* acs.jcim.0c00936 (2020) doi:10.1021/acs.jcim.0c00936.

6. Oprea, T. I. & Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **3**, 157–166 (2001).

7. Gaytán-Hernández, D. *et al.* Art driven by visual representations of chemical space. *Journal of Cheminformatics* **15**, 100 (2023).

8. Sattarov, B. *et al.* De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J. Chem. Inf. Model.* **59**, 1182–1196 (2019).

9. Bort, W. *et al.* Discovery of novel chemical reactions by deep generative recurrent neural network. *Sci Rep* **11**, 3178 (2021).

10. Bonachera, F., Marcou, G., Kireeva, N., Varnek, A. & Horvath, D. Using self-organizing maps to accelerate similarity search. *Bioorganic & Medicinal Chemistry* **20**, 5396–5409 (2012).

11. Espadoto, M., Martins, R. M., Kerren, A., Hirata, N. S. T. & Telea, A. C. Toward a Quantitative Survey of Dimension Reduction Techniques. *IEEE Transactions on Visualization and Computer Graphics* **27**, 2153–2173 (2021).

12. Wang, K. *et al.* Comparative analysis of dimension reduction methods for cytometry by time-of-flight data. *Nat Commun* **14**, 1836 (2023).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

13. Vikram, M., Pavan, R., Dineshbhai, N. D. & Mohan, B. Performance Evaluation of Dimensionality Reduction Techniques on High Dimensional Data. in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* 1169–1174 (2019). doi:10.1109/ICOEI.2019.8862526.

14. Maaten, L. van der & Hinton, G. E. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

15. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018).

16. Tr, T. Dimensionality Reduction: A Comparative Review.

17. Alsenan, S. A., Al-Turaiki, I. M. & Hafez, A. M. Feature Extraction Methods in Quantitative Structure–Activity Relationship Modeling: A Comparative Study. *IEEE Access* **8**, 78737–78752 (2020).

18. Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics* **12**, 12 (2020).

19. Villares, M., Saunders, C. M. & Fey, N. Comparison of Dimensionality Reduction Techniques for the Visualisation of Chemical Space in Organometallic Catalysis. *Artificial Intelligence Chemistry* 100055 (2024) doi:10.1016/j.aichem.2024.100055.

20. Osolodkin, D. I. *et al.* Progress in visual representations of chemical space. *Expert Opinion on Drug Discovery* **10**, 959–973 (2015).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

21. Medina-Franco, J. L., Martinez-Mayorga, K., Giulianotti, M. A., Houghten, R. A. & Pinilla, C. Visualization of the Chemical Space in Drug Discovery. *Current Computer - Aided Drug Design* **4**, 322–333 (2008).

22. Horvath, D., Marcou, G. & Varnek, A. Generative topographic mapping in drug design. *Drug Discovery Today: Technologies* S1740674920300044 (2020) doi:10.1016/j.ddtec.2020.06.003.

23. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, D1100–D1107 (2012).

24. Wilkinson, L., Anand, A. & Grossman, R. Graph-theoretic scagnostics. in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* 157–164 (2005). doi:10.1109/INFVIS.2005.1532142.

25. Horvath, D. & Jeandenans, C. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces-a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J Chem Inf Comput Sci* **43**, 680–690 (2003).

26. Sidorov, P., Gaspar, H., Marcou, G., Varnek, A. & Horvath, D. Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *J Comput Aided Mol Des* **29**, 1087–1108 (2015).

27. Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D. & Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Molecular Informatics* **34**, 348–356 (2015).

28. Casciuc, I. *et al.* Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J. Chem. Inf. Model.* **59**, 564–572 (2019).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

29. Lin, A., Horvath, D., Marcou, G., Beck, B. & Varnek, A. Multi-task generative topographic mapping in virtual screening. *J Comput Aided Mol Des* **33**, 331–343 (2019).

30. Albergante, L., Bac, J. & Zinovyev, A. Estimating the effective dimension of large biological datasets using Fisher separability analysis. at http://arxiv.org/abs/1901.06328 (2019).

31. *RDKit: Open-Source Cheminformatics; Http://Www.Rdkit.Org*.

32. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **5**, 107–113 (1965).

33. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280 (2002).

34. Coupry, D. E. & Pogány, P. Application of deep metric learning to molecular graph similarity. *Journal of Cheminformatics* **14**, 11 (2022).

35. Poličar, P. G., Stražar, M. & Zupan, B. openTSNE: A Modular Python Library for t-SNE Dimensionality Reduction and Embedding. *Journal of Statistical Software* **109**, 1–30 (2024).

36. Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D. & Varnek, A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* **55**, 84–94 (2015).

37. Zhang, Y., Shang, Q. & Zhang, G. pyDRMetrics - A Python toolkit for dimensionality reduction quality assessment. *Heliyon* **7**, e06199 (2021).

38. Wilkinson, L. & Wills, G. Scagnostics Distributions. *Journal of Computational and Graphical Statistics* **17**, 473–491 (2008).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

39. Lehmann, D. J., Hundt, S. & Theisel, H. A study on quality metrics vs. human perception: Can visual measures help us to filter visualizations of interest? *it - Information Technology* **57**, 11–21 (2015).

40. scagnostics: Compute scagnostics - scatterplot diagnostics.

41. Gove, R., Cadalzo, L., Leiby, N., Singer, J. M. & Zaitzeff, A. New guidance for using t-SNE: Alternative defaults, hyperparameter selection automation, and comparative evaluation. *Visual Informatics* **6**, 87–97 (2022).

42. Bac, J., Mirkes, E. M., Gorban, A. N., Tyukin, I. & Zinovyev, A. Scikit-Dimension: A Python Package for Intrinsic Dimension Estimation. *Entropy* **23**, 1368 (2021).

43. Mittal, R. R., McKinnon, R. A. & Sorich, M. J. Comparison Data Sets for Benchmarking QSAR Methodologies in Lead Optimization. *J. Chem. Inf. Model.* **49**, 1810–1820 (2009).

44. Tian, T. *et al.* Benchmarking compound activity prediction for real-world drug discovery applications. *Commun Chem* **7**, 1–19 (2024).

45. Rohrer, S. G. & Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **49**, 169–184 (2009).

46. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **7**, 20 (2015).

47. Olier, I., Vellido, A. & Giraldo, J. Kernel Generative Topographic Mapping. *Computational Intelligence* (2010).

48. Wei, V., Ivkin, N., Braverman, V. & Szalay, A. Sketch and Scale: Geo-distributed tSNE and UMAP. at http://arxiv.org/abs/2011.06103 (2020).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

49. Lin, A. *et al.* Parallel Generative Topographic Mapping: An Efficient Approach for Big Data Handling. *Molecular Informatics* **39**, 2000009 (2020).

50. Zheng, Q. *et al.* From Whole to Part: Reference-Based Representation for Clustering Categorical Data. *IEEE Trans. Neural Netw. Learning Syst.* **31**, 927–937 (2020).

51. Whitaker, R. & Hotz, I. Transformations, Mappings, and Data Summaries. in *Foundations of Data Visualization* (eds. Chen, M., Hauser, H., Rheingans, P. & Scheuermann, G.) 121–157 (Springer International Publishing, Cham, 2020). doi:10.1007/978-3-030-34444-3_6.

52. Cihan Sorkun, M., Mullaj, D., Koelman, J. M. V. A. & Er, S. ChemPlot, a Python Library for Chemical Space Visualization**. *Chemistry–Methods* **2**, e202200005 (2022).

53. *Foundations of Data Visualization*. (Springer International Publishing, Cham, 2020). doi:10.1007/978-3-030-34444-3.

54. Filipowicz, A. *et al.* Visual Elements and Cognitive Biases Influence Interpretations of Trends in Scatter Plots. at http://arxiv.org/abs/2310.15406 (2023).

55. Pandey, A. V., Krause, J., Felix, C., Boy, J. & Bertini, E. Towards Understanding Human Similarity Perception in the Analysis of Large Sets of Scatter Plots. in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* 3659–3669 (Association for Computing Machinery, New York, NY, USA, 2016). doi:10.1145/2858036.2858155.

56. Etemadpour, R., Shintree, S. & Shereen, A. D. Brain Activity is Influenced by How High Dimensional Data are Represented: An EEG Study of Scatterplot Diagnostic (Scagnostics) Measures. *J Healthc Inform Res* **8**, 19–49 (2024).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

57. Amoroso, N. *et al.* Making sense of chemical space network shows signs of criticality. *Sci Rep* **13**, 21335 (2023).

58. Scalfani, V. F., Patel, V. D. & Fernandez, A. M. Visualizing chemical space networks with RDKit and NetworkX. *Journal of Cheminformatics* **14**, 87 (2022).

59. Velkoborsky, J. & Hoksza, D. Scaffold analysis of PubChem database as background for hierarchical scaffold-based visualization. *Journal of Cheminformatics* **8**, 74 (2016).

60. Llompart, P. *et al.* Harnessing Medicinal Chemical Intuition from Collective Intelligence.

61. Teso, S., Alkan, Ö., Stammer, W. & Daly, E. Leveraging explanations in interactive machine learning: An overview. *Front Artif Intell* **6**, 1066049 (2023).

62. Wassenaar, P., Guetschel, P. & Tangermann, M. Approximate UMAP allows for high-rate online visualization of high-dimensional data streams. at http://arxiv.org/abs/2404.04001 (2024).

63. *Principal Manifolds for Data Visualization and Dimension Reduction*. (Springer, Berlin Heidelberg, 2008).

64. Park, H., Onwuli, A., Butler, K. & Walsh, A. Mapping inorganic crystal chemical space. *Faraday Discuss.* (2024) doi:10.1039/D4FD00063C.

65. Bishop, C. M., Svensén, M. & Williams, C. K. I. Developments of the generative topographic mapping. *Neurocomputing* **21**, 203–224 (1998).

66. Chan, D. M., Rao, R., Huang, F. & Canny, J. F. T-SNE-CUDA: GPU-Accelerated T-SNE and its Applications to Modern Data. in *2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)* 330–338 (2018). doi:10.1109/CAHPC.2018.8645912.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

67. Nolet, C. J. *et al.* Bringing UMAP Closer to the Speed of Light with GPU Acceleration.

68. Karlov, D. S., Sosnin, S., Tetko, I. V. & Fedorov, M. V. Chemical space exploration guided by deep neural networks. *RSC Adv.* **9**, 5151–5157 (2019).

69. Varnek, A. & Baskin, I. I. Chemoinformatics as a Theoretical Chemistry Discipline. *Molecular Informatics* **30**, 20–32 (2011).

70. Tossou, P., Wognum, C., Craig, M., Mary, H. & Noutahi, E. Real-World Molecular Out-Of-Distribution: Specification and Investigation. *J. Chem. Inf. Model.* **64**, 697–711 (2024).

71. Liu, J. *et al.* Towards Out-Of-Distribution Generalization: A Survey. at http://arxiv.org/abs/2108.13624 (2023).

72. Warr, W. A., Nicklaus, M. C., Nicolaou, C. A. & Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **62**, 2021–2034 (2022).

73. VanHorn, K. C. & Çobanoğlu, M. C. Haisu: Hierarchically supervised nonlinear dimensionality reduction. *PLOS Computational Biology* **18**, e1010351 (2022).

74. Marcílio-Jr, W. E., Eler, D. M., Paulovich, F. V. & Martins, R. M. HUMAP: Hierarchical Uniform Manifold Approximation and Projection. at http://arxiv.org/abs/2106.07718 (2023).

75. Tino, P. & Nabney, I. Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 639–656 (2002).

76. Avellaneda, M. Hierarchical PCA and Applications to Portfolio Management. at http://arxiv.org/abs/1910.02310 (2019).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

77. Zabolotna, Y. *et al.* Chemspace Atlas: Multiscale Chemography of Ultralarge Libraries for Drug Discovery. *J. Chem. Inf. Model.* **62**, 4537–4548 (2022).

78. Hajderanj, L., Weheliye, I. & Chen, D. A New Supervised t-SNE with Dissimilarity Measure for Effective Data Visualization and Classification. in *Proceedings of the 8th International Conference on Software and Information Engineering* 232–236 (Association for Computing Machinery, New York, NY, USA, 2019). doi:10.1145/3328833.3328853.

79. Ghojogh, B., Ghodsi, A., Karray, F. & Crowley, M. Uniform Manifold Approximation and Projection (UMAP) and its Variants: Tutorial and Survey. at http://arxiv.org/abs/2109.02508 (2021).

80. Zahoránszky-Kőhalmi, G., Wan, K. K. & Godfrey, A. G. Hilbert-curve assisted structure embedding method. *Journal of Cheminformatics* **16**, 87 (2024).

81. Marx, V. Seeing data as t-SNE and UMAP do. *Nat Methods* 1–4 (2024) doi:10.1038/s41592-024-02301-x.

82. Ravuri, A. & Lawrence, N. D. Towards One Model for Classical Dimensionality Reduction: A Probabilistic Perspective on UMAP and t-SNE. at http://arxiv.org/abs/2405.17412 (2024).

83. Ravuri, A., Vargas, F., Lalchand, V. & Lawrence, N. D. Dimensionality Reduction as Probabilistic Inference. at http://arxiv.org/abs/2304.07658 (2023).

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Supplementary Information
From High Dimensions to Human Comprehension: Dimensionality Reduction in Chemical Space Exploration

Alexey A. Orlov, Tagir N. Akhmetshin, Dragos Horvath, Gilles Marcou, Alexandre Varnek*

key words: dimensionality reduction, chemical libraries, chemical space, chemography, principal component analysis, t-distributed Stochastic Neighbor Embedding, Uniform Manifold Approximation and Projection, Generative Topographic Mapping

Laboratory of Chemoinformatics, UMR 7140 CNRS, University of Strasbourg, 4, Blaise Pascal Str., 67000 Strasbourg, France

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

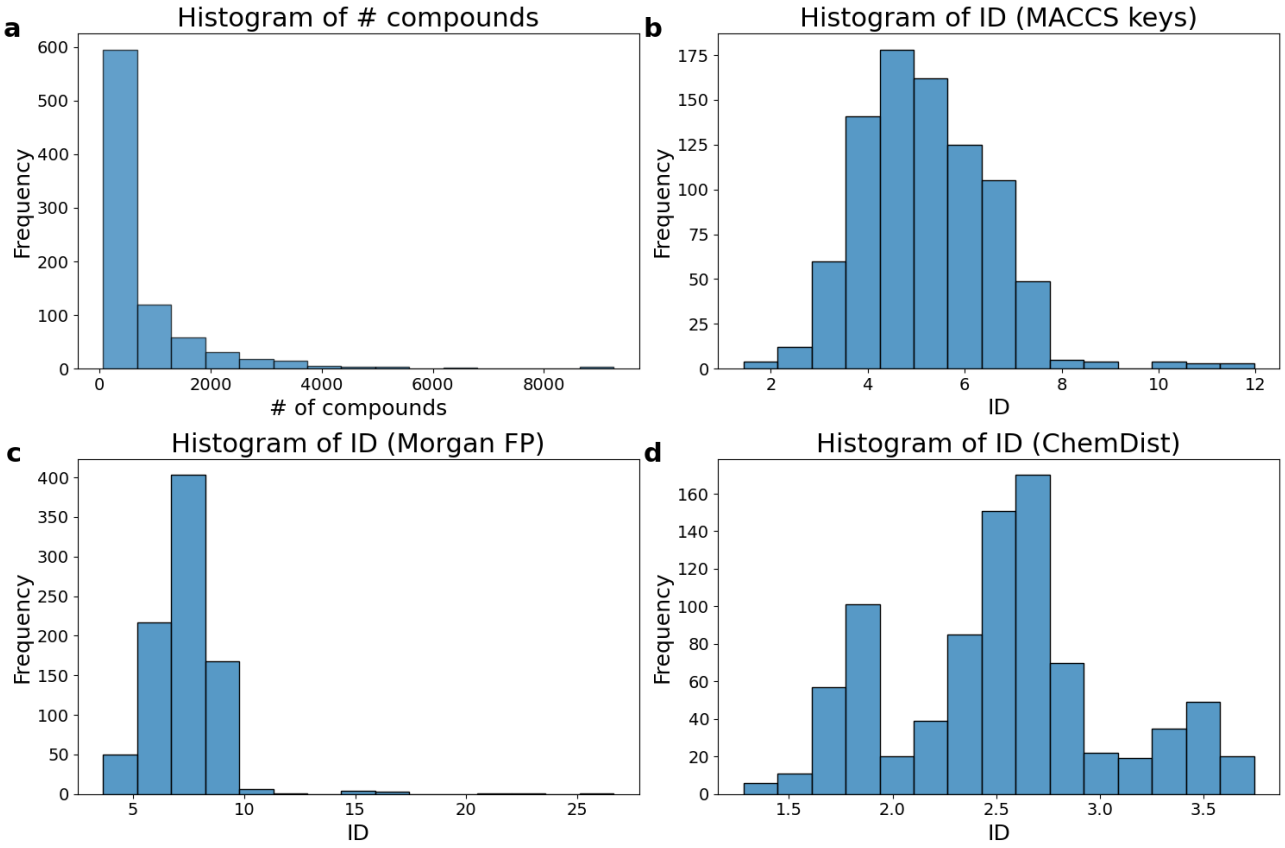
Supplementary Table ST1. Neighborhood and distance preservation metrics. Ed – Euclidean distance, Jd – Jaccard distance. LOLO – leave-one library-out set up.

Dataset	Method	Features														
		Morgan fingerprints (radius 2, fp size 1024)					MACCS keys					ChemDist embeddings				
		P_{20NN} (%)	AUC_{QNN}	kmax	Q_{local}	Q_{global}	P_{20NN} (%)	AUC_{QNN}	kmax	Q_{local}	Q_{global}	P_{20NN} (%)	AUC_{QNN}	kmax	Q_{local}	Q_{global}
All target-related ChEMBL subsets (Ed neighbors)	PCA	19±8	0.56±0.03	62±111	0.19±0.06	0.58±0.06	28±9	0.69±0.04	168±129	0.36±0.08	0.75±0.04	28±9	0.69±0.04	168±129	0.36±0.08	0.75±0.04
	t-SNE	42±10	0.58±0.03	7±5	0.46±0.05	0.58±0.03	63±8	0.71±0.03	7±5	0.64±0.07	0.71±0.03	63±8	0.71±0.03	7±5	0.64±0.07	0.71±0.03
	UMAP	40±9	0.56±0.03	11±10	0.37±0.05	0.56±0.03	61±7	0.7±0.03	16±9	0.53±0.05	0.7±0.03	61±7	0.7±0.03	16±9	0.53±0.05	0.7±0.03
	GTM	40±8	0.57±0.03	13±7	0.36±0.06	0.57±0.03	60±8	0.68±0.06	20±8	0.54±0.07	0.68±0.06	60±8	0.68±0.06	20±8	0.54±0.07	0.68±0.06
All target-related ChEMBL subsets (Jd neighbors)	PCA	26±11	0.64±0.06	121±112	0.31±0.11	0.68±0.07	29±10	0.71±0.05	178±135	0.39±0.08	0.77±0.06	-[1]	-	-	-	-
	t-SNE	47±12	0.63±0.06	19±25	0.44±0.08	0.64±0.07	59±8	0.7±0.04	10±8	0.58±0.07	0.71±0.04	-	-	-	-	-
	UMAP	46±11	0.62±0.06	21±23	0.4±0.08	0.63±0.06	58±9	0.69±0.04	18±10	0.51±0.06	0.7±0.04	-	-	-	-	-
	GTM	52±11	0.65±0.06	22±12	0.45±0.08	0.65±0.06	59±9	0.69±0.06	24±11	0.53±0.08	0.7±0.06	-	-	-	-	-
Low-ID Merged[2] ChEMBL subsets (16287 compounds, Ed neighbors)	PCA	5±0	0.56±0.0	63±5	0.1±0.0	0.58±0.0	10±0	0.67±0.0	306±5	0.29±0.0	0.73±0.0	27±0	0.9±0.0	400±8	0.65±0.01	0.95±0.0
	t-SNE	65±0	0.58±0.0	3±0	0.59±0.01	0.58±0.0	67±0	0.65±0.0	3±0	0.63±0.01	0.65±0.0	68±0	0.82±0.0	1±0	0.67±0.01	0.82±0.0
	UMAP	56±0	0.57±0.0	4±0	0.49±0.0	0.57±0.0	59±0	0.64±0.0	4±0	0.51±0.01	0.64±0.0	59±0	0.85±0.0	144±197	0.56±0.04	0.87±0.02
	GTM	40±0	0.52±0.0	9±0	0.4±0.01	0.52±0.0	46±0	0.52±0.0	12±0	0.46±0.0	0.52±0.0	41±0	0.63±0.0	9±1	0.39±0.01	0.63±0.0
Random ChEMBL subsets average (Ed neighbors)	PCA	3±2	0.56±0.0	954±550	0.36±0.02	0.87±0.01	13±7	0.65±0.0	190±104	0.26±0.01	0.71±0.0	34±15	0.88±0.0	267±142	0.61±0.01	0.93±0.0
	t-SNE	10±2	0.54±0.01	2±1	0.16±0.04	0.54±0.01	38±2	0.68±0.02	1±0	0.48±0.07	0.68±0.02	56±6	0.84±0.01	156±152	0.59±0.01	0.86±0.01
	UMAP	8±2	0.52±0.01	2±0	0.11±0.03	0.52±0.01	34±4	0.68±0.0	50±49	0.33±0.01	0.69±0.01	51±9	0.83±0.01	185±124	0.58±0.02	0.86±0.01
	GTM	9±1	0.52±0.02	38±75	0.1±0.08	0.55±0.09	37±5	0.59±0.06	18±6	0.32±0.04	0.59±0.06	44±12	0.68±0.05	29±7	0.41±0.08	0.69±0.05
Low-ID LOLO (Ed neighbors)	PCA	18±6	0.61±0.04	118±63	0.26±0.1	0.67±0.07	29±6	0.68±0.04	137±131	0.37±0.09	0.74±0.07	55±7	0.9±0.01	124±64	0.66±0.03	0.95±0.01
	t-SNE	17±8	0.53±0.04	31±25	0.15±0.07	0.54±0.04	27±7	0.67±0.04	138±80	0.37±0.1	0.73±0.06	50±6	0.85±0.01	138±74	0.6±0.03	0.9±0.02
	UMAP	12±7	0.58±0.04	139±82	0.22±0.09	0.67±0.06	30±6	0.69±0.03	129±88	0.39±0.08	0.75±0.05	57±6	0.88±0.01	101±49	0.64±0.02	0.91±0.01
	GTM	28±6	0.58±0.04	48±34	0.27±0.07	0.61±0.06	36±6	0.6±0.04	58±72	0.36±0.06	0.63±0.06	39±3	0.64±0.02	12±12	0.37±0.03	0.65±0.03

[1] Tanimoto similarity was not used for the optimization of models built using ChemDist embeddings.

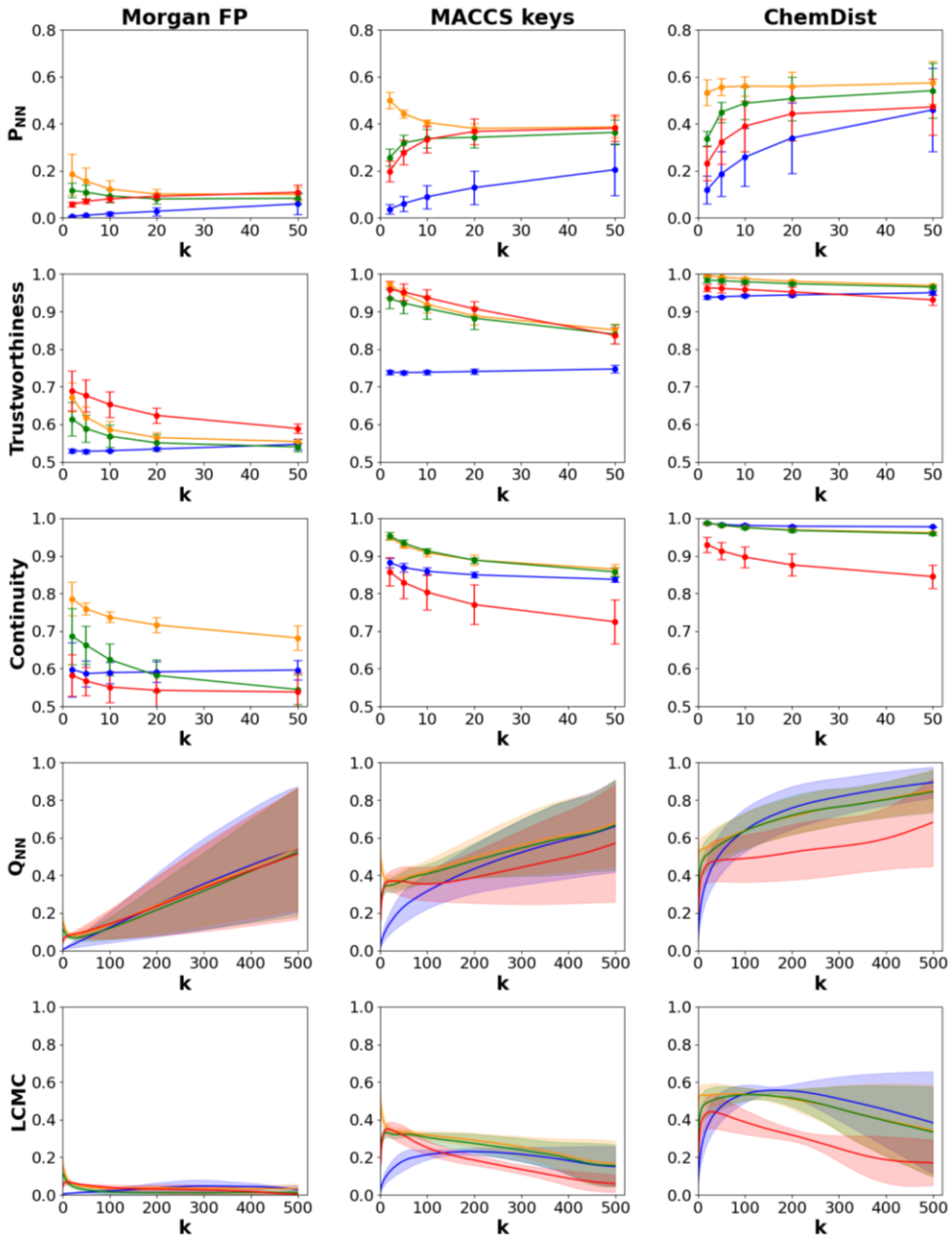
[2] For datasets with more than 2,500 compounds, a subset of 2,500 compounds was randomly selected to calculate the AUC, Q_{local}, Q_{global}, trustworthiness, and continuity values. This procedure was repeated 3 times.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0



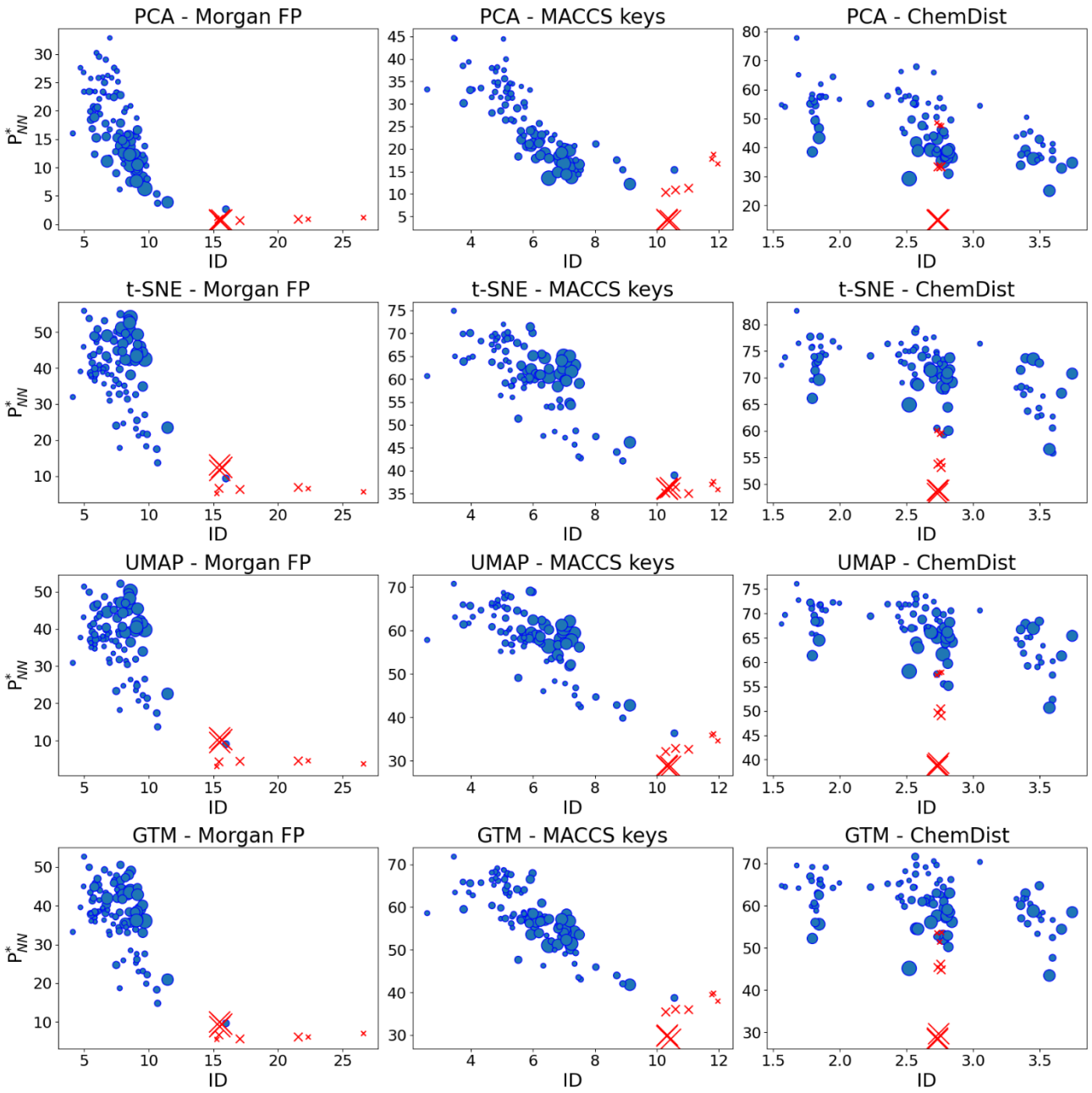
Supplementary Figure SF1. Distribution of dataset sizes (**a**) and intrinsic dimension values calculated by Fisher method values among selected ChEMBL datasets with Morgan count fingerprints (**b**), MACCS keys (c), ChemDist (**d**) used as features.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0



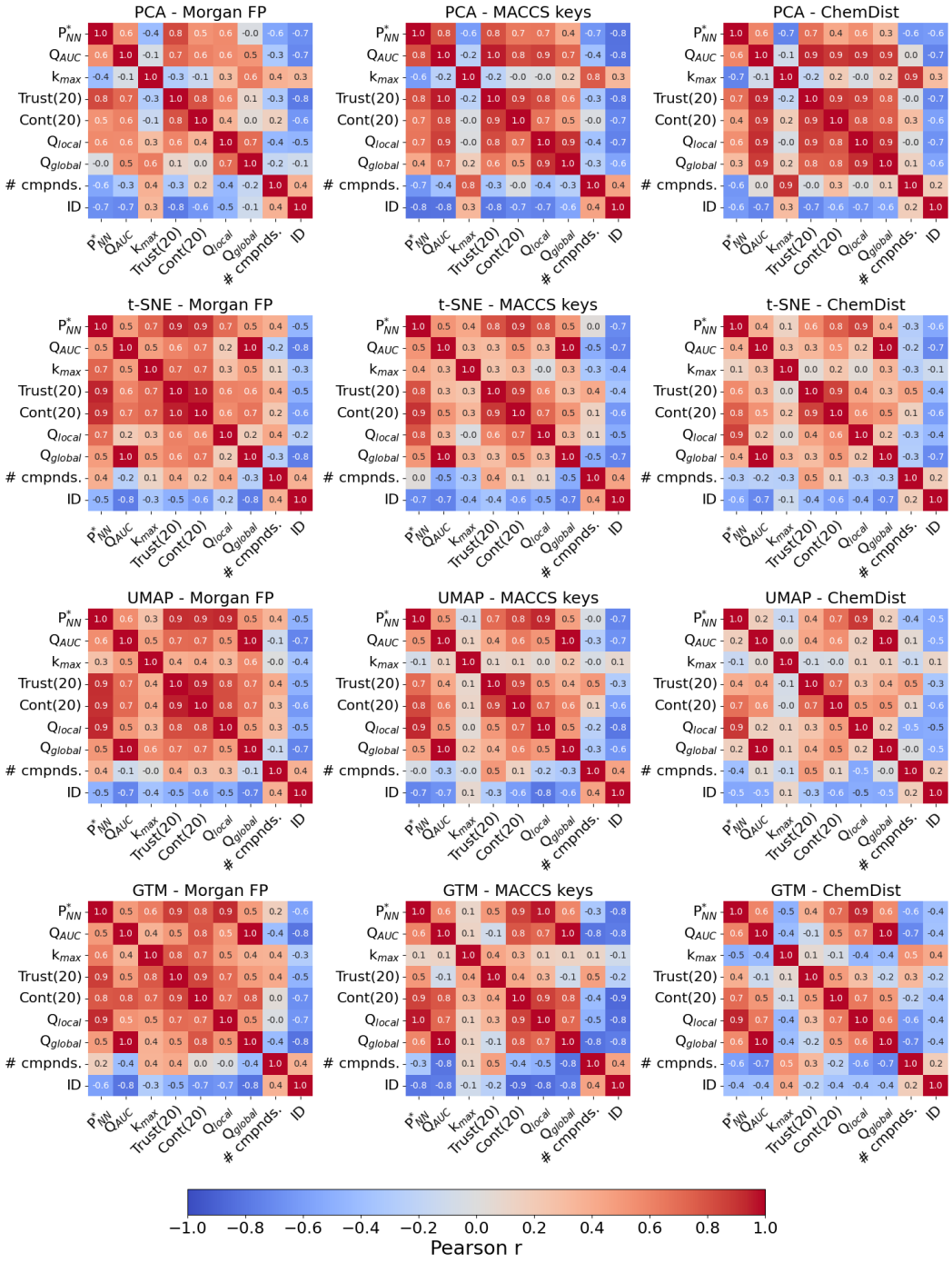
Supplementary Figure SF2. Average neighborhood preservation metrics (calculated using Euclidean distance) for nine random subsets (3x500 compounds, 3x1500 compounds, 3x9269 compounds) from ChEMBL. The models' hyperparameters were selected to maximize the preservation of neighbors among the 20 nearest ones Euclidean distance in the original space was used). Color scheme: PCA – blue, t-SNE – orange, UMAP – green, GTM – red. The ratio of nearest neighbors (PNN) preserved at different k-values, trustworthiness, continuity, co-k-nearest neighbor size (QNN), and Local Continuity Meta Criterion (LCMC) as functions of the k-nearest neighbors are shown. Standard deviation values calculated across datasets are shown as bars or filled areas. Corresponding AUC, Qlocal, Qglobal, k-max values can be found in Supplementary Table ST1.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0



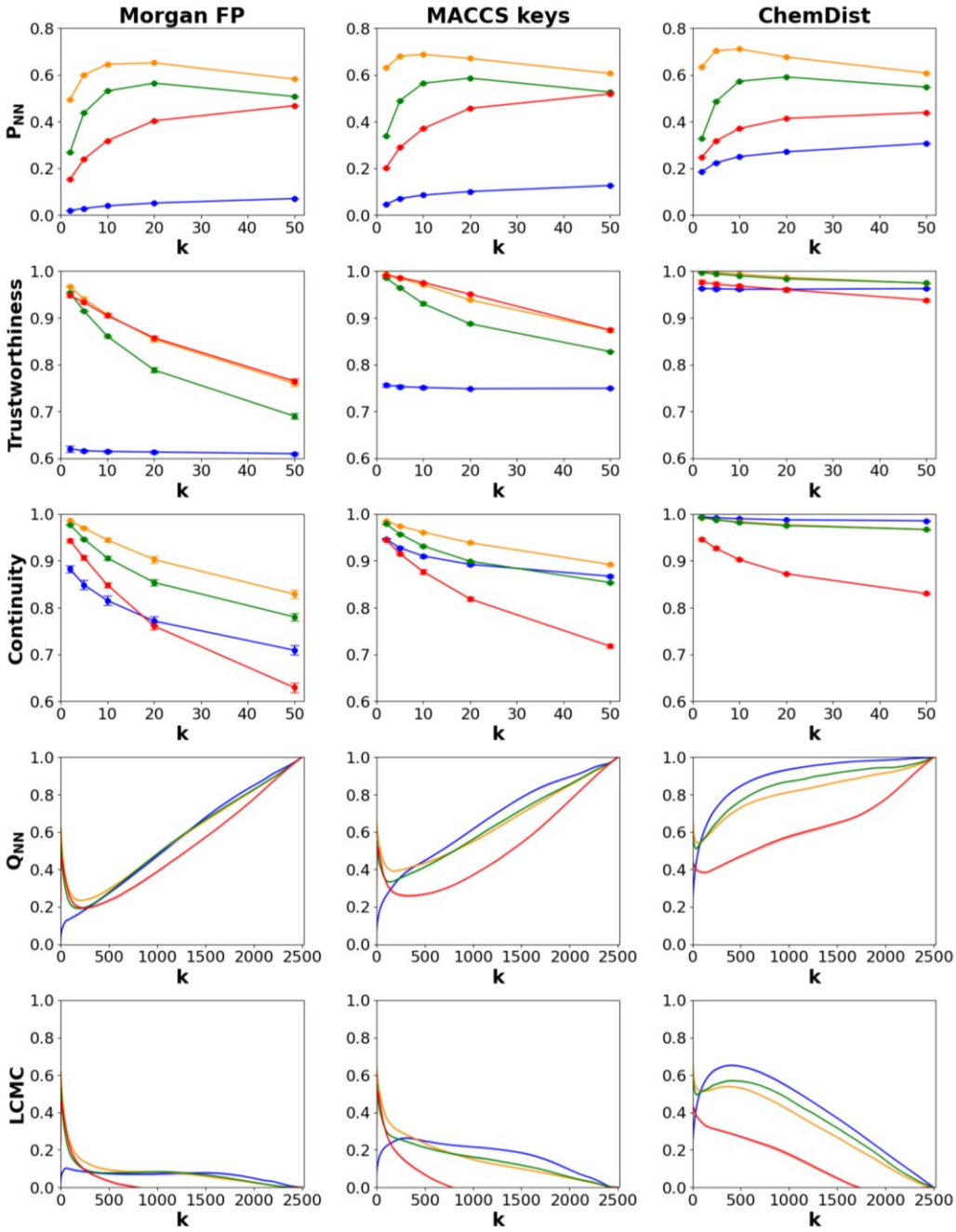
Supplementary Figure SF3. The adjusted ratio of 20 nearest preserved neighbors (P_{NN}) as a function of intrinsic dimension (ID) calculated by Fisher's algorithm across different dimensionality reduction techniques (PCA, t-SNE, UMAP, GTM) and features (Morgan fingerprints, MACCS keys, ChemDist embeddings). The size of the data points reflects the number of compounds in the dataset. The random ChEMBL subsets are shown as red crosses.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0



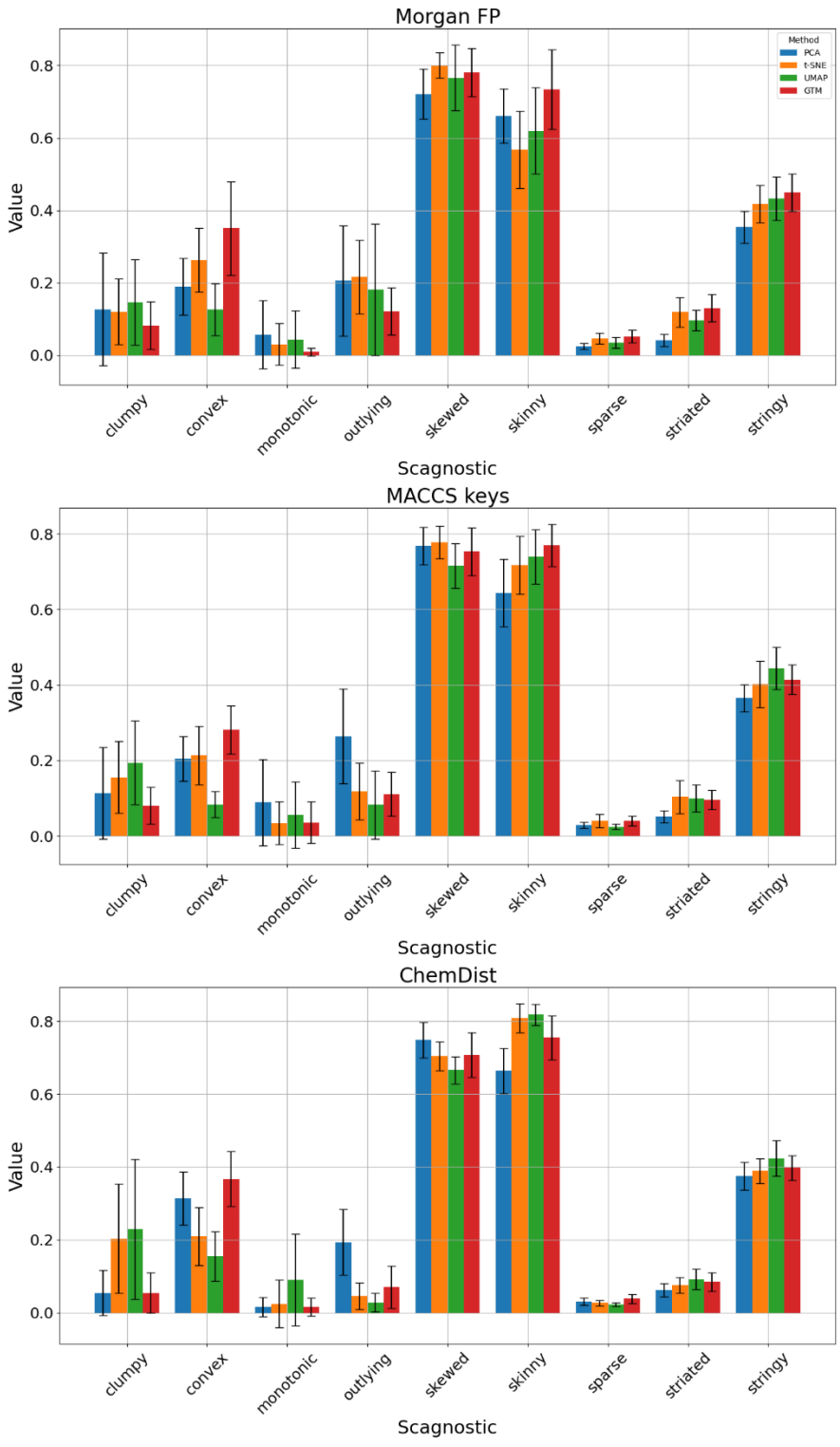
Supplementary Figure SF4. The heatmaps showing Pearson correlation coefficients between various neighborhood preservation metrics the adjusted ratio of 20 nearest preserved neighbors (P^*_{NN}), an area under co-k-nearest neighbor size curve (AUC Q_{NN}), Q_{local}, Q_{global}, k-max values, trustworthiness (Trust(20)) and continuity (Cont(20)) values for 20 nearest neighbors), number of compounds in the dataset (# of cmpnds.), and intrinsic dimension (ID) calculated by Fisher method for different dimensionality reduction techniques (PCA, t-SNE, UMAP, GTM) using various features (Morgan fingerprints, MACCS keys, ChemDist embeddings). A color scale from blue (-1) to red (1) illustrates the strength and direction of correlations.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0



Supplementary Figure SF5. Neighborhood preservation metrics for the optimized models built on 16,287 compounds combining 18 low-ID ChEMBL targets. The models' hyperparameters were selected to maximize the preservation of neighbors among the 20 nearest ones Euclidean distance in the original space was used). Color scheme: PCA – blue, t-SNE – orange, UMAP – green, GTM – red. The ratio of nearest neighbors (P_{NN}) preserved at different k-values, trustworthiness, continuity, co-k-nearest neighbor size (Q_{NN}), and Local Continuity Meta Criterion (LCMC) as functions of the k-nearest neighbors are shown. Standard deviation values calculated across datasets are shown as bars or filled areas. Corresponding AUC, Q_{local}, Q_{global}, k-max values can be found in Supplementary Table ST1.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0



Supplementary Figure SF6. The figure presents the distribution of scagnostics (scatter plot diagnostics) for low-dimensional visualizations across 94 datasets from ChEMBL, optimized to preserve the 20 closest neighbors. The subplots correspond to different types of molecular representations: Morgan count fingerprints, MACCS keys, and ChemDist embeddings. Color scheme: PCA – blue, t-SNE – orange, UMAP – green, GTM – red.

https://doi.org/10.26434/chemrxiv-2024-pxb4f **ORCID:** https://orcid.org/0000-0003-1886-925X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0