MULTIPARAMETER PERSISTENT HOMOLOGY FOR MOLECULAR PROPERTY PREDICTION

Andac Demir, Bulent Kiziltan
AI & Computational Sciences (AICS), Novartis Institutes for BioMedical Research (NIBR)
Cambridge, MA, 02139, USA
{andac.demir,bulent.kiziltan}@novartis.com

ABSTRACT

In this study, we present a novel molecular fingerprint generation method based on multiparameter persistent homology. This approach reveals the latent structures and relationships within molecular geometry, and detects topological features that exhibit persistence across multiple scales along multiple parameters, such as atomic mass, partial charge, and bond type, and can be further enhanced by incorporating additional parameters like ionization energy, electron affinity, chirality and orbital hybridization. The proposed fingerprinting method provides fresh perspectives on molecular structure that are not easily discernible from single-parameter or single-scale analysis. Besides, in comparison with traditional graph neural networks, multiparameter persistent homology has the advantage of providing a more comprehensive and interpretable characterization of the topology of the molecular data. We have established theoretical stability guarantees for multiparameter persistent homology, and have conducted extensive experiments on the Lipophilicity, FreeSolv, and ESOL datasets to demonstrate its effectiveness in predicting molecular properties.

1 INTRODUCTION

Drug discovery process is the early phase of the pharmaceutical R&D pipeline that takes $10-15$ years end-to-end and costs in excess of ~ 2 billion US dollars (Berdigaliyev & Aljofan, 2020). Early phases of drug discovery process begin with identifying possible drug targets (gene/protein) and their role in the disease progression. Target identification is followed by high-throughput screening (HTS) experiments (via either structure- or ligand-based virtual screening methods) to identify the drug candidates within large compound libraries that affect the target in the desired way, also called as "hits" or "leads". In the final phases of the R&D pipeline, drug candidates have to pass a series of rigorous controlled tests in clinical trials to be considered for regulatory approval by FDA. Overall, drug discovery is a complex and error-prone process. Specifically, in oncology only 4% of therapies entering phase 1 clinical trials ultimately gains provisional approval from FDA (Mullard, 2016) due to experimental disproof of the promised therapeutic efficacy or unforeseen side effects. Another challenging problem which arises in this domain is that the known chemical space including public databases and proprietary data owned by private organizations probably contains on the order 100 million molecules, while the chemical space might contain as many as 10^{60} compounds obeying Lipinski's rule-of-five for oral bioavailability (Reymond & Awale, 2012).

Molecular property prediction has received substantial interest in recent years to accelerate the drug discovery process (Stokes et al., 2020) and predict the 3D structure of proteins (Jumper et al., 2021), with models showing potential to solve contemporary problems in materials science (Schmidt et al., 2019) and quantum chemistry (Dral, 2020). Most early studies utilized canonical compound representations such as SMILES (Zheng et al., 2019) Morgan fingerprints (Zhang et al., 2019) or eigenspectrum of Coulomb matrices (Montavon et al., 2012) as low level molecular descriptors. More recently, variants of Graph Neural Networks (GNN) such as GIN (Xu et al., 2018; Peng et al., 2020), GAT (Velickovic et al., 2017; Wang et al., 2021) and MPNN (Yang et al., 2019) exhibited state-of-the-art performance for molecular representation learning across several compound datasets on bioactivity and molecular property prediction. Despite their success, utilization of GNNs for molecular property prediction suffers from 2 severe limitations:

1. Neighborhood information is typically aggregated by permutation invariant, but non-injective operations such as an average, sum or max. This leads to an oversmoothing problem, where node embeddings converge to similar values and the information-to-noise ratio of the message received by the nodes decreases (Chen et al., 2020; Ogawa et al., 2021).
2. GNNs operating on the chemical structure of molecules hold the underlying assumption that atoms are connected together via chemical electronic bonds (ionic and covalent bonds). In fact, spatial arrangement of atoms in a molecule is also affected by the ubiquitous but subtle van der Waals forces between atoms. Although the van der Waals forces are comparatively weaker forces than ionic and covalent bonds, they play a fundamental role in specifying the molecular geometry, also known as conformation. A number of studies have shown that using $3D$ molecular conformers significantly improves the accuracy of molecular property prediction (Schütt et al., 2017; Gasteiger et al., 2020; Liu et al., 2021; Gasteiger et al., 2021). However, generating several low energy stable $3D$ molecular conformers is computationally infeasible for large scale applications (Xu et al., 2021; Shi et al., 2021; Ganea et al., 2021). Some alternative methods extract more $3D$ information by using bond lengths (Chen et al., 2019), bond angles (Gasteiger et al., 2020) or torsion angles (Gasteiger et al., 2021) as edge features.

Our main objective is to investigate whether topological data analysis (TDA) tools, in particular persistent homology (PH), can overcome the limitations of GNNs for molecular representation learning. In this paper, we propose utilizing multiparameter persistent homology to produce novel topological fingerprints of molecules and evaluate their performance as well as suitability for property prediction on benchmark datasets: Lipophilicity, FreeSolv and ESOL. Persistent homology analyzes the topological features of a graph at multiple scales, represented by a filtration of the graph. A filtration is a sequence of nested subspaces of the data, where each subspace includes the previous ones and adds new elements. The topological features are then calculated for each subspace in the filtration, and the ones that persist across multiple scales are considered to be significant.

We use the Vietoris-Rips (VR) graph filtration to construct a sequence of simplical complexes of a molecular graph. VR graph filtration is a simple and effective way, as it only requires specifying a distance threshold and does not require any prior knowledge of the topology of the data besides the shortest path distances of all node pairs. In this approach, the number of bonds between two vertices (atoms) is used as a measure of the distance between them. For example, if there is a direct edge (a bond) between two vertices, the distance between them is considered to be 1. If there is no direct edge between two vertices, the distance between them is considered to be infinity. In the VR graph filtration, a molecular graph is constructed by connecting vertices that are within a specified distance threshold from each other. This graph represents the first subspace in the filtration. The next subspace is then obtained by increasing the distance threshold and adding new connections to the graph. This process is repeated to obtain a sequence of nested graphs, which form the filtration. This results in a sequence of graphs that capture the topological features of the original graph at different scales.

This paper extends our previous work, namely **ToDD**, for ligand based virtual screening, which predicts the binding affinity between a target protein and a small molecule with no prior information of the $3D$ structure of the target protein (Demir et al., 2022).

The key contributions of this paper are:

1. We develop a novel method to generate molecular fingerprints using persistent homology, which reveals hidden structures and relationships in the molecular geometry and uncovers topological features that persist across multiple scales. The most important advantage of persistent homology over GNNs is that it provides a more comprehensive and interpretable characterization of the topology of the data.
2. We extend traditional persistent homology by analyzing topological features that persist across multiple scales along multiple parameters (atomic mass, partial charge and bond type). Additionally, these parameters can be augmented using periodic properties such as ionization energy and electron affinity as well as molecular information such as chirality, orbital hybridization, number of Hydrogen bonds or number of conjugated bonds at the cost of computational complexity. Our fingerprinting method based on multiparameter persistent homology reveals new insights into the molecular structure that are not easily apparent from a single parameter or scale analysis.

3. We establish theoretical guarantees for the stability of compound fingerprints extracted by multiparameter persistent homology.
4. We perform extensive experiments on Lipophilicity, FreeSolv and ESOL datasets to demonstrate the effectiveness of the proposed compound fingerprints for molecular property prediction.

2 Related Work

There are 2 main approaches to predict the physical and chemical properties of molecules: 1) Applying well-known models like support vector machines or gradient boosting regression trees to expert-engineered descriptors or molecular fingerprints and 2) optimizing the model architecture of GNNs.

In the first approach, the models are applied to molecular fingerprints, such as the Dragon descriptors (Mauri et al., 2006) or Morgan (ECFP) fingerprints (Rogers & Hahn, 2010). One direction of improvement in this approach is to use domain expertise and augment the feature representation of nodes (atoms) with more chemical information. Additionally, some studies have used explicit $3D$ atomic coordinates to further improve performance (Schütt et al., 2017; Kondor et al., 2018; Faber et al., 2017; Feinberg et al., 2018).

In the second approach, the focus is on optimizing the model architecture and improving neighborhood aggregation. One such model is the Graph Convolutional Neural Network (GCN), which learns the compound's feature representation by the convolution operations performed in the spectral domain of the compound's $2D$ graph, which is obtained by transforming the graph into a set of eigenvectors and eigenvalues. GCN has been shown to be flexible and capable of capturing complex relationships (Wu et al., 2018). Another one, the Message Passing Neural Network (MPNN) framework presented in (Gilmer et al., 2017) operates by passing messages between nodes in the graph, updating the representations of the nodes and edges in the process. The message passing process is performed multiple times, allowing the model to build up a more complex and informative representation of the graph.

Recently, conformer generation has become an important step in property prediction because the physical and chemical properties of a molecule depend on its $3D$ structure. For example, the solubility of a molecule depends on its ability to dissolve in a solvent, and this ability can be influenced by the shape and orientation of the molecule. Similarly, the reactivity of a molecule can be influenced by the orientation of its functional groups, which can affect its ability to participate in chemical reactions. Therefore, conformer generation allows us to make a more accurate prediction of molecular properties (Axelrod & Gomez-Bombarelli, 2020). This is especially important for molecules that have flexible or non-rigid structures, as they can adopt different conformations in different environments. $3D$ Infomax (Stärk et al., 2022) pre-trains a $2D$ network by maximizing the mutual information (MI) between its representation of a molecular graph and a $3D$ representation produced from the molecules' conformers. The weights of $2D$ network are then fine-tuned to predict properties.

3 Persistent Homology

Here, we introduce single parameter persistent homology. In essence, the process of persistent homology consists of three steps. The first step is *graph decomposition* by masking vertices based on either ascending or descending order of their values, which breaks down a graph into many smaller subgraphs. In the second step, the *persistent homology* machinery tracks the changes (on each subgraph separately) in topological features such as birth and death times as they occur in a sequence of simplicial complexes. Finally, in the *vectorization* step, these records can be transformed into a vector that can be utilized in machine learning models.

Graph Decomposition: For a given unweighted graph (compound) $\mathcal{G} = (\mathcal{V}, \mathbb{E})$ with $\mathcal{V} = \{v_1, \ldots, v_m\}$ the set of nodes (atoms) and $\mathbb{E} = \{e_{rs}\}$ the set of edges (bonds), we decompose $\mathcal{G}$ into many subgraphs using a function $f : \mathcal{V} \to \mathbb{R}$ with threshold sets $\mathcal{I} = \{\alpha_i\}_{i=1}^m$, where $\alpha_1 = \min_{v\in\mathcal{V}} f(v) < \alpha_2 < \ldots < \alpha_m = \max_{v\in\mathcal{V}} f(v)$. For $\alpha_i \in \mathcal{I}$, let $\mathcal{V}_i = \{v_r \in \mathcal{V} \mid f(v_r) \leq \alpha_i\}$ (sublevel sets for f). This defines a hierarchy $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset \mathcal{V}_m = \mathcal{V}$ among the nodes with respect to the function f and yields a nested sequence of subgraphs as illustrated in Figure 1. For

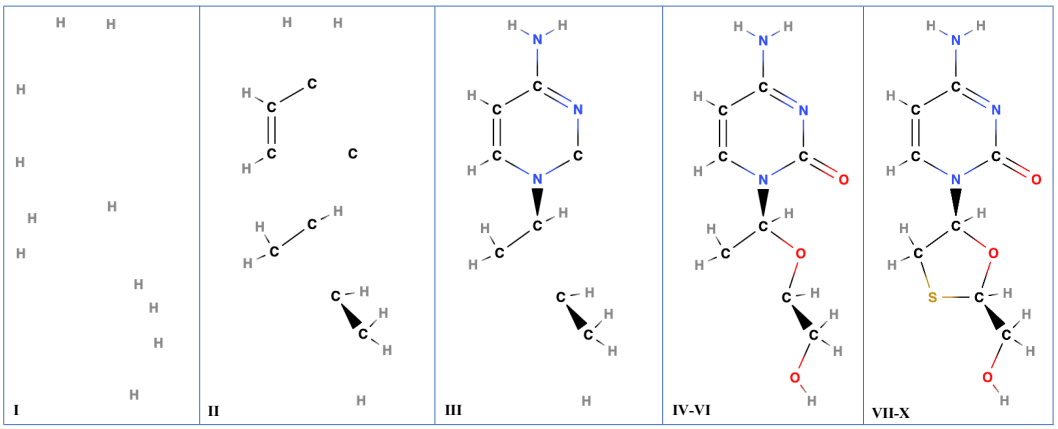

Figure 1: **Graph Decomposition** masks some vertices based on the values of the vertices' parameters and considers only the remaining vertices and edges. In this dataset, 10 unique atoms (H, C, N, O, F, P, S, Cl, Br, I) were present. Hence, there are 10 subgraphs created (some subgraphs can be identical in this process). The original compound shown in the last column is lamivudine, an antiviral used to treat hepatitis B. Vertices with the highest atomic mass are masked first and they are added with respect to ascending values of atomic mass. Atoms are coded by their color: Gray=Hydrogen, Black=Carbon, Blue=Nitrogen, Red=Oxygen, Yellow=Sulfur. Figure 1 is representative of the graph evolution along the y-axis in the initial column of Figure 2. Our framework establishes the Vietoris-Rips complexes for every subgraph as demonstrated along the rows of Figure 2. Subsequently, it computes the rank of the homology groups of dimension 0, 1 for each sequence of simplicial complexes. We repeat the same process using parameters: partial charge (with respect to the ascending order of decile groups in partial charge histogram) and bond types (initially, vertices that constitute a ring structure are added, followed by the addition of vertices linked by triple bonds, double bonds, and ultimately, single bonds). This process of masking vertices and creating subgraphs provides valuable insights into the relationships between the atom/bond properties and the topological features of the dataset.

molecular machine learning applications, this filtering function f can be atomic mass, partial charge, bond type, electron affinity, ionization energy or another important function representing chemical properties of the atoms or bonds. One can also use the natural graph induced functions like node degree, betweenness, etc.

Vietoris-Rips (VR) Filtration: This section outlines our technique for constructing VR simplicial complexes for $2D$ multipersistence, which can be generalized to $3D$ or higher dimensions, though such an extension falls beyond the scope of this paper.

Before constructing VR simplicial complexes, we compute the distances between each node in graph $\mathcal{G}$, i.e., $d(v_r, v_s) = d_{rs}$ is the length of the shortest path from v_r to v_s where each edge has length 1. Let $K = \max d_{rs}$. Then, for each $1 \leq i_0 \leq m$, define VR-filtration for the vertex set $\mathcal{V}_{i_0}$ with the distances $d(v_r, v_s) = d_{rs}$, i.e., $\Delta_{i_0 0} \subseteq \Delta_{i_0 1} \subseteq \ldots \subseteq \Delta_{i_0 K}$ (See Figure 2). This gives $m \times (K+1)$ simplicial complexes $\{\Delta_{ij}\}$ where $1 \leq i \leq m$ and $0 \leq j \leq K$. This is called the *bipersistence module*. One can imagine increasing sequence of $\{\mathcal{V}_i\}$ as vertical direction, and induced VR-complexes $\{\Delta_{ij}\}$ as the horizontal direction. In our construction, we fix the slicing direction as the horizontal direction (VR-direction) in the bipersistence module, and obtain the persistence diagrams in these slices.

The toy example in Figure 2 shows a small graph $\mathcal{G}$ instead of a real compound to keep the exposition simple. Our sublevel filtration (vertical direction) comes from the degree function. Degree of a node is the number of edges incident to it. In the first column, we simply see the single sublevel filtration of $\mathcal{G}$ by the degree function. In each row, we develop VR-filtration of the subgraph by using the graph distances between the nodes. Here, graph distance between nodes means the length of the shortest path (geodesic) in the graph where each edge is taken as length 1. Then, in the second column, we add the edges for the nodes whose graph distance is equal to 2. In the third column, we add the (blue) edges for the nodes whose graph distance is equal to 3. Finally, in the last column, we add the (red) edges for the nodes whose graph distance is equal to 4. By construction, all the graphs in the last column must be a complete graph as there is no more edge to add.

After getting the bifiltration, for each $1 \leq i_0 \leq m$, we obtain a single filtration $\mathcal{V}_{i_0} = \Delta_{i_0 0} \subseteq \Delta_{i_0 1} \subseteq \ldots \subseteq \Delta_{i_0 K}$ in horizontal direction. Each threshold level of VR filtration provides a persistence diagram $PD(\mathcal{V}_{i_0})$. Hence, we obtain m persistence diagrams $\{PD(\mathcal{V}_i)\}$. Then we apply a vectorization, φ, to each persistence diagram and obtain m row vectors of fixed size r, i.e. $\vec{\varphi}_i = \varphi(PD(\mathcal{V}_i))$. This generates a $2D$-vector $\mathbf{M}_\varphi$ (a matrix) of size $m \times (K+1)$.

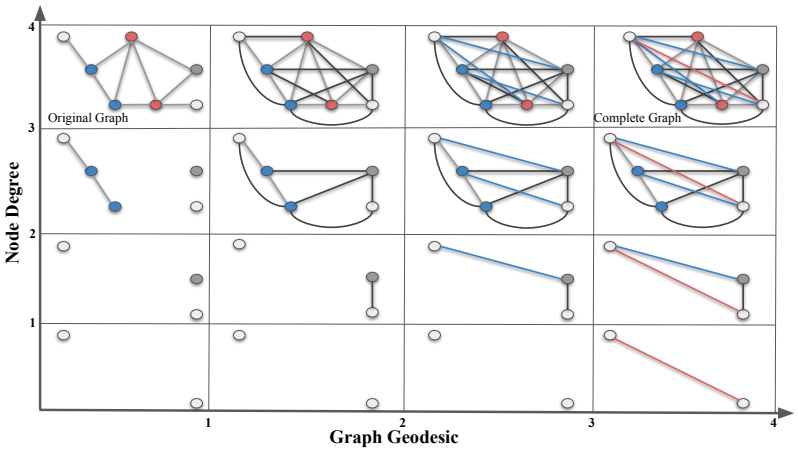

Figure 2: **Vietoris-Rips (VR) Simplicial Filtrations.** In this illustration, a bifiltration is provided that combines a sublevel (vertical) and a Vietoris-Rips (VR) filtration (horizontal) of a simple graph $\mathcal{G}$ (top box in the first column). The vertical direction employs a sublevel filtration based on node degree with thresholds of $1, 2, 3$ and 4. The horizontal direction employs a VR-filtration based on graph distance (geodesic length). The first column displays gray edges between nodes with a graph distance of 1. The second column shows black edges between nodes with a graph distance of less than or equal to 2. The third column displays blue edges between nodes with a graph distance of 3, and the last column shows red edges between nodes with a graph distance of 4.

Persistence Diagrams: After VR filtration, we systematically keep track of the evolution of topological patterns in the sequence of simplicial complexes $\{\widehat{\mathcal{G}}_i\}_{i=1}^N$. A k-dimensional topological feature (or k-hole) may represent connected components (0-dimension), loops (1-dimension) and cavities (2-dimension). For each k-dimensional topological feature σ, persistent homology records its first appearance in the filtration sequence, say $\widehat{\mathcal{G}}_{b_\sigma}$, first disappearance in later complexes, $\widehat{\mathcal{G}}_{d_\sigma}$ with a unique pair (b_σ, d_σ), where $1 \leq b_\sigma < d_\sigma \leq N$. We call b_σ *the birth time* of σ and d_σ *the death time* of σ and $d_\sigma - b_\sigma$ *the life span* (or persistence) of σ. Persistence diagram records all these birth and death times of the topological features. Let $0 \leq k \leq D$ where D is the highest dimension in the simplicial complex $\widehat{\mathcal{G}}_N$. Then k^{th} persistence diagram $\mathrm{PD_k}(\mathcal{G}) = \{(b_\sigma, d_\sigma) \mid \sigma \in H_k(\widehat{\mathcal{G}}_i) \text{ for } b_\sigma \leq i < d_\sigma\}$. Here, $H_k(\widehat{\mathcal{G}}_i)$ represents the kth *homology group* of $\widehat{\mathcal{G}}_i$ which keeps the information of the k-holes in the simplicial complex $\widehat{\mathcal{G}}_i$. We use 0 and 1 dimensional homology features, i.e., $PD_0(\mathcal{G})$ and $PD_1(\mathcal{G})$ in our implementation.

Vectorizations (Fingerprinting): While PH extracts hidden shape patterns from data as persistence diagrams (PD), PDs being collection of points in $\mathbb{R}^2$ by itself are not practical for statistical and machine learning (ML) purposes. Instead, the common techniques are by faithfully representing PDs as kernels (Kriege et al., 2020) or vectorizations (Hensel et al., 2021). One can consider this step as converting PDs into a useful format to be used in ML process as fingerprints of the data. We use Betti curve vectorization (Chung & Lawson, 2022) to transform Persistent Homology (PH) information represented as Persistent Diagrams (PDs) into a feature vector.

4 Multiparameter Persistence (MP) Fingerprints

4.1 Stability of MP Fingerprints

Stability of Single Persistence Vectorizations: A specific persistence diagram vectorization, denoted as φ, can be thought of as a mapping from the space of persistence diagrams to the space of functions. The concept of stability refers to the smoothness of this transformation. Essentially, it assesses whether a slight perturbation in the persistence diagram results in a significant change in the vectorization. To make this assessment meaningful, it is necessary to establish a metric in the space of persistence diagrams that defines what constitutes a "slight perturbation". The most commonly used metric for this purpose is the Wasserstein distance, also known as the matching distance, which is defined as follows.

Let $PD(\mathcal{X}^+)$ and $PD(\mathcal{X}^-)$ be persistence diagrams two datasets $\mathcal{X}^+$ and $\mathcal{X}^-$ (We omit the dimensions in PDs). Let $PD(\mathcal{X}^+) = \{q_j^+\} \cup \Delta^+$ and $PD(\mathcal{X}^-) = \{q_l^-\} \cup \Delta^-$ where $\Delta^\pm$ represents the

diagonal (representing trivial cycles) with infinite multiplicity. Here, $q_j^+ = (b_j^+, d_j^+) \in PD(\mathcal{X}^+)$ represents the birth and death times of a topological feature σ_j in $\mathcal{X}^+$. Let $\phi : PD(\mathcal{X}^+) \to PD(\mathcal{X}^-)$ represent a bijection (matching). With the existence of the diagonal $\Delta^\pm$ in both sides, we make sure the existence of these bijections even if the cardinalities $|\{q_j^+\}|$ and $|\{q_l^-\}|$ are different.

Definition 4.1 *Let $PD(\mathcal{X}^\pm)$ be persistence diagrams of the datasets $\mathcal{X}^\pm$, and $\mathbf{M} = \{\phi\}$ represent the space of matchings as described above. Then, the pth Wasserstein distance $\mathcal{W}_p$ defined as*

$$\mathcal{W}_p(PD(\mathcal{X}^+), PD(\mathcal{X}^-)) = \min_{\phi \in \mathbf{M}} \left(\sum_j \|q_j^+ - \phi(q_j^+)\|_\infty^p \right)^{\frac{1}{p}}, \quad p \in \mathbb{Z}^+.$$

Now, let's define the stability of vectorizations. A vectorization can be viewed as a mapping from the space of persistence diagrams, $\mathbf{P}$, to the space of functions or vectors $\mathbf{Y}$, for example, $\Psi : \mathbf{P} \to \mathbf{Y}$. In particular, if Ψ is the persistence landscape, then $\mathbf{Y} = \mathcal{C}([0, K], \mathbb{R})$ and if Ψ is the Betti summary, then $\mathbf{Y} = \mathbb{R}^m$. The stability of the vectorization Ψ refers to the continuity of Ψ as a mapping. Let $\mathrm{d}(.,.)$ be a suitable metric on the space of vectorizations. The stability of Ψ can then be defined as follows:

Definition 4.2 *Let $\Psi : \mathbf{P} \to \mathbf{Y}$ be a vectorization for single persistence diagrams. Let $\mathcal{W}_p, \mathrm{d}$ be the metrics on $\mathbf{P}$ and $\mathbf{Y}$ respectively as described above. Let $\psi^\pm = \Psi(PD(\mathcal{X}^\pm)) \in \mathbf{Y}$. Then, Ψ is called stable if*

$$\mathrm{d}(\psi^+, \psi^-) \leq C \cdot \mathcal{W}_{p_\Psi}(PD(\mathcal{X}^+), PD(\mathcal{X}^-))$$

In this context, the constant $C > 0$ is independent of $\mathcal{X}^\pm$. The stability inequality states that the changes in the vectorizations are limited by the changes in persistence diagrams. The proximity of two persistence diagrams is reflected in the proximity of their respective vectorizations. A vectorization φ is referred to as *stable* if it satisfies this stability inequality for a given d and $\mathcal{W}_p$ (Atienza et al., 2020).

Now, we are ready to show the stability of Multiparameter Persistent Fingerprints. Consider two graphs, $\mathcal{G}^+ = (\mathcal{V}^+, \mathbb{E}^+)$ and $\mathcal{G}^- = (\mathcal{V}^-, \mathbb{E}^-)$. A stable SP vectorization is represented by φ, and it satisfies the stability equation,

$$\mathrm{d}(\varphi(PD(\mathcal{G}^+)), \varphi(PD(\mathcal{G}^-))) \leq C_\varphi \cdot \mathcal{W}_{p_\varphi}(PD(\mathcal{G}^+), PD(\mathcal{G}^-)) \tag{1}$$

for some $1 \leq p_\varphi \leq \infty$. Here, $\varphi(\mathcal{G}^\pm)$ represent the corresponding vectorizations for $PD(\mathcal{G}^\pm)$ and $\mathcal{W}_p$ represents Wasserstein-p distance as defined in Definition 4.1.

Now, let $f : \mathcal{V}^\pm \to \mathbb{R}$ be a filtering function with threshold set $\{\alpha_i\}_{i=1}^m$. Then, define the sublevel vertex sets $\mathcal{V}_i^\pm = \{v_r \in \mathcal{V}^\pm \mid f(v_r) \leq \alpha_i\}$. For each $\mathcal{V}_i^\pm$, construct the induced VR-filtration $\Delta_{i0}^\pm \subset \Delta_{i1}^\pm \subset \cdots \subset \Delta_{iK}^\pm$ as before. For each $1 \leq i_0 \leq m$, we will have persistence diagram $PD(\mathcal{V}_{i_0}^\pm)$ of the filtration $\{\Delta_{i_0 k}^\pm\}$.

The induced matching distance between multiple persistence diagrams is defined as follows,

$$\mathbf{D}_{p,f}(\mathcal{G}^+, \mathcal{G}^-) = \sum_{i=1}^m \mathcal{W}_p(PD(\mathcal{V}_i^+), PD(\mathcal{V}_i^-)). \tag{2}$$

Now, we define the distance between induced MP Fingerprints as,

$$\mathfrak{D}_f(\mathbf{M}_\varphi(\mathcal{G}^+), \mathbf{M}_\varphi(\mathcal{G}^-)) = \sum_{i=1}^m \mathrm{d}(\varphi(PD(\mathcal{V}_i^+)), \varphi(PD(\mathcal{V}_i^-))) \tag{3}$$

Theorem 4.1 *Let φ be a stable SP vectorization. Then, the induced MP Fingerprint $\mathbf{M}_\varphi$ is also stable, i.e., with the notation above, there exists $\widehat{C}_\varphi > 0$ such that for any pair of graphs $\mathcal{G}^+$ and $\mathcal{G}^-$, we have the following inequality.*

$$\mathfrak{D}(\mathbf{M}_\varphi(\mathcal{G}^+), \mathbf{M}_\varphi(\mathcal{G}^-)) \leq \widehat{C}_\varphi \cdot \mathbf{D}_{p_\varphi}(\{PD(\mathcal{G}^+)\}, \{PD(\mathcal{G}^-)\})$$

Proof: As φ is a stable SP vectorization, by Equation 1, for any $1 \leq i \leq m$, we have $\mathrm{d}(\varphi(PD(\mathcal{V}_i^+)), \varphi(PD(\mathcal{V}_i^+))) \leq C_\varphi \cdot \mathcal{W}_{p_\varphi}(PD(\mathcal{V}_i^+), PD(\mathcal{V}_i^-))$ for some $C_\varphi > 0$, where $\mathcal{W}_{p_\varphi}$ is Wasserstein-p distance. Notice that the constant $C_\varphi > 0$ is independent of i. Hence,

$$\begin{aligned}
\mathfrak{D}(\mathbf{M}_\varphi(\mathcal{G}^+), \mathbf{M}_\varphi(\mathcal{G}^-)) &= \sum_{i=1}^{m} \mathrm{d}(\varphi(PD(\mathcal{V}_i^+)), \varphi(PD(\mathcal{V}_i^-))) \\
&\leq \sum_{i=1}^{m} C_\varphi \cdot \mathcal{W}_{p_\varphi}(PD(\mathcal{V}_i^+), PD(\mathcal{V}_i^-)) \\
&= C_\varphi \sum_{i=1}^{m} \mathcal{W}_{p_\varphi}(PD(\mathcal{V}_i^+), PD(\mathcal{V}_i^-)) \\
&= C_\varphi \cdot \mathbf{D}_{p_\varphi}(\mathcal{G}^+, \mathcal{G}^-)
\end{aligned}$$

where the first and last equalities are due to Equation 2 and Equation 3, while the inequality follows from Equation 1 which is true for any i.

4.2 Computational Complexity of MP Fingerprints

The computational complexity (CC) of the MP Fingerprint $\mathbf{M}_\psi^d$ depends on the vectorization technique ψ used and the number of filtering functions d. For a single persistence diagram PD_k, CC is $\mathcal{O}(\mathcal{N}^3)$, where $\mathcal{N}$ is the number of k-simplices (Otter et al., 2017). If r is the resolution size of the multipersistence grid, then $CC(\mathbf{M}_\psi^d) = \mathcal{O}(r^d \cdot \mathcal{N}^3 \cdot C_\psi(m))$ where $C_\psi(m)$ is CC for ψ and m is the number of barcodes in PD_k, e.g., if ψ is Persistence Landscape, then $C_\psi(m) = m^2$ (Bubenik, 2015) and hence CC for MP Landscape with three filtering functions ($d = 3$) is $\mathcal{O}(r^3 \cdot \mathcal{N}^3 \cdot m^2)$. Alternatively, for MP Betti summaries, the computation of persistence diagrams is not required. Instead, the rank of homology groups in the MP module must be determined. As a result, the computational complexity for MP Betti summary is significantly reduced by utilizing minimal representations (Lesnick & Wright, 2019; Kerber & Rolle, 2021). The feature extraction process is parallelized across the eight cores of an Intel Core i7 CPU (equipped with 100GB of RAM) through the use of multiprocessing. Further evaluation of the computation time for MP fingerprint extraction from datasets can be found in Table 2. In comparison to graph-based models that encode compounds by discovering common molecular fragments (known as motifs) (Jin et al., 2020), all ToDD models necessitate fewer computational resources during their training phase.

5 Experiments

5.1 Datasets & Baselines

Lipophilicity[1] is a crucial characteristic of drug molecules that impacts both permeability through membranes and solubility. The dataset, sourced from the ChEMBL database, contains experimental results for the octanol/water distribution coefficient (logD at pH 7.4) of 4200 compounds.

FreeSolv[2] is a compilation of both calculated and experimentally determined hydration free energies for 642 small molecules in water.

ESOL[3] is a collection of 1128 chemical compounds and their corresponding water solubility values.

We thoroughly evaluate the performance of our methods against the 11 state-of-the-art baselines: Weave (Kearnes et al., 2016), SchNet (Schütt et al., 2017), Node-MPN (Gilmer et al., 2017), Edge-MPN (Yang et al., 2019) D-MPNN (Yang et al., 2019), MGCN (Lu et al., 2019), OT-GNN (Bécigneul et al., 2020), MV-GNN (Ma et al., 2020), StructGNN (Lukashina et al., 2020), GRAPHCL (You et al., 2020) and 3D Infomax (Stärk et al., 2022) on benchmark datasets.

[1] https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/Lipophilicity.csv:

[2] https://deepchemdata.s3.us-west-1.amazonaws.com/datasets/freesolv.csv.gz:

[3] https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/delaney-processed.csv

5.2 EXPERIMENTAL RESULTS

We train a Gradient Boosting Regression Tree (GBRT) model with 1000 boosting stages using MP Fingerprints, namely ToDD (Demir et al., 2022) as the input. The maximum depth of the trees was optimized to 6 through tuning. The minimum required number of samples to split a node within the tree was 2. The Friedman loss function and a learning rate of 0.1 were utilized for model optimization. The effectiveness of the proposed method was rigorously evaluated through a 5-fold cross-validation, demonstrating its competitiveness against state-of-the-art GNN models as shown in Table 1. ToDD is significantly better than the best MoleculeNet models on FreeSolv and ESOL, and is not substantially different on Lipophilicity. These findings suggest that ToDD surpasses the best MoleculeNet models while avoiding the need for training large-scale GNNs or generating $3D$ conformations. Furthermore, both ToDD and Morgan fingerprints exhibit a marked improvement over all GNN baselines on FreeSolv. However, it is important to note, that FreeSolv is a small scale dataset and the performance of GNNs can be impacted due to the limited number of training samples. Hypothetically, ToDD can demonstrate a significant advantage in leveraging domain knowledge to enhance property prediction scores with relative ease. For instance, it is a well-known observation that introducing nonpolar groups, such as methyl groups, as new substituents into a molecule can increase its lipophilicity. ToDD can effectively leverage bond polarity as an additional parameter to extract MP fingerprints, thereby integrating vital domain information for improved performance in lipophilicity prediction.

Table 1: Comparison of RMSE performance between baseline models and ToDD on molecular property prediction task using MoleculeNet (Wu et al., 2018), a large scale benchmark for molecular machine learning. Best score is highlighted in bold, and the best baseline is underlined.

Model	Lipophilicity	FreeSolv	ESOL
Morgan (radius=4)	0.817±0.045	0.478±0.033	1.255±0.051
Weave	0.813±0.042	2.398±0.250	1.158±0.055
SchNet	0.909±0.098	3.215±0.755	1.045±0.064
D-MPNN	0.646±0.041	1.010 ± 0.064	0.980±0.258
MGCN	1.113±0.041	3.349±0.097	1.266±0.147
Node-MPN	0.672±0.051	2.185±0.952	1.167±0.430
Edge-MPN	0.653±0.046	2.177±0.914	0.980±0.258
MV-GNN	**0.599±0.016**	1.840±0.019	0.805±0.036
GRAPHCL	0.714±0.011	3.744±0.292	0.959±0.047
3D Infomax	0.695±0.012	2.337±0.227	0.894±0.028
ToDD	0.738±0.025	**0.354±0.053**	**0.612±0.083**

Table 2: RMSE scores for single parameter and multiparameter PH fingerprints and clock time performance to extract Vietoris-Rips persistent homology features.

	Atomic Mass		Partial Charge		Bond Type		All Parameters	
Dataset	**RMSE**	**Time**	**RMSE**	**Time**	**RMSE**	**Time**	**RMSE**	**Time**
Lipophilicity	0.989±0.027	55 sec	0.989±0.034	42 sec	0.995±0.049	17 sec	**0.738±0.025**	114 sec
FreeSolv	0.605±0.124	7 sec	0.557±0.099	7 sec	0.894±0.180	2 sec	**0.354±0.053**	16 sec
ESOL	0.929±0.066	11 sec	0.997±0.103	11 sec	1.340±0.078	4 sec	**0.612±0.083**	26 sec

5.3 ABLATION STUDIES

We investigate the impact of incorporating additional information about the domain on the model's performance. Table 2 shows the results of evaluating the performance of single parameter persistence using only atomic mass, partial charge, and bond type. Afterwards, the Betti vectorizations from all the modalities are combined. Our findings indicate that the significance of single parameter persistence varies across the parameters, but combining the topological fingerprints learned from each modality into a single representation leads to a significant improvement in the RMSE scores, due to the complementary nature of the information sources.

6 CONCLUSION

A major challenge in the field of molecular machine learning is the conversion of molecules into concise fixed-length vectors. To address this issue, ToDD employs multiparameter persistent homology to construct a hierarchical topological representation of molecules. Our results indicate that this method has produced promising outcomes in predicting bioactivity and molecular properties, with the potential for further improvement through the integration of additional chemical parameters.

REFERENCES

Nieves Atienza, Rocío González-Díaz, and Manuel Soriano-Trigueros. On the stability of persistent entropy and new summary functions for topological data analysis. *Pattern Recognition*, 107: 107509, 2020.

Simon Axelrod and Rafael Gomez-Bombarelli. Molecular machine learning with conformer ensembles. *arXiv preprint arXiv:2012.08452*, 2020.

Gary Bécigneul, Octavian-Eugen Ganea, Benson Chen, Regina Barzilay, and Tommi S Jaakkola. Optimal transport graph neural networks. 2020.

Nurken Berdigaliyev and Mohamad Aljofan. An overview of drug discovery and development. *Future Medicinal Chemistry*, 12(10):939–947, 2020.

P. Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77–102, 2015.

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3438–3445, 2020.

Pengfei Chen, Weiwen Liu, Chang-Yu Hsieh, Guangyong Chen, and Shengyu Zhang. Utilizing edge features in graph neural networks via variational information maximization. *arXiv preprint arXiv:1906.05488*, 2019.

Yu-Min Chung and Austin Lawson. Persistence curves: A canonical framework for summarizing persistence diagrams. *Advances in Computational Mathematics*, 48(1):6, 2022.

Andac Demir, Baris Coskunuzer, Ignacio Segovia-Dominguez, Yuzhou Chen, Yulia Gel, and Bulent Kiziltan. Todd: Topological compound fingerprinting in computer-aided drug discovery. *bioRxiv*, pp. 2022–11, 2022.

Pavlo O Dral. Quantum chemistry in the age of machine learning. *The journal of physical chemistry letters*, 11(6):2336–2347, 2020.

Felix A Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S Schoenholz, George E Dahl, Oriol Vinyals, Steven Kearnes, Patrick F Riley, and O Anatole von Lilienfeld. Machine learning prediction errors better than dft accuracy. *arXiv preprint arXiv:1702.05532*, 2017.

Evan N Feinberg, Debnil Sur, Zhenqin Wu, Brooke E Husic, Huanghao Mai, Yang Li, Saisai Sun, Jianyi Yang, Bharath Ramsundar, and Vijay S Pande. Potentialnet for molecular property prediction. *ACS central science*, 4(11):1520–1530, 2018.

Octavian Ganea, Lagnajit Pattanaik, Connor Coley, Regina Barzilay, Klavs Jensen, William Green, and Tommi Jaakkola. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Advances in Neural Information Processing Systems*, 34:13757–13769, 2021.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:52, 2021.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *ICML*, pp. 4839–4848. PMLR, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30: 595–608, 2016.

Michael Kerber and Alexander Rolle. Fast minimal presentations of bi-graded persistence modules. In *ALENEX*, pp. 207–220. SIAM, 2021.

Risi Kondor, Hy Truong Son, Horace Pan, Brandon Anderson, and Shubhendu Trivedi. Covariant compositional networks for learning graphs. *arXiv preprint arXiv:1801.02144*, 2018.

Nils M Kriege, Fredrik D Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 5(1):1–42, 2020.

Michael Lesnick and Matthew Wright. Computing minimal presentations and bigraded betti numbers of 2-parameter persistent homology. *arXiv preprint arXiv:1902.05708*, 2019.

Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013*, 2021.

Chengqiang Lu, Qi Liu, Chao Wang, Zhenya Huang, Peize Lin, and Lixin He. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1052–1060, 2019.

Nina Lukashina, Alisa Alenicheva, Elizaveta Vlasova, Artem Kondiukov, Aigul Khakimova, Emil Magerramov, Nikita Churikov, and Aleksei Shpilman. Lipophilicity prediction with multitask learning and molecular substructures representation. *arXiv preprint arXiv:2011.12117*, 2020.

Hehuan Ma, Yatao Bian, Yu Rong, Wenbing Huang, Tingyang Xu, Weiyang Xie, Geyan Ye, and Junzhou Huang. Multi-view graph neural networks for molecular property prediction. *arXiv preprint arXiv:2005.13607*, 2020.

Andrea Mauri, Viviana Consonni, Manuela Pavan, and Roberto Todeschini. Dragon software: An easy approach to molecular descriptor calculations. *Match*, 56(2):237–248, 2006.

Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole Lilienfeld, and Klaus-Robert Müller. Learning invariant representations of molecules for atomization energy prediction. *Advances in neural information processing systems*, 25, 2012.

Asher Mullard. Parsing clinical success rates. *Nature Reviews Drug Discovery*, 15(7):447–448, 2016.

Yuya Ogawa, Seiji Maekawa, Yuya Sasaki, Yasuhiro Fujiwara, and Makoto Onizuka. Adaptive node embedding propagation for semi-supervised classification. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pp. 417–433. Springer, 2021.

Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6:1–38, 2017.

Yuzhong Peng, Yanmei Lin, Xiao-Yuan Jing, Hao Zhang, Yiran Huang, and Guang Sheng Luo. Enhanced graph isomorphism network for molecular admet properties prediction. *IEEE Access*, 8: 168344–168360, 2020.

Jean-Louis Reymond and Mahendra Awale. Exploring chemical space for drug discovery using the chemical universe database. *ACS chemical neuroscience*, 3(9):649–657, 2012.

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

Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5 (1):83, 2019.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning*, pp. 9558–9568. PMLR, 2021.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pp. 20479–20502. PMLR, 2022.

Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

Haiyang Wang, Guangyu Zhou, Siqi Liu, Jyun-Yu Jiang, and Wei Wang. Drug-target interaction prediction with graph attention networks. *arXiv preprint arXiv:2107.06099*, 2021.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Minkai Xu, Shitong Luo, Yoshua Bengio, Jian Peng, and Jian Tang. Learning neural generative dynamics for molecular conformation generation. *arXiv preprint arXiv:2102.10240*, 2021.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8): 3370–3388, 2019.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.

Jin Zhang, Daniel Mucs, Ulf Norinder, and Fredrik Svensson. Lightgbm: An effective and scalable algorithm for prediction of chemical toxicity–application to the tox21 and mutagenicity data sets. *Journal of chemical information and modeling*, 59(10):4150–4158, 2019.

Shuangjia Zheng, Xin Yan, Yuedong Yang, and Jun Xu. Identifying structure–property relationships through smiles syntax analysis with self-attention mechanism. *Journal of chemical information and modeling*, 59(2):914–923, 2019.