Yair Schiff, Vijil Chenthamarakshan,
Samuel C. Hoffman, Karthikeyan
Natesan Ramamurthy, Payel Das

Augmenting Molecular Deep Generative Models with Topological Data Analysis Representations

Paper Number: 4439

icassp 2022 Singapore

Background

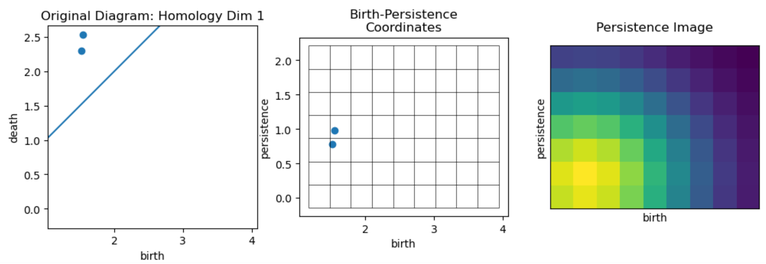
- Deep generative models in computer vision and natural language have inspired research in molecular generation.
- Popular input representations for molecules are strings and 2D graphs.
- Ignoring 3D positions of atoms during generation discards valuable information connected to their structure and target properties.
- **Goal:** Can we encode 3D information robustly and efficiently?

Core idea

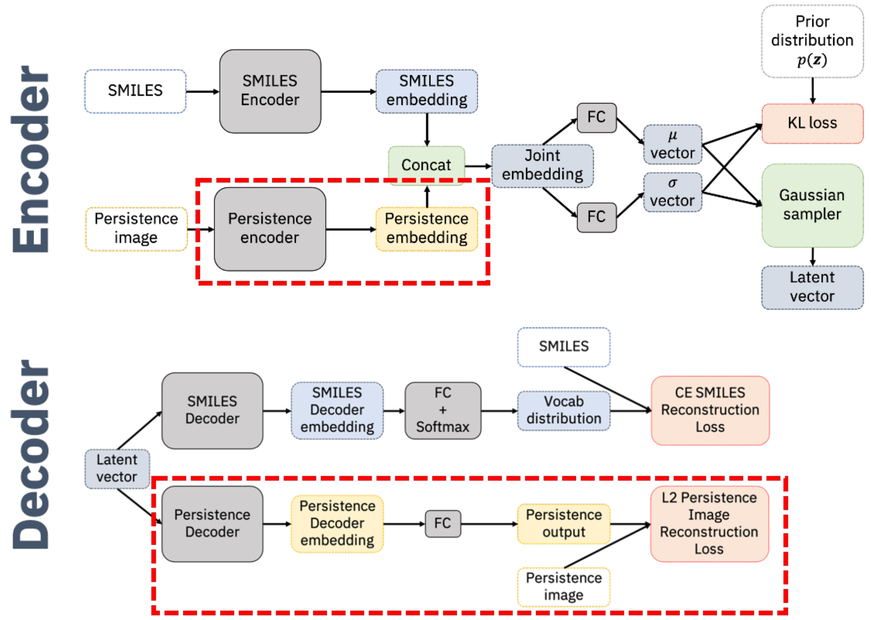
- Leverage persistent homology, a tool in Topological Data Analysis (TDA) to extract translation, rotation, and node permutation invariant, global 3D shape information about molecules.
- Learn a more informative latent representation of molecules by encoding and decoding both *SMILES* and TDA representations (*persistence images*).
- Use the latent space enhanced with this information to generate higher quality (novel, valid, geometrically consistent) molecules.

TDA: Analyzing the *shape of data*

- Represents global topological information and is robust to coordinate system used to represent molecules.
- Not overly sensitive to noise, helping generalization in real-world scenarios where molecules exhibit conformational dynamics.
- Persistence images are vector representations that can be pre-computed offline for efficient incorporation into ML models.



Architecture and Loss function



Objective function

$$\mathcal{L}_{\text{VAE+TDA}}(\theta, \phi; \beta, \lambda) =$$
$$\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z})]$$
$$- \beta D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))]$$
$$+ \lambda \mathbb{E}_{p(\mathbf{y})}[||\hat{\mathbf{y}} - \mathbf{y}||_2]$$

Our contribution

Empirical results

	QM9	SMILES	3D	3D + q	GVAE*	CGVAE†	MPGVAE*	MolGAN*	G-SchNet†
Validity	1.000	0.819	0.840	0.852	0.810	**1.000**	0.91	0.98	0.771
Atomic composition									
F	0.025	0.033	**0.019**	0.018	0.235	–	0.127	–	–
O	1.404	**1.406**	1.295	1.303	1.017	1.528	1.457	0.861	1.786
N	1.044	1.308	1.243	1.235	0.998	1.111	0.675	0.469	**1.071**
C	6.323	6.041	6.273	**6.282**	6.750	6.898	6.740	7.454	6.064
Sum	8.796	**8.789**	8.829	8.837	9.000	–	9.000	–	–
χ^2	–	0.002	**0.001**	**0.001**	0.014	–	0.009	–	–
Sum (No F)	8.771	8.755	8.810	8.819	**8.765**	9.537	8.872	8.784	8.921
χ^2 (No F)	–	0.002	0.001	0.001	0.004	**0.000**	0.005	0.025	0.003
Ring size									
R3	0.470	0.479	0.462	**0.470**	0.560	0.430	0.552	0.385	0.623
R4	0.586	0.490	0.561	**0.582**	0.333	0.692	0.647	0.247	0.657
R5	0.495	0.409	0.482	**0.483**	0.218	0.902	0.526	0.325	0.430
R6	0.158	0.169	0.155	**0.157**	0.110	0.649	0.104	0.115	0.133
Sum	1.709	1.600	**1.731**	1.734	1.222	2.673	1.828	1.072	1.843
χ^2	–	0.003	**0.000**	**0.000**	0.040	0.056	0.005	0.017	0.008

Table 1: Chemical validity, atomic composition, and ring size distribution of generated molecules. Sum total and χ^2 distances between QM9 ground truth and generated histograms are also provided for compositional analyses. Number next to "R" refers to the ring size. Best and second best results are indicated by bold and underline, respectively. Baseline values taken from [30] = *, [23] = †.

TDA-augmented VAEs better capture both atom and ring counts

	PI	Fingerprint
Baseline	0.214	0.306
3D coords	0.376	0.323
3D coords + q	**0.406**	**0.350**

Table 2: Pearson correlation coefficients between Euclidean distance in latent space and distances computed with two different structural similarity metrics for input data, estimated over all pairs of molecules in QM9 test set. In the *PI* column, we use Euclidean distance of 3D coordinate persistence images. In the *Fingerprint* column, we use the 1− Tanimoto similarity on MACCS keys.

Latent spaces of TDA-augmented VAEs better encode the structural information

Scan code to view full paper:
https://arxiv.org/abs/2106.04464

Poster presented at the IEEE International Conference on Acoustics, Speech and Signal Processing, 2022.