Deep Learning and Chemistry

Chris Koch

Scale of Industrial Chemical Industry

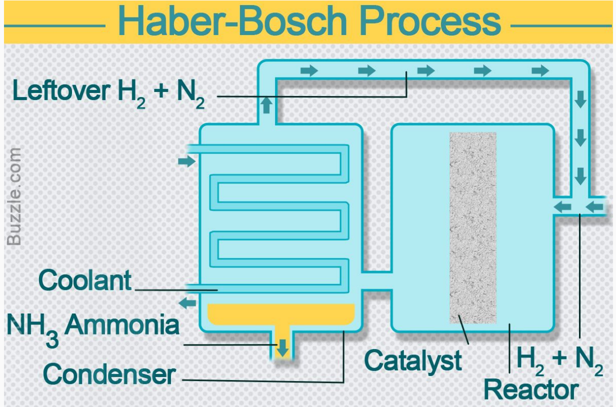
The chemical industry has revenues of $3 trillion annually.

It uses 12% of global industrial energy.

Half of the Nitrogen atoms in your body have been through a single, man-made reaction at one point, known as the Haber-Bosch process.

The Haber-Bosch process is considered a large factor in the global population increase from 1.6 billion people in 1900 to about 7.7 billion people today.

Cement is responsible for 7% of annual CO_2 emissions.



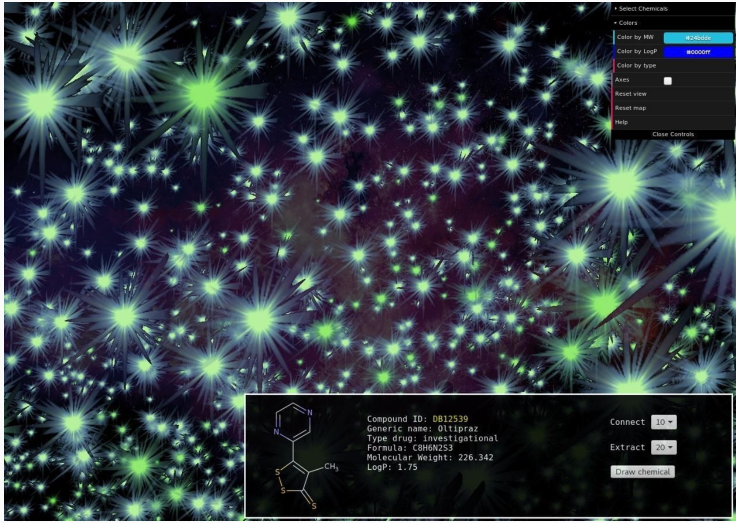
Searching Chemical Space

The number of possible small, drug-like chemical compounds has been estimated to be between 10^{23} and 10^{60}.

There have only been 10^{8} compounds synthesized in human history.

For chemical development, the number of possible small chemical compounds is estimated to be around 10^{200}.

Traditional methods cannot perform unguided searches of chemical space.



Color by MW #24bdde
Color by LogP #0000ff

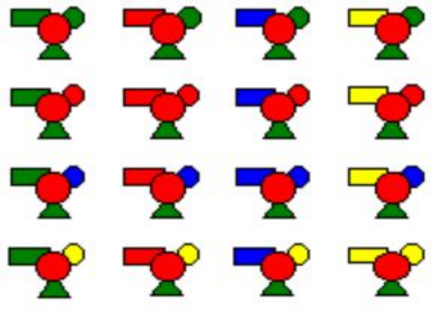
Compound ID: DB12539
Generic name: Oltipraz
Type drug: investigational
Formula: C8H6N2S3
Molecular Weight: 226.342
LogP: 1.75

Connect 10
Extract 20

Traditional Chemical Development Process

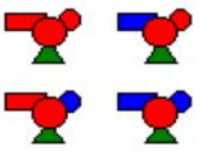
Chemical Libraries are Built Using Scaffolding

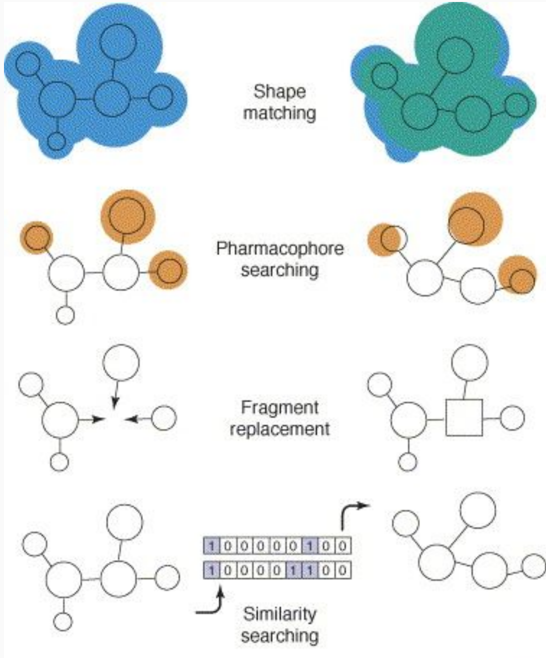
Massive chemical libraries are built by appending chemical “building blocks” (functional groups and chemical fragments) to existing chemicals with desirable properties (“scaffold” molecules)



Chemical Libraries Are Screened

Heuristics and resource intensive simulations are used to narrow down pool of candidate compounds.

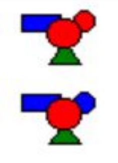




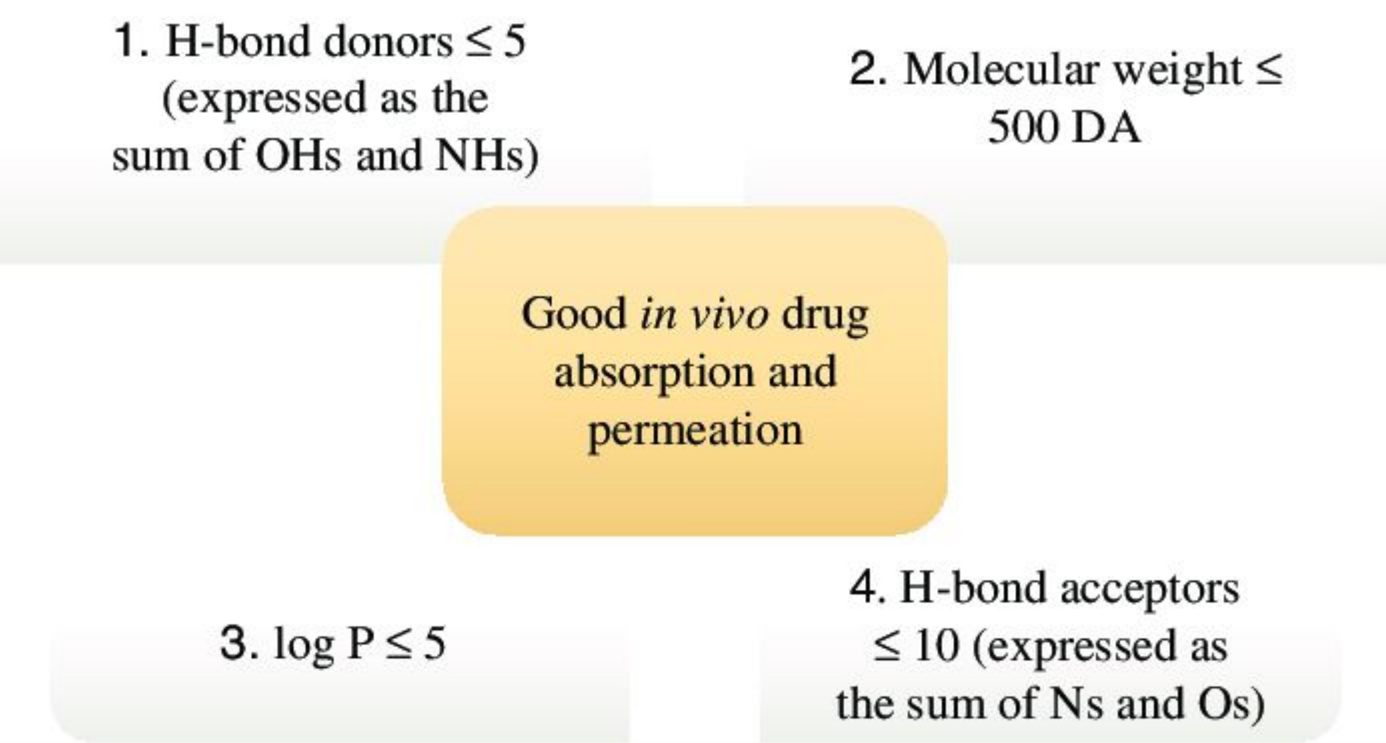
Drug Discovery Today: Technologies

Promising Compounds Undergo Further Optimization

Chemical engineers determine the right processes needed to synthesize these compounds, then they are physically tested.



Traditional Heuristics for Drug Development



1. H-bond donors ≤ 5 (expressed as the sum of OHs and NHs)

2. Molecular weight ≤ 500 DA

Good *in vivo* drug absorption and permeation

3. log P ≤ 5

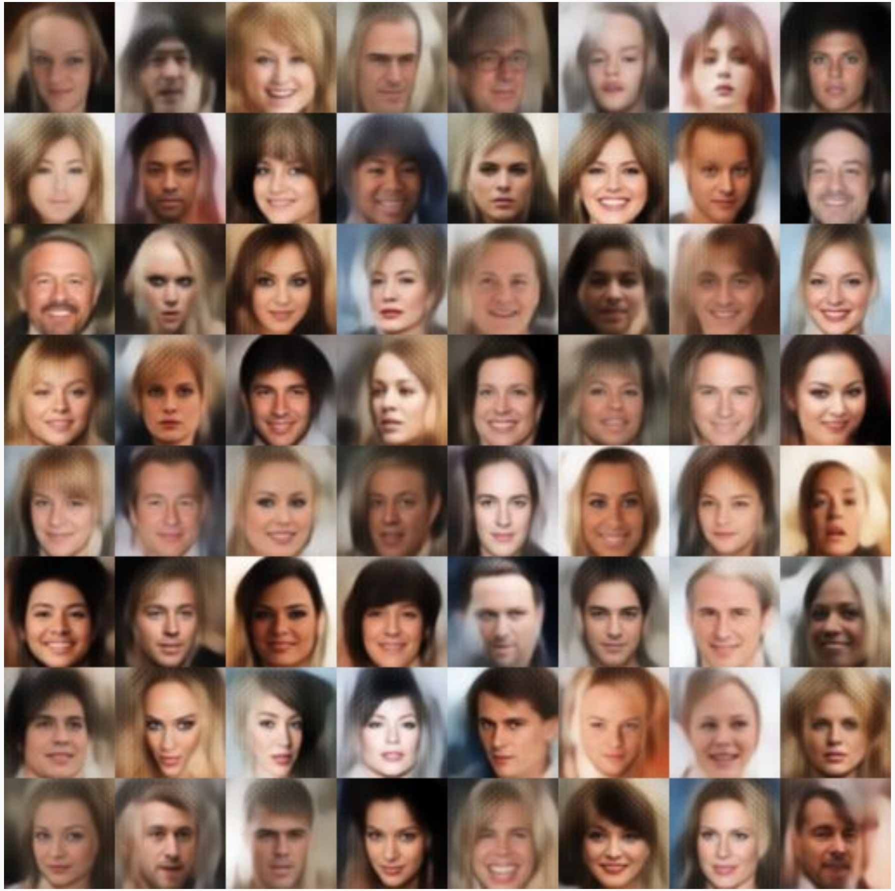
4. H-bond acceptors ≤ 10 (expressed as the sum of Ns and Os)

Lipinski’s Rule of Five: A heuristic used to predict oral bioavailability of a compound.

Heuristics can narrow down a search space, but rules based only on our knowledge of existing chemistry can miss novel structures and compound classes or produce suboptimal results.

For many chemical properties, there aren’t robust heuristics yet.

Generated Celebrity Faces using a Variational Autoencoder



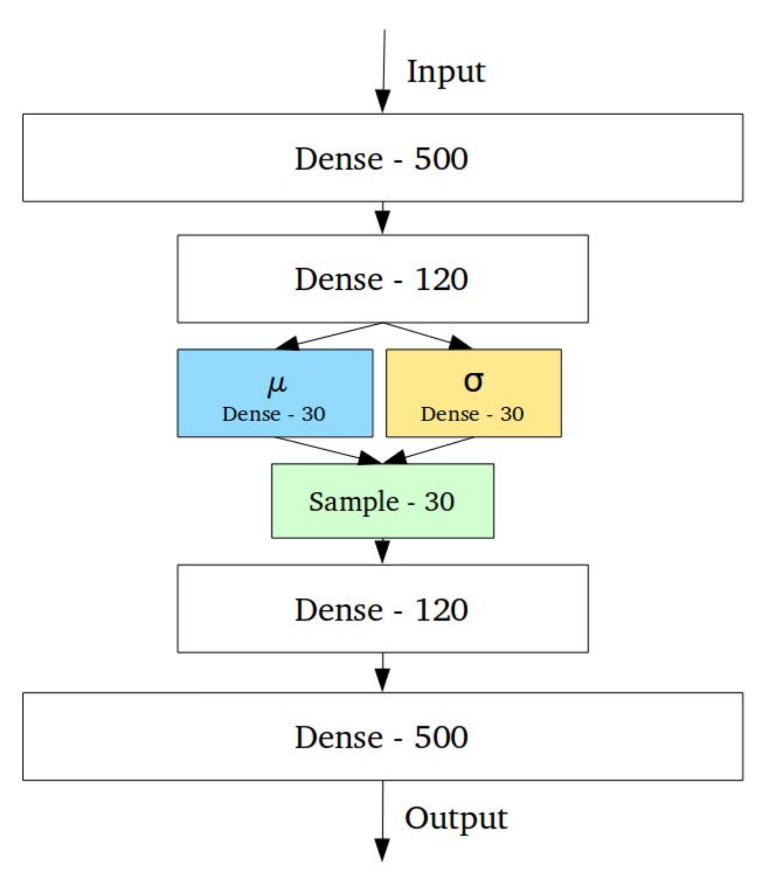
Variational Autoencoders

One neural network, the Encoder, compresses input data into a small, continuous vector.

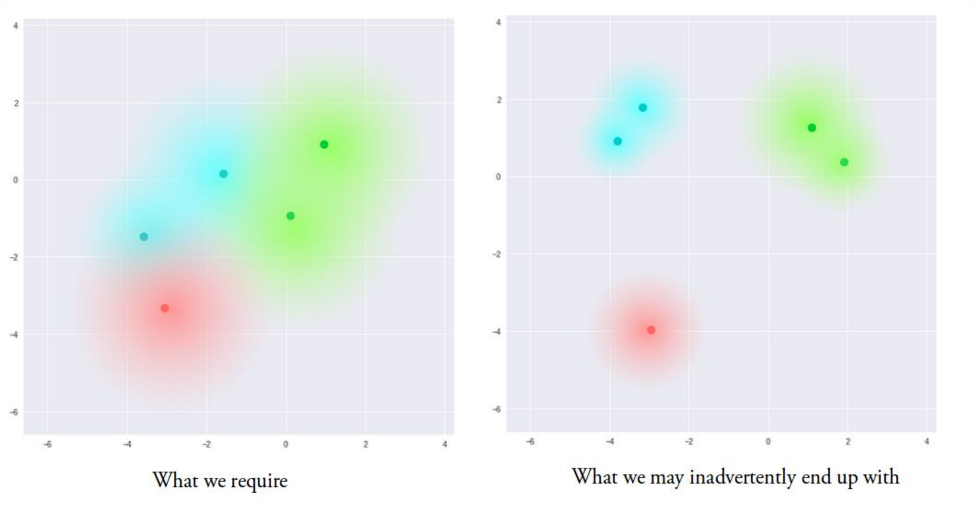
This continuous vector is known as the Code or encoding.

The Decoder attempts to reconstruct the input data from the Code. It is trained using reconstruction loss - the difference between the input and the output.

A Variational Autoencoder (VAE) involves random sampling of means and variances of the output activations from the encoder.

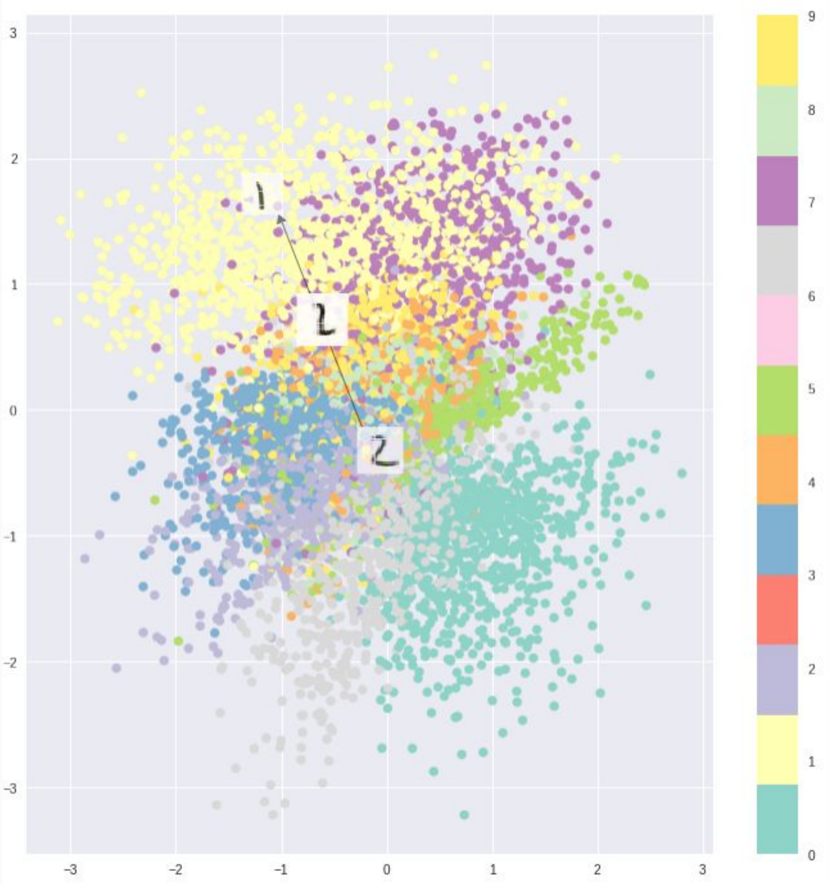


VAE and Latent Space



What we require

What we may inadvertently end up with

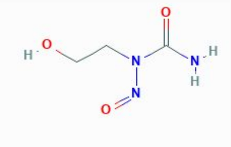


$$\sum_{i=1}^{n} \sigma_i^2+\mu_i^2-\log(\sigma_i)-1$$

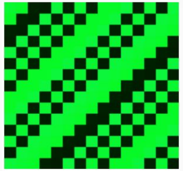
Kullback–Leibler Divergence: Measures divergence of two distributions. In the equation above, the standard normal is the second distribution.

Molecular Encodings

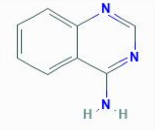
$C_3H_7N_3O$



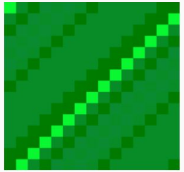
NC(=O)N(CCO)N=O



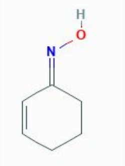
$C_8H_7N_3$



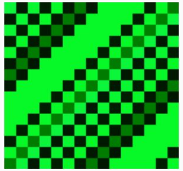
Nc1ncnc2ccccc12



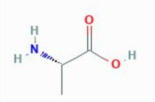
C_6H_9NO



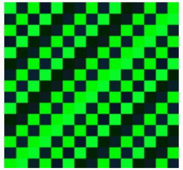
O\N=C\1/CCCC=C1



$C_3H_7NO_2$



C[C@H](N)C(=O)O



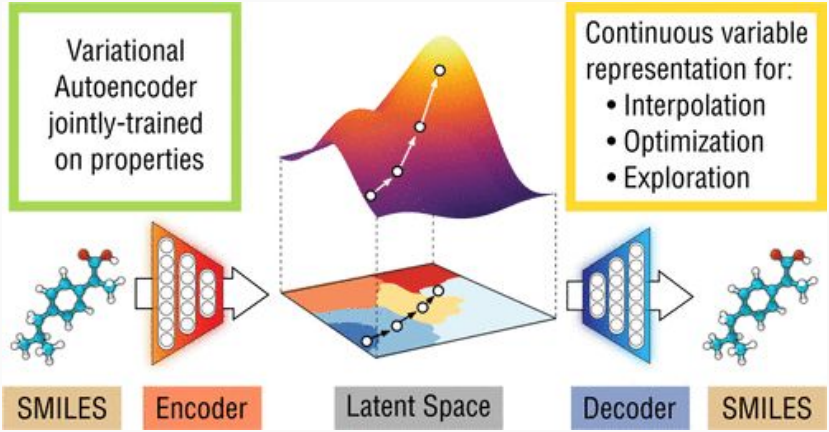
Variational Autoencoders for Chemical Development

“Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules” by Bombarelli et al. 2017

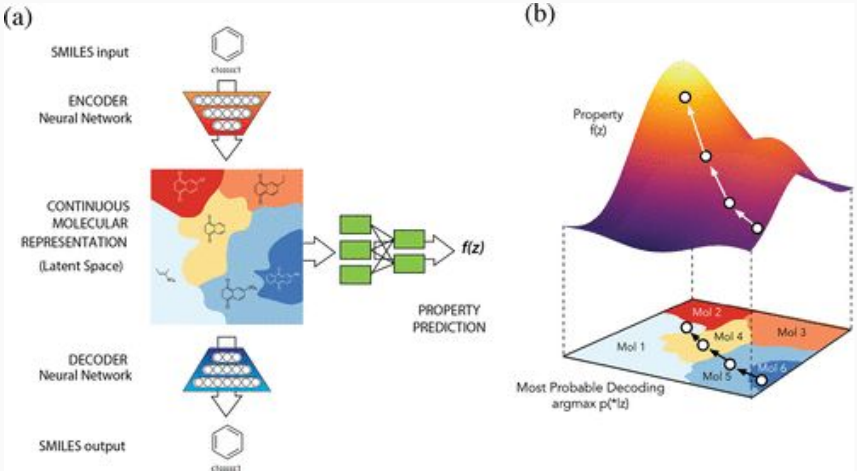
Introduced the use of Variational Autoencoders for chemical generation.

VAEs produce a continuous representation which can be used for interpolation and exploration in the latent space.

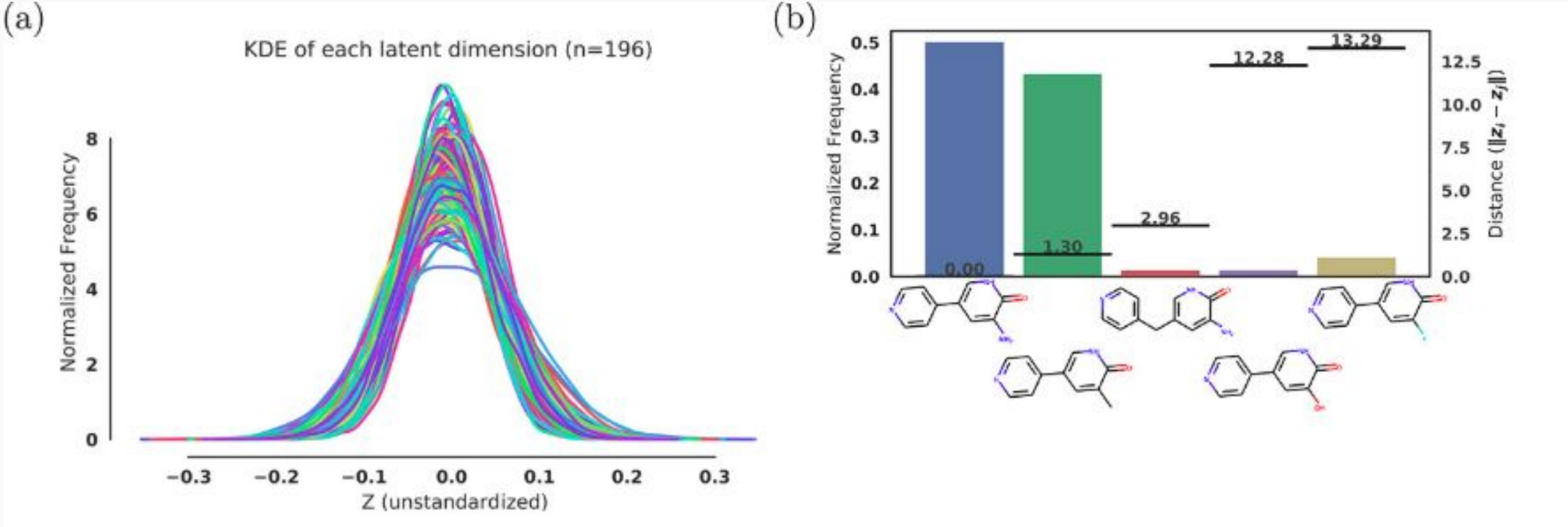
Property prediction network trained on log P, SAS, and QED. This property prediction network is jointly trained with the autoencoder.



Variational Autoencoder jointly-trained on properties
Continuous variable representation for:
• Interpolation
• Optimization
• Exploration
SMILES
Encoder
Latent Space
Decoder
SMILES

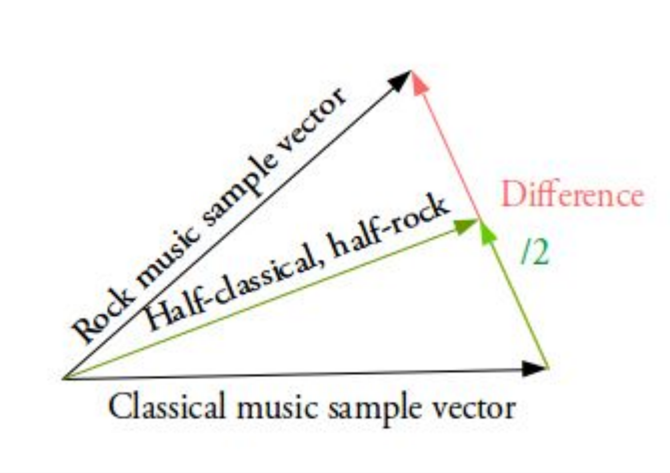


Density and Nondeterministic Decoding

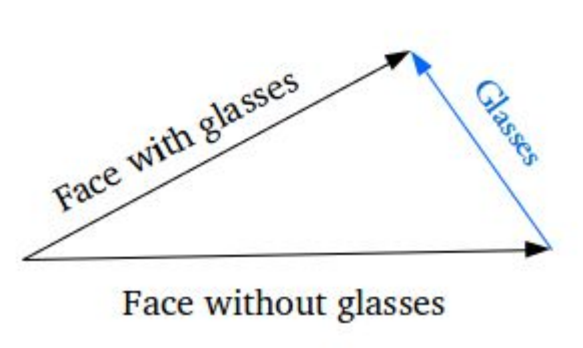


(a) Kernel Density Estimation (KDE) of each latent dimension of the autoencoder, i.e., the distribution of encoded molecules along each dimension of our latent space representation; (b) histogram of sampled molecules for a single point in the latent space; the distances of the molecules from the original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

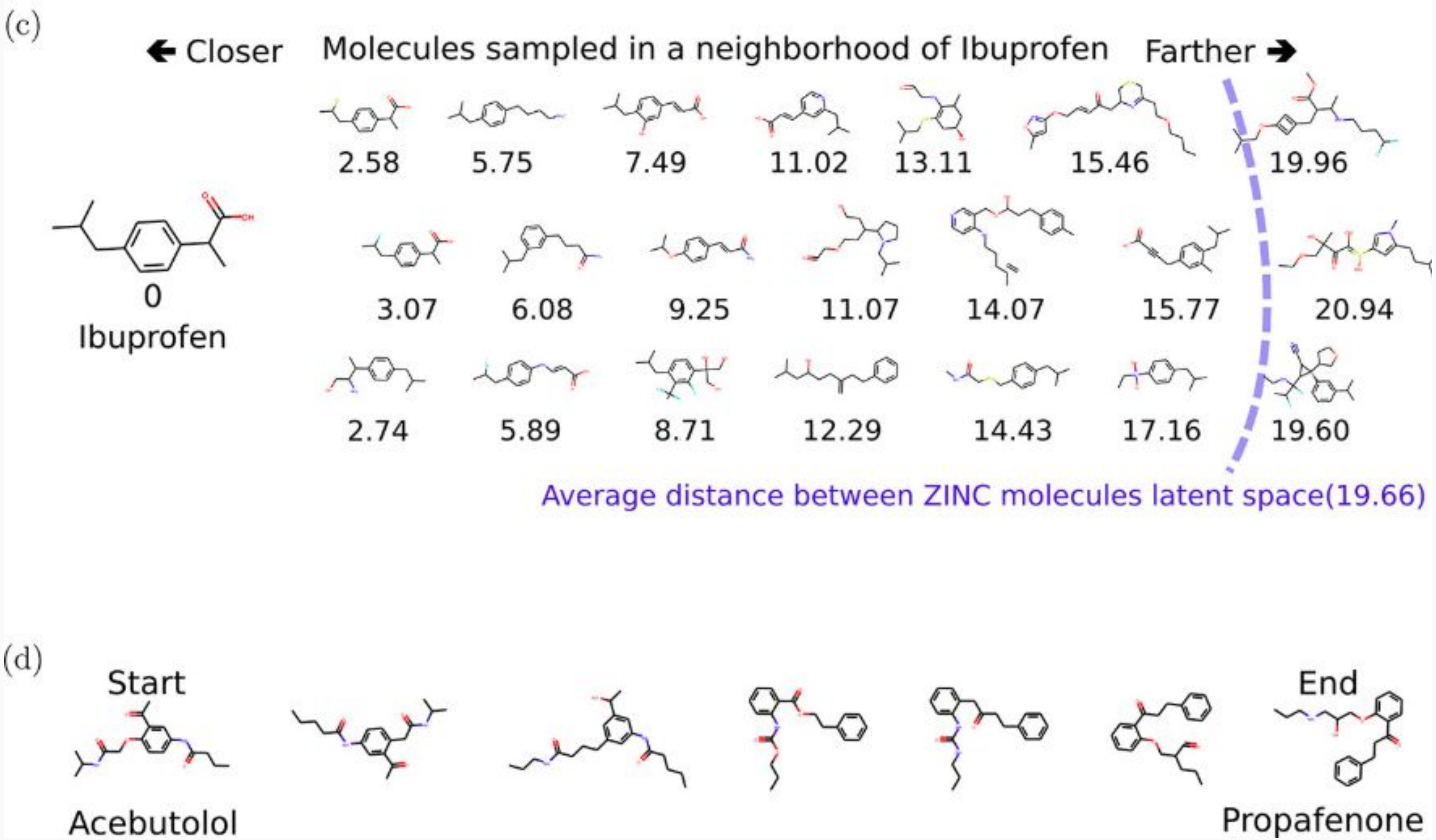
Interpolating and Adding Features



Interpolation between two different molecules.



Feature addition: The difference of two vectors is used to calculate a third vector which is the feature that is present in vector A but not in vector B.



(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.

Organization of Latent Space

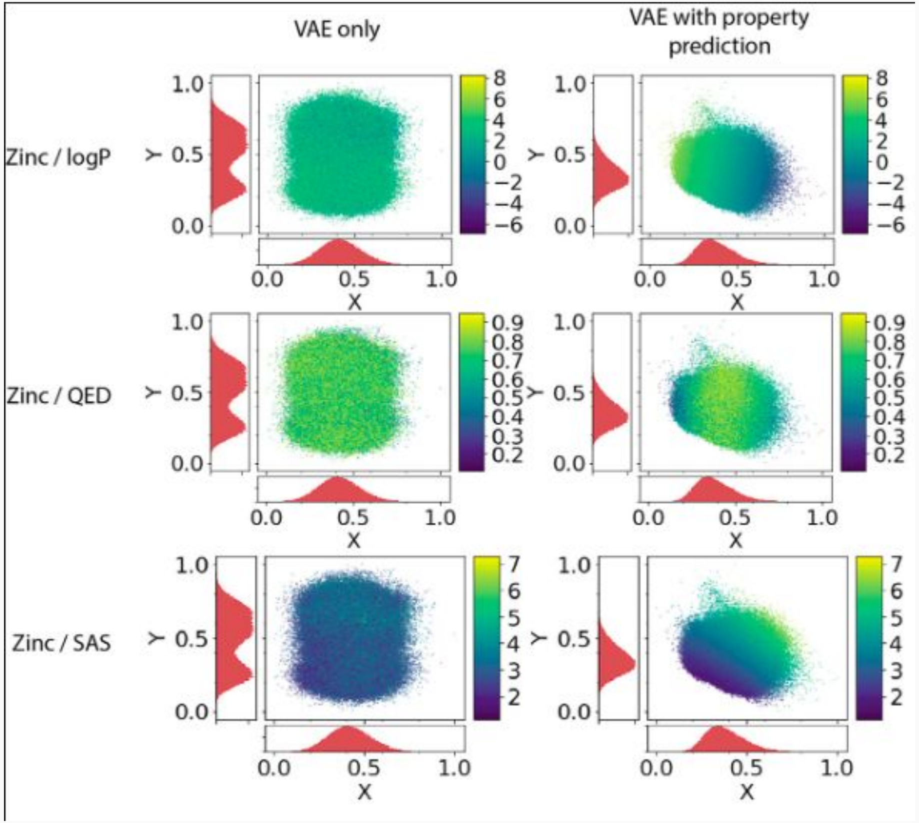


Figure 3. Two-dimensional PCA analysis of latent space for variational autoencoder. The two axis are the principle components selected from the PCA analysis; the color bar shows the value of the selected property. The first column shows the representation of all molecules from the listed data set using autoencoders trained without joint property prediction. The second column shows the representation of molecules using an autoencoder trained with joint property prediction. The first three rows show the results of training on molecules from the ZINC data set for the logP, QED, and SAS properties;

Molecular Optimization

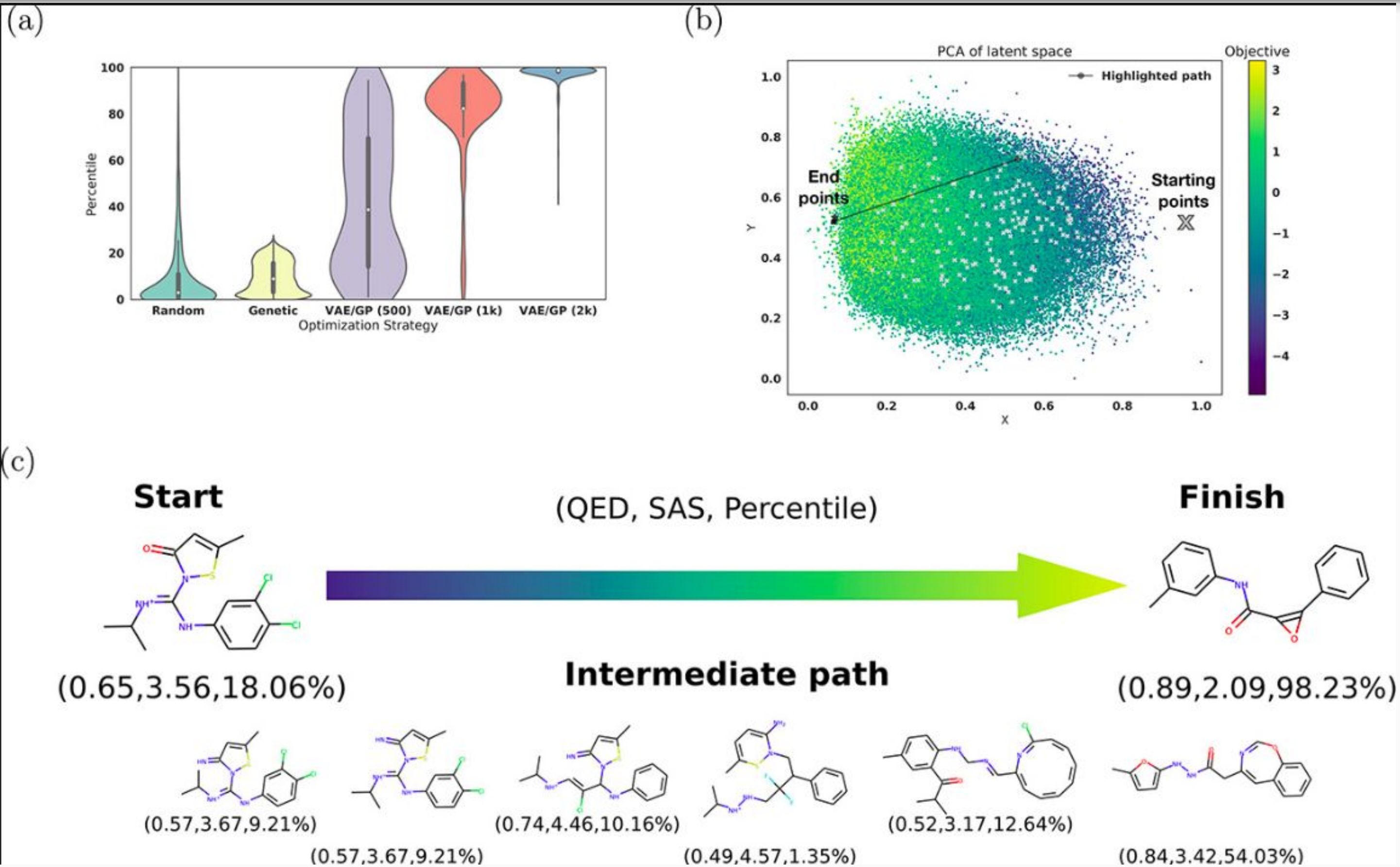
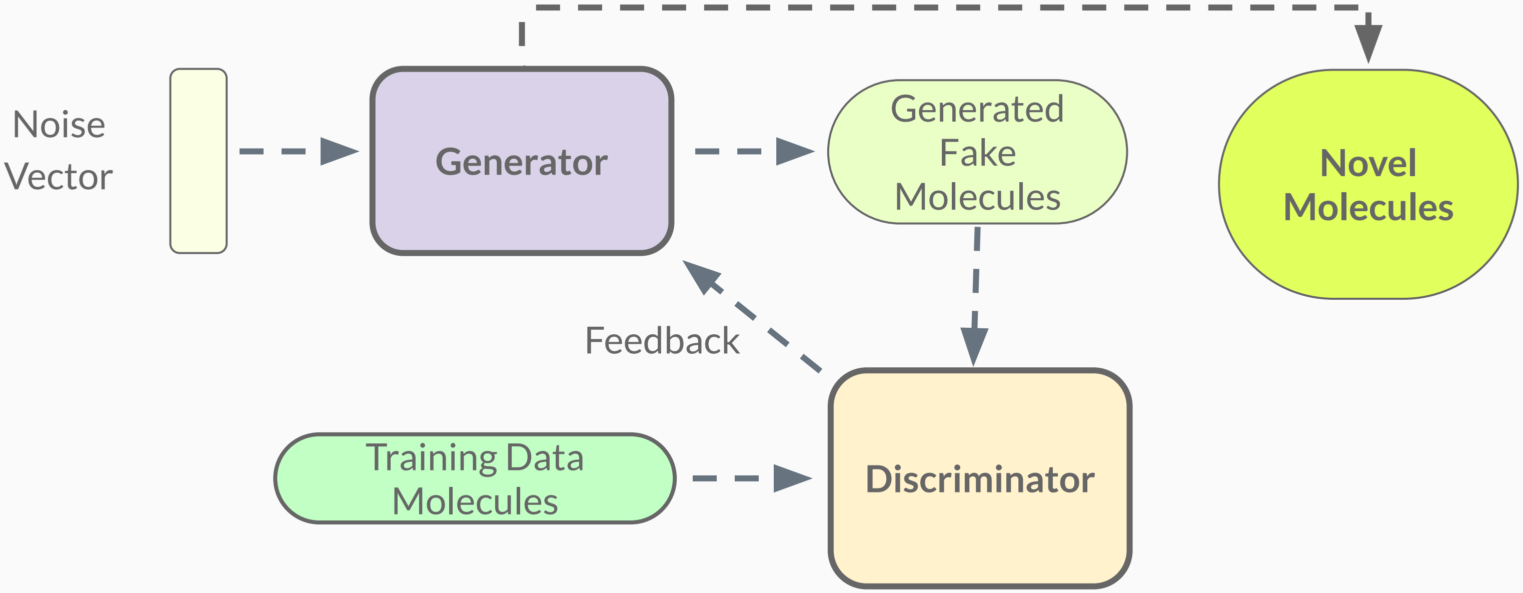


Figure 4. Optimization results for the jointly trained autoencoder using 5 × QED − SAS as the objective function. (a) shows a violin plot which compares the distribution of sampled molecules from normal random sampling, SMILES optimization via a common chemical transformation with a genetic algorithm, and from optimization on the trained Gaussian process model with varying amounts of training points.... This graph shows the combined results of four sets of trials. (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). (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.

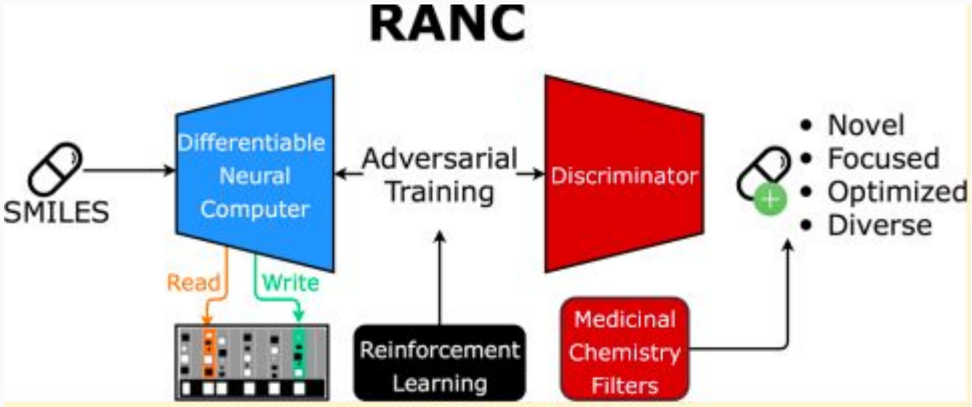
Structure of a GAN



GANs for Chemical Discovery

ORGAN (Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models): A Wasserstein GAN for Chemical Compound Generation

RANC (Reinforced Adversarial Neural Computer for de Novo Molecular Design): A DNC based GAN for Chemical Compound Generation



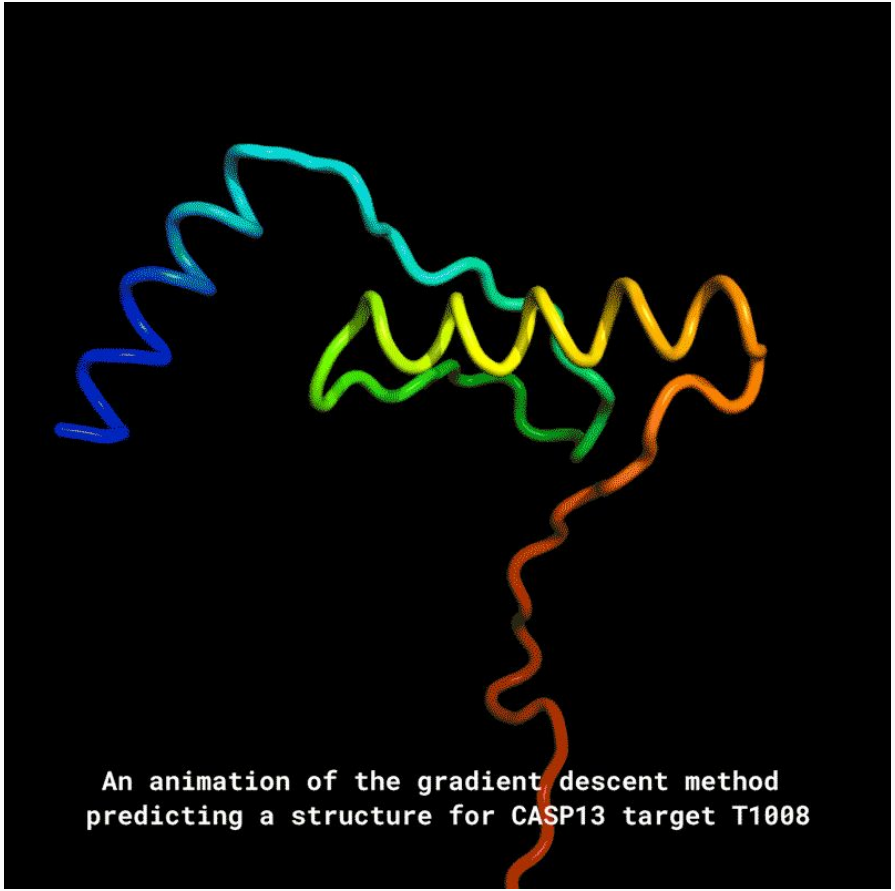
RL Agent: A RNN trained to predict the next character in the sequence (in this case SMILES strings).

Reward function: QED and SAS of the molecules output by the RNN.

Actions: The addition of characters to a SMILES string

Policy: The learned probability distribution of the RNN.

AlphaFold



An animation of the gradient descent method predicting a structure for CASP13 target T1008

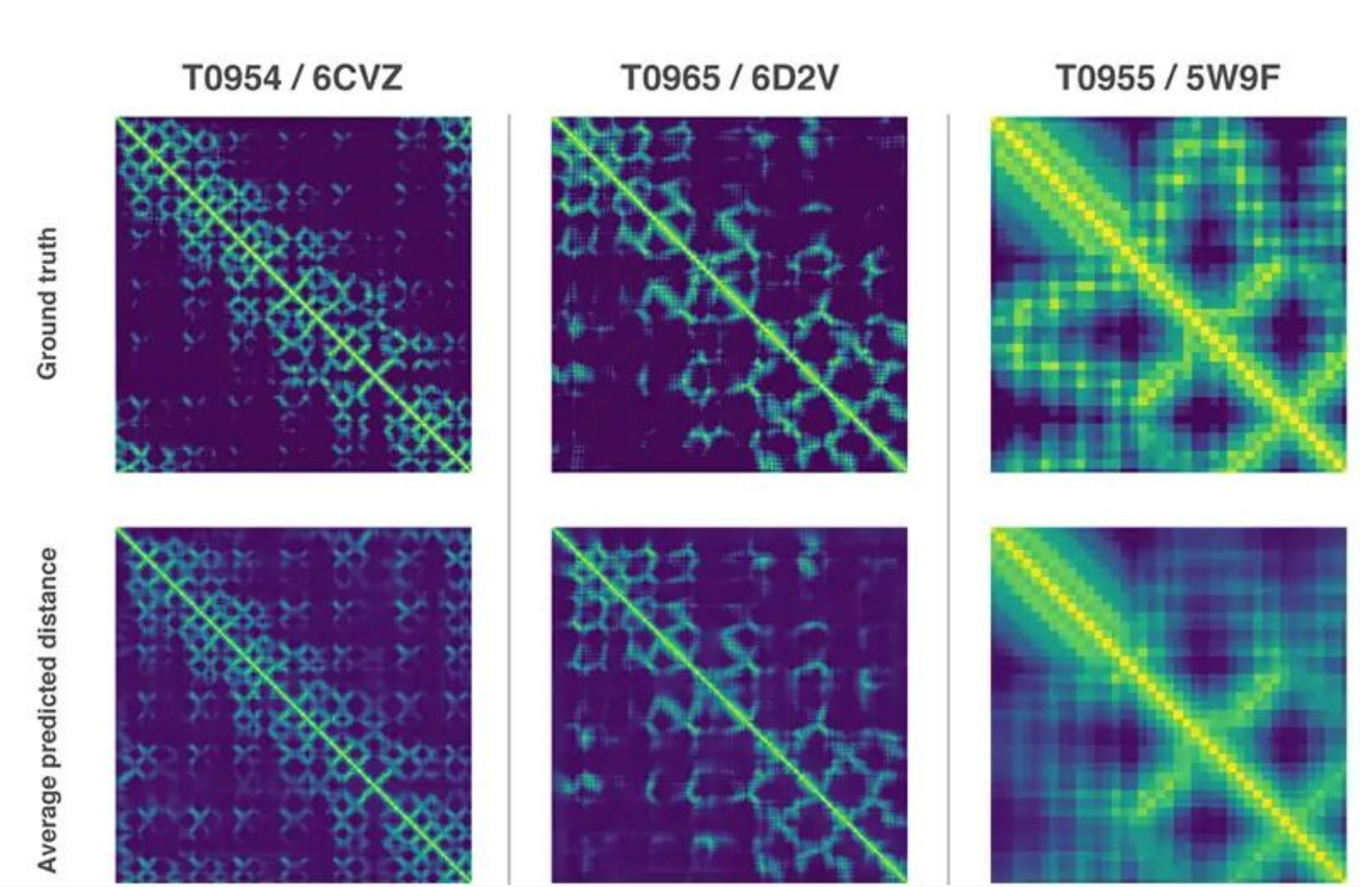
Protein folding: One of the most important problems in biology. Proteins have use because of their shape; takes a lot of lab work to determine a protein's shape. The basis of pharmaceuticals is optimizing drug-protein interactions.

Protein folding disorders include Alzheimer's, Parkinson's, cystic fibrosis, Huntington's.

Protein engineering also has numerous applications.

13 Nobel prizes involving protein structure research have been awarded between 1946-2009

AlphaFold



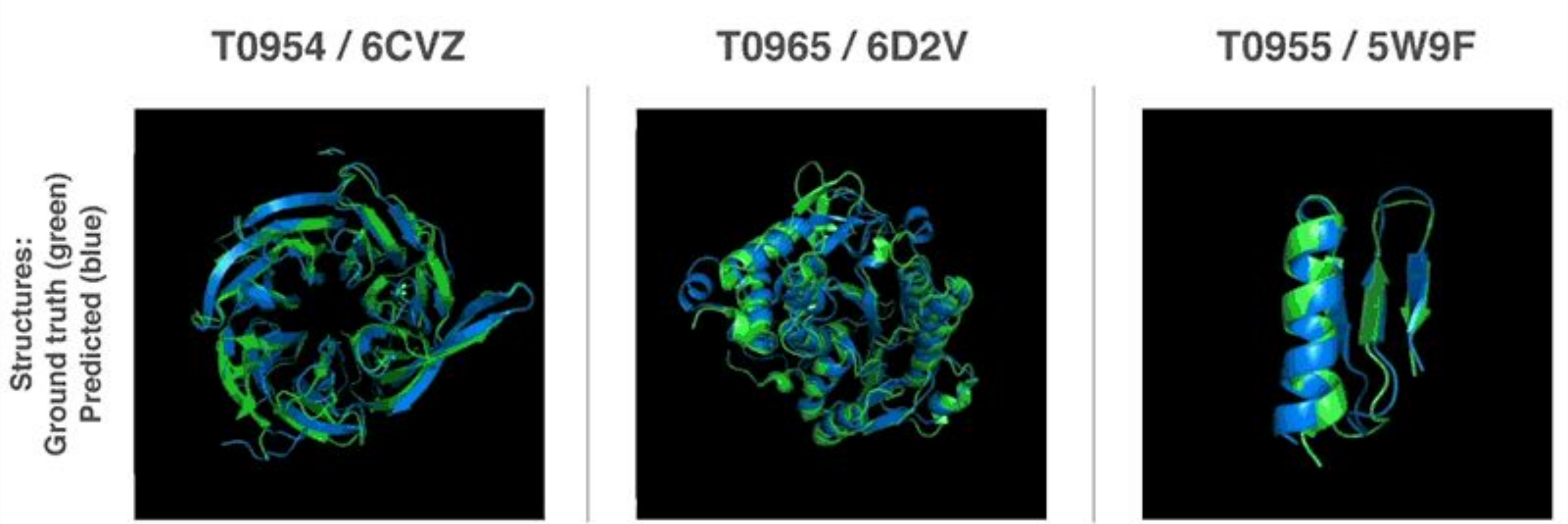
The properties our networks predict are: (a) the distances between pairs of amino acids and (b) the angles between chemical bonds that connect those amino acids. The first development is an advance on commonly used techniques that estimate whether pairs of amino acids are near each other.

AlphaFold

1st place out of 97 competitors in the 13th Critical Assessment of Structure Prediction (CASP)

AlphaFold made the most accurate prediction 25 times out of 45 tests.

The second place algorithm made the most accurate prediction on 3 of the 43 tests.



T0954 / 6CVZ

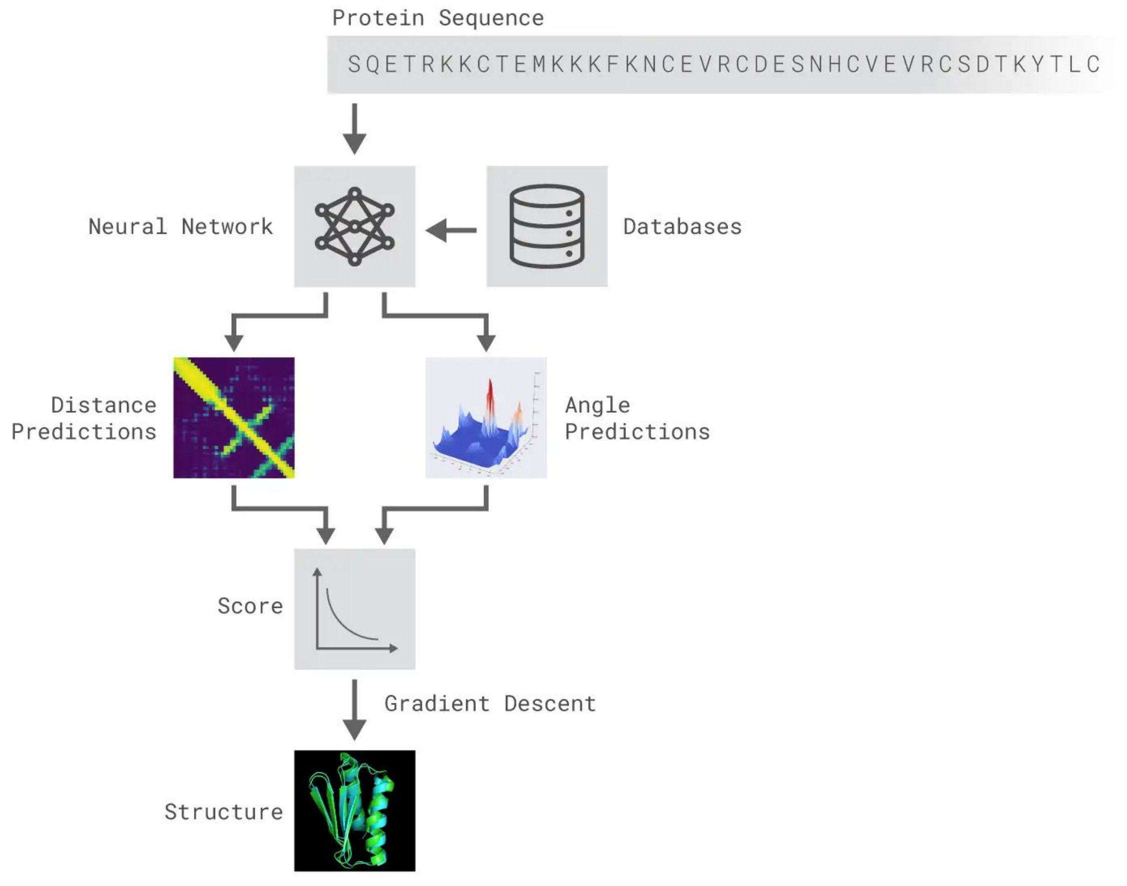
T0965 / 6D2V

T0955 / 5W9F

AlphaFold

“We trained a neural network to predict a separate distribution of distances between every pair of residues in a protein. These probabilities were then combined into a score that estimates how accurate a proposed protein structure is. We also trained a separate neural network that uses all distances in aggregate to estimate how close the proposed structure is to the right answer.

The second method optimised scores through gradient descent … which resulted in highly accurate structures. This technique was applied to entire protein chains rather than to pieces that must be folded separately before being assembled, reducing the complexity of the prediction process.”



Time Permitting: Other Interesting Applications

- Deep Neural Networks for approximating the Schrodinger equation
- CycleGANs for molecular optimization
- Applying Neural Networks recursively to predict a reaction sequence

Sources

1. Almi, Imane & Belaidi, Salah & Nadjib, Melkemi & Bouzidi, Djemoui. (2018). Chemical Reactivity, Drug-Likeness and Structure Activity/Property Relationship Studies of 2,1,3-Benzoxadiazole Derivatives as Anti-Cancer Activity. Journal of Bionanoscience. 12. 10.1166/jbns.2018.1503.
2. Gaurav Sahni. “Structural Biology Related Nobel Prizes.” *Ebi.Ac.Uk*, 2009, www.ebi.ac.uk/pdbe/docs/nobel/nobels.html. Accessed 25 Nov. 2019.
3. Gómez-Bombarelli, Rafael, et al. “Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules.” *ACS Central Science*, vol. 4, no. 2, 12 Jan. 2018, pp. 268–276, 10.1021/acscentsci.7b00572. Accessed 10 Nov. 2019.
4. Google’s DeepMind aces protein folding. “Google’s DeepMind Aces Protein Folding.” *Science* | *AAAS*, 6 Dec. 2018, www.sciencemag.org/news/2018/12/google-s-deepmind-aces-protein-folding. Accessed 14 Aug. 2019.
5. Irhum Shafkat. “Intuitively Understanding Variational Autoencoders.” *Medium*, Towards Data Science, 4 Feb. 2018, towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf. Accessed 25 Nov. 2019.
6. Ritter, Steven K. “The Haber-Bosch Reaction: An Early Chemical Impact On Sustainability | August 18, 2008 Issue - Vol. 86 Issue 33 | Chemical & Engineering News.” *Acs.Org*, 2019, cen.acs.org/articles/86/i33/Haber-Bosch-Reaction-Early-Chemical.html. Accessed 25 Nov. 2019.
7. Senior, Andrew, et al. “AlphaFold: Using AI for Scientific Discovery.” *Deepmind*, 2018, deepmind.com/blog/article/alphafold.