Mapping Chemical Space: Topological Data Analysis of Chemical Latent Space with Mapper

Dhruv Meduri
University of Utah, Salt Lake City, USA

Chuan-Shen Hu
Nanyang Technological University, Singapore

Cong Shen
National University of Singapore, Singapore

Kelin Xia
Nanyang Technological University, Singapore

Bei Wang[1]
University of Utah, Salt Lake City, USA

Abstract

The vast chemical space, encompassing virtually innumerable molecules and materials, presents both immense opportunities and significant challenges. The design and discovery of novel drugs and functional materials may be viewed as a search within this space; however, the sheer scale of potential candidates renders exhaustive exploration infeasible. To address this, we introduce *Chemical Mapper*, a framework that integrates topological data analysis with deep learning to enable the visual exploration and analysis of chemical latent spaces. At its core, *Chemical Mapper* employs mapper, a widely used tool in topological data analysis, to investigate the organizational principles of chemical latent spaces defined by molecular representations learned by geometric deep learning models. In doing so, *Chemical Mapper* not only highlights groups of molecular representations but also uncovers the relationships among them through linkages and branching structures. Our results show that *Chemical Mapper* reveals intrinsic patterns associated with molecular scaffolds, functional groups, and chemical properties, as well as the structural and functional evolutions of the molecules.

2012 ACM Subject Classification Applied computing → Chemistry; Mathematics of computing → Algebraic topology

Keywords and phrases Practice of computational topology, topological data analysis, applications in chemistry, mapper algorithm, high-dimensional data analysis, chemical spaces, geometric deep learning, latent space geometry

Supplementary Material https://doi.org/10.5281/zenodo.19241865

Funding *Dhruv Meduri*: NSF IIS-2205418 and DMS-2134223
Kelin Xia: Tier 1 grant RG16/23 and Tier 2 grant MOE-T2EP20125-0004
Bei Wang: NSF IIS-2205418 and DMS-2134223

1 Introduction

Navigating chemical space, which encompasses the immense number of compounds formed by all possible atomic combinations, is both challenging and full of promise [10, 24, 42, 41]. For example, the GDB-17 dataset, which includes molecules composed of up to 17 atoms of C, N, O, S, and halogens, already enumerates 166.4 billion structures [41]. Estimates suggest that more than 10^{60} molecules could satisfy Lipinski's rule-of-five and thus represent

[1] Corresponding author

potential drug candidates. Even restricting attention to drug candidates of up to 30 atoms yields a chemical space of about 10^{20}–10^{24} compounds. This staggering diversity provides an effectively unlimited reservoir of opportunities for novel drugs, functional materials, and green technologies [24, 42]. Indeed, chemical space is widely regarded as the ultimate frontier for drug discovery, as all approved drugs originate from exhaustive screening and refinement of candidates within this vast molecular universe [42].

Understanding and navigating the vast landscape of chemical space is essential for advancing drug discovery, materials science, and chemical engineering [3, 13, 24, 29, 44, 50, 52]. Broadly, there are two principal approaches for the visualization and analysis of chemical space: *global clustering* and *local topology*.

Global clustering models are frequently built on dimensionality reduction techniques [14, 23]. For instance, the *Chemical Space Project* [41] employs PCA [34] to illustrate chemical diversity, whereas *ChemPlot* [46], a Python library for chemical space visualization, supports PCA, t-SNE [47], as well as UMAP [28]. *Drug Discovery Maps* (DDM) [20] applies t-SNE to visualize the molecular similarity of experimental drugs in chemical space. More recently, AI-driven approaches have been adopted to study the chemical latent space, which embeds molecule structures into a mathematical space defined by molecular features. This chemical latent space captures structural diversity within chemical libraries and supports the exploration of broader chemical space for designing novel drug candidates [32]. Variational autoencoders (VAEs) have been widely used for both learning [15] and constructing chemical latent spaces [32]. The overarching idea is to transform chemical structures into low-dimensional or latent representations that preserve key structural and functional properties, thereby enabling visualization, clustering, and analysis. A variety of clustering methods [18, 19, 33] have been applied in this context. For example, Cheng et al. [9] used unsupervised clustering with Gaussian mixture models (GMMs) to partition chemical space, while Hadipour et al. [16] applied K-means [26] and BIRCH [51]—a scalable hierarchical clustering algorithm—to group VAE-based molecular representations. A recent extension, BitBIRCH [35], was shown to cluster 450,000 molecules in around 2.2 minutes.

In contrast, local topology models highlight connections at the level of individual compounds [25, 27, 36, 43]. For example, interaction networks among proteins, ligands, RNAs, DNAs, and other molecules are widely used to model chemical space. Beyond direct interactions, molecular graphs can also be constructed based on structural similarity. While these graph models can also support visualization and clustering, their key strength lies in preserving local molecular connectivity, which is particularly valuable for molecular search, optimization, and modification under structural or functional constraints.

The mapper construction [45] is a central tool in topological data analysis, offering a topological summary that balances global clustering with local structure. As a discrete analogue of the Reeb graph [2, 40], mapper provides a versatile framework for exploring and analyzing high-dimensional data [4, 7, 30]. Unlike conventional clustering methods, it not only identifies clusters but also encodes their interrelationships. Crucially, mapper preserves the intrinsic topology of the data space, enabling the detection of subtle structural features such as branches, chains, and loops that are often overlooked by dimensionality reduction techniques. In its simplest form, the *mapper graph* represents clusters as nodes and their overlaps as edges. Recently, mapper graphs have emerged as powerful tools for interpreting the structure of deep learning activation spaces [5, 37, 38, 54, 55] and exploring the topology of word embeddings [39].

We present *Chemical Mapper*, a mapper framework for the visual exploration and analysis of chemical latent space. The central idea is to employ mapper graphs to investigate the

organizational principles of chemical latent spaces learned by deep learning models. Specifically, we focus on chemical latent space defined by molecular representations learned from geometric deep learning models such as GEM (Geometry-Enhanced Molecular representation learning, also referred to as the Geometry Embedding Model) [13], GraphMAE [17], and Mole-BERT [48]. From these latent representations, we construct a mapper graph in which each node denotes a cluster of molecule representations and each edge encodes the overlap between clusters. Our analysis shows that edges in the mapper graph capture shared motifs, substructures, or scaffold types, while topological patterns such as branches and chains reveal systematic variations in functional groups and chemical properties.

In addition, *Chemical Mapper* enables the characterization of how molecular representations evolve during the pre-training and fine-tuning phases of deep learning models. By comparing mapper graphs across training stages, we trace the progression of molecular representations and assess their ability to generalize chemical properties. This provides new insight into the latent chemical structures learned by deep models.

Overall, this work advances topology-driven molecular analysis, highlighting the value of mapper graphs for both chemical space exploration and the interpretation of deep learning models. Our findings emphasize the importance of integrating topological data analysis with machine learning to achieve structured, interpretable molecular representations, thereby paving the way for future developments in computational chemistry, materials discovery, and AI-driven chemical modeling.

The source code and datasets associated with *Chemical Mapper* are publicly available at https://doi.org/10.5281/zenodo.19241865, along with supplementary videos that demonstrate its interactive exploration capabilities.

2 Materials and Methods

2.1 Datasets

To demonstrate the adaptability of our *Chemical Mapper* framework in exploring diverse chemical latent spaces, we employ three chemical datasets in this study: (1) PubChem, a large-scale molecular repository; (2) a toxicity dataset containing rat oral acute toxicity data; and (3) a water solubility dataset with $logS$ values that quantify how readily compounds dissolve in water.

First, to illustrate the representational capability of mapper in capturing molecular distributions, we focus on a randomly selected subset of two million chemical structures from the PubChem database [21], a comprehensive repository of small molecules. Second, to highlight the ability of *Chemical Mapper* to capture both molecular structures and associated physicochemical properties, we incorporate two additional datasets from Deep-PK [31], which focus on toxicity and water solubility. The toxicity dataset contains 10,130 molecules and measures acute oral toxicity in rats, typically expressed as the lethal dose for 50% of test animals (LD_{50}). The water solubility dataset comprises 9,964 molecules, each labeled with the logarithm of water solubility (logS) at 20-25°C, expressed in log mol/L.

2.2 AI-Enhanced Chemical Structural Representations

For structural analyses using *Chemical Mapper* (Section 3.2), we employ GEM [13] as the primary model for generating molecular structural representations. GEM utilizes a geometry-based graph neural network (GNN) architecture enhanced with self-supervised learning strategies designed to capture fine-grained molecular geometry. Molecular representations

are produced using a pre-trained GEM model for two million molecules, each represented as a 32-dimensional real-valued vector.

To further examine the training behavior of other geometric deep learning models, we also evaluate GraphMAE [17] and Mole-BERT [48] under pre-trained and fine-tuned conditions (Section 3.3). GraphMAE, inspired by masked autoencoding techniques in natural language processing, learns effective graph representations through self-supervised learning by masking portions of node features and training the network to reconstruct them. In contrast, Mole-BERT adopts a transformer-based architecture to capture rich structural and chemical information. Drawing inspiration from BERT, it employs masking strategies to model contextual relationships within molecular graphs, enabling more flexible and comprehensive molecular feature learning.

2.3 Mapper Algorithm

The mapper construction is a powerful topological data analysis technique that captures the intrinsic shape of high-dimensional datasets [45]. Its 1D form, the mapper graph, provides a simplified graphical summary of complex point clouds, where nodes represent clusters and edges capture overlapping relationships among them. When applied to chemical space, the mapper graph enables hierarchical exploration of molecular scaffolds, functional groups, and properties, revealing meaningful topological patterns and chemical insights.

The mapper algorithm takes as input a high-dimensional point cloud $\mathbb{X}$ equipped with a real-valued function $f : \mathbb{X} \to \mathbb{R}$, referred to as a *filter function*. It produces a graphical summary of $\mathbb{X}$ in the form of a *mapper graph*, where each node represents a cluster interpreted as a topological neighborhood, and an edge connects two nodes if their corresponding neighborhoods overlap.

We illustrate the construction of a mapper graph using a 1D point cloud $\mathbb{X}$ sampled from the letter $\mathcal{Y}$, rendered in a mathematical script font. As shown in Figure 1 (left), $\mathbb{X}$ is equipped with a height function f, which serves as the filter function. To obtain a topological summary of $\mathbb{X}$, we begin with a finite cover $\mathcal{V} = \{V_1, V_2, \dots\}$ of $f(\mathbb{X})$, formed by a set of partially overlapping intervals such that $f(\mathbb{X}) \subseteq \bigcup_j V_j$ (see Figure 1, middle). For each interval V_j, we consider the clusters induced by points in $f^{-1}(V_j)$; these clusters form a finite cover $\mathcal{U} = \{U_1, U_2, \dots\}$ of $\mathbb{X}$ such that $\mathbb{X} \subseteq \bigcup_j U_j$, highlighted by rectangles. For example, points in $f^{-1}(V_1)$ form two clusters, U_1 and U_2, while those in $f^{-1}(V_2)$ yield four. The 1D *nerve* of $\mathcal{U}$ defines the *mapper graph* in Figure 1 (right), where each cluster U_i corresponds to a node i, and an edge connects nodes i and j whenever clusters U_i and U_j have a nonempty intersection. For example, an edge connects the nodes representing clusters U_2 and U_3. The resulting mapper graph captures key topological features of the letter $\mathcal{Y}$, including its branching and looping structures.

In our study, we use the l^2-norm as the filter function, which quantifies how strongly a model is activated by an input molecule and has been shown to yield meaningful insights when analyzing latent representations in deep learning [38, 39]. Given a fixed filter function, the mapper graph construction is determined by: (1) the number of intervals and their overlap, which control the resolution of the cover; and (2) the parameters of the clustering algorithm. The overlap among cover elements is crucial for maintaining connectivity among topological neighborhoods at a given resolution, thereby revealing both local and global topological structure. In practice, we employ DBSCAN [12] for clustering, with parameters ϵ and `minPts`, where ϵ defines the neighborhood radius determining how close points must be to belong to the same cluster, and `minPts` specifies the minimum number of points required to form a dense cluster. See Section 4.1 for a discussion on parameter turning.

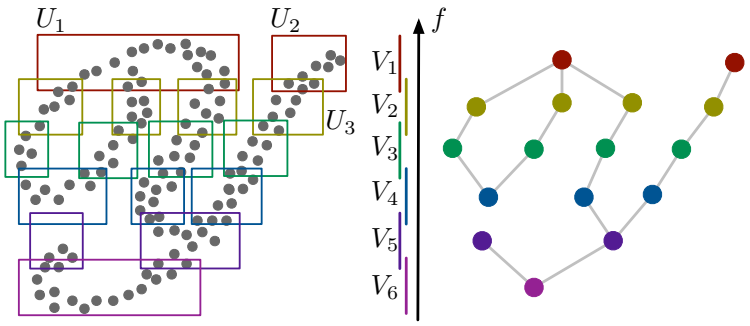

Figure 1 A simple illustration of mapper graph construction. Left: a point cloud $\mathbb{X}$ is equipped with a height function f. Middle: a cover $\mathcal{U} = \{U_1, U_2 \dots\}$ of $\mathbb{X}$ is induced by a cover $\mathcal{V} = \{V_1, V_2, \dots\}$ of $f(\mathbb{X})$ consisting of 6 intervals with 25% overlap. Right: the 1D nerve of $\mathcal{U}$ yields the mapper graph.

3 Results

3.1 Topological Data Analysis with Mapper

The mapper construction offers a key advantage in its ability to integrate topological structure into clustering analysis. As an initial study, we compare t-SNE [47] and mapper [45] for visualizing the chemical space of molecular representations.

Our study uses two million chemical structures from the PubChem database [22], comprising ten distinct scaffolds (i.e., core molecular frameworks), each represented by 200,000 molecules. Each chemical graph is converted into a 32-dimensional feature vector using the pre-trained GEM model. Detailed experimental settings are provided in Section 2.

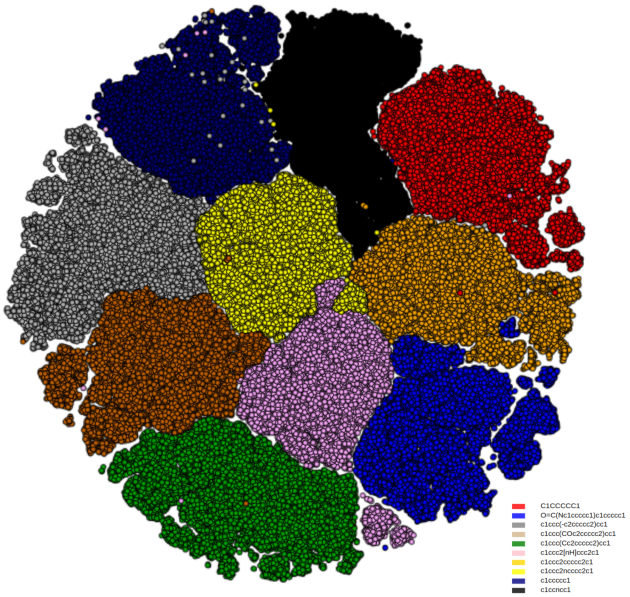

Figure 2 Application of t-SNE to two million chemical structures from the PubChem database. The embedding reveals scaffold-based clusters but offers limited insight into both intra- and inter-cluster relationships.

In the t-SNE visualization shown in Figure 2, each data point is colored by its chemical scaffold. In contrast, the mapper graphs used throughout this work represent each node as a group of molecular representations, visualized as a pie chart whose color proportions reflect the distribution of scaffolds within the node (see Figures 10 and 11), except in Section 3.2.3. Nodes rendered in a single color indicate high purity, meaning all molecules share the same scaffold (e.g., the mapper graph in Figure 5 contains only one impure node out of 5,104 total nodes). Although nodes are displayed with a uniform size by default, they may contain varying numbers of molecules, and their sizes can be adjusted if needed. The corresponding scaffold colormap, along with representative molecular structures, is shown in Figure 3.

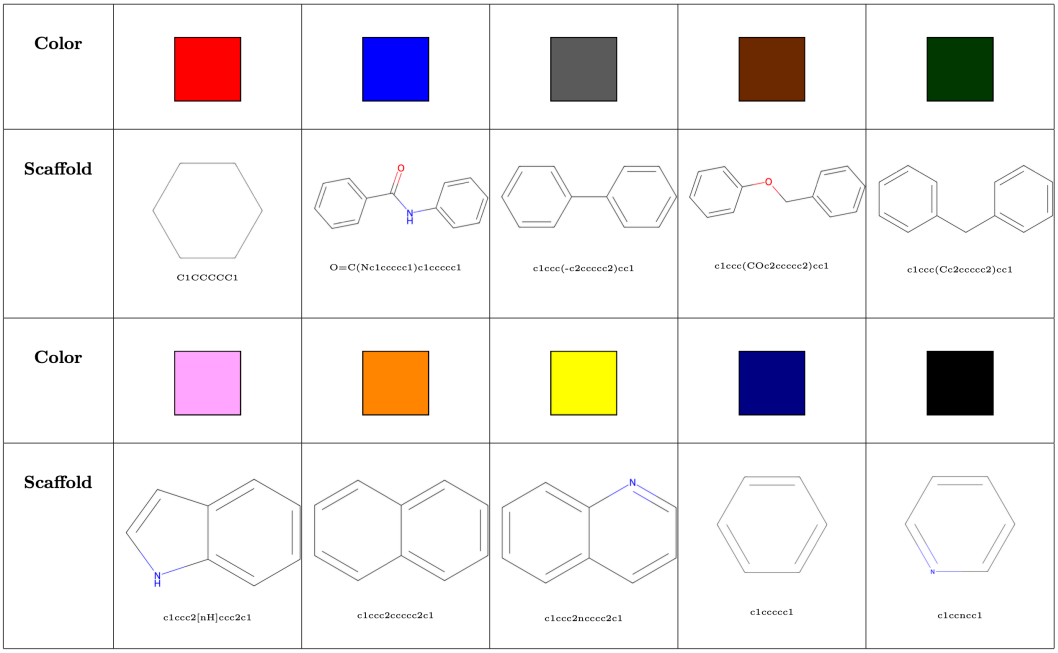

Figure 3 Scaffold colormap with representative molecular structures.

While t-SNE organizes molecular representations into ten well-separated clusters, it does not capture relationships among them. In contrast, the mapper graph in Figure 5 reveals both inter- and intra-cluster connectivity and exposes finer-grained topological features—such as branches, chains, and loops—that capture meaningful structural distinctions within the chemical space. By capturing both local and global organization, the mapper graph offers a more comprehensive view of molecular structure, surpassing the representational capacity of conventional clustering methods. In subsequent sections, our results demonstrate that the mapper graph effectively uncovers topological organization within the chemical space, revealing clusters and their relationships and providing deeper insights into molecular structure and function (see Sections 3.2 and 3.3 for details).

3.2 Mapping GEM Chemical Latent Space with Mapper

We integrate molecular representations learned from geometric deep learning models into the mapper algorithm to create an exploratory framework, *Chemical Mapper*. As illustrated in Figure 4, the workflow begins with molecular datasets from which latent representations are extracted using deep learning models. These representations form a high-dimensional point cloud that is analyzed using the mapper algorithm. The resulting mapper graph

captures relationships among scaffolds, functional groups, and molecular properties. Downstream analyses leverage this graph to reveal structural patterns and property distributions, highlighting the framework's capacity to uncover organizational principles within chemical latent space.

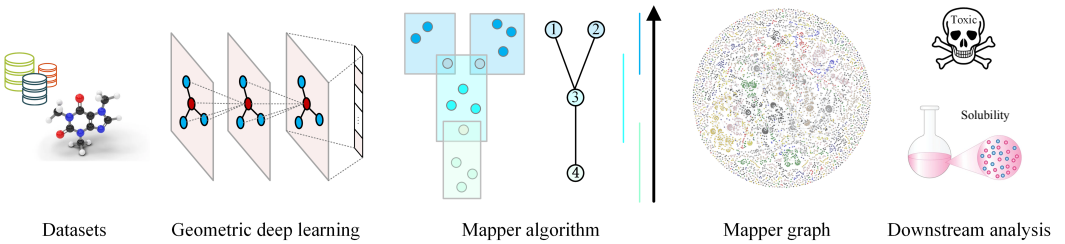

Figure 4 The workflow of the AI-enhanced *Chemical Mapper* for chemical space analysis. From left to right: input of molecular structures; extraction of molecular representations using geometric deep learning models; application of the mapper algorithm to the resulting chemical latent space to construct a mapper graph; and downstream analysis of chemical structures and properties (such as solubility and toxicity) informed by the topology of the mapper graph.

We analyze a large-scale dataset of two million molecules spanning ten scaffolds from the PubChem database [22]. Each molecule is embedded into a 32-dimensional latent space using a pre-trained GEM model. Figure 5 shows the mapper graph constructed from these representations using 100 intervals with 50% overlap, $\epsilon = 1$, and `minPts` $= 10$.

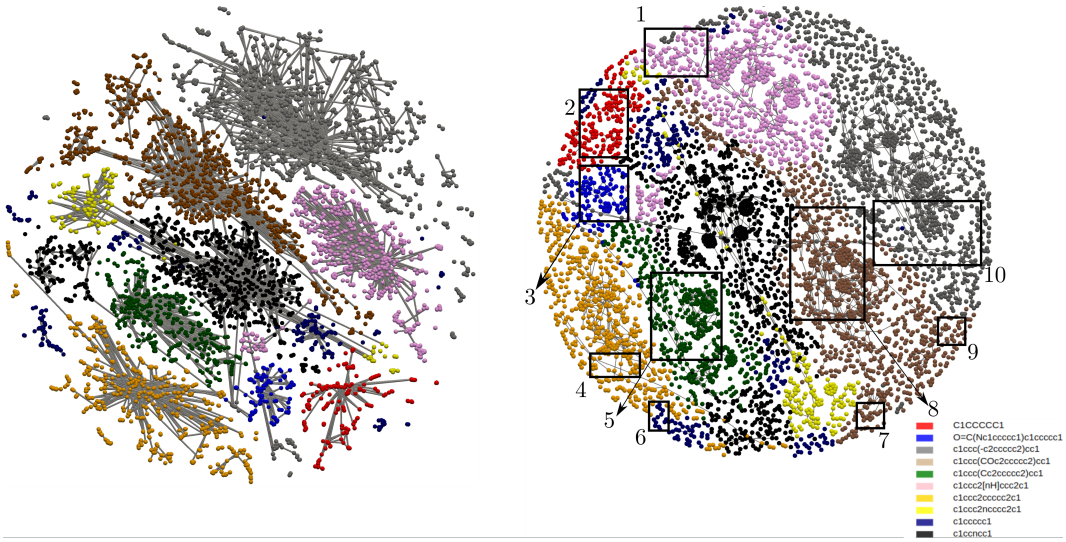

Figure 5 Mapper graph of two million molecular representations from a pre-trained GEM model. The left visualization uses an anchored layout, while the right shows the same graph with a force-directed layout. Numbered regions on the right are examined in detail in subsequent sections.

In Figure 5(left), the mapper graph is drawn using an anchored layout, where each node is initialized at the mean of the 2D t-SNE coordinates of its assigned data points, effectively grounding the graph in the underlying embedding. Figure 5(right) shows the corresponding force-directed layout, which refines this initialization by iteratively adjusting node positions through attractive forces along edges and repulsive forces between nodes to

improve separation and reduce visual clutter. For the remainder of the paper, we use the force-directed layout, as it yields clearer and more interpretable visualizations.

We analyze the topological structure of this graph and its relationship to variations in the chemical latent space, focusing on connected components (Section 3.2.1) and branches (Section 3.2.2). We further examine how the graph organization correlates with chemical functions and properties such as toxicity and solubility (Section 3.2.3).

3.2.1 Exploring Connected Components in Chemical Latent Space

As illustrated in Figure 5, the mapper graph consists of multiple connected components, each corresponding to a group of molecules sharing a common scaffold in the latent space. Within a given scaffold, molecules further partition into components of varying sizes, revealing finer-grained structural distinctions.

Figure 6 highlights four connected components, each containing representative molecules drawn from a shared scaffold but distinguished by their functional groups. The top-left component is enriched with molecules featuring amide bonds (O=C–NH–), whereas the bottom-left component consists exclusively of molecules containing sulfate groups (O=S=O). On the right, the top component contains molecules that consistently feature a combination of carboxylic acid (O=C–OH), epoxide (–C–O–C–), and halide substituents. The bottom-right component similarly comprises molecules that simultaneously contain nitro groups (O–N(=O)–O), carboxylic acid, and halide substitutions. This demonstrates that a component can be characterized by a set of functional groups rather than a single one. Together, these topological patterns indicate that the GEM model has learned not only to separate molecules by their underlying scaffolds but also to organize them hierarchically according to their functional groups.

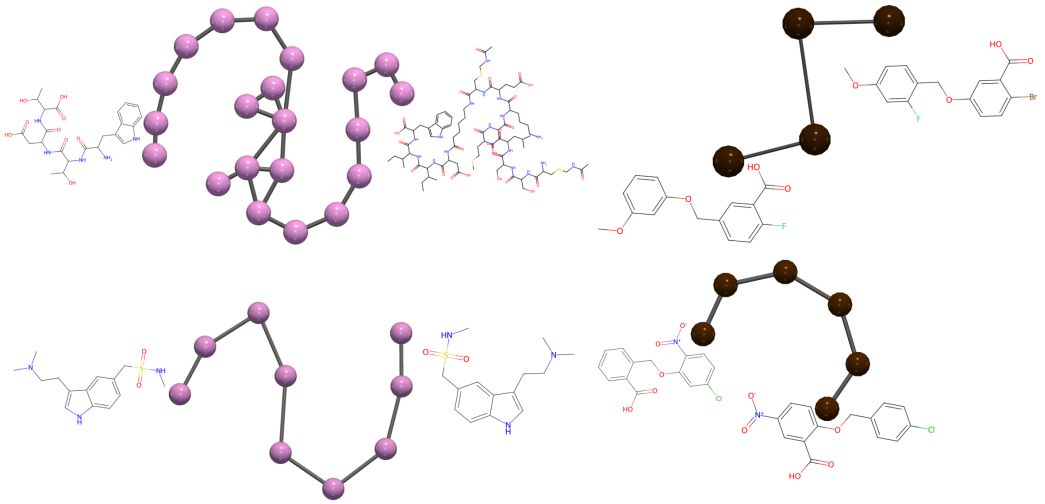

Figure 6 Four chain-like connected components with representative molecules identified by *Chemical Mapper* on the pre-trained GEM model. The components on the left are extracted from region 1 in Figure 5 (right), while the components on the right correspond to regions 9 (top) and 7 (bottom), respectively.

3.2.2 Exploring Chains and Branches in Chemical Latent Space

A *chain* in a mapper graph refers to a linear arrangement of nodes forming a sequential, chain-like structure, whereas *branches* are chains that emerge from a common source node. We investigate the evolution of chemical structures across mapper edges in the latent space by analyzing these chains and branches in the mapper graph.

The components shown in Figure 6 exhibit chain-like topologies, each consisting of a nearly linear sequence of nodes. In the top-left component, the number of amide bonds (O=C–NH–) within the molecules gradually increases from one end of the chain to the other. The GEM model not only groups molecules with a dominant presence of amide bonds into the same component but also differentiates them according to the number of such bonds. This represents a specific instance of a broader pattern in the chemical space, where similar trends are observed across multiple regions of the mapper graph.

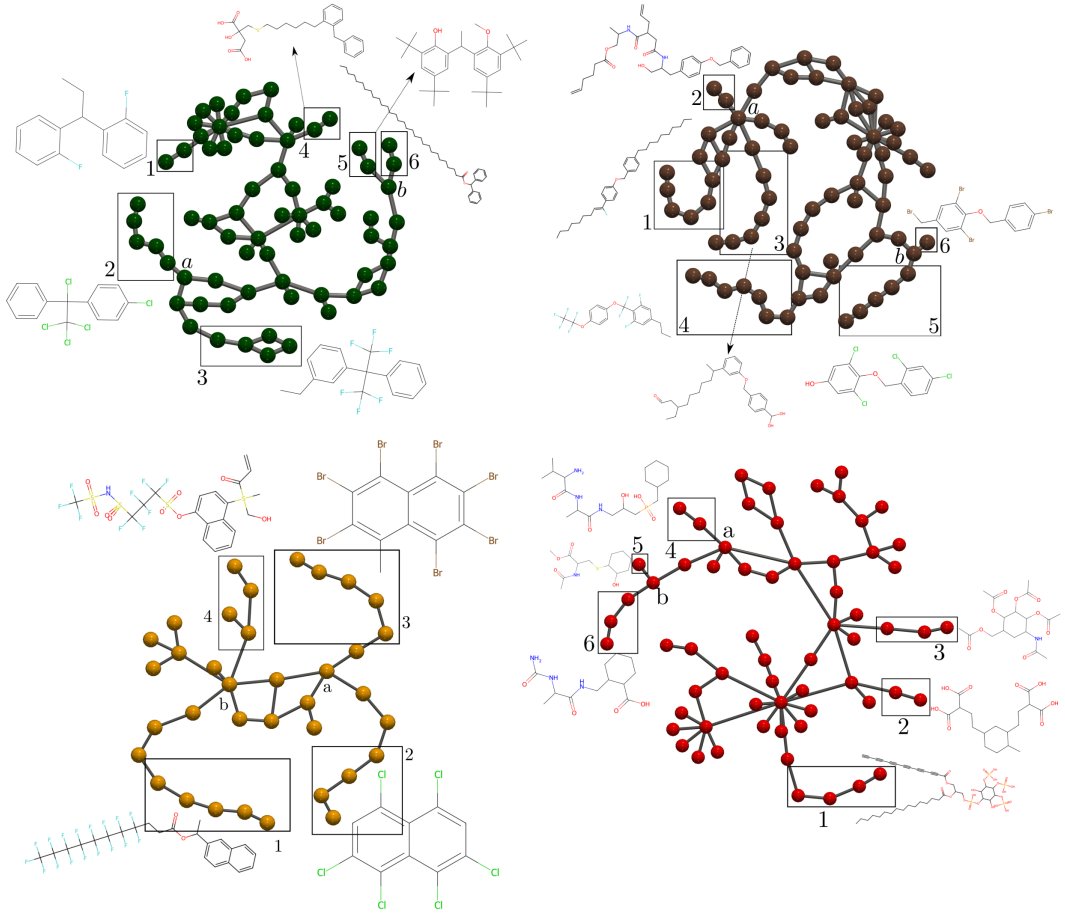

Figure 7 Four connected components with prominent branching structures identified by *Chemical Mapper* on the GEM model. The components are extracted from region 5 (top-left), region 8 (right), region 4 (bottom-left), and region 2 (bottom-right) in Figure 5 (right). Branches are highlighted using numbered boxes, and representative molecules are displayed adjacent to their corresponding branches or indicated with arrows.

In contrast, branches in the mapper graph represent bifurcations that correspond to distinct functional groups. As shown in Figure 7, the numbered branches (highlighted by boxes) reveal fine-scale structural variations among molecular clusters. The top-left

component features multiple branches that share the same scaffold (c1ccc(Cc2ccccc2)cc1) but differ in their functional groups. Specifically, branch 1 contains molecules with fluoride groups (–F), whereas branch 4 includes molecules with carboxylic groups (COOH). Branches 5 and 6 emerge from the branching node *b*, with branch 5 comprising molecules rich in quaternary carbon groups and branch 6 containing molecules with long carbon chains.

These bifurcations not only distinguish functional groups with markedly different chemical structures but also separate those with closely related frameworks that differ only by atomic composition. For instance, from the branching node *a*, branch 2 primarily contains molecules with chloride (–Cl) and trichloride ($-CCl_3$) groups, whereas branch 3 consists of molecules featuring trifluoride ($-CF_3$) groups.

Similarly, the top-right component in Figure 7 displays multiple branches that reflect comparable bifurcation patterns of functional groups. Branches 1, 2, and 3 originate from node *a*, containing molecules characterized by long carbon chains, amides (O=C–NH–), and alcohols (–OH), respectively. Branches 5 and 6 bifurcate from node *b*, with branch 5 including molecules containing chlorides (–Cl) and branch 6 those containing bromides (–Br). Branch 4 comprises molecules with a prominent presence of fluoride (e.g., –F, $-CF_5$) groups.

Similarly, the bottom-left component of Figure 7 is composed of molecules with a high halogen content, and it is notable that all halogen-bearing molecules cluster together within a single component. Within this component, the mapper graph indicates a hierarchical separation of molecules according to the specific halogen substituent. Branch 1 consists of molecules featuring fluorine (–F), branch 2 contains molecules with chloride (–Cl), and branch 3 includes those with bromine (–Br). Notably, branches 2 and 3 both emerge from node *a*. Although they differ in the halogen substituent, their molecules remain structurally similar because the halogens are directly attached to the underlying aromatic scaffold. Branch 4 comprises molecules that contain both fluorine (–F) and sulfone (O=S=O) groups. Branches 1 and 4 originate from node *b*, reflecting their shared fluorine-containing molecules.

A similar pattern of organization appears in the bottom-right component of Figure 7. Branches 1, 2, and 3 contain molecules enriched in phosphate (PO_4) and long carbon chains, carboxylic acid (O=C–OH), and aldehyde (O=CH) functional groups, respectively. In contrast, branches 4, 5, and 6 all contain amide functional groups (O=C–NH–). Rather than forming a single branch, these three branches diverge hierarchically, reflecting additional chemical variations among the molecules. Branch 4 diverges first at node *a*, indicating the presence of a prominent phosphoryl functional group (PO_2), in addition to the amide group. From the remaining branch emerging from node *a*, a further split occurs at node *b*. This subsequent bifurcation reflects an additional layer of chemical variation: branch 5 contains molecules featuring sulfur substituents (–S–), whereas branch 6 includes molecules enriched in carboxylic acid functional groups (O=C–OH). This hierarchical branching structure shows how the mapper graph progressively separates molecules based on both their dominant and secondary functional groups.

3.2.3 Mapper for Chemical Function Analysis

Given the strong relationship between molecular structure and chemical function, mapper graphs provide a powerful framework for interpreting chemical properties. We use the water solubility and toxicity datasets from Deep-PK [31], comprising 9,964 molecules with solubility labels and 10,130 molecules with toxicity labels. Both properties are reported on a logarithmic scale. Using these datasets, we train the GEM model [13] to predict properties for the two million molecules analyzed in this study. Specifically, 90% of the labeled data are used for training and the remaining 10% for validation, with model parameters optimized

based on validation performance. Once the GEM model reaches its optimal accuracy, it is employed to predict water solubility and toxicity for unlabeled molecules, thereby assigning property labels to previously uncharacterized chemical structures for downstream analysis.

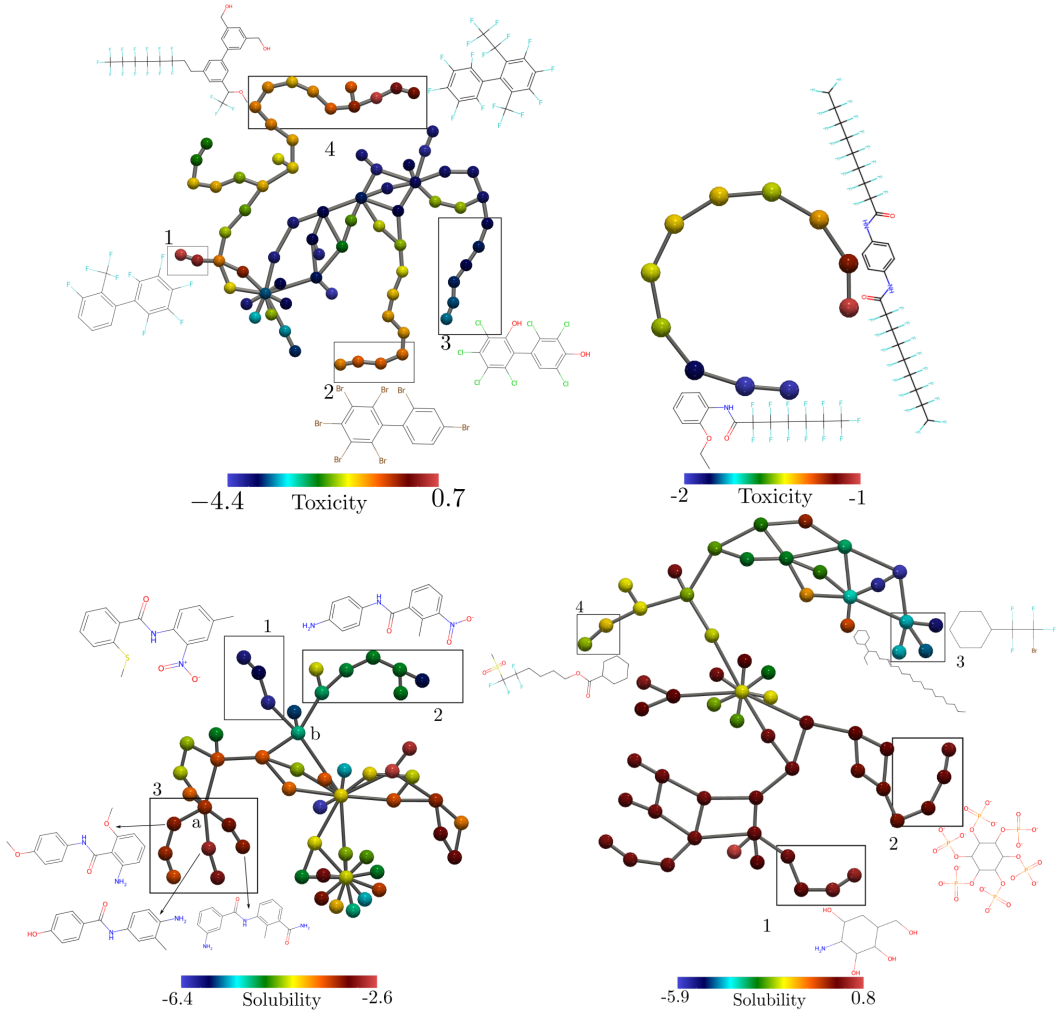

Figure 8 Four connected components with prominent branching structures or chains identified by *Chemical Mapper* and colored according to averaged chemical properties. The components are extracted from region 10 (top-left), region 6 (top-right), region 3 (bottom-left), and region 2 (bottom-right) in Figure 5 (right). Branches are highlighted with numbered boxes, and representative molecules are displayed adjacent to their corresponding branches or indicated with arrows.

Our key finding is that the mapper graph effectively visualizes the organization of molecules with similar chemical properties into cohesive regions, while molecules spanning different property ranges are distributed across distinct branches. This organization aligns closely with structural characteristics: the branching patterns in the mapper graph reflect variations in functional groups, the primary determinants of chemical behavior.

Figure 8 illustrates four connected components from the mapper graph in Figure 5. In the top panel, nodes are colored by molecular toxicity, and in the bottom, by water solubility; colormaps are adjusted for each component to better highlight diversity. In both cases, nodes within each branch share similar colors or exhibit gradual color transitions, indicating smooth variation in the corresponding property. In Figure 8 (top-left), branch 1 contains highly toxic

molecules, likely due to the strong presence of fluoride groups (–F). Branch 2, which shows higher toxicity than branch 3, is enriched with bromide-containing (–Br) molecules, while branch 3 is characterized by chloride (–Cl) substitution. Branch 4 is particularly notable: toxicity increases progressively along the branch. Although all molecules in this branch contain fluoride groups, the less toxic end includes additional functional groups, whereas the highly toxic end consists solely of fluoride-bearing (–F) molecules. Representative molecular structures are displayed alongside each branch.

In Figure 8 (top-right), the component forms a single linear chain in which toxicity increases steadily along its length. This gradual rise in toxicity parallels the chemical trend observed in the structures: all molecules in the chain contain fluorine substituents (–F), but the number of fluorine functional groups increases progressively from one end of the chain to the other. As a result, molecules at the high-toxicity end exhibit a substantially heavier fluorine presence compared to those at the low-toxicity end.

A similar pattern is observed in the component in Figure 8 (bottom-right). Within each branch, nodes display uniform or gradually changing colors. Branch 1 consists of highly soluble molecules, primarily due to the presence of alcohol (–OH) and amine (–NH_2) groups. Branch 2 also contains highly soluble molecules, likely attributable to phosphate (–PO_4) groups. This is interpretable, as phosphate-containing molecules (–PO_4) appear in high-solubility regions, due to their charged and polar nature. These functional groups correspond to identifiable subgraphs (C–O–H and P–O–O patterns) that the GEM model learns to encode, forming distinct branches in the mapper graph. In Area 3, the branches are dominated by molecules with halide substitutions (e.g., –Br, –F) or very long carbon chains (–$(CH_2)_n$–), which generally correspond to low water solubility. Branch 4 contains molecules of moderate solubility, as they appear to contain sulfone (–SO_2–) groups that increase solubility, but also halides that reduce it, resulting in an intermediate solubility profile.

A comparable pattern emerges in Figure 8 (bottom-left) component, where each branch contains molecules with either uniform or smoothly varying solubility levels. Branch 1 consists of molecules with low solubility, primarily due to the presence of nitro groups (–NO_2). Branch 2, which also contains nitro-bearing (–NO_2) molecules, exhibits moderately higher solubility because these molecules additionally feature amine functional groups (–NH_2), which enhance aqueous solubility. Note, branches 1 and 2 diverge from node b, reflecting their shared nitro substitution. Area 3 contains three branches, each composed of molecules with high solubility. These molecules predominantly feature epoxide (–C–O–C–), alcohol (–OH), or amide (O=C–NH–) functional groups. All three branches emerge from node a, consistent with the fact that their molecular structures are closely related, differing mainly in the specific solubility-enhancing functional group present.

Finally, for completeness, we present the full mapper graph of the chemical latent space from the pre-trained GEM model, where nodes are colored according to predicted toxicity (left) and solubility (right) in Figure 9.

3.3 Exploring Chemical Latent Spaces from Different Learning Models

In this section, we demonstrate how mapper graphs can serve as a powerful tool for characterizing and evaluating molecular representations learned by geometric deep learning models such as GraphMAE and Mole-BERT. We further show that mapper graphs enable qualitative comparisons of the learning capabilities between pre-trained and fine-tuned models.

Following the procedure for the GEM model (Section 3.2), we apply *Chemical Mapper* to the molecular representations of the same two million molecules from the PubChem

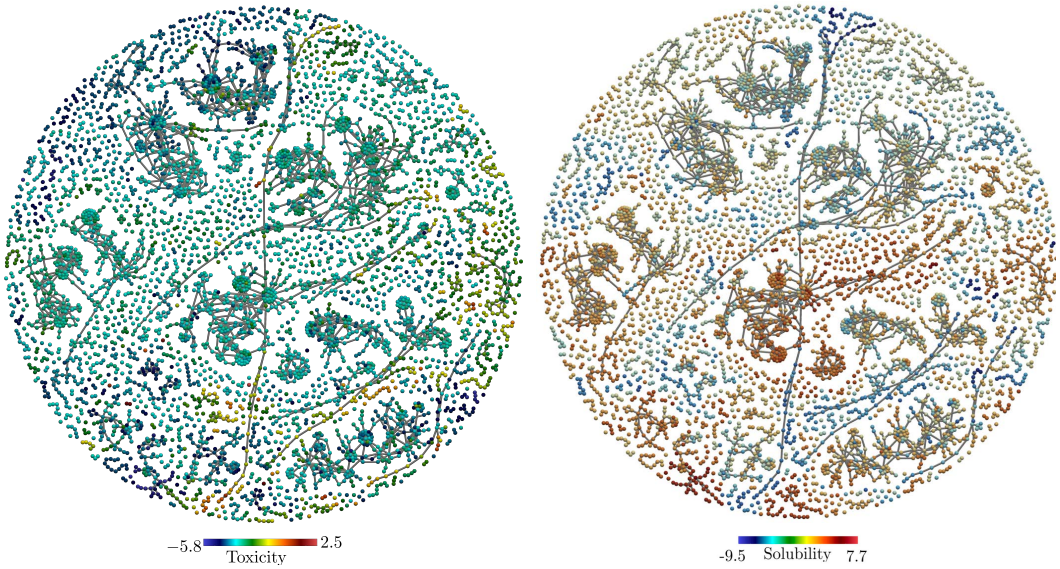

Figure 9 Mapper graphs of molecular representations from the pre-trained GEM model, computed using the same experimental settings as in Figure 5. Nodes are colored according to molecular toxicity (left) and solubility (right), as discussed in Section 3.2.3 and Section 2.

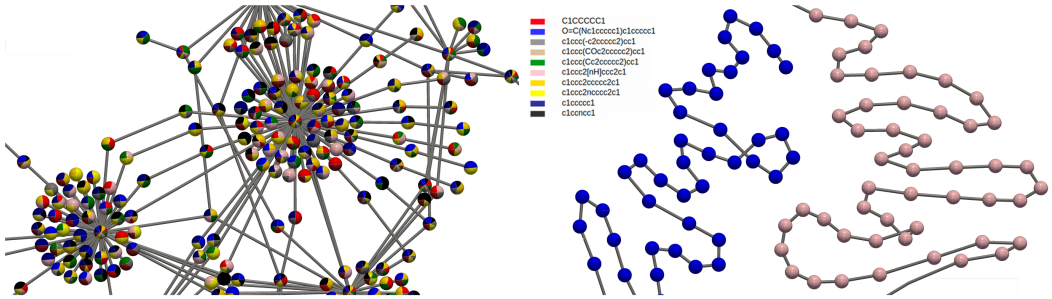

Figure 10 Partial views of mapper graphs derived from latent representations of the pre-trained (left) and fine-tuned (right) GraphMAE models. Fine-tuning refines and restructures the latent space, revealing its evolution during training. Both graphs are computed with 500 intervals, 40% overlap, and DBSCAN parameters $\epsilon = 5$ and `minPts` $= 5$.

database [22], learned by the pre-trained GraphMAE model. A subgraph of the resulting mapper graph is shown in Figure 10 (left), where each node is represented as a pie chart indicating the distribution of molecular scaffolds within that node. Each node clearly contains a mixture of scaffolds, suggesting that the pre-trained GraphMAE model does not separate the chemical space according to scaffold structure (cf. Section 3.2.1). Moreover, the mapper graph exhibits no clear branching patterns, in contrast to the structured branches observed for the GEM model (Section 3.2.2).

Subsequently, the pre-trained GraphMAE model is fine-tuned on the task of predicting molecular scaffolds, and representations for the same two million molecules are extracted from the fine-tuned model. Using identical parameters, a new mapper graph is constructed. Figure 10 (right) shows a portion of this graph, where the nodes are noticeably more homogeneous, with molecules sharing the same scaffold clustering together. Fine-tuning the

GraphMAE model, therefore, induces a clearer partitioning of the chemical space according to molecular scaffolds. The resulting mapper graph contains 1,606 nodes—a substantial increase from the 420 nodes in the mapper graph of the pre-trained model—indicating that fine-tuning expands the representational space and more effectively separates molecules with different scaffolds.

Furthermore, the mapper graph derived from the fine-tuned model consists predominantly of chains and exhibits few branches. As discussed in Section 3.2.2, branching in the chemical space reflects distinctions among functional groups. Because the fine-tuning task focuses solely on scaffold prediction and does not consider functional groups, the lack of branching suggests that the chemical space is organized primarily around scaffold structure rather than functional group variation.

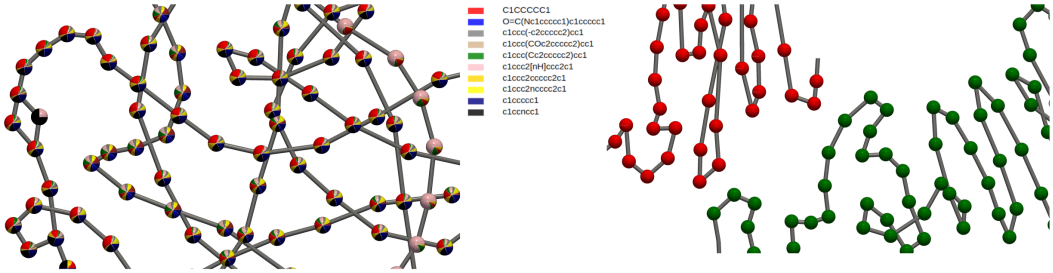

Figure 11 Partial views of mapper graphs of molecular representations from the pre-trained (left) and fine-tuned (right) Mole-BERT model. Nodes are colored by scaffolds.

We conduct a similar experiment with Mole-BERT on the same two million PubChem molecules, obtaining results qualitatively consistent with GraphMAE. Figure 11 compares mapper graphs from the pre-trained (left) and fine-tuned (right) Mole-BERT models for the scaffold classification task.

For completeness, we may also compare the pre-trained and fine-tuned GEM model. Figure 12 shows the complete mapper graph of the chemical latent space derived from the GEM model fine-tuned on the scaffold classification task, with nodes colored by molecular scaffolds. We perform a comparative analysis between the pre-trained and fine-tuned GEM models. As shown in Figure 5 and Figure 12, the mapper graph of the fine-tuned GEM model exhibits the same qualitative trends observed in the pre-trained model (Section 3.2): molecules with different scaffolds remain clearly separated into distinct connected components, and branching structures continue to reflect variations in functional groups.

Because the chemical space produced by the pre-trained GEM model already partitions molecules according to scaffolds (as shown in Figure 5), fine-tuning on this task does not substantially affect the purity of the cluster nodes. However, comparing the mapper graphs in Figure 5 and Figure 12 reveals notable topological changes. Certain scaffold classes become more dispersed, exhibiting a larger number of isolated nodes and fewer interconnections (e.g., c1ccccc1), whereas others contract significantly (e.g., c1ccc(-c2ccccc2)cc1 and c1ccc(COc2ccccc2)cc1). Investigating the causes and implications of these topological changes represents an important direction for future research.

4 Conclusion, Discussion and Future Work

In conclusion, the immense size and complexity of chemical space present substantial challenges for analysis and discovery. In this work, we introduce *Chemical Mapper*, a mapper

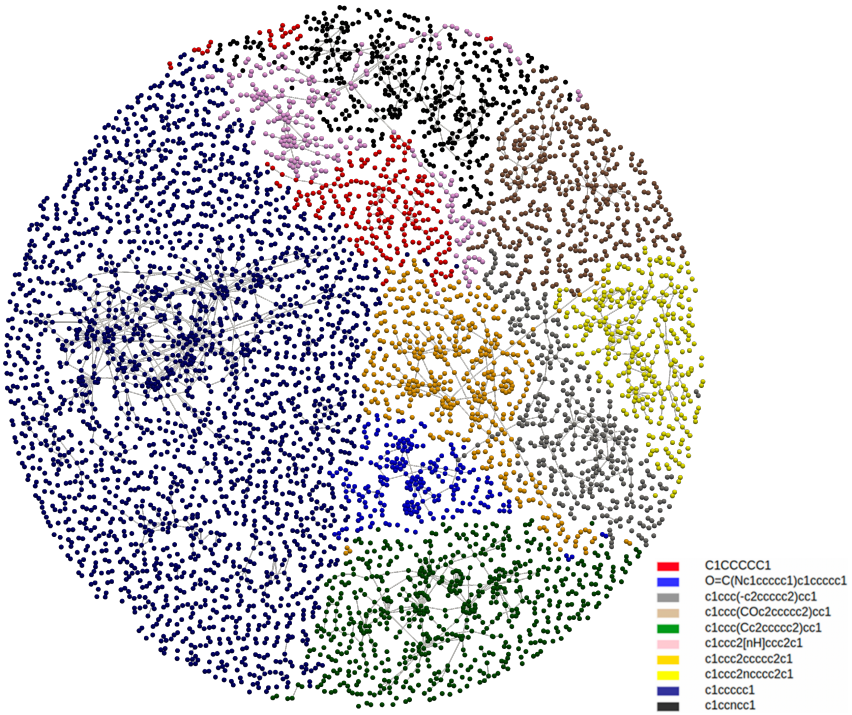

Figure 12 Mapper graph of molecular representations from the fine-tuned GEM model, colored by molecular scaffolds.

framework that enables effective visualization, search, exploration, and interpretation of chemical latent spaces. Our approach offers several key advantages. First, similar to traditional clustering methods, the mapper graph organizes molecules with similar properties into coherent groups. Second, beyond simple clustering, the graph's topological structures, such as connected components and branches, capture the continuous variations and evolution among molecular structures. Third, by jointly encoding both structural clustering and topological organization, mapper graphs provide a powerful lens for understanding chemical functions and properties. Finally, *Chemical Mapper* serves not only as an analysis tool but also as a diagnostic framework for evaluating chemical representations: it allows qualitative assessment of how well deep learning models capture meaningful chemical patterns.

Our framework provides a principled and interpretable approach for navigating chemical latent space, revealing its intrinsic organization and supporting both the development and evaluation of molecular deep learning methods. We conclude with a discussion of parameter tuning, stability, extensibility, and opportunities for enhanced human–AI collaboration.

4.1 Parameter Tuning

The *Chemical Mapper* framework involves several key parameters: the number of intervals m, the percentage of overlap p, and the DBSCAN clustering parameters, including the neighborhood radius ϵ and the minimum number of points `minPts` required to form a core point. These parameters are collectively denoted as $(m, p, \epsilon, \texttt{minPts})$. In practice, they are often hand-tuned (as in this study), with best practice being to select values within ranges where the mapper graph structures (such as branches, loops, and connected components) remain stable. Prior works have investigated automated parameter selection for mapper

graphs [1, 6, 7, 8], and ϵ can be estimated using the elbow method [12, 53].

4.2 Stability

Theoretical studies have investigated the structural stability of mapper graphs [4, 7] and their multiscale variants [11]. In particular, mapper graphs are stable under perturbations of the data $(\mathbb{X}, f)$, with stability depending on how the cover $\mathcal{U}$ aligns with the critical values of f [4]. However, these theoretical guarantees rely on statistical assumptions that are often not satisfied in real-world datasets. Motivated by this gap, we provide empirical evidence of stability in Figure 13.

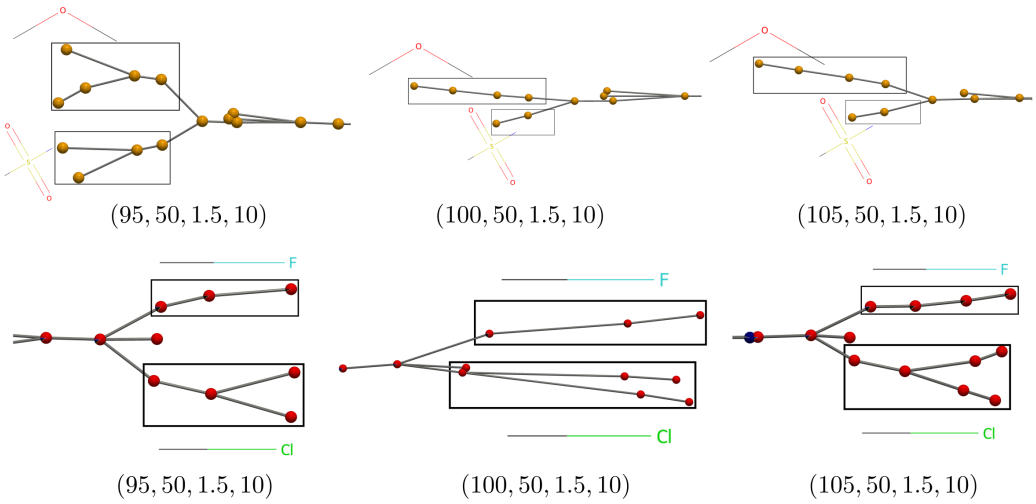

Figure 13 Two examples of local branches in mapper graphs generated under varying parameter settings $(m, p, \epsilon, \texttt{minPts})$, where p, ϵ, and `minPts` are fixed and m varies from 95 to 105. In both top and bottom cases, a component with the same branching structure and the same set of molecules persists across different m values, illustrating the stability of the mapper construction.

Building on these findings, we examine the stability of mapper graphs under varying parameter configurations using two million molecular representations from PubChem, generated by the pre-trained GEM model (same as the dataset used in Section 3). As shown in Figure 13, local branches in the mapper graphs remain consistent across suitable ranges of m, where each branch corresponds to molecules sharing a functional group, indicated alongside the branch. In particular, the first row in Figure 13 highlights connected components containing molecules with –O– and –(O=S=O)– functional groups, exhibiting the same bifurcation pattern across three values of m. Similarly, the second row in Figure 13 shows a bifurcation between molecules containing –F and –Cl groups across different interval numbers.

Overall, this experiment provides an initial demonstration of the robustness of the mapper framework in capturing chemically meaningful topological structures across parameter settings. Although parameter variations affect the granularity of branching, the key separations corresponding to distinct functional groups remain stable. Future work should investigate how chemical semantics relate to parameter choices, guiding principled parameter selection and further enhancing the mapper's interpretability and applicability in cheminformatics.

4.3 Extendability and Human-AI Collaboration

We plan to extend *Chemical Mapper* to study structural motifs beyond scaffolds, including functional groups as well as aromatic and reactivity motifs. The recent Explainable Mapper framework [49] provides an effective strategy for analyzing high-dimensional latent spaces by integrating topological representations with agentic large language model (LLM)–based explanations. In the chemical domain—where latent spaces may encode relationships among molecular structure, reactivity, and function—this approach opens up rich opportunities for human–AI collaboration. In such a setting, rather than relying solely on human experts to explain mapper elements (Section 3.2), agents powered by LLMs or VLMs (Vision–Language Models) can propose distinguishing descriptors (e.g., fingerprints, substructures, or property ranges) and generate hypotheses about the chemical attributes that unify a given mapper element—such as the connected components, chains, and branches described in Section 3.2. Human experts can then validate or refine these hypotheses using their domain knowledge. We leave this direction for future work.

References

1 Enrique Alvarado, Robin Belton, Emily Fischer, Kang-Ju Lee, Sourabh Palande, Sarah Percival, and Emilie Purvine. G-Mapper: Learning a cover in the mapper construction. *SIAM Journal on Mathematics of Data Science*, 7(2):572–596, 2025.

2 Silvia Biasotti, Daniela Giorgi, Michela Spagnuolo, and Bianca Falcidieno. Reeb graphs for shape analysis and applications. *Theoretical Computer Science*, 392(1-3):5–22, 2008.

3 Maria Boulougouri, Pierre Vandergheynst, and Daniel Probst. Molecular set representation learning. *Nature Machine Intelligence*, 6(7):754–763, 2024.

4 Adam Brown, Omer Bobrowski, Elizabeth Munch, and Bei Wang. Probabilistic convergence and stability of random mapper graphs. *Journal of Applied and Computational Topology*, 5(1):99–140, 2021.

5 Rickard Brüel Gabrielsson and Gunnar Carlsson. Exposition and interpretation of the topology of neural networks. In *18th IEEE International Conference On Machine Learning And Applications*, pages 1069–1076, 2019.

6 Quang-Thinh Bui, Bay Vo, Hoang-Anh Nguyen Do, Nguyen Quoc Viet Hung, and Vaclav Snasel. F-Mapper: A fuzzy mapper clustering algorithm. *Knowledge-Based Systems*, 189(C), 2020.

7 Mathieu Carriere and Steve Oudot. Structure and stability of the one-dimensional Mapper. *Foundations of Computational Mathematics*, 18:1333–1396, 2018.

8 Nithin Chalapathi, Youjia Zhou, and Bei Wang. Adaptive covers for mapper graphs using information criteria. In *IEEE International Conference on Big Data*, pages 3789–3800, 2021.

9 Lixue Cheng, Jiace Sun, and Thomas F. III Miller. Accurate molecular-orbital-based machine learning energies via unsupervised clustering of chemical space. *Journal of Chemical Theory and Computation*, 18(8):4826–4835, 2022.

10 Connor W Coley. Defining and exploring chemical spaces. *Trends in Chemistry*, 3(2):133–145, 2021.

11 Tamal K. Dey, Facundo Mémoli, and Yusu Wang. Multiscale mapper: Topological summarization via codomain covers. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 997–1013, 2016.

12 Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

13 Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

14 Andrej Gisbrecht and Barbara Hammer. Data visualization by nonlinear dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(2):51–73, 2015.

15 Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

16 Hamid Hadipour, Chengyou Liu, Rebecca Davis, Silvia T. Cardona, and Pingzhao Hu. Deep clustering of small molecules at large-scale via variational autoencoder embedding and k-means. *BMC Bioinformatics*, 23(4):132, 2022.

17 Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604, 2022.

18 Anil K Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

19 Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323, 1999.

20 Antonius P. A. Janssen, Sebastian H. Grimm, Ruud H. M. Wijdeven, Eelke B. Lenselink, Jacques Neefjes, Constant A. A. van Boeckel, Gerard J. P. van Westen, and Mario van der Stelt. Drug Discovery Maps, a machine learning model that visualizes and predicts kinome–inhibitor interaction landscapes. *Journal of Chemical Information and Modeling*, 59(3):1221–1229, 2019.

21 Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023.

22 Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. PubChem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, 2016.

23 John A Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

24 Christopher Lipinski and Andrew Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432(7019):855–861, 2004.

25 Yu Liu, Cole Mathis, Michał Dariusz Bajczyk, Stuart M. Marshall, Liam Wilbraham, and Leroy Cronin. Exploring and mapping chemical space with molecular assembly trees. *Science Advances*, 7(39):eabj2465, 2021.

26 Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

27 Gerald M. Maggiora and Jürgen Bajorath. Chemical space networks: a powerful new paradigm for the description of chemical space. *Journal of Computer-Aided Molecular Design*, 28:795–802, 2014.

28 Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

29 Oscar Méndez-Lucio, Mazen Ahmad, Ehecatl Antonio del Rio-Chanona, and Jörg Kurt Wegner. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3(12):1033–1039, 2021.

30 Elizabeth Munch and Bei Wang. Convergence between categorical representations of Reeb space and mapper. In Sándor Fekete and Anna Lubiw, editors, *32nd International Symposium on Computational Geometry*, volume 51 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 53:1–53:16, Dagstuhl, Germany, 2016. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

31 Yoochan Myung, Alex G.C. de Sá, and David B. Ascher. Deep-PK: deep learning for small molecule pharmacokinetic and toxicity prediction. *Nucleic Acids Research*, page gkae254, 2024.

32 Toshiki Ochiai, Tensei Inukai, Manato Akiyama, Kairi Furui, Masahito Ohue, Nobuaki Matsumori, Shinsuke Inuki, Motonari Uesugi, Toshiaki Sunazuka, Kazuya Kikuchi, et al. Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity. *Communications Chemistry*, 6(1):249, 2023.

33 Gbeminiyi John Oyewole and George Alex Thopil. Data clustering: application and trends. *Artificial Intelligence Review*, 56(7):6439–6475, 2023.

34 Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

35 Kenneth López Pérez, Vicky Jung, Lexin Chen, Kate Huddleston, and Ramón Alain Miranda-Quintana. BitBIRCH: efficient clustering of large molecular libraries. *Digital Discovery*, 4:1042–1051, 2025.

36 Daniel Probst and Jean-Louis Reymond. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics*, 12(1):12, 2020.

37 Emilie Purvine, Davis Brown, Brett Jefferson, Cliff Joslyn, Brenda Praggastis, Archit Rathore, Madelyn Shapiro, Bei Wang, and Youjia Zhou. Experimental observations of the topology of convolutional neural network activations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9470–9479, 2023.

38 Archit Rathore, Nithin Chalapathi, Sourabh Palande, and Bei Wang. TopoAct: Visually exploring the shape of activations in deep learning. *Computer Graphics Forum*, 40(1):382–397, 2021.

39 Archit Rathore, Yichu Zhou, Vivek Srikumar, and Bei Wang. TopoBERT: Exploring the topology of fine-tuned word representations. *Information Visualization*, 22(3):186–208, 2023.

40 Georges Reeb. Sur les points singuliers d'une forme de pfaff completement integrable ou d'une fonction numerique [on the singular points of a completely integrable pfaff form or of a numerical function]. *Comptes Rendus Acad. Sciences Paris*, 222:847–849, 1946.

41 Jean-Louis Reymond. The chemical space project. *Accounts of Chemical Research*, 48(3):722–730, 2015.

42 Jean-Louis Reymond, Ruud van Deursen, Lorenz C. Blum, and Lars Ruddigkeit. Chemical space as a source for new drugs. *Medicinal Chemistry Communications*, 1:30–38, 2010.

43 Vincent F Scalfani, Vishank D Patel, and Avery M Fernandez. Visualizing chemical space networks with RDKit and NetworkX. *Journal of Cheminformatics*, 14(1):87, 2022.

44 Cong Shen, Jiawei Luo, and Kelin Xia. Molecular geometric deep learning. *Cell Reports Methods*, 3(11), 2023.

45 Gurjeet Singh, Facundo Mémoli, and Gunnar E. Carlsson. Topological methods for the analysis of high dimensional data sets and 3D object recognition. *Eurographics Symposium on Point-Based Graphics*, 2:091–100, 2007.

46 Murat Cihan Sorkun, Dajt Mullaj, J. M. Vianney A. Koelman, and Süleyman Er. ChemPlot, a Python library for chemical space visualization. *Chemistry-Methods*, 2(7):e202200005, 2022.

47 Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

48 Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. Mole-BERT: Rethinking pre-training graph neural networks for molecules. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

49 Xinyuan Yan, Rita Sevastjanova, Sinie van der Ben, Mennatallah El-Assady, and Bei Wang. Explainable Mapper: Charting LLM embedding spaces using perturbation-based explanation and verification agents. arXiv preprint arXiv:2507.18607, 2025.

50 Shuwen Yang, Ziyao Li, Guojie Song, and Lingsheng Cai. Deep molecular representation learning via fusing physical and chemical information. *Advances in Neural Information Processing Systems*, 34:16346–16357, 2021.

51 Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997.

52 Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3D molecular representation learning framework. In *Proceedings of the 11th International Conference on Learning Representations*. International Conference on Learning Representations, 2023.

53 Youjia Zhou, Nithin Chalapathi, Archit Rathore, Yaodong Zhao, and Bei Wang. Mapper Interactive: A scalable, extendable, and interactive toolbox for the visual exploration of high-dimensional data. In *IEEE 14th Pacific Visualization Symposium*, pages 101–110, 2021.

54 Youjia Zhou, Helen Jenne, Davis Brown, Madelyn Shapiro, Brett Jefferson, Cliff Joslyn, Gregory Henselman-Petrusek, Brenda Praggastis, Emilie Purvine, and Bei Wang. Comparing mapper graphs of artificial neuron activations. In *Topological Data Analysis and Visualization*, pages 41–50, 2023.

55 Youjia Zhou, Yi Zhou, Jie Ding, and Bei Wang. Visualizing and analyzing the topology of neuron activations in deep adversarial training. In Timothy Doster, Tegan Emerson, Henry Kvinge, Nina Miolane, Mathilde Papillon, Bastian Rieck, and Sophia Sanborn, editors, *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning*, volume 221 of *Proceedings of Machine Learning Research*, pages 134–145. PMLR, 2023.