

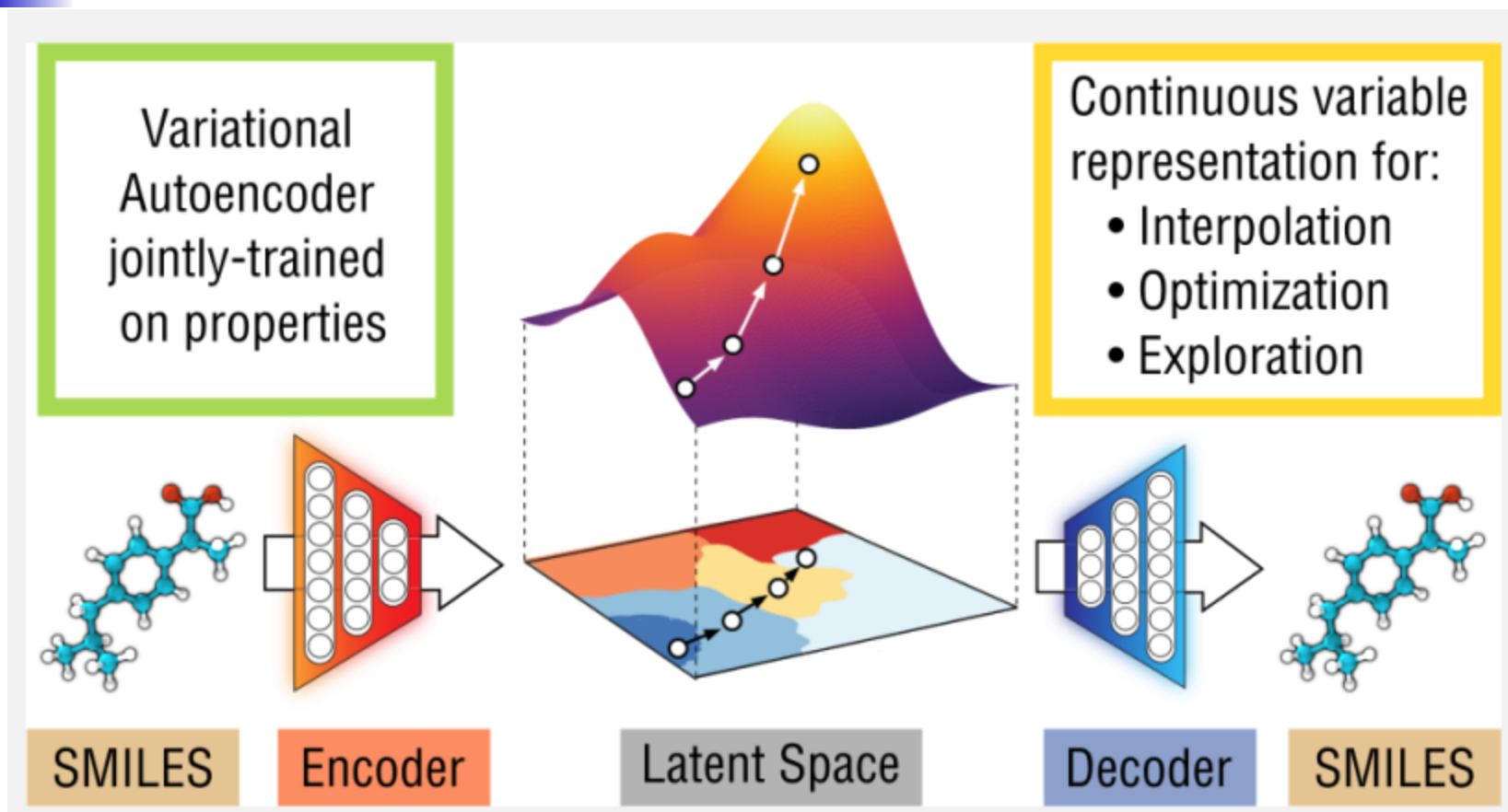


# Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

---

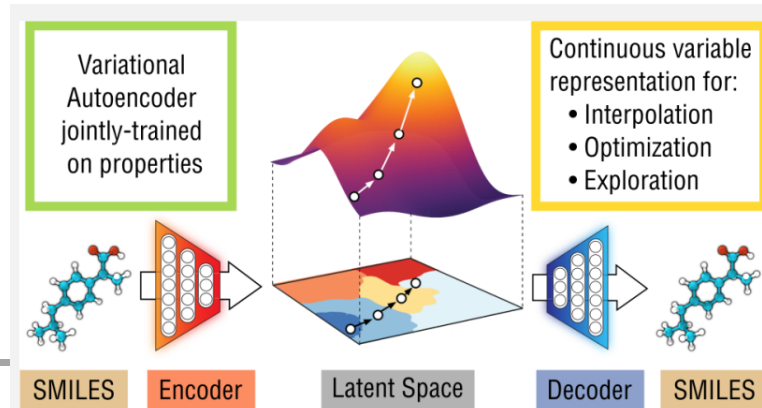
Reference: Rafael Gómez-Bombarelli , Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzelt, Ryan P. Adams, and Alán Aspuru-Guzik, “Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules”, ACS Cent. Sci. 2018, 4, 2, 268-276

# System Architecture



1  
2  
3

# System Architecture



This model allows to generate new molecules for efficient exploration and optimization through open-ended spaces of chemical compounds.

A deep neural network was trained on hundreds of thousands of existing chemical structures to construct three coupled functions: an encoder, a decoder, and a predictor.

The encoder converts the discrete representation of a molecule into a real-valued continuous vector, and the decoder converts these continuous vectors back to discrete molecular representations.

The predictor estimates chemical properties from the latent continuous vector representation of the molecule.

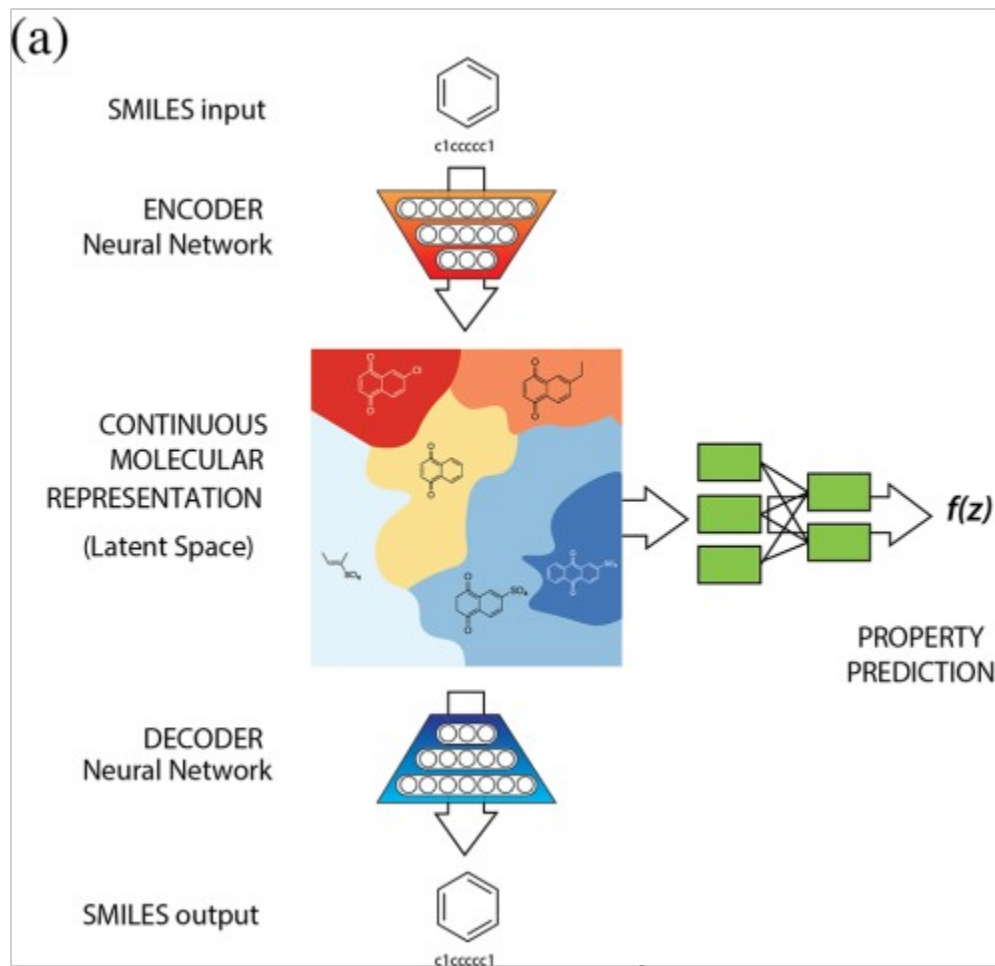
Continuous representations of molecules allow to automatically generate novel chemical structures by performing simple operations in the latent space, such as decoding random vectors, perturbing known chemical structures, or interpolating between molecules.

Figure 1. (a) A diagram of the autoencoder used for molecular design, including the joint property prediction model

Starting from a discrete molecular representation, such as a SMILES string, the encoder network converts each molecule into a vector in the latent space, which is effectively a continuous molecular representation.

Given a point in the latent space, the decoder network produces a corresponding SMILES string.

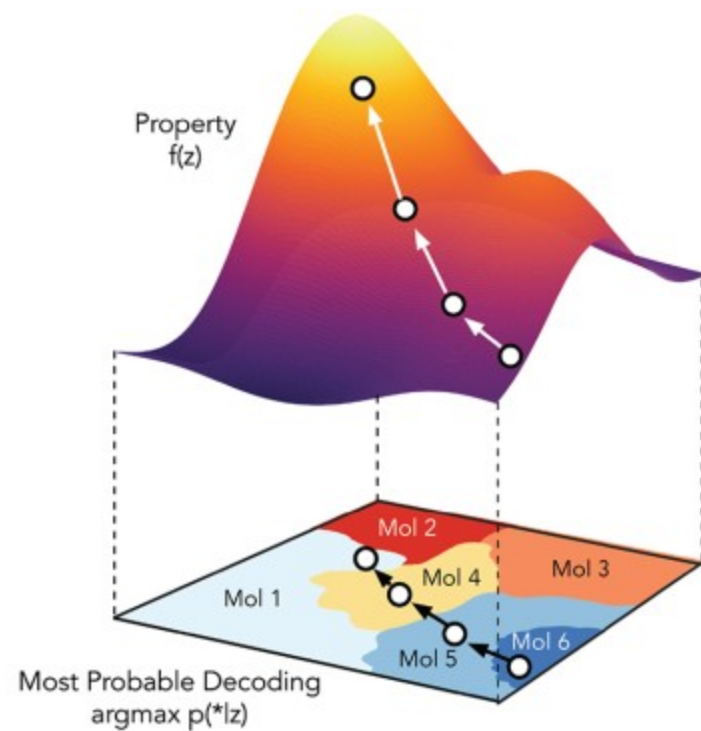
A multilayer perceptron network estimates the value of target properties associated with each molecule.



# Figure 1. (b) Gradient-based optimization in continuous latent space.

After training a surrogate model  $f(z)$  to predict the properties of molecules based on their latent representation  $z$ , we can optimize  $f(z)$  with respect to  $z$  to find new latent representations expected to have high values of desired properties.

These new latent representations can then be decoded into SMILES strings, at which point their properties can be tested empirically.





# Training system

---

Two autoencoder systems were trained: one with 108 000 molecules from the QM9 data set of molecules with fewer than 9 heavy atoms 31 and another with 250 000 drug-like commercially available molecules extracted at random from the ZINC database

The latent space representations for the QM9 and ZINC data sets had 156 dimensions and 196 dimensions, respectively.

SMILES-based text encoding used a subset of 35 different characters for ZINC and 22 different characters for QM9.

For ease of computation, they encoded strings up to a maximum length of 120 characters for ZINC and 34 characters for QM9, although in principle there is no hard limit to string length.

Shorter strings were padded with spaces to this same length.



# RESULTS AND DISCUSSION

---

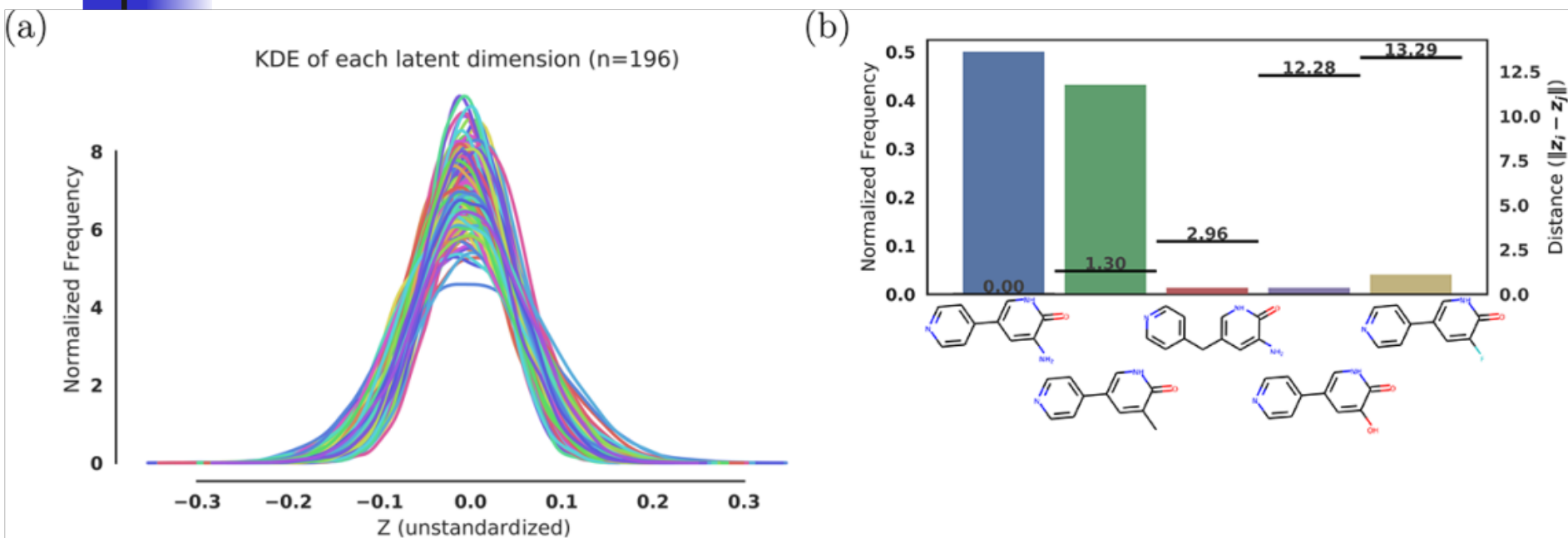
Representation of Molecules in  
Latent Space

Property Prediction of Molecules

Optimization of Molecules via  
Properties

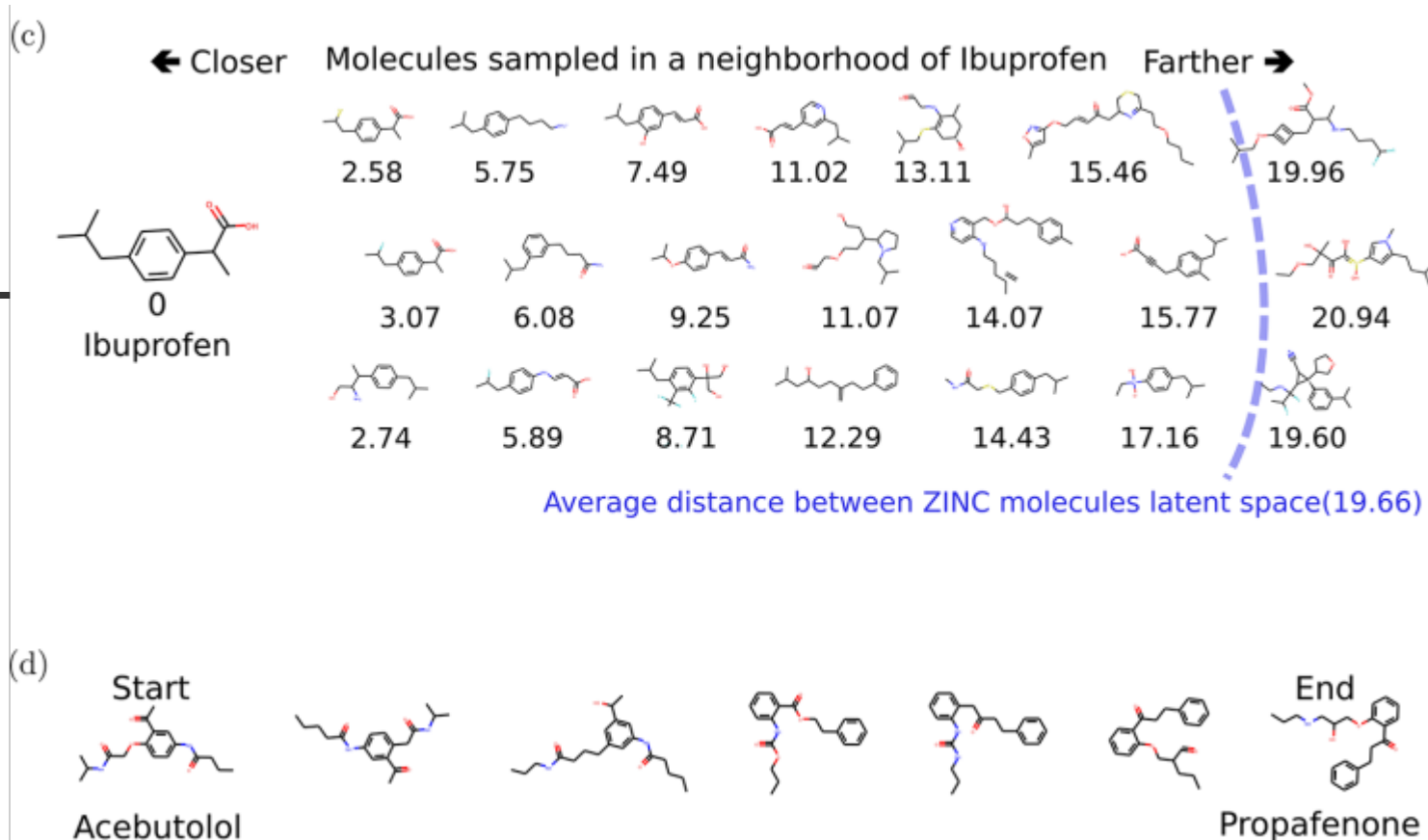
# Representation of Molecules in Latent Space

Figure 2. Representations of the sampling results from the variational autoencoder



- (a) Kernel Density Estimation (KDE) of each latent dimension of the autoencoder, i.e., the distribution of encoded molecules along each dimension of our latent space representation;
- (b) histogram of sampled molecules for a single point in the latent space; the distances of the molecules from the original query are shown by the lines corresponding to the right axis;

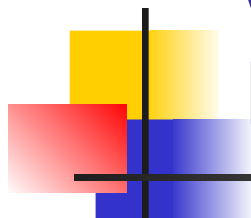




(c) molecules sampled near the location of ibuprofen in latent space. The values below the molecules are the distance in latent space from the decoded molecule to ibuprofen;

(d) slerp interpolation between two molecules in latent space using six steps of equal distance.

# Table 1. Comparison of Molecule Generation Results to Original Datasets

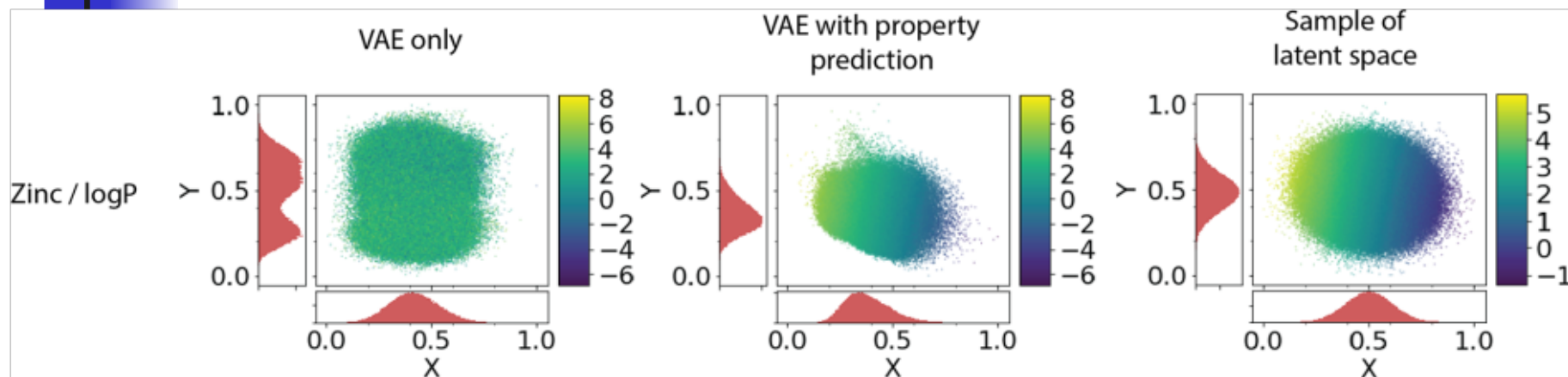


source <sup>a</sup>	data set <sup>b</sup>	samples <sup>c</sup>	logP <sup>d</sup>	SAS <sup>e</sup>	QED <sup>f</sup>	% in ZINC <sup>g</sup>	% in emol <sup>h</sup>
Data	ZINC	249k	2.46 (1.43)	3.05 (0.83)	0.73 (0.14)	100	12.9
GA	ZINC	5303	2.84 (1.86)	3.80 (1.01)	0.57 (0.20)	6.5	4.8
VAE	ZINC	8728	2.67 (1.46)	3.18 (0.86)	0.70 (0.14)	5.8	7.0
Data	QM9	134k	0.30 (1.00)	4.25 (0.94)	0.48 (0.07)	0.0	8.6
GA	QM9	5470	0.96 (1.53)	4.47 (1.01)	0.53 (0.13)	0.018	3.8
VAE	QM9	2839	0.30 (0.97)	4.34 (0.98)	0.47 (0.08)	0.0	8.9

- a. Describes the source of the molecules: data refers to the original data set, GA refers to the genetic algorithm baseline, and VAE to our variational autoencoder trained without property prediction.
- b. Shows the data set used, either ZINC or QM9.
- c. Shows the number of samples generated for comparison, for data, this value simply reflects the size of the data set. Columns d–f show the mean and, in parentheses, the standard deviation of selected properties of the generated molecules and compares that to the mean and standard deviation of properties in the original data set.
- d. Shows the water–octanol partition coefficient (logP).<sup>36</sup> eShows the synthetic accessibility score (SAS).<sup>37</sup>
- f. Shows the Qualitative Estimate of Drug-likeness (QED),<sup>38</sup> ranging from 0 to 1. We also examine how many of the molecules generated by each method are found in two major molecule databases: gZINC; hE-molecules<sup>39</sup>, and compare these values against the original data set.

# Property Prediction of Molecules

Figure 3. Two-dimensional PCA analysis of latent space for variational autoencoder



Two-dimensional PCA analysis of latent space for variational autoencoder.

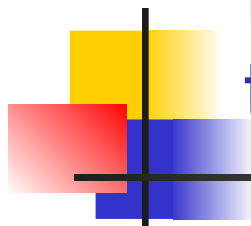
The two axis are the principle components selected from the PCA analysis; the color bar shows the value of the selected property.

The first column shows the representation of all molecules from the listed data set using autoencoders trained without joint property prediction.

The second column shows the representation of molecules using an autoencoder trained with joint property prediction.

The third column shows a representation of random points in the latent space of the autoencoder trained with joint property prediction; the property values predicted for these points are predicted using the property predictor network.

# Table 2. MAE Prediction Error for Properties Using Various Methods on the ZINC and QM9 Datasets



database/property	mean <sup>a</sup>	ECFP <sup>b</sup>	CM <sup>b</sup>	GC <sup>b</sup>	1-hot SMILES <sup>c</sup>	Encoder <sup>d</sup>	VAE <sup>e</sup>
ZINC250k/logP	1.14	0.38		0.05	0.16	0.13	0.15
ZINC250k/QED	0.112	0.045		0.017	0.041	0.037	0.054
QM9/HOMO, eV	0.44	0.20	0.16	0.12	0.12	0.13	0.16
QM9/LUMO, eV	1.05	0.20	0.16	0.15	0.11	0.14	0.16
QM9/Gap, eV	1.07	0.30	0.24	0.18	0.16	0.18	0.21

a. Baseline, mean prediction.

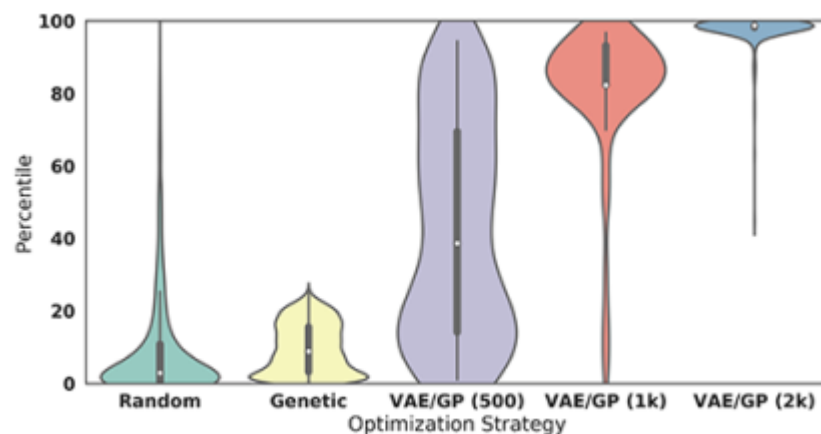
b. As implemented in Deepchem benchmark (MoleculeNet), 40 ECFP-circular fingerprints, CM-coulomb matrix, GCgraph convolutions.

c. 1-hot-encoding of SMILES used as input to property predictor.

d. The network trained without decoder loss. eFull variational autoencoder network trained for individual properties.

# Optimization of Molecules via Properties

Figure 4. Optimization results for the jointly trained autoencoder using  $5 \times \text{QED} - \text{SAS}$  as the objective function

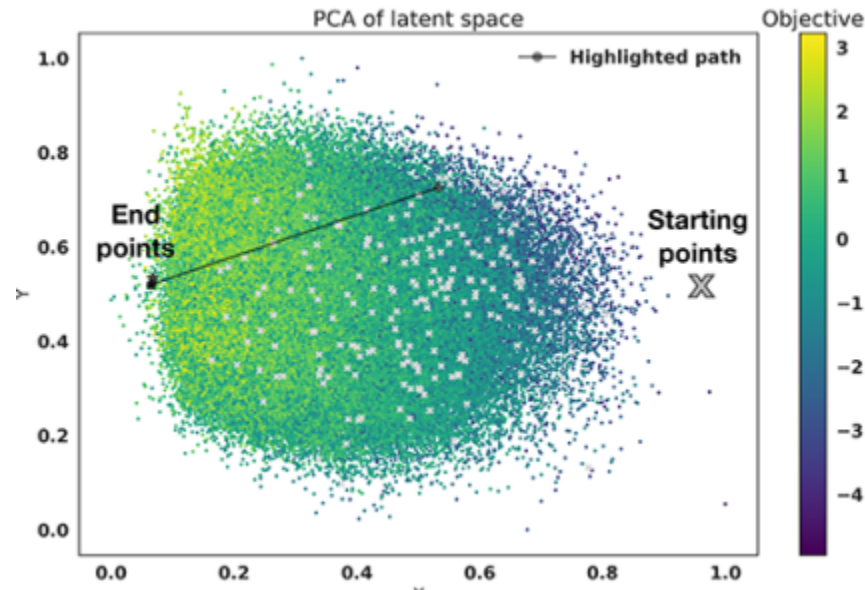


(a) shows a violin plot which compares the distribution of sampled molecules from normal random sampling, SMILES optimization via a common chemical transformation with a genetic algorithm, and from optimization on the trained Gaussian process model with varying amounts of training points.

To offset differences in computational cost between the random search and the optimization on the Gaussian process model, the results of 400 iterations of random search were compared against the results of 200 iterations of optimization.

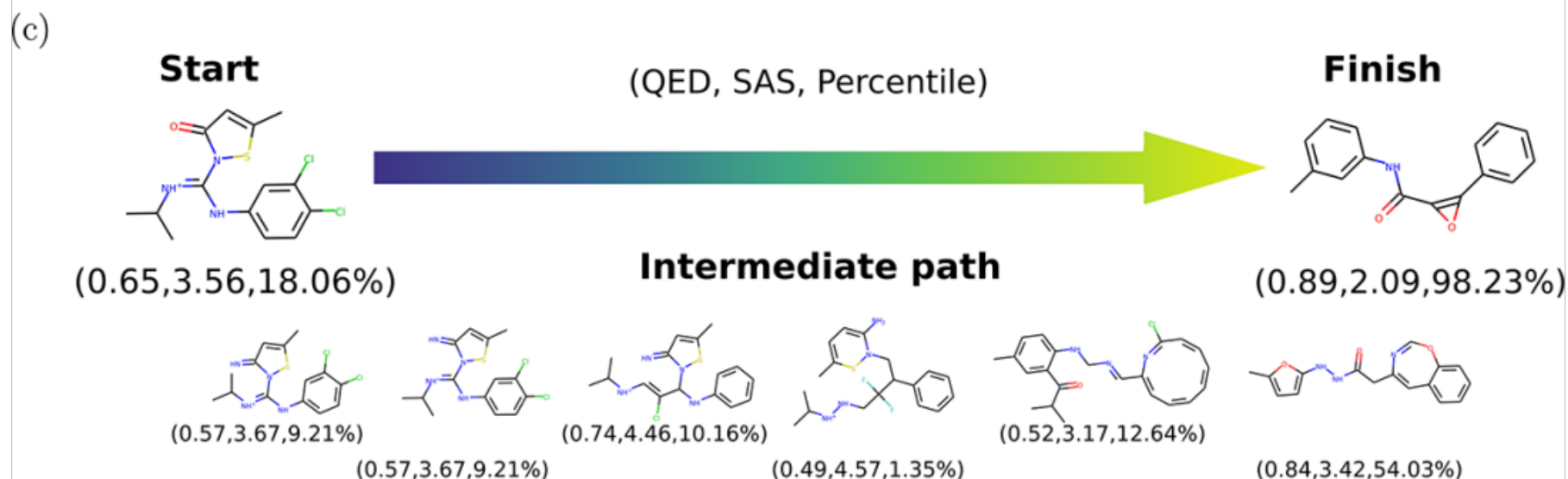
This graph shows the combined results of four sets of trials.

Figure 4. Optimization results for the jointly trained autoencoder using  $5 \times \text{QED} - \text{SAS}$  as the objective function



(b) shows the starting and ending points of several optimization runs on a PCA plot of latent space colored by the objective function. Highlighted in black is the path illustrated in part (c).

Figure 4. Optimization results for the jointly trained autoencoder using  $5 \times \text{QED} - \text{SAS}$  as the objective function



(c) shows a spherical interpolation between the actual start and finish molecules using a constant step size. The QED, SAS, and percentile score are reported for each molecule.



# Conclusions

---

Propose a new family of methods for exploring chemical space based on continuous encodings of molecules.

These methods eliminate the need to hand-craft libraries of compounds and allow a new type of directed gradient-based search through chemical space. In autoencoder model, observed high fidelity in reconstruction of SMILES strings and the ability to capture characteristic features of a molecular training set.

The autoencoder exhibited good predictive power when training jointly with a property prediction task, and the ability to perform gradient-based optimization of molecules in the resulting smoothed latent space.