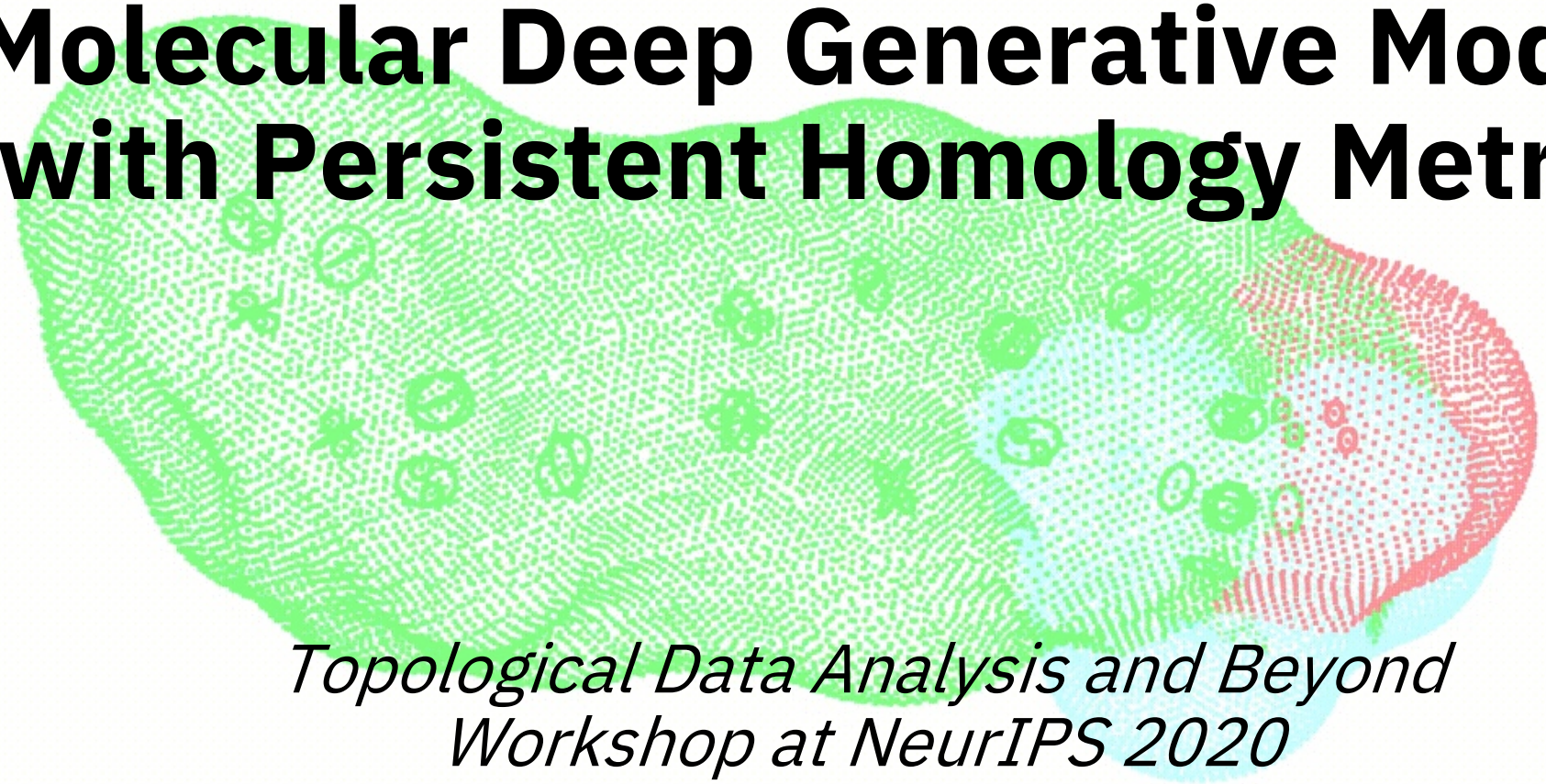
Characterizing the Latent Space of Molecular Deep Generative Models with Persistent Homology Metrics



Topological Data Analysis and Beyond
Workshop at NeurIPS 2020

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

IBM **Research**

Karthikeyan Natesan Ramamurthy

Yair Schiff

Vijil Chenthamarakshan

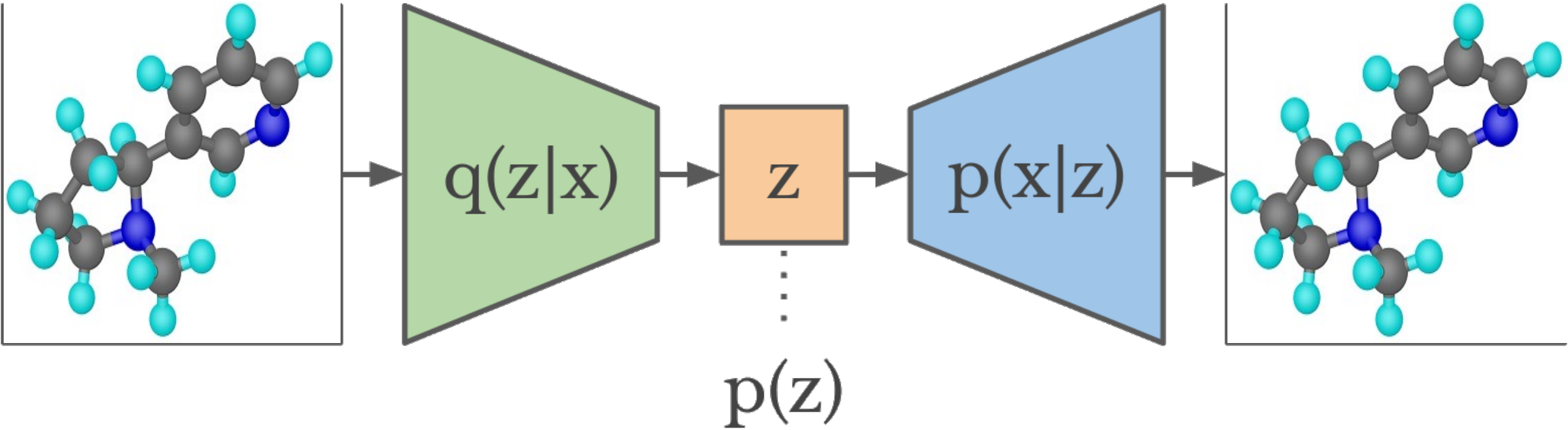
Payel Das

Molecule image credit: https://www.freeiconspng.com/img/9334

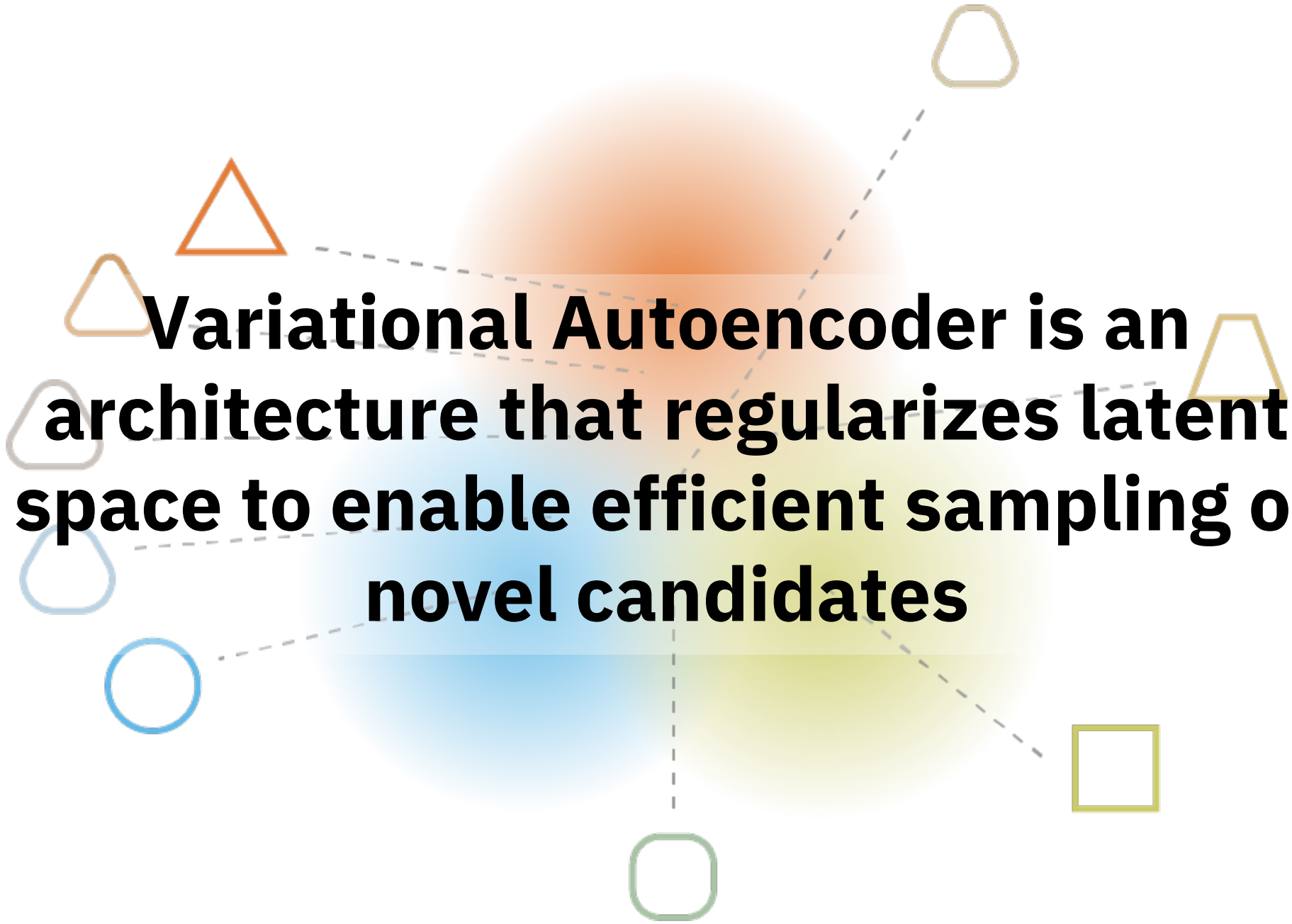
Drug development pipeline currently costs billions of dollars and spans decades

Image credit: https://villagehealthpartners.com/our-blog/what-you-need-to-know-about-lab-work

Deep generative models are accelerating this process



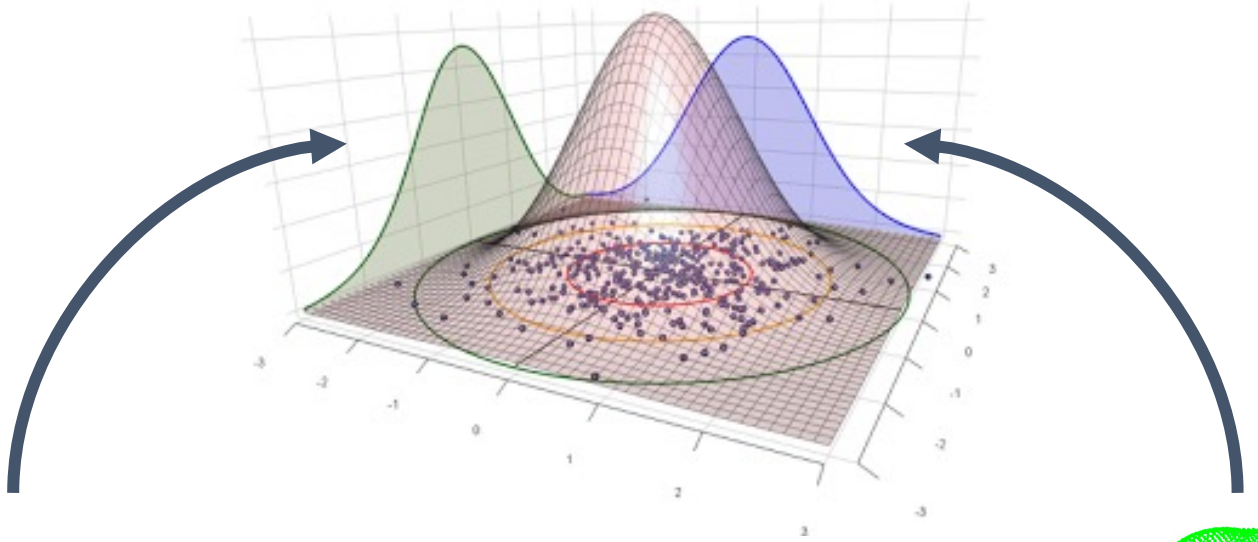
Model image credit: https://danijar.com/building-variational-auto-encoders-in-tensorflow/; Molecule image credit: https://keydifferences.com/difference-between-atom-and-molecule.html



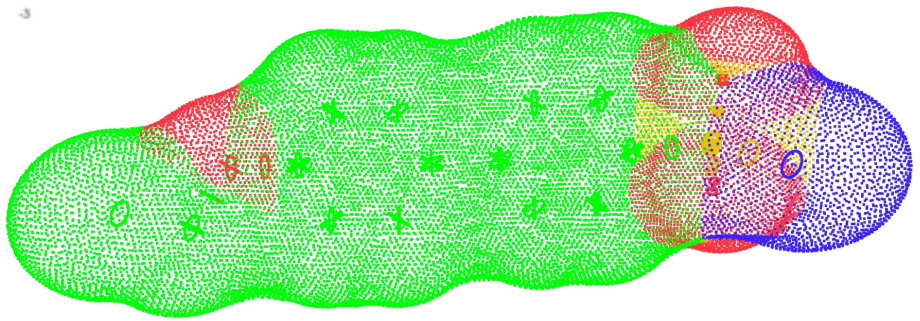
Variational Autoencoder is an architecture that regularizes latent space to enable efficient sampling of novel candidates

Image credit: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

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

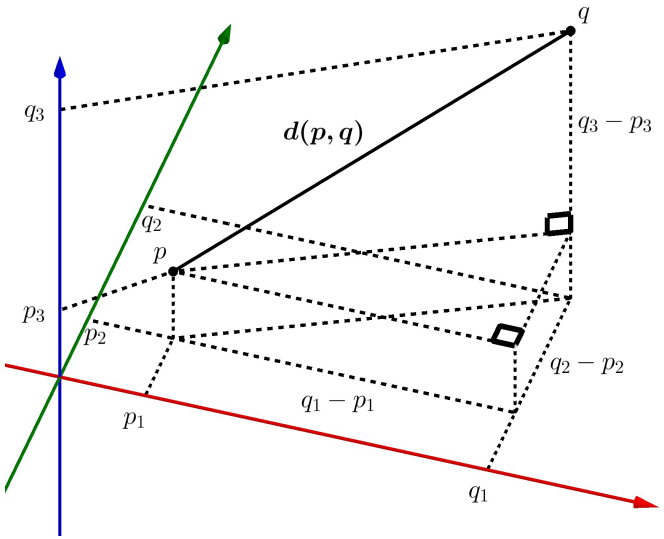


```
CCOC1=CC=C(C=C1)C2=C
C=C(C=C2)S(=O)(=O)N
```



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

Euclidean distance of latent vectors



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

vs.

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

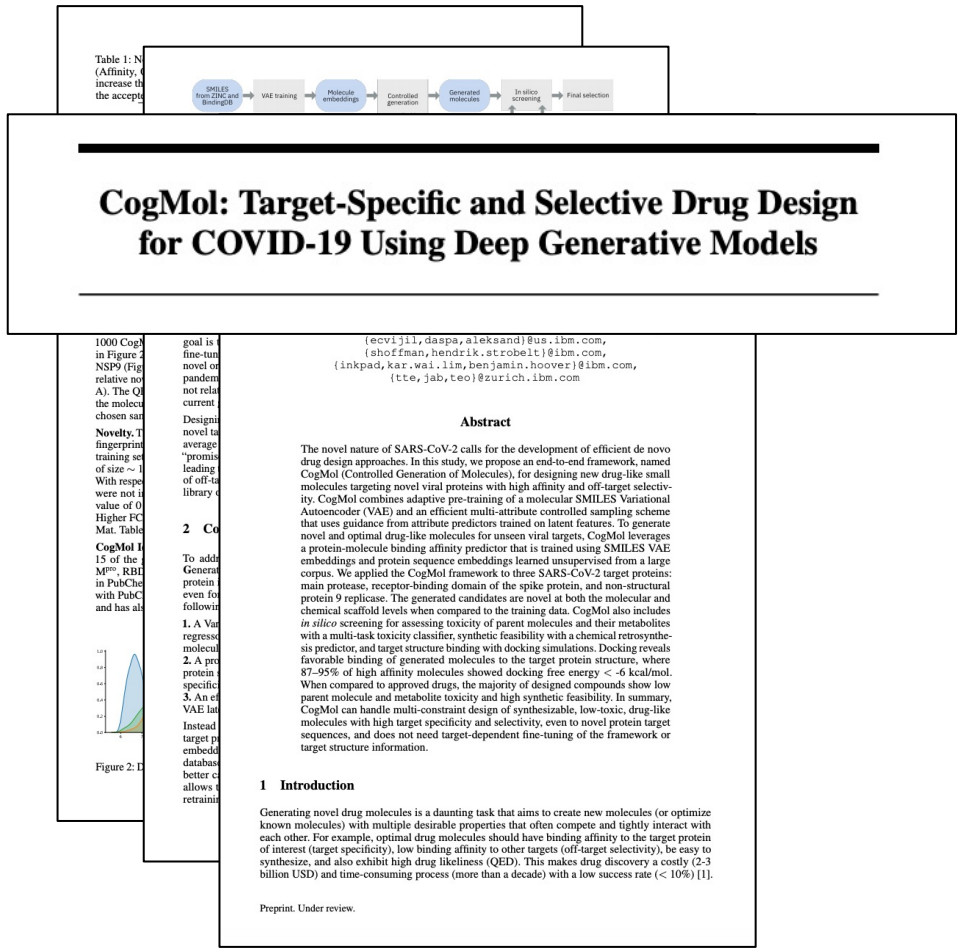
$$\mathrm{Hil}_F^i(a) := \beta_i(F_a)$$

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

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

Tanimoto distance on Fingerprint representations

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



CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models

{ecvijil,daspa,aleksand}@us.ibm.com,
{shoffman,hendrik.strobelt}@ibm.com,
{inkpad,kar.wai.lim,benjamin.hoover}@ibm.com,
{tte,jab,teo}@zurich.ibm.com

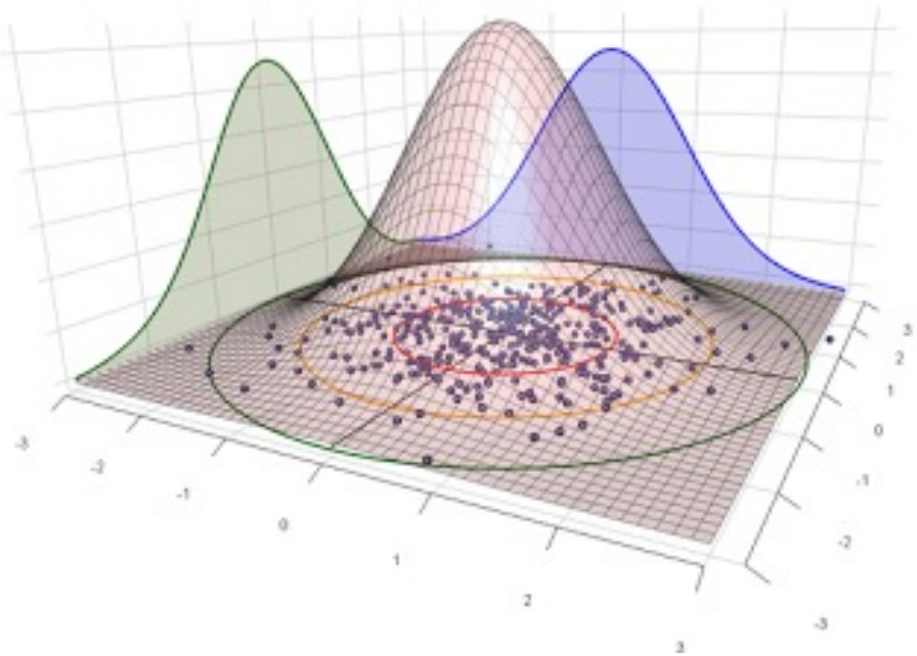
Abstract

The novel nature of SARS-CoV-2 calls for the development of efficient de novo drug design approaches. In this study, we propose an end-to-end framework, named CogMol (Controlled Generation of Molecules), for designing new drug-like small molecules targeting novel viral proteins with high affinity and off-target selectivity. CogMol combines adaptive pre-training of a molecular SMILES Variational Autoencoder (VAE) and an efficient multi-attribute controlled sampling scheme that uses guidance from attribute predictors trained on latent features. To generate novel and optimal drug-like molecules for unseen viral targets, CogMol leverages a protein-molecule binding affinity predictor that is trained using SMILES VAE embeddings and protein sequence embeddings learned unsupervised from a large corpus. We applied the CogMol framework to three SARS-CoV-2 target proteins: main protease, receptor-binding domain of the spike protein, and non-structural protein 9 replicase. The generated candidates are novel at both the molecular and chemical scaffold levels when compared to the training data. CogMol also includes *in silico* screening for assessing toxicity of parent molecules and their metabolites with a multi-task toxicity classifier, synthetic feasibility with a chemical retrosynthesis predictor, and target structure binding with docking simulations. Docking reveals favorable binding of generated molecules to the target protein structure, where 87–95% of high affinity molecules showed docking free energy < -6 kcal/mol. When compared to approved drugs, the majority of designed compounds show low parent molecule and metabolite toxicity and high synthetic feasibility. In summary, CogMol can handle multi-constraint design of synthesizable, low-toxic, drug-like molecules with high target specificity and selectivity, even to novel protein target sequences, and does not need target-dependent fine-tuning of the framework or target structure information.

1 Introduction

Generating novel drug molecules is a daunting task that aims to create new molecules (or optimize known molecules) with multiple desirable properties that often compete and tightly interact with each other. For example, optimal drug molecules should have binding affinity to the target protein of interest (target specificity), low binding affinity to other targets (off-target selectivity), be easy to synthesize, and also exhibit high drug likeliness (QED). This makes drug discovery a costly (2-3 billion USD) and time-consuming process (more than a decade) with a low success rate (< 10%) [1].

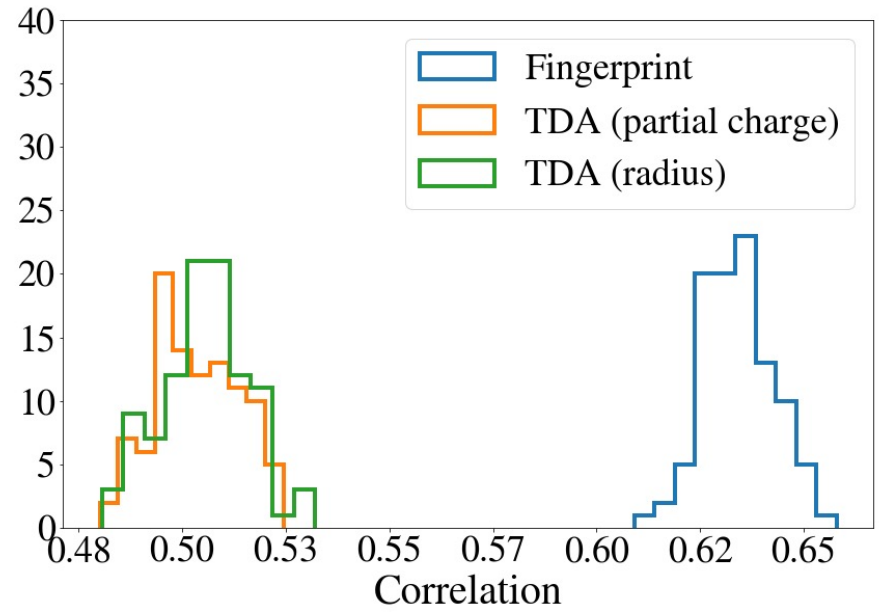
Preprint. Under review.

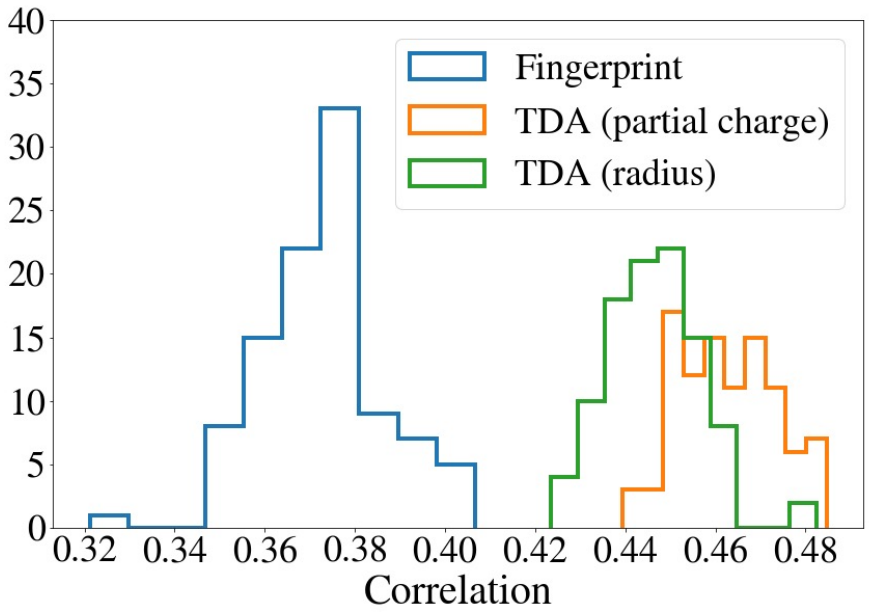


Applying our approach…

Article screenshots from: “CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models” (https://arxiv.org/abs/2004.01215); Distribution image credit: https://medium.com/@nhuthai08/pattern-recognition-chapter-2-normal-distribution-f26d456db2a4

Correlation analyses





	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

New visualizations

Incorporate TDA metrics directly into training

Image credit: https://www.hemophilia.org/Newsroom/NHF-Community-News/A-New-Path-Forward-in-Research; Analytics icon credit: https://www.onlinewebfonts.com/icon/569476; Target icon credit: https://uxwing.com/target-icon/


Thank you!