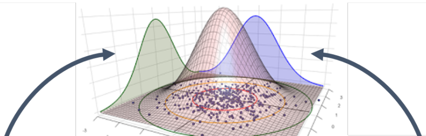
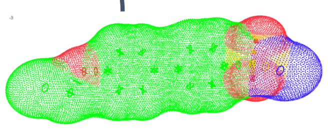
Characterizing the Latent Space of Molecular Deep Generative Models with Persistent Homology Metrics

Motivation

How much relevant semantic information is captured in the latent space of molecular generative models?



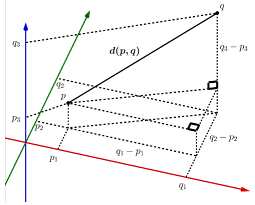
CCOC1=CC=C(C=C1)C2=C
C=C(C=C2)S(=O)(=O)N



Distribution image credit: https://medium.com/@nhuthai08/pattern-recognition-chapter-2-normal-distribution-f26d456db2a4

Approach

Euclidean distance of latent vectors



L2 distance of Restricted Hilbert function of 2-parameter persistence diagrams

$$\mathrm{Hil}^i_F(a) := \beta_i(F_a)$$

$$\mathrm{RH}^i_F(a) := \begin{cases} \mathrm{Hil}^i_F(a) & \text{for } a \in R_i(F), \\ 0 & \text{otherwise.} \end{cases}$$

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

vs.

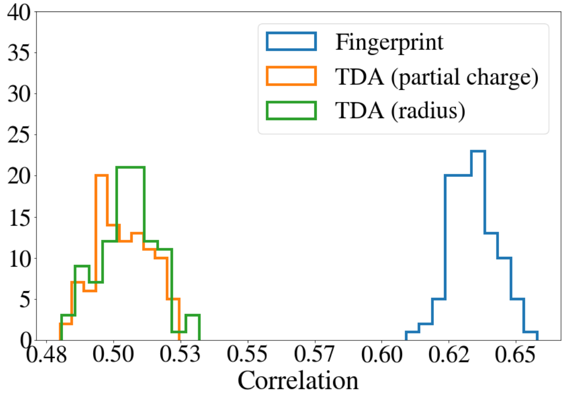
Tanimoto distance on Fingerprint representations

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

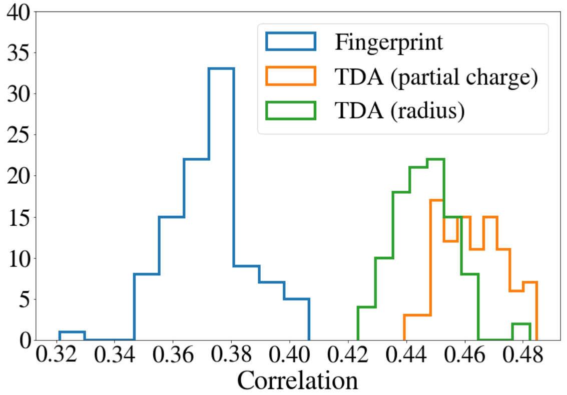
Euclidean distance image credit: https://commons.wikimedia.org/wiki/File:Euclidean_distance_3d_1_cropped.png

Results

Metric correlation analysis on a recent molecular VAE reveals that topological features are consistently preserved across that model's latent space



	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



NeurIPS | 2020

Yair Schiff, Vijil Chenthamarakshan, Karthikeyan Natesan Ramamurthy, and Payel Das

IBM Research