Characterizing the Latent Space of Molecular Deep Generative Models with Persistent Homology Metrics

Yair Schiff[*]
IBM
Yorktown Heights, NY 10598
yair.schiff@ibm.com

Vijil Chenthamarakshan
IBM Research
ecvijil@us.ibm.com

Karthikeyan Natesan Ramamurthy
IBM Research
knatesa@us.ibm.com

Payel Das[*]
IBM Research
daspa@us.ibm.com

Abstract

Deep generative models are increasingly becoming integral parts of the *in silico* molecule design pipeline and have dual goals of learning the chemical and structural features that render candidate molecules viable while also being flexible enough to generate novel designs. Specifically, Variational Auto Encoders (VAEs) are generative models in which encoder-decoder network pairs are trained to reconstruct training data distributions in such a way that the latent space of the encoder network is smooth. Therefore, novel candidates can be found by sampling from this latent space. However, the scope of architectures and hyperparameters is vast and choosing the best combination for *in silico* discovery has important implications for downstream success. Therefore, it is important to develop a principled methodology for distinguishing how well a given generative model is able to learn salient molecular features. In this work, we propose a method for measuring how well the latent space of deep generative models is able to encode structural and chemical features of molecular datasets by correlating latent space metrics with metrics from the field of topological data analysis (TDA). We apply our evaluation methodology to a VAE trained on SMILES strings and show that 3D topology information is consistently encoded throughout the latent space of the model.

1 Introduction

The discovery process for novel molecules, from ideation to market, typically takes over a decade and costs billions of dollars. There is therefore a growing need to accelerate this timeline to tackle crises, like COVID-19 or climate change. Recently, there has been a surge of machine learning (ML), particularly deep learning, methods for learning molecule representations for various applications in chemistry, biology, and materials science. One specific line of work involves the use of deep generative models to design novel molecules that satisfy certain properties.

These generative models can be trained on a wide variety of inputs, such as image, feature vector, 3D coordinate, text, or graph representations. The choice of input representation has significant ramifications on performance and success of downstream tasks. It also impacts how well different information (e.g. chemical, sub-structural, 3D topological) is captured in a generative model's latent

[*] Corresponding authors

Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

space. Comparing latent vector representations with a more human-readable and well-established molecular metric can help in estimating the information content of the learned molecular embeddings.

For example, many Natural Language Processing-inspired ML methods for molecular generation train on string representations of molecules known as SMILES [6]. While these 1D string representations have been successfully used in several prediction and design applications, their biggest detractor is that they do not explicitly encode 3D structure, which is important in determining function, such as protein binding affinity or catalyst efficiency. Finding a way to quantify how much information about 3D structure these generative models can encode, despite having been trained on 1D string representations, will be an important tool in their evaluation and comparison to other models.

In this work, we examine the latent space of a generative model trained on molecular SMILES representations. We specifically look at the latent space of the VAE model from Chenthamarakshan et al. [3], which recently demonstrated success in developing novel, chemically valid, drug-like molecules with potential anti-COVID activity and selectivity. We compare euclidean distance on latent space vectors from this model and Tanimoto distance on fingerprint representations for the corresponding molecules to see how well molecular sub-structure information is captured. The novelty of our work comes from our evaluation of how well the VAE's latent space encodes 3D topological structure of molecules. We perform this evaluation by correlating latent space euclidean distance (referred to throughout as z distance) with a metric on two parameter persistent homology of molecules. We explore two parameter persistent homology both because of its recent success in virtual molecular screening [7] and because of its ability to simultaneously capture structural and chemical features, by way of the second parameter. We find that for the VAE from Chenthamarakshan et al. [3], 3D information is captured in a uniform manner across the latent space. In contrast, information from the more common bit vector fingerprint representation, is well captured in regions of the latent space that are near the training distribution, but not consistently across the entire latent space. We believe that this work provides a means of connecting generative models trained on inputs that do not explicitly encode 3D information with representations of 3D molecular topology.

2 Methodology

The basic tenet of our methodology is to quantify how well the latent space of a generative model encodes semantic information about the underlying data. We achieve this by correlating the z distance with metrics on other molecular representations, such as persistence diagrams and fingerprints.

The persistent homology of a distance function of a molecule can be calculated by treating the constituent atoms' 3D coordinates as a discrete point cloud. A second parameter (e.g. atomic partial charge) can be used to control the number of atoms in the point cloud, resulting in two parameter persistent homology. We compute the Restricted Hilbert function [7], which evaluates the Betti numbers at each step of the bifiltration. We use the ℓ_2 distance between the Restricted Hilbert functions of two molecules as a distance metric between their two parameter persistence diagrams. For more details, please see Appendix A, and for more technical details about the relevant terminology and calculations, please see Keller et al. [7]. Our pipeline for calculating two parameter persistent homology on molecules is as follows: first, the RDKit [8, 9] and OpenBabel [1, 12] libraries are used to convert SMILES representations to 3D atomic point clouds and to extract parameters, such as partial charge and atomic radius. For each molecule, its two parameter persistence diagram and Restricted Hilbert function are calculated with the RIVET tool [10, 14].

To have a benchmark for how well the latent space encodes information for other representations, we compare it with the Tanimoto distance on fingerprint representations of molecules. Fingerprints are widely used bit vector representations that encode information about sub-structures present in a molecule [6]. We use the RDKit library to generate molecular fingerprints and to calculate Tanimoto distance, which measures alignment between the bit vectors of two molecules (see Appendix B for more information).

While our approach for evaluating the latent space of generative models is not tied to any specific architecture or implementation, for the purposes of this work we use the VAE model and training detailed in Chenthamarakshan et al. [3]. One of the strongest indications that the latent space of this model is able to encode 3D structure to some extent is the success of the model in generating molecules capable of binding to the 3D pocket of target protein structures [3]. We use both the latent embeddings of a subset of the training data, as well as randomly sampled vectors from the latent space distribution, which are in turn decoded back to SMILES strings. For the training data embeddings, we simply use their corresponding SMILES for the persistent homology pipeline described above. For the random vectors drawn from the latent distribution, we pass them through the VAE decoder network to obtain corresponding SMILES strings and use the RDKit library to discard non-meaningful samples. The purpose of also analyzing random vectors drawn from the latent distributions is to see how well the latent space encodes information in regions that are not near training data.

To capture an overview of the latent space of this model as a whole, we follow an approach similar to that described in Maragakis et al. [11]. First, we calculate all pairwise z distances for N random data points. This produces $\mathcal{O}(N^2)$ pairwise distances. These distances are quantized into B bins of equal width. We then sample, n pairwise distances from each of the bins, which gives us $B * n$ random latent embedding pairwise distances. For each of the molecules pairs that correspond to the $B * n$ distances, we also calculate the ℓ_2 Restricted Hilbert function and Tanimoto distances. Plotting these metrics and calculating their correlation coefficient with the z distances provides an overview of how well the latent space is able to encode different types of structural information from the molecules.

3 Related work

Correlating metrics on outputs and embeddings of deep generative models with metrics on other molecule representations is a widely used technique and is a natural approach for quantifying and providing a graphical representation of how well a model captures semantic information. However, to the best of our knowledge, in the domain of molecular generative models, most works rely on metrics on the 1D fingerprint representations of molecules, such as Jaccard or Tanimoto distance.

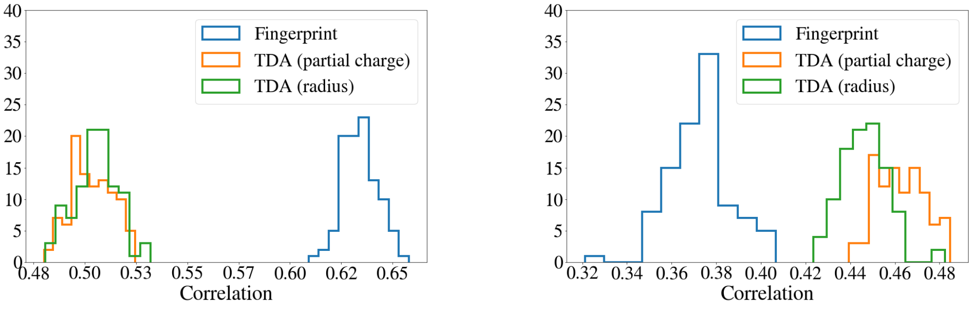
For example, in Duvenaud et al. [5], where the authors create a differentiable fingerprint representation for molecules, they employ this approach by seeing how Tanimoto distance on canonical circular fingerprint representations of molecules correlates with this same metric on their neural network generated fingerprint representation. Similarly, a graph like the ones presented in Appendix D comes from Maragakis et al. [11] where the authors compare latent space euclidean distance with Jaccard distance on fingerprint representations. Our approach is novel in that we compare latent space distances to a more informative and useful measure, namely distances between two parameter persistence diagrams.

An explicit use of persistent homology to evaluate generative models in other domains, such as tabular data, is presented in Charlier et al. [2]. However, the analysis in that work is quite different from ours, since Charlier et al. [2] compare persistent homology of data distributions, training vs. generated, whereas we calculate the persistent homology of 3D atomic coordinates of individual molecules, following Keller et al. [7].

4 Results

As mentioned in Section 2, in our analysis we use both samples from the training data and randomly sampled points from the latent space. For the training data, we start with 10,000 training molecules. We calculate the ∼50 million pairwise z distances for these 10,000 points and bin these distances into 10 equal-width groups. From each of these bins, we randomly select 400 molecule pairs. The 10th bin is excluded, since it does contain at least 400 points, giving us 3,600 pairwise distances. For each pair of molecules, we then calculate the ℓ_2 distance on Restricted Hilbert functions and Tanimoto distance. Restricted Hilbert functions are calculated for two separate bifiltrations, one where atomic partial charge is used as the second parameter, and one where atomic radius is the second parameter. For these 3,600 molecule pairs, we calculate the *Pearson r* correlation coefficient between distances.

We repeat this process for random vectors sampled from the VAE latent distribution to see how well information is encoded in the latent space away from the training data. Specifically, we sample 5,000 vectors from the VAE latent distribution. Of these, about 4,500 have valid SMILES string decodings. As above, we calculate all the pairwise z distances for the 4,500 latent space vectors and create ten equal-width bins for these ∼10 million distances. From each distance bin, we randomly choose up to 400 pairs and repeat the correlation calculations between latent and fingerprint / topological representation metrics. Again, the 10th bin is excluded, since it does contain at least 400 points, giving us 3,600 pairwise distances.



(a) Training data points

(b) Random samples latent distribution

Figure 1: Distributions of correlation coefficients for 100 random draws of 400 pairs from each distance bin for (a) training molecules and (b) synthesized SMILES strings corresponding to molecules sampled randomly using the latent space distribution. For each of the 100 random draws, correlations are calculated between the 3,600 pairwise z distances and Tanimoto distances, ℓ_2 Restricted Hilbert function (TDA) distances when partial charge is used as the second parameter, and ℓ_2 Restricted Hilbert function (TDA) distances when atomic radius is used as the second parameter.

Table 1: Statistics for distributions in Figure 1.

	Training data			*Random latent sample*		
	Fingerprint	TDA (partial charge)	TDA (radius)	Fingerprint	TDA (partial charge)	TDA (radius)
Median	0.635	0.503	0.507	0.377	0.464	0.449
Mean	0.636	0.504	0.506	0.377	0.465	0.449
Std. dev.	0.008	0.010	0.010	0.014	0.010	0.011

To get robust estimates of this correlation, we repeat the process of randomly drawing 400 pairs from each distance bin 100 times. For each of these draws, we calculate the correlation coefficient between z distances and metrics for the other molecule representations. The distributions of these correlations are presented in Figure 1 and their corresponding statistics are recorded in Table 1.

Figure 1(a) reveals that for the training data points, z distance shows stronger correlation with Tanimoto distance, when compared to distance estimated using two parameter persistent homology representation. This result is not surprising, since the VAE model is trained to minimize reconstruction loss for the training distribution. For training samples, sub-structure information that is explicitly encoded by both the SMILES string and fingerprint representations is better preserved in the latent space. However, while investigating the random samples generated from the latent space, we find that the correlation between z distance and Tanimoto distance is significantly weaker. In contrast, the correlation between z distance and persistent homology-based distance is consistently around 0.5 for both training molecules and random samples generated from the latent distribution. This result indicates that the 3D topological information is moderately, but uniformly, captured throughout the latent space, when compared to Tanimoto distance. Figure 1 further shows that the correlations with ℓ_2 Restricted Hilbert function distance are consistent for both of the second parameters we examine, partial charge and atomic radius, indicating that the VAE latent space is able to consistently capture the 3D topology of molecules, irrespective of the choice of the second parameter. Figures 6, 7, and 8 in Appendix D show one example of a molecule pair for which the z distance correlates better with the persistent homology-based distance, when compared to Tanimoto distance.

5 Conclusion and future directions

TDA tools are being adopted across ML and other domains of science, and we believe this is an exciting direction for *in silico* molecular screening with deep generative models. Our approach is useful for comparing the semantics of latent spaces of different architectures and, although not presented here, for tracking how the latent space of a particular model evolves throughout training in its ability to learn 3D molecular structures. One immediate extension of our work includes developing new visualizations that compare latent metrics to metrics on topological structures. We will also extend this analysis to investigate the effect of using different fingerprints and different generative models on the final results. More broadly, we believe that incorporating topological metrics directly into the generative model training pipeline can significantly improve the models. However, this will be a non-trivial effort requiring significant research due to the computationally intensive and non-differentiable pipeline for computing two parameter persistent homology.

References

[1] The open babel package. URL http://www.openbabel.org.

[2] J. Charlier, R. State, and J. Hilger. Phom-gem: Persistent homology for generative models. In *2019 6th Swiss Conference on Data Science (SDS)*, pages 87–92. IEEE, 2019.

[3] V. Chenthamarakshan, P. Das, I. Padhi, H. Strobelt, K. W. Lim, B. Hoover, S. C. Hoffman, and A. Mojsilovic. Target-specific and selective drug design for covid-19 using deep generative models. *arXiv preprint arXiv:2004.01215*, 2020.

[4] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia. MeshLab: an Open-Source Mesh Processing Tool. In V. Scarano, R. D. Chiara, and U. Erra, editors, *Eurographics Italian Chapter Conference*. The Eurographics Association, 2008. ISBN 978-3-905673-68-5. doi: 10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136.

[5] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[6] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, 4(4): 828–849, 2019.

[7] B. Keller, M. Lesnick, and T. L. Willke. Persistent homology for virtual screening. *ChemRxiv*, 2018.

[8] G. Landrum. Rdkit: Open-source cheminformatics. URL https://www.rdkit.org.

[9] G. Landrum. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling, 2013.

[10] M. Lesnick and M. Wright. Interactive visualization of 2-d persistence modules. *arXiv preprint arXiv:1512.00180*, 2015.

[11] P. Maragakis, H. Nisonoff, B. Cole, and D. E. Shaw. A deep-learning view of chemical space designed to facilitate drug discovery. *arXiv preprint arXiv:2002.02948*, 2020.

[12] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33, 2011.

[13] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[14] The RIVET Developers. Rivet. URL https://github.com/rivetTDA/rivet.

[15] Q.-Y. Zhou, J. Park, and V. Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.

A Two parameter persistence homology

We briefly describe persistent homology in the context of molecular data. A more detailed presentation can be found in Keller et al. [7]. First, the 3D atomic coordinates of a molecule are treated as a discrete point cloud. Using a filtration, such as Vietoris-Rips, sequences of simplicial complexes are created by successively connecting atoms that are further away from each other. For each step in the filtration, the homology of dimensions zero, one, and two can be thought of as representing the connected components, holes, and voids (respectively) of the simplicial complex, and the the rank or Betti numbers $\mathcal{B}_i$ for $i \in 0, 1, 2$, can be thought of as a count of these different dimensional holes. In two parameter persistent homology, the second parameter acts as a threshold for which points are considered in the point cloud. That is, as the threshold for the second parameter varies, more and more atoms are included in the simplicial complex filtration. The Restricted Hilbert function evaluates $\mathcal{B}_i$ at each step of the bifiltration. ℓ_2 distance for two functions is defined as:

$$\ell_2(f,g) = \sqrt{\int (f-g)^2 dA}$$

In this work, we evaluate only 0th and 1st dimension homologies both for computational considerations and because, as noted in Keller et al. [7], the majority of the energy for ℓ_2 distances on Restricted Hilbert functions for molecules comes from lower dimension homologies.

B Fingerprint representations

The fingerprint representations and related metrics used in this work are evaluated using the RDKit library. First SMILES strings, which use ASCII characters to encode molecules' graph structure [6], are converted to bit vectors know as Molecular ACCess System (MACCS) keys fingerprints. In these MACCS fingerprint signatures, each bit records a binary structural feature of the molecule [6].

Tanimoto similarity measures the alignment between two bit vector signatures by taking a ratio commonly known as 'intersection over union' in other contexts. Specifically, it is calculated as the bit-wise `AND` ($\cap$) divided by bit-wise `OR` ($\cup$) of two bit vectors:

$$Tanimoto(A,B) = \frac{A \cap B}{A \cup B}$$

Subtracting this quantity from 1 gives a distance rather than a similarity metric.

C VAE Model

In this section, we provide a brief overview of the VAE model that we examine, which comes from Chenthamarakshan et al. [3]. In Chenthamarakshan et al. [3], the authors begin with semi-supervised training of the encoder-decoder VAE networks on $\sim$2 million molecules. VAE loss is comprised of an L2 reconstruction term and KL divergence annealing to ensure smoothness of the latent space. For both the encoder and decoder, the authors use Gated Recurrent Unit architectures. Latent space vectors are 128-dimensional. In addition to optimizing reconstruction loss and KL loss, the latent embeddings are jointly trained on two property prediction tasks. To generate molecule candidates, the authors use an efficient sampling scheme to search for molecules that satisfy several, often conflicting, specifications. For more information about the architecture, hyperparameter selection, training protocol, and results, please see Chenthamarakshan et al. [3].

We also provide more detail here about how training samples and randomly sampled latent vectors are used in our evaluation pipeline. Because our analysis compares metrics for multiple molecule representations, for each sample, we needed its latent vector, bit vector fingerprint, and two parameter persistence diagram representations.

For the training samples, for each data point, we start with the SMILES string. This string is passed through the VAE's encoder network to obtain the corresponding latent vector. The SMILES string is also processed by the the RDKit library to produce its MACCS fingerprint signature. Finally, the OpenBabel library is used to extract 3D coordinates of the molecule's constituent atoms, which are saved as a point cloud and passed through the RIVET tool to produce persistence diagrams.

For the randomly sampled latent vectors, we begin by first drawing high-dimensional vectors from the encoder's learned latent distribution. These vectors are passed through the VAE's decoder network to produce SMILES strings from which we can produce corresponding fingerprint and persistence diagram representations.

D Supplementary correlation plots

In Figures 2 and 3, we present the distribution of pairwise z distances for both the training and random latent points. As mentioned in Section 4, the final bin in both cases is excluded from the correlation coefficient calculations since it does not contain sufficient observations.

In Figures 4 and 5, we plot one random draw, of the 100 displayed in Figure 1, and the correlation comparison for z distances with Tanimoto distances and ℓ_2 Restricted Hilbert function distances.

Finally, in Figures 6, 7, and 8, we display two molecules decoded from random samples of the latent distribution that have small z distance (in the 10th percentile of z distance) and similar persistence diagrams (in the 10th percentile of Hilbert function ℓ_2 distances), but large fingerprint distances (in the 90th percentile of Tanimoto distances). We choose this molecule pair as a representative example of an instance where proximity in 3D topology representations is well captured in the latent space, but where distance in fingerprint representations do not correlate well with latent distance. For each molecule, we display a 2D drawing that was created using the RDKit library (Figure 6), a 3D point cloud mesh grid that was created using the MeshLab [4], Open3D [15], and PyMol [13] libraries (Figure 7), and the two parameter persistence diagram created using the RIVET graphical user interface (Figure 8).

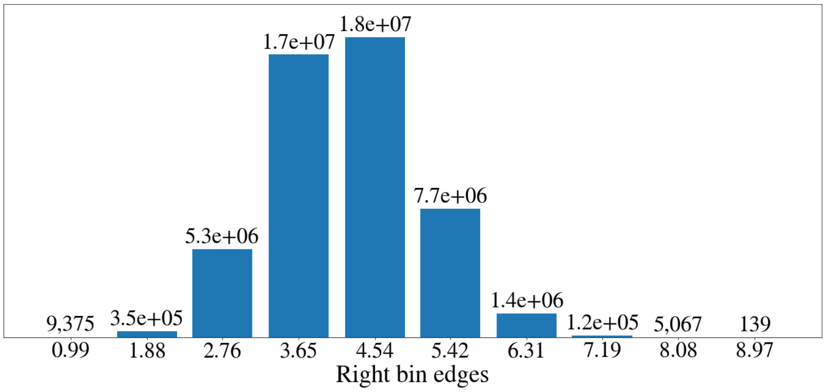


Figure 2: Number of pairwise z distances within each bin for 10,000 training molecules. Horizontal axis marks represent the right edge of each bin. Data labels represent the count of pairwise distances that fall within each bin.

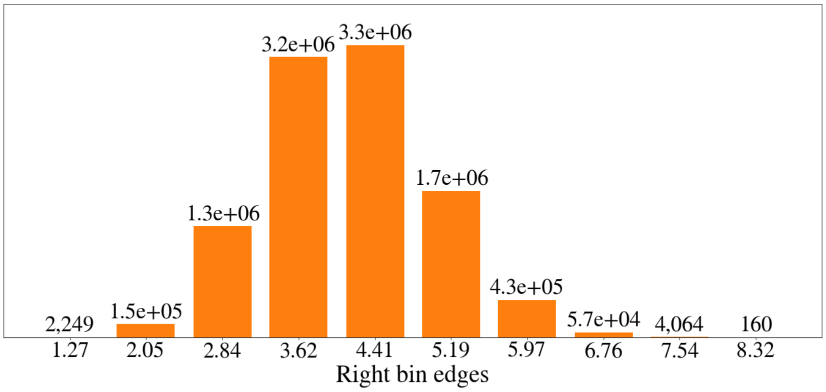
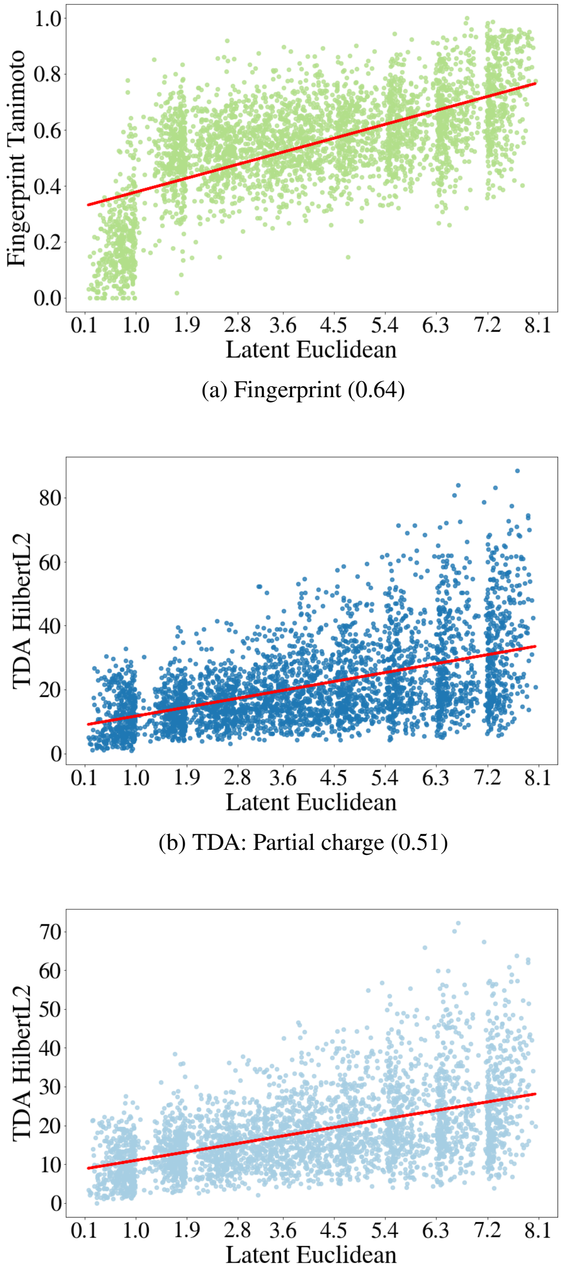


Figure 3: Number of pairwise z distances within each bin for 4,500 random latent embeddings. Horizontal axis marks represent the right edge of each bin. Data labels represent the count of pairwise distances that fall within each bin.

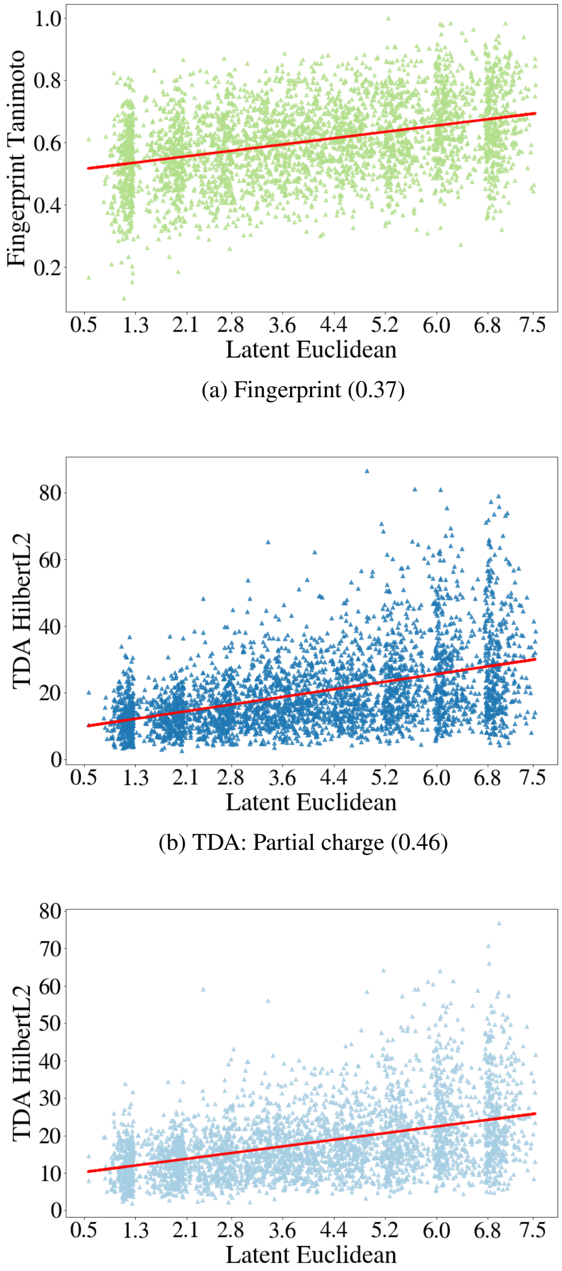


(a) Fingerprint (0.64)

(b) TDA: Partial charge (0.51)

(c) TDA: Radius (0.51)

Figure 4: Comparison of distances for a single random draw of 3,600 molecule pairs (400 from each of the 9 distance bins considered). We compare z distances for **random training data points** with (a) Tanimoto distances, (b) Hilbert function ℓ_2 distances for two parameter persistence with partial charge as second parameter, and (c) Hilbert function ℓ_2 distances for two parameter persistence with atomic radius as second parameter. *Pearson r* coefficients for each metric comparison are in parentheses next to each sub-figure title.

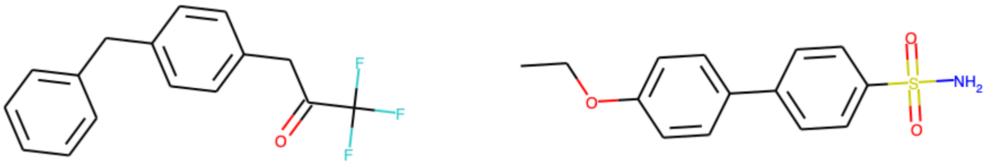


(a) Fingerprint (0.37)

(b) TDA: Partial charge (0.46)

(c) TDA: Radius (0.44)

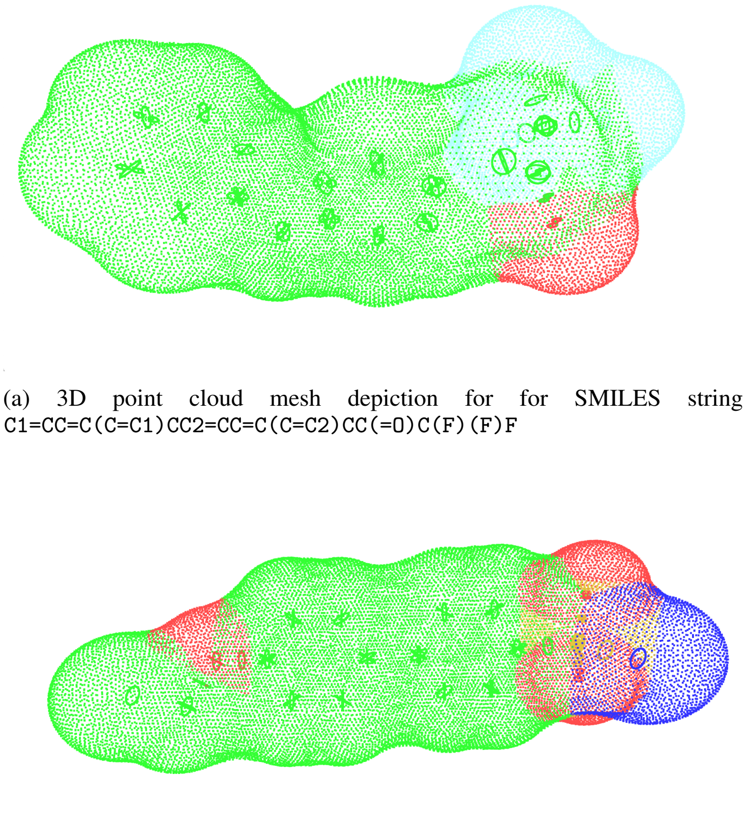
Figure 5: Comparison of distances for a single random draw of 3,600 molecule pairs (400 from each of the 9 distance bins considered). We compare z distances for **random samples generated from the latent space** with (a) Tanimoto distances, (b) Hilbert function ℓ_2 distances for two parameter persistence with partial charge as second parameter, and (c) Hilbert function ℓ_2 distances for two parameter persistence with atomic radius as second parameter. *Pearson r* coefficients for each metric comparison are in parentheses next to each sub-figure title.



(a) 2D molecular drawing for SMILES string `C1=CC=C(C=C1)CC2=CC=C(C=C2)CC(=O)C(F)(F)F`

(b) 2D molecular drawing for SMILES string `CCOC1=CC=C(C=C1)C2=CC=C(C=C2)S(=O)(=O)N`

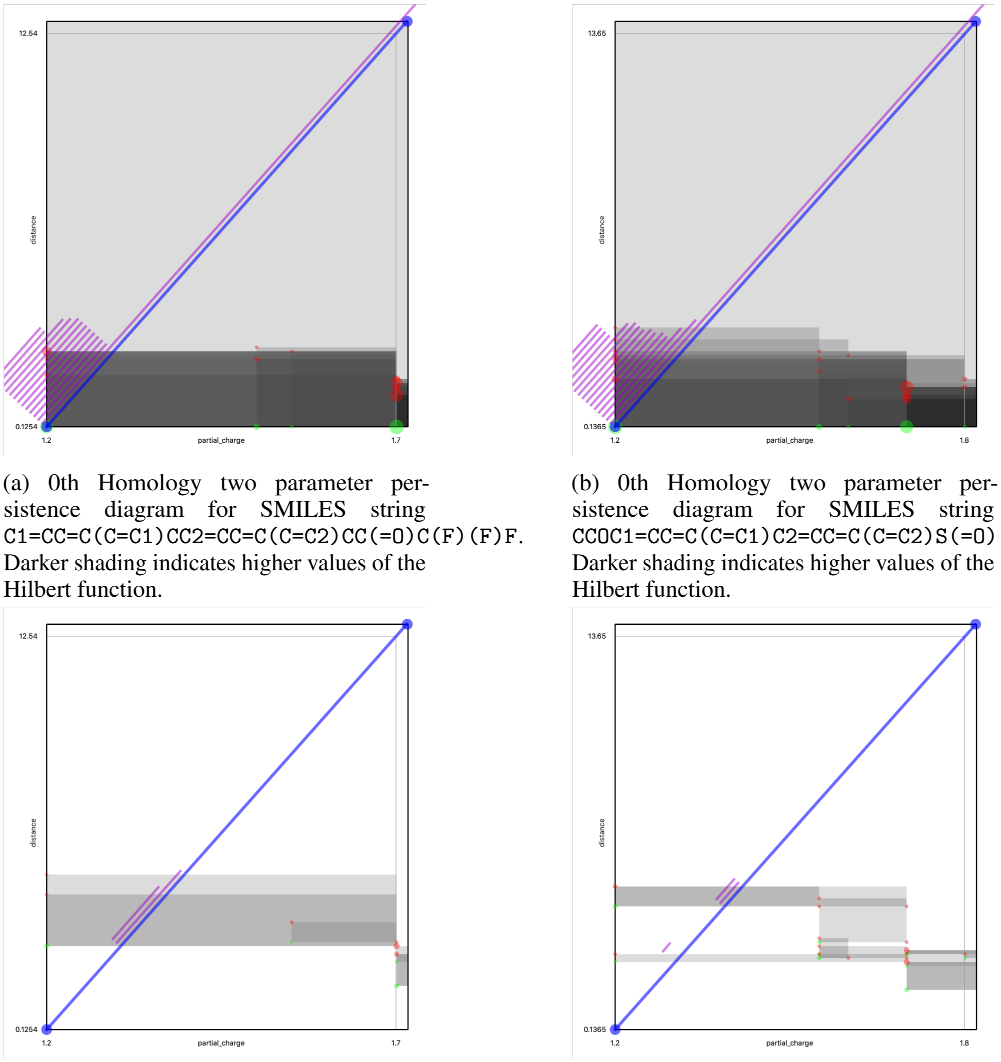
Figure 6: 2D drawing for two molecules that demonstrated higher correlation between latent and persistence diagram metrics than latent and fingerprint metrics.



(a) 3D point cloud mesh depiction for for SMILES string `C1=CC=C(C=C1)CC2=CC=C(C=C2)CC(=O)C(F)(F)F`

(b) 3D point cloud mesh depiction for SMILES string `CCOC1=CC=C(C=C1)C2=CC=C(C=C2)S(=O)(=O)N`

Figure 7: 3D point cloud mesh depictions for two molecules that demonstrated higher correlation between latent and persistence diagram metrics than latent and fingerprint metrics.



(a) 0th Homology two parameter persistence diagram for SMILES string `C1=CC=C(C=C1)CC2=CC=C(C=C2)CC(=O)C(F)(F)F`. Darker shading indicates higher values of the Hilbert function.

(b) 0th Homology two parameter persistence diagram for SMILES string `CCOC1=CC=C(C=C1)C2=CC=C(C=C2)S(=O)(=O)N`. Darker shading indicates higher values of the Hilbert function.

(c) 1st Homology two parameter persistence diagram for SMILES string `C1=CC=C(C=C1)CC2=CC=C(C=C2)CC(=O)C(F)(F)F`. Darker shading indicates higher values of the Hilbert function.

(d) 1st Homology two parameter persistence diagram for SMILES string `CCOC1=CC=C(C=C1)C2=CC=C(C=C2)S(=O)(=O)N`. Darker shading indicates higher values of the Hilbert function.

Figure 8: 0th and 1st Homology two parameter persistence diagrams for two molecules that demonstrated higher correlation between latent and persistence diagram metrics than latent and fingerprint metrics.