CSC412:
Variational Autoencoders

David Duvenaud

Class Projects

- Focus is on research skills, providing context for results, evidence for claims
- Excuse to start larger project

Variational Inference

- Directly optimize the parameters phi of an approximate distribution q(z|x, phi) to match p(z|x, theta)

- What if there is a local latent variable per-datapoint, and some global parameters? e.g. Bayesian PCA, generative image models, topic models

- Directly optimize the parameters phi_i of each approximate distribution q(z_i|x_i, phi_i) to match p(z_i|x_i, theta)

SVI Algorithm:

- 1. Sample z from q(z|x, phi)
- 2. Return log p(z, x | theta) - log q(z | x, phi)

- In this setting, only one set of latent params z, as in a Bayesian neural net

SVI w/ per-datapoint latents:

1. Sample x_i from dataset
2. Optimize phi_i to minimize KL(q(z_i | x_i, phi_i)|| p(z_i | x_i, theta))
3. Sample z from q(z_i|x_i, phi_i)
4. Return log p(z_i, x_i | theta) - log q(z_i | x_i, phi_i)

Variational Autoencoder:

1. Sample x_i from dataset

2. Compute phi_i as a function of x, and recognition params phi_r: phi_i = f(x, phi_r)

3. Sample z from q(z_i|x_i, phi_i)

4. Return log p(z_i, x_i | theta) - log q(z_i | x_i, phi_i)

Consequences of using a recognition network

- Don’t need to re-optimize phi_i each time theta changes - much faster
- Recognition net won’t necessary give optimal phi_i
- Can have fast test-time inference (vision)
- Can train recognition net with gradient descent
  - Could also differentiate through optimization

Simple but not obvious

- It took a long time get here!
  - Independently developed as denoising autoencoders (Bengio et al.) and amortized inference (many others)
  - Helmholtz machine - same idea in 1995 but used discrete latent variables

The Helmholtz Machine

Peter Dayan
Geoffrey E. Hinton
Radford M. Neal
Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 1A4, Canada

Richard S. Zemel
CNL, The Salk Institute, PO Box 85800, San Diego, CA 92186-5800 USA

Discovering the structure inherent in a set of patterns is a fundamental aim of statistical inference or learning. One fruitful approach is to build a parameterized stochastic generative model, independent draws from which are likely to produce the patterns. For all but the simplest generative models, each pattern can be generated in exponentially many ways. It is thus intractable to adjust the parameters to maximize

Autoencoder Motivation

- Want compact representation of data
- x = dec(enc(x))
- Need to prevent enc = dec = identity
- So, add noise to encoding
- Gives VAE bound but with a free parameter

Benefits of compact latent code

- http://www.dpkingma.com/sgvb_mnist_demo/demo.html

- Nearby z’s give similar x

- Recent work on ‘disentangling’ latent rep

Code examples

- Show VAE code

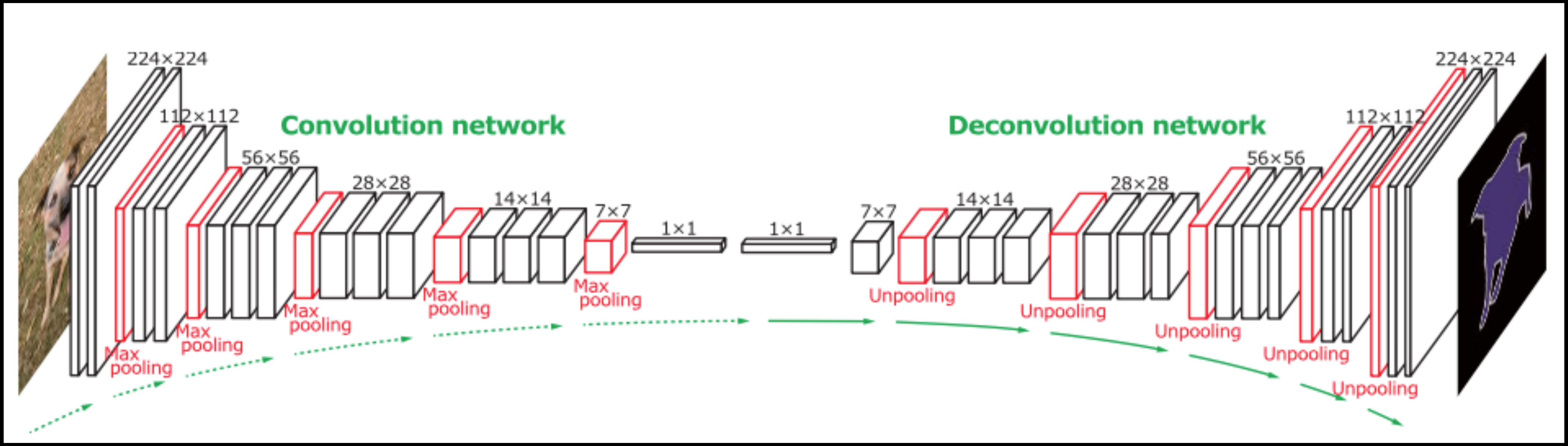
Variations: Decoder

- Orginally, $p(x|z) = N(x \mid dec(z, theta), sigma\ I)$
- Final step has independence assumption, causes blurry samples
- $p(x|z)$ can be anything: rnn, pixelRNN, real NVP, de_convolutional net

Figure 6: Samples from hierarchical PixelVAE on the 64x64 ImageNet dataset.

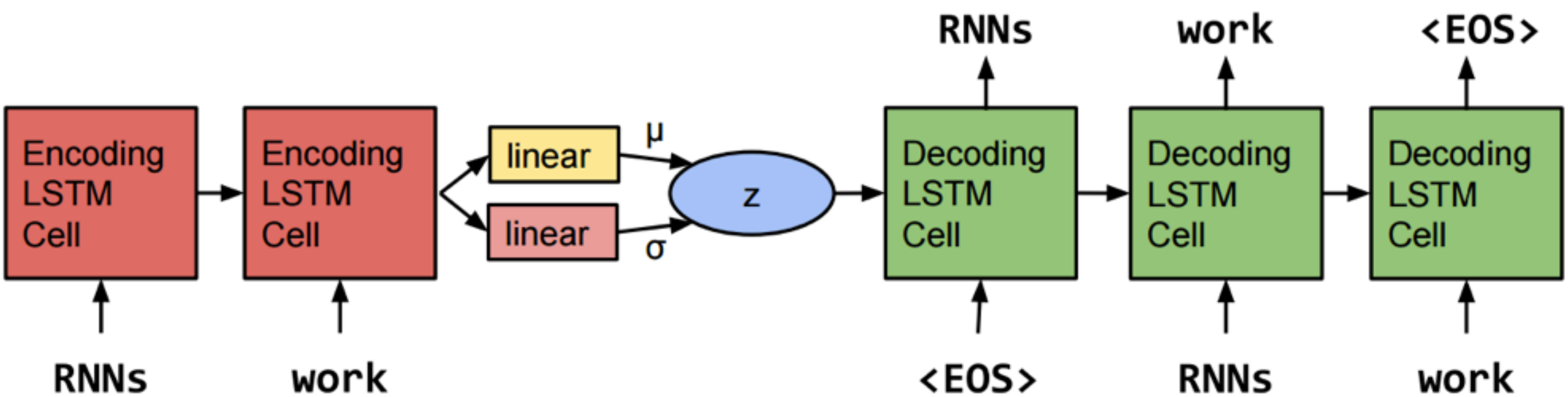
Variations

- Decoder often looks like inverse of encoder
- Encoders can come from supervised learning



Learning Deconvolution Network for Semantic Segmentation
http://arxiv.org/abs/1505.04366.

Text autoencoders



- *Generating Sentences from a Continuous Space.* Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, Samy Bengio

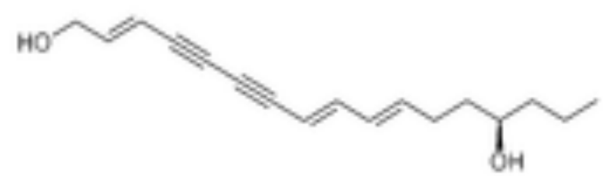
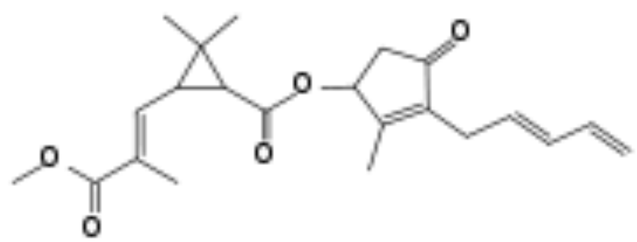
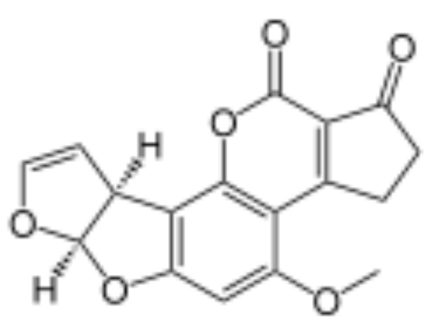
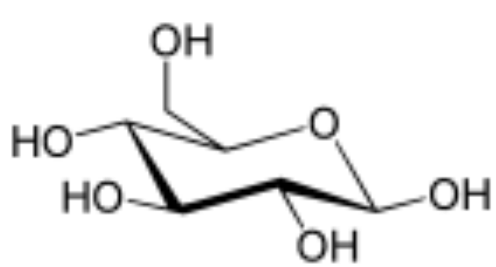
Text VAE - **Interpolation**

“ i want to talk to you . ”
“i want to be with you . ”
“i do n’t want to be with you . ”
i do n’t want to be with you .
she did n’t want to be with him .

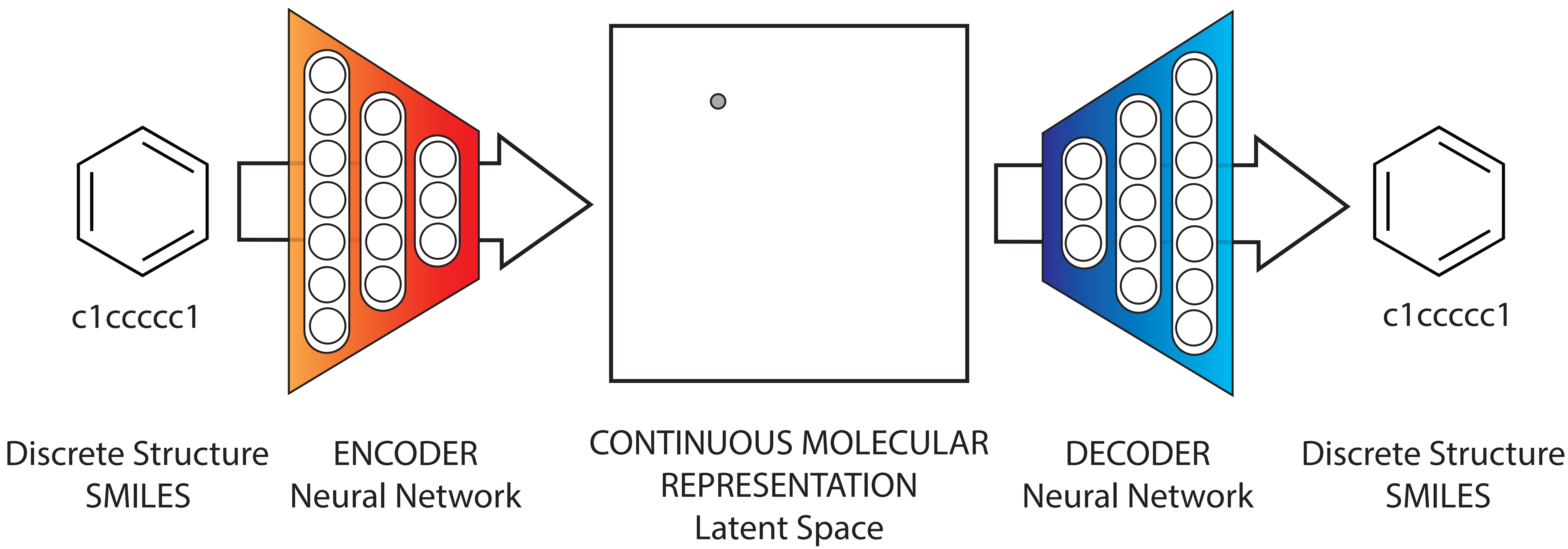
it made me want to cry .
no one had seen him since .
it made me feel uneasy .
no one had seen him .
the thought made me smile .
the pain was unbearable .
the crowd was silent .
the man called out .
the old man said .
the man asked .

he was silent for a long moment .
he was silent for a moment .
it was quiet for a moment .
it was dark and cold .
there was a pause .
it was my turn .

What is a molecule?

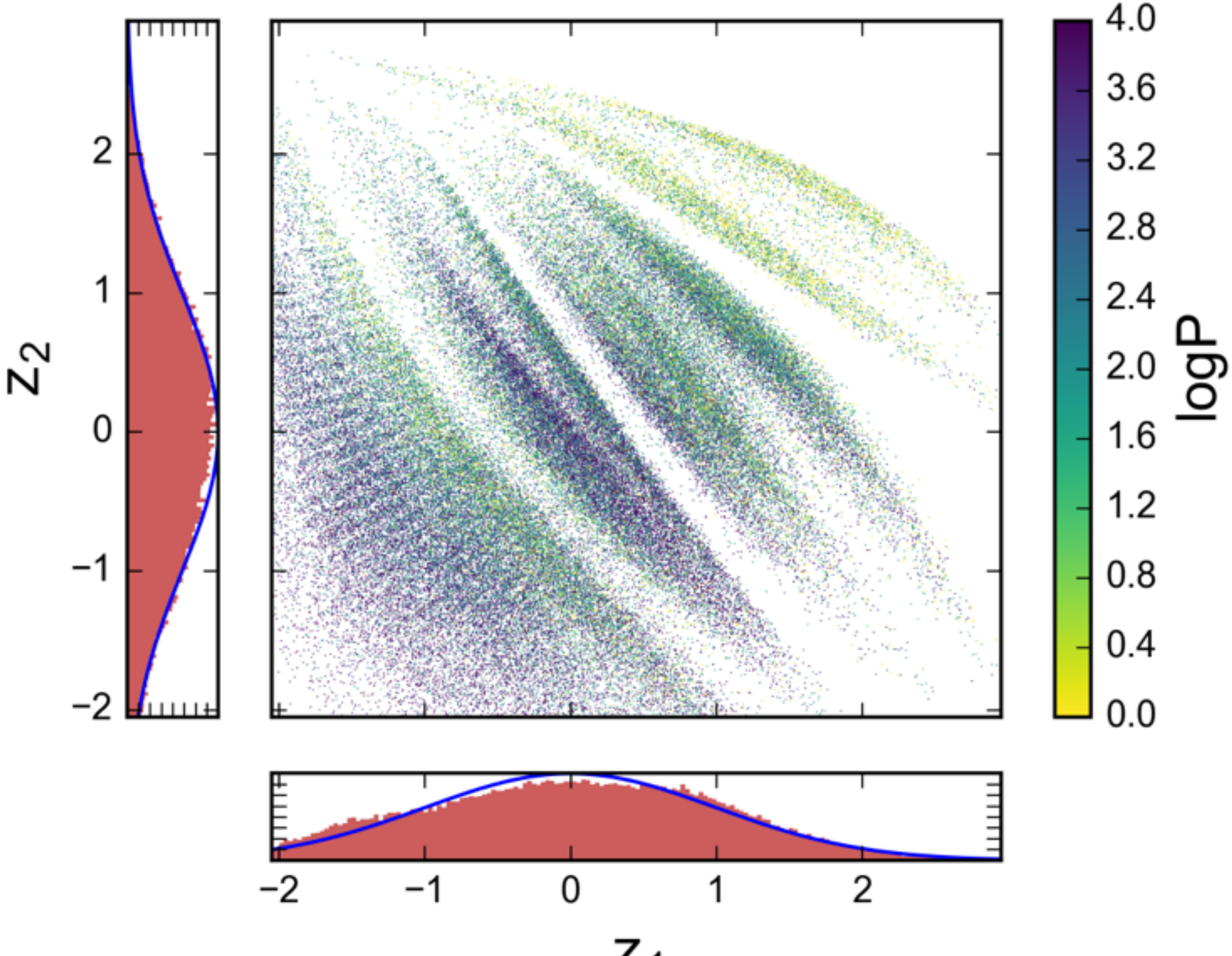
Graph	SMILES string
	CCC[C@@H](O)CC\C=C\C=C\C#CC#C\C=C\CO
	COC(=O)C(\C)=C\C1C(C)(C)[C@H]1C(=O)O[C@@H]2C(C)=C(C(=O)C2)CC=CC=C
	O1C=C[C@H]([C@H]1O2)c3c2cc(OC)c4c3OC(=O)C5=C4CCC(=O)5
	OC[C@@H](O1)[C@@H](O)[C@H](O)[C@@H](O)[C@@H](O)1

Repurposing text autoencoders

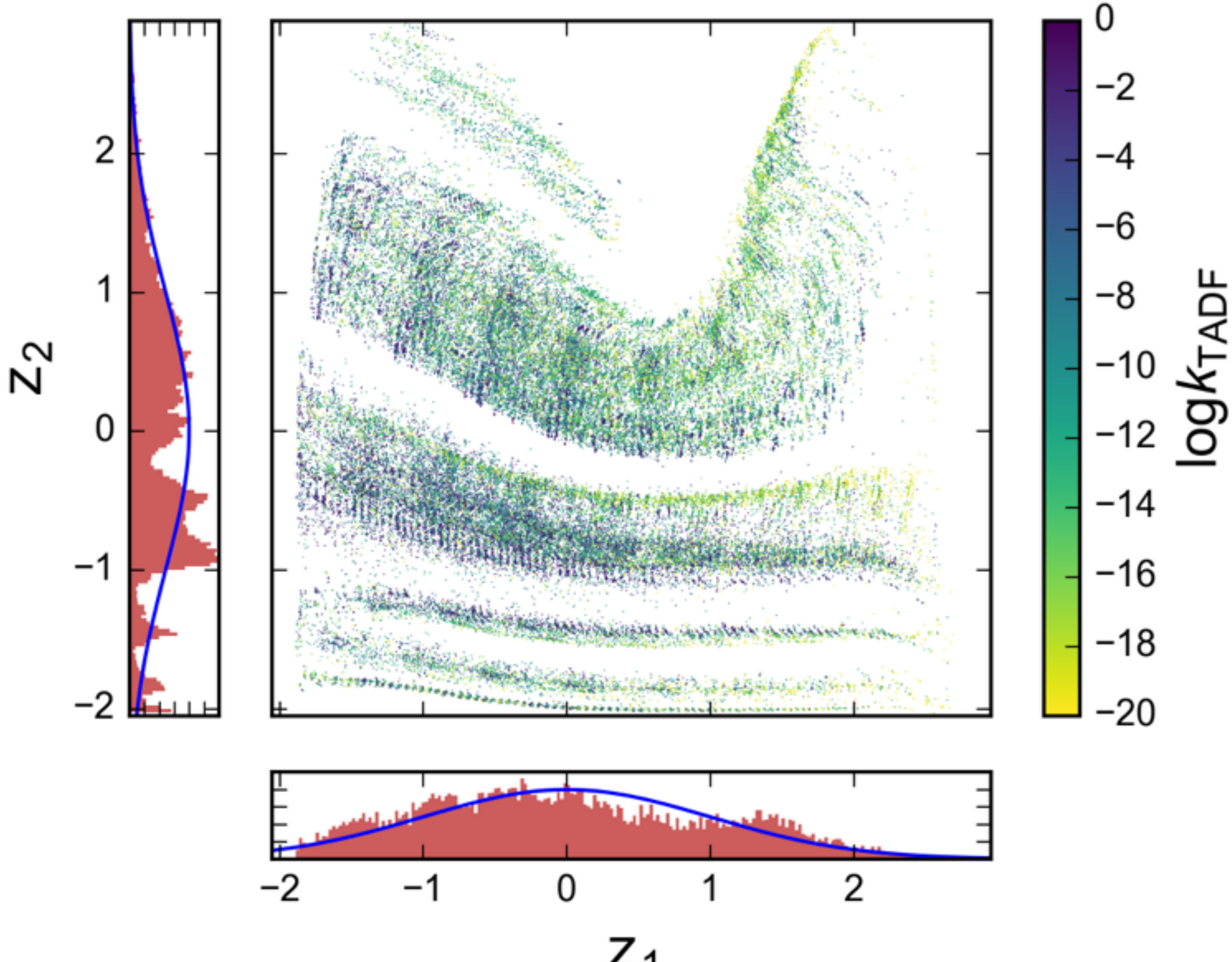


Can be trained on unlabeled data

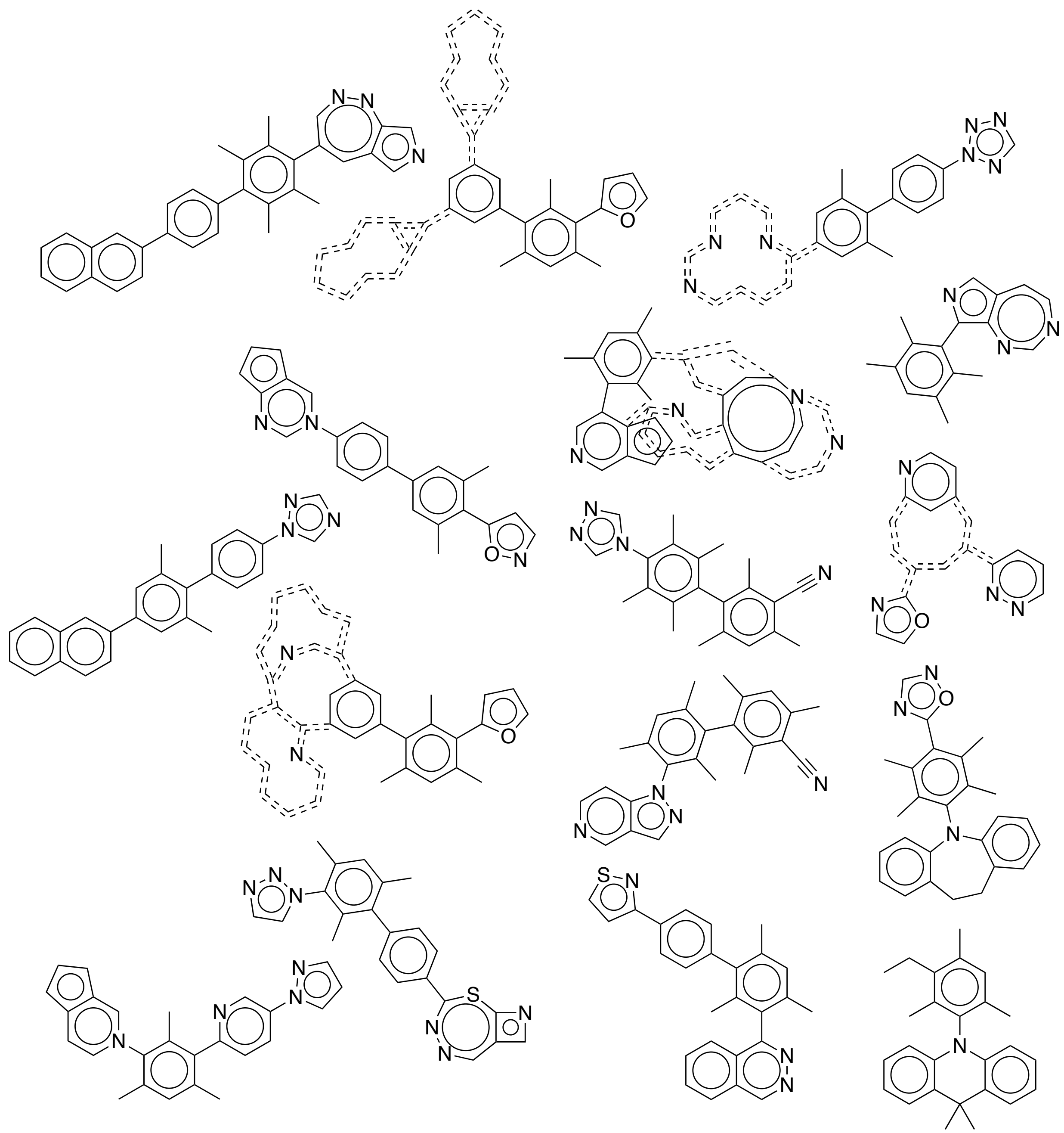
Map of 220,000 Drugs



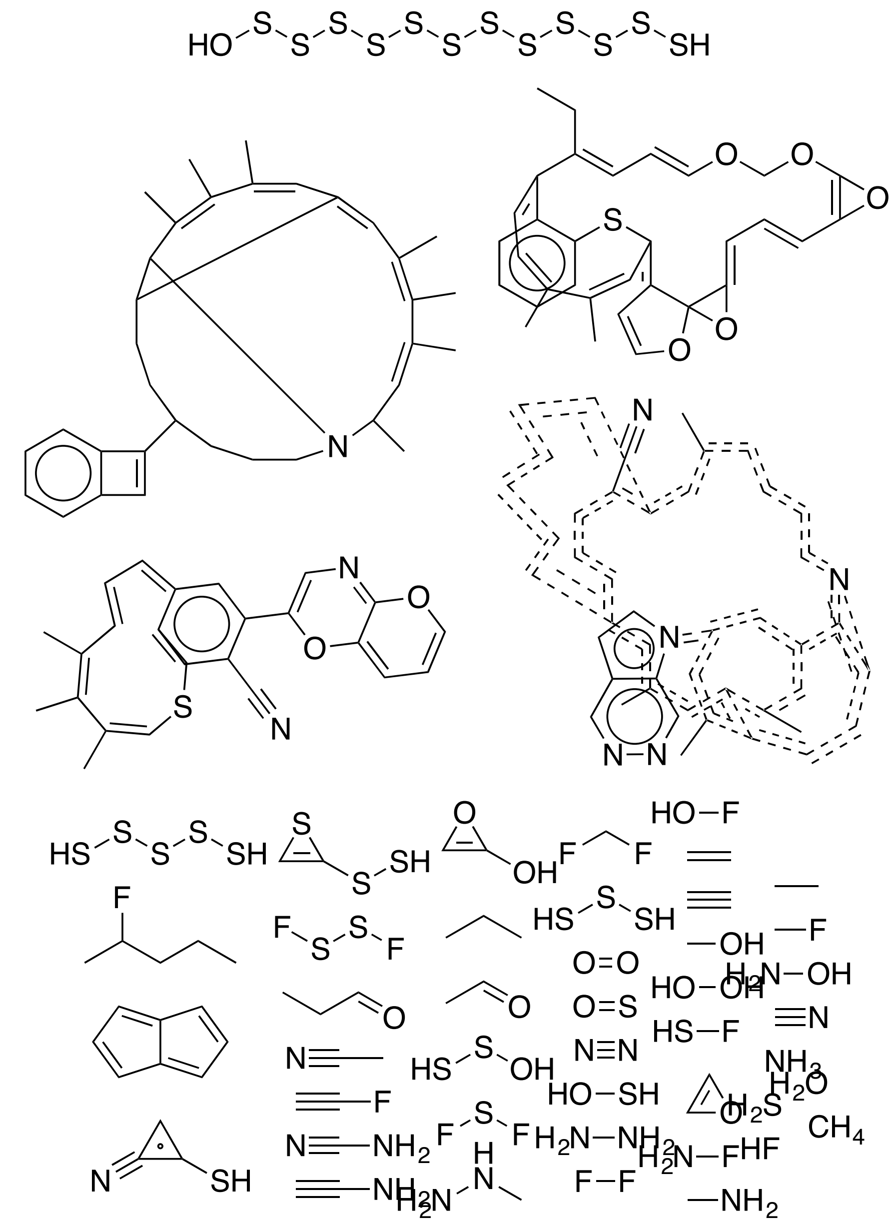
Map of 100,000 OLEDs



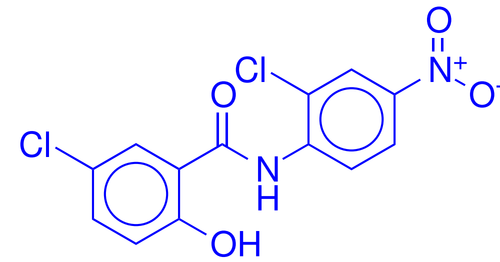
Random Organic LEDs

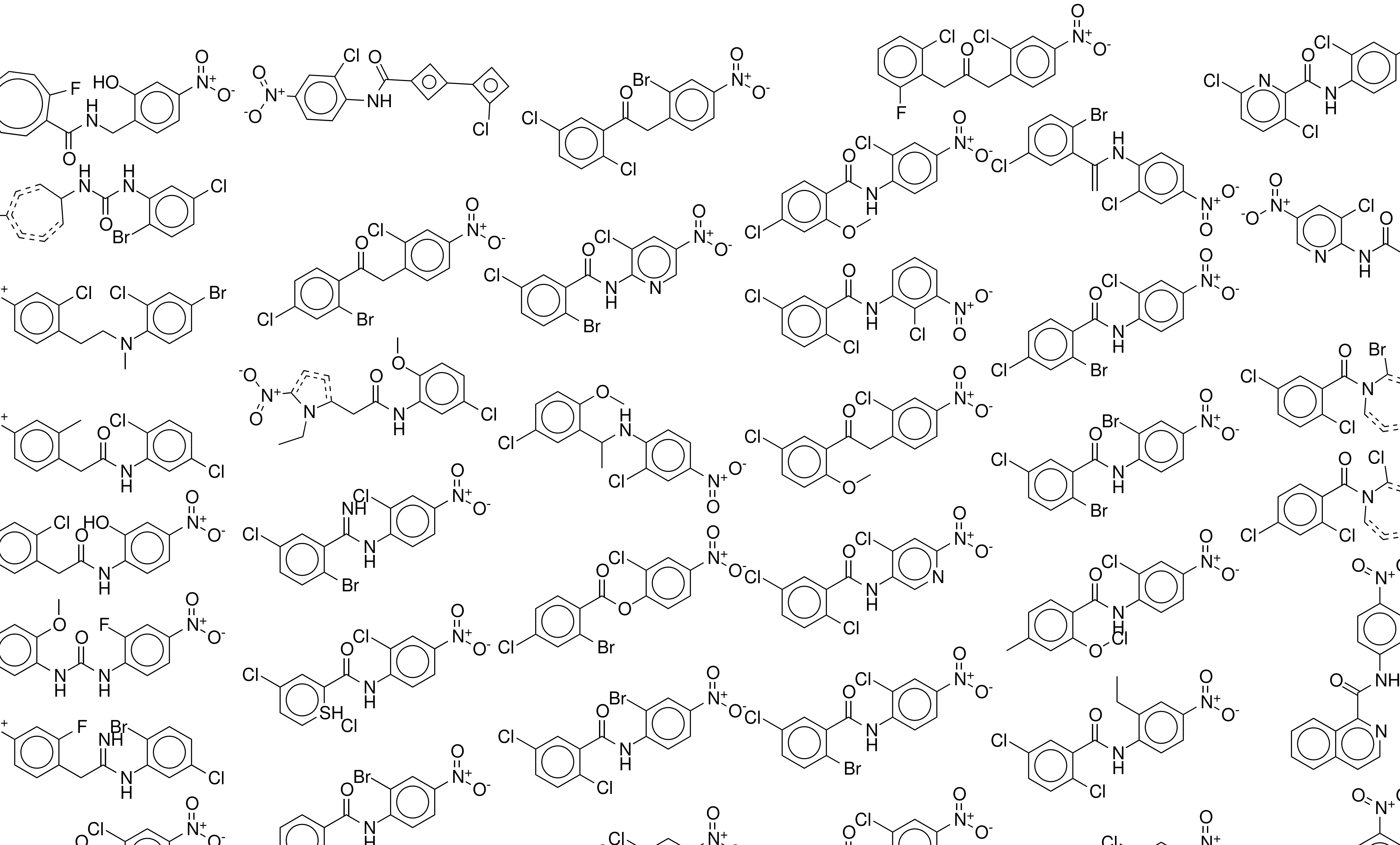


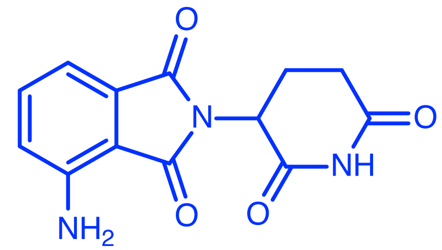
Variational autoencoder

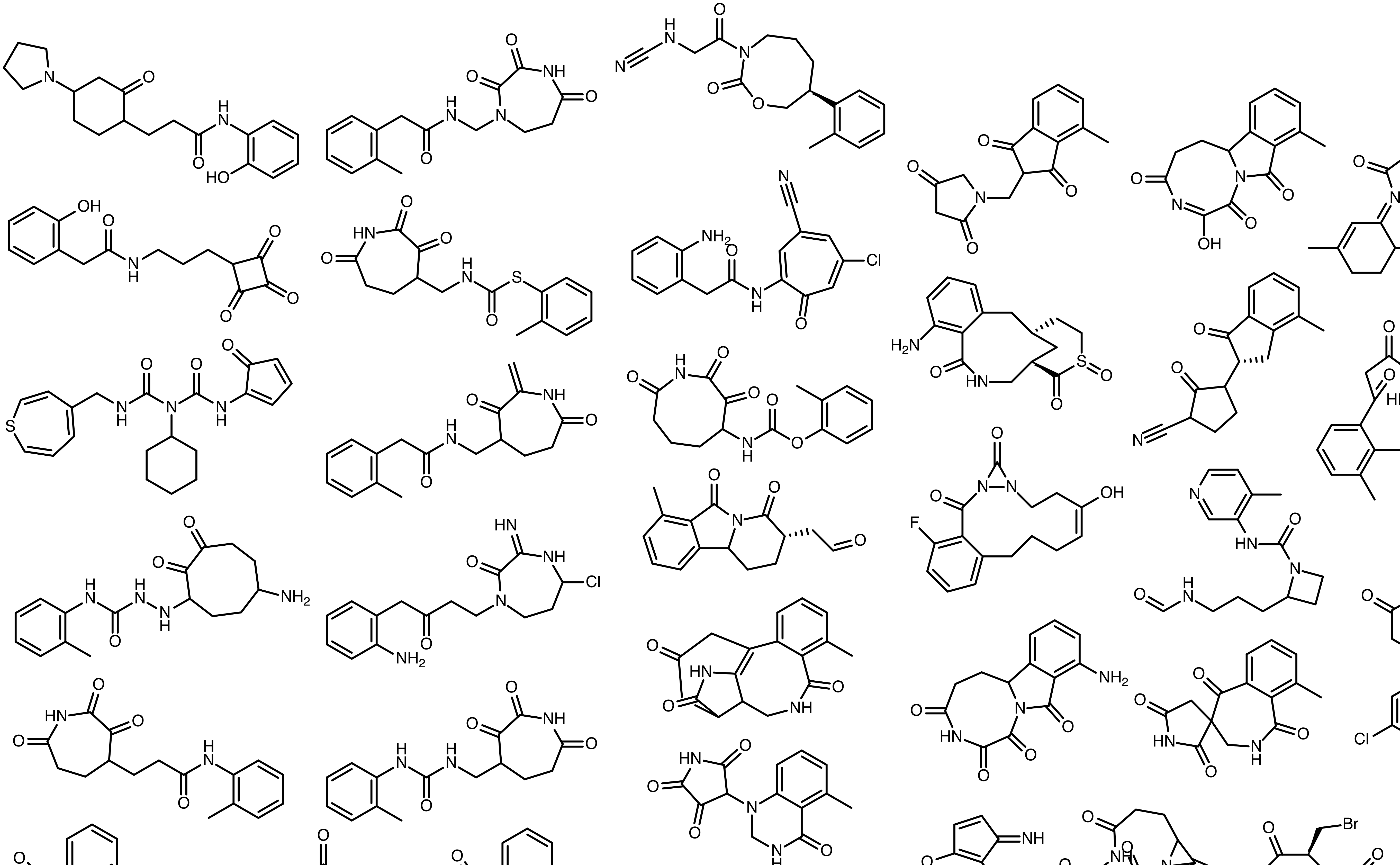


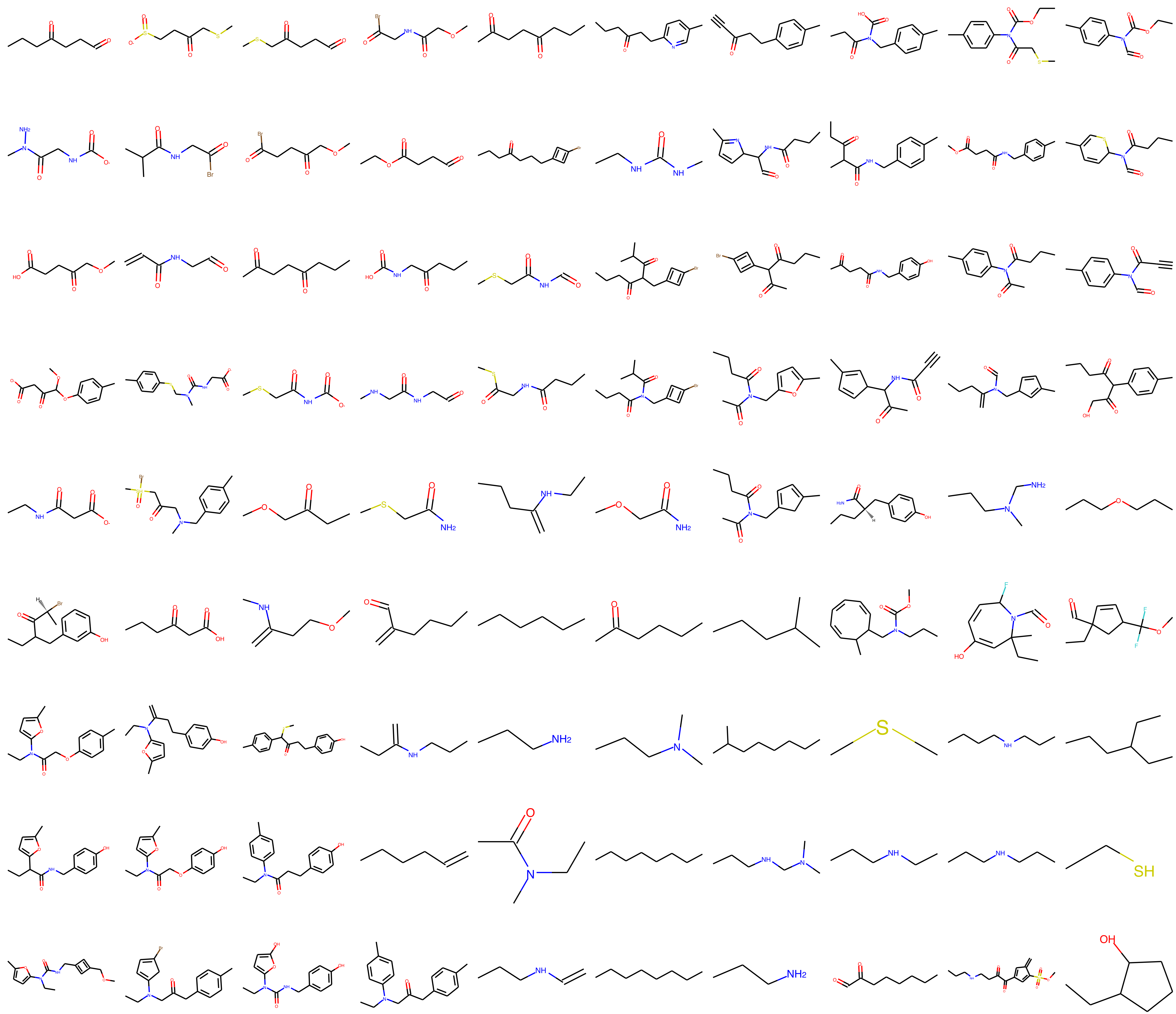
Standard autoencoder

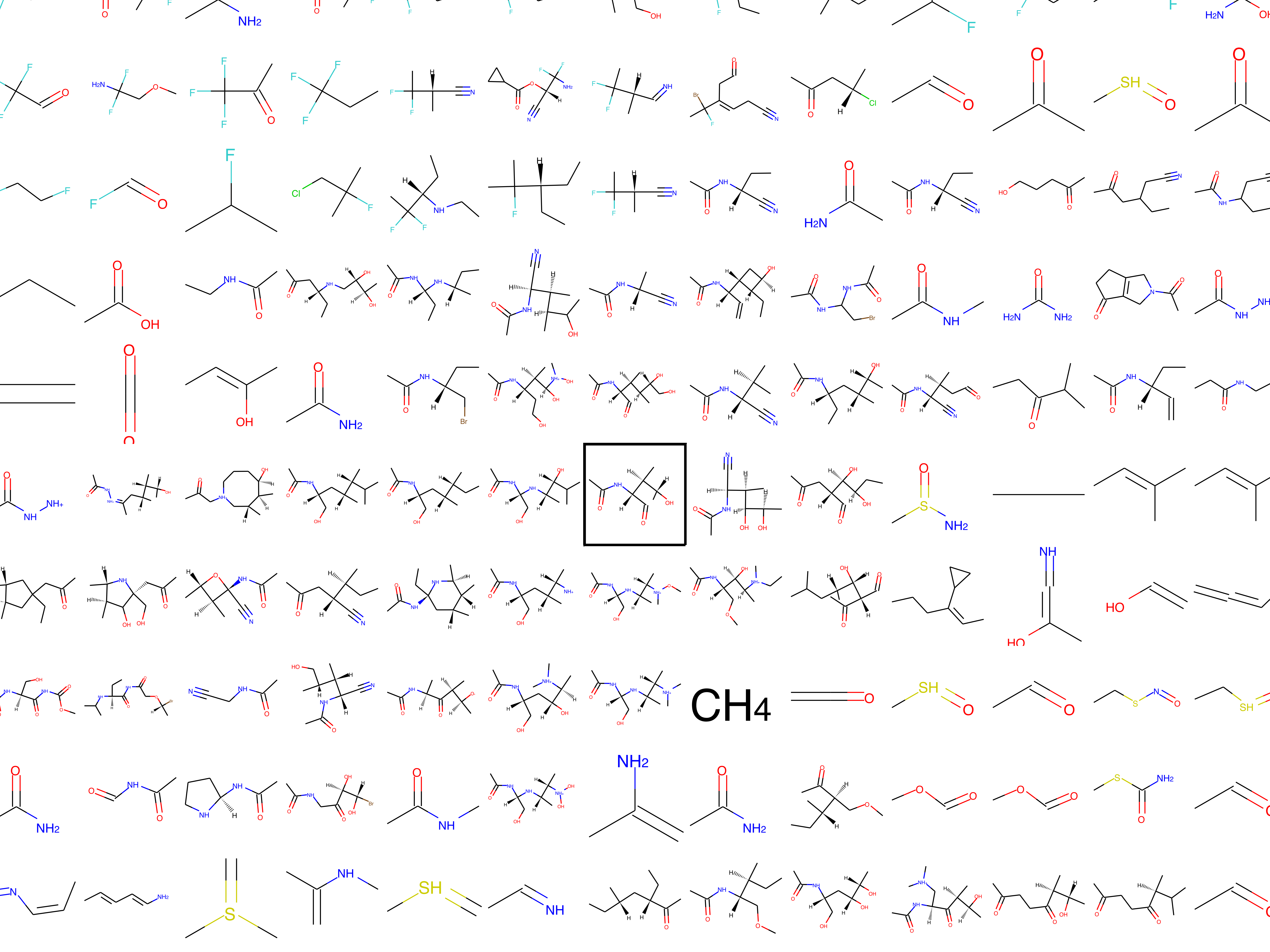
Molecules near 

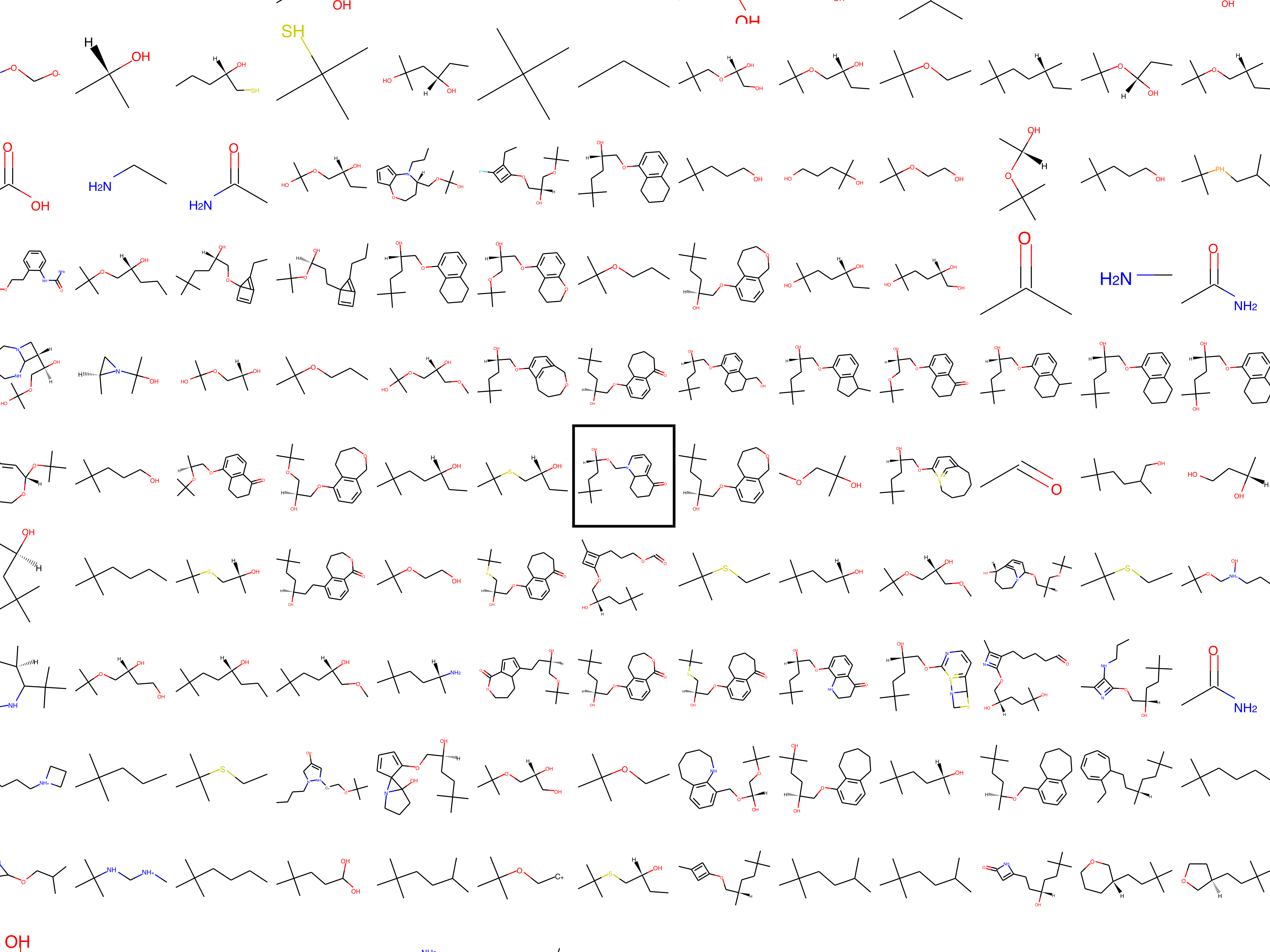


Molecules near 

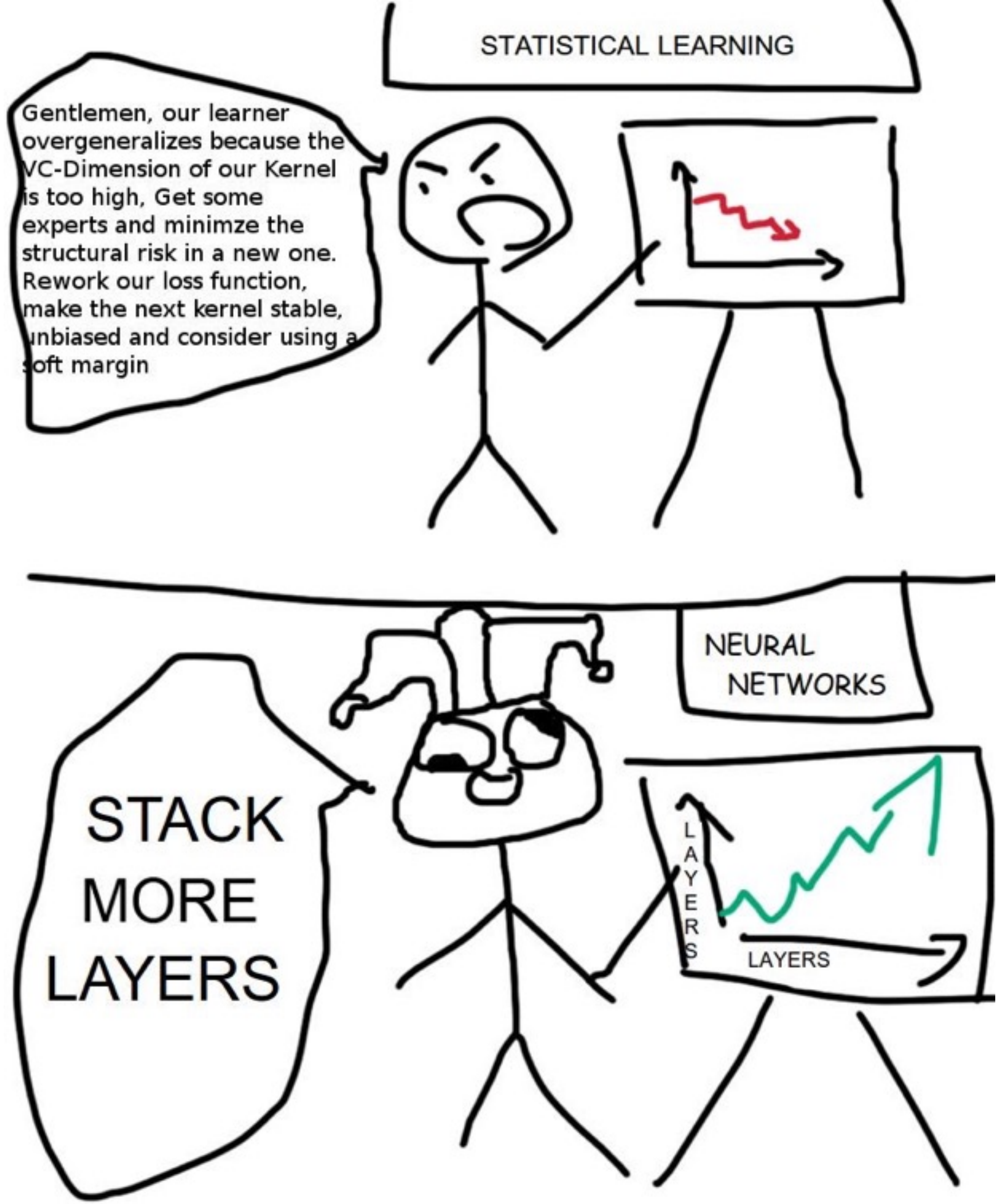




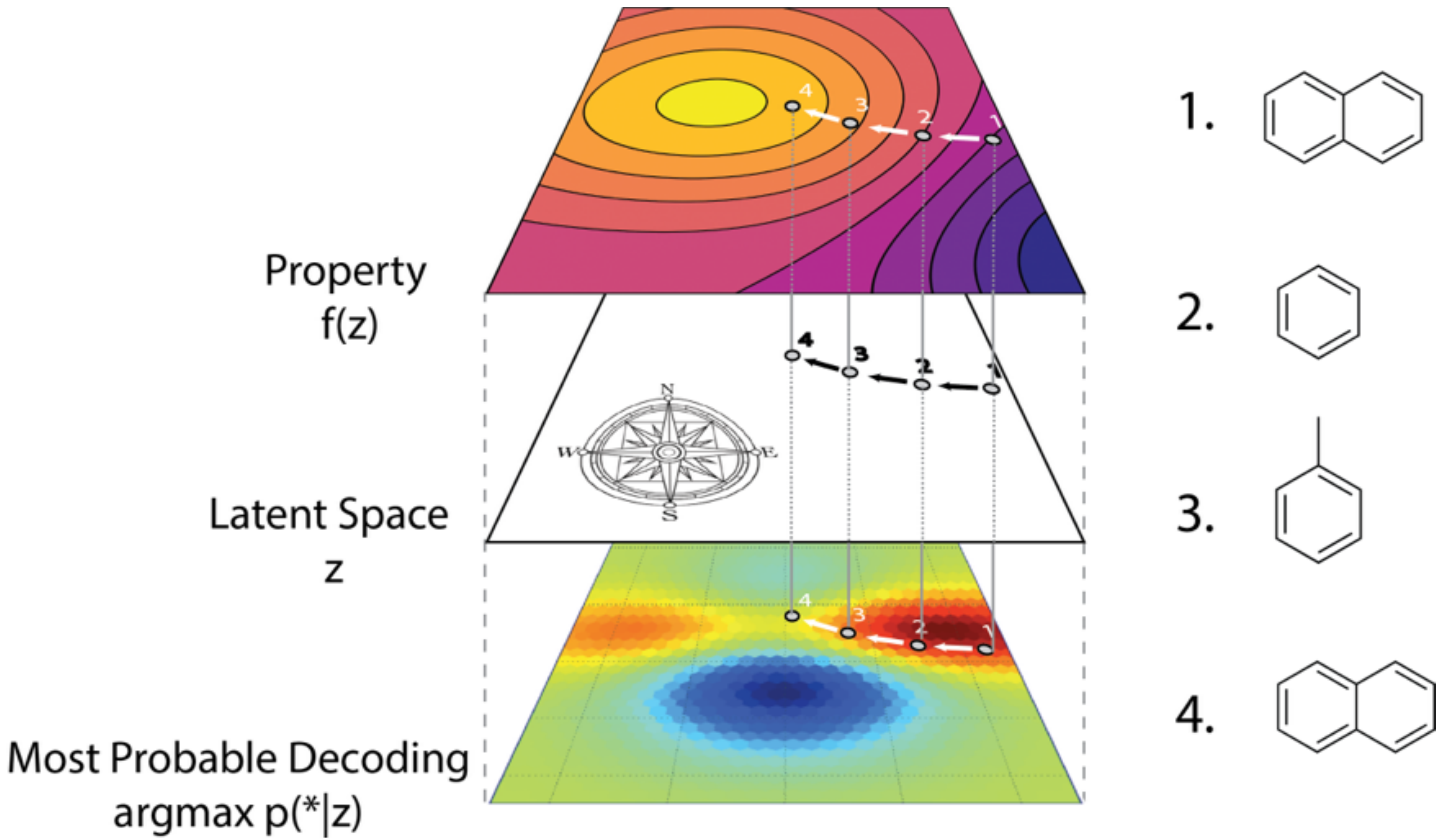




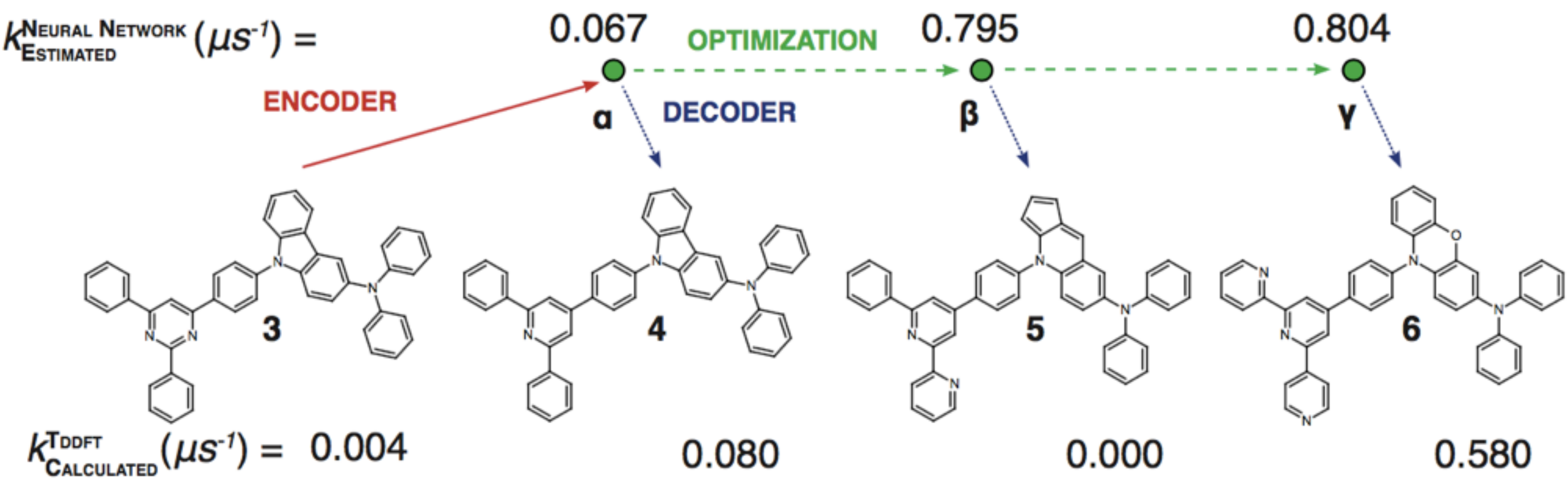
No chemistry-specific design!



Gradient-based optimization



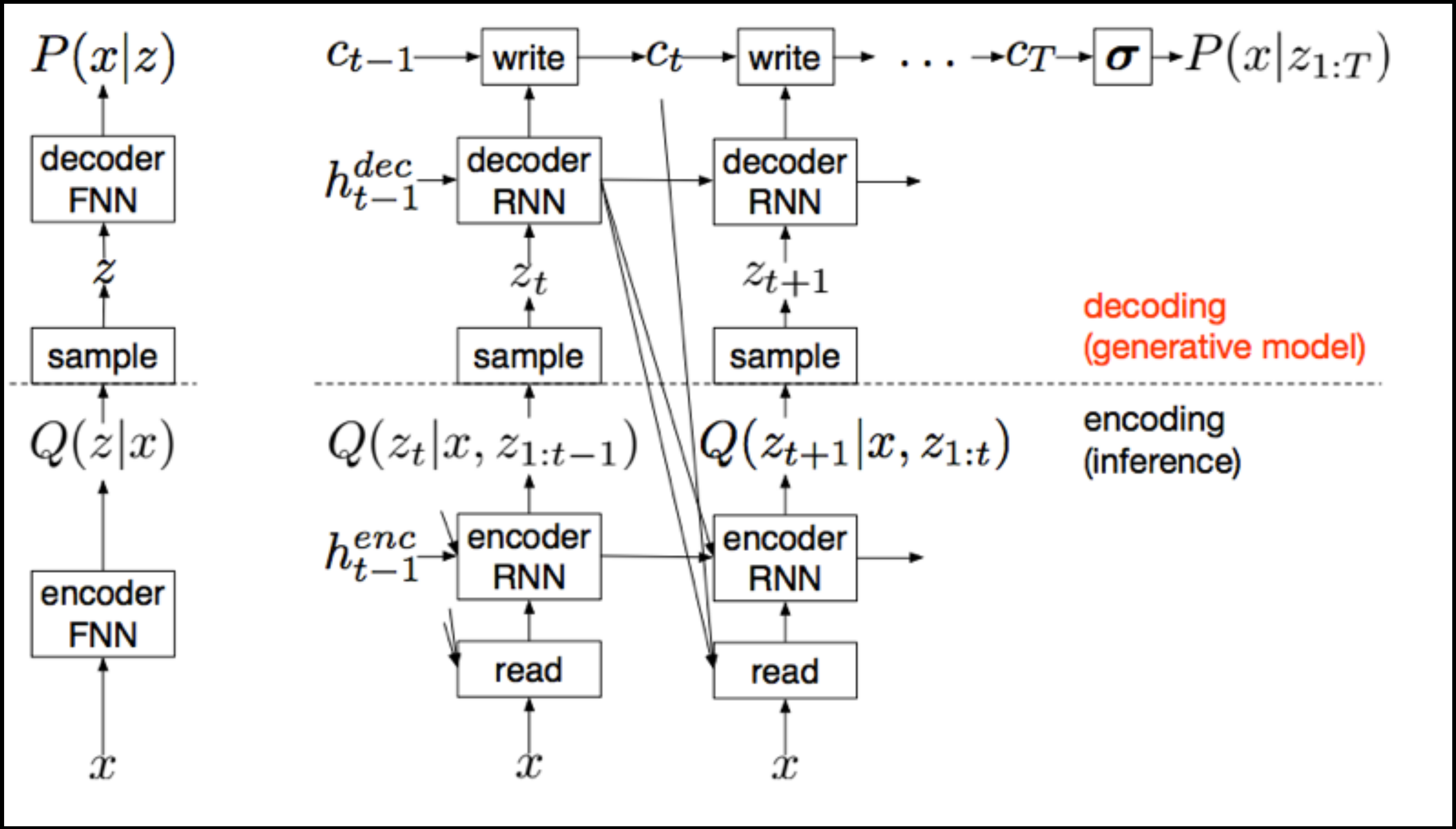
Gradient-based optimization



- Can't necessarily start from given molecule, need to encode/decode
- Can't go too far from start, wander into 'holes' or empty regions

Encoder can look at decoder

- https://www.youtube.com/watch?v=Zt-7MI9eKEo



Recent Extensions

- Importance-Weighted Autoencoders (IWAE) Burda, Grosse, Salakhudtinov
- Mixture distributions in posterior
- GAN-style ideas to avoid evaluating $q(z|x)$, $p(x|z)$, even $p(z)$
- Normalizing flows: Produce arbitrarily-complicated $q(z|x)$
- Incorporate HMC or local optimization to define $q(z|x)$

Generative Adversarial Networks

- Also a latent-variable model: $x = f(z)$, z from $N(0,I)$
- Trained adversarially, can also optimize $p(x)$
- Recent work on adversarially-trained VAEs:
- Match $p(z, x)$ to $q(z, x)$
  - Sample z from $p(z)$, x from $p(x|z)$
  - Sample x from data, z from $q(z|x)$