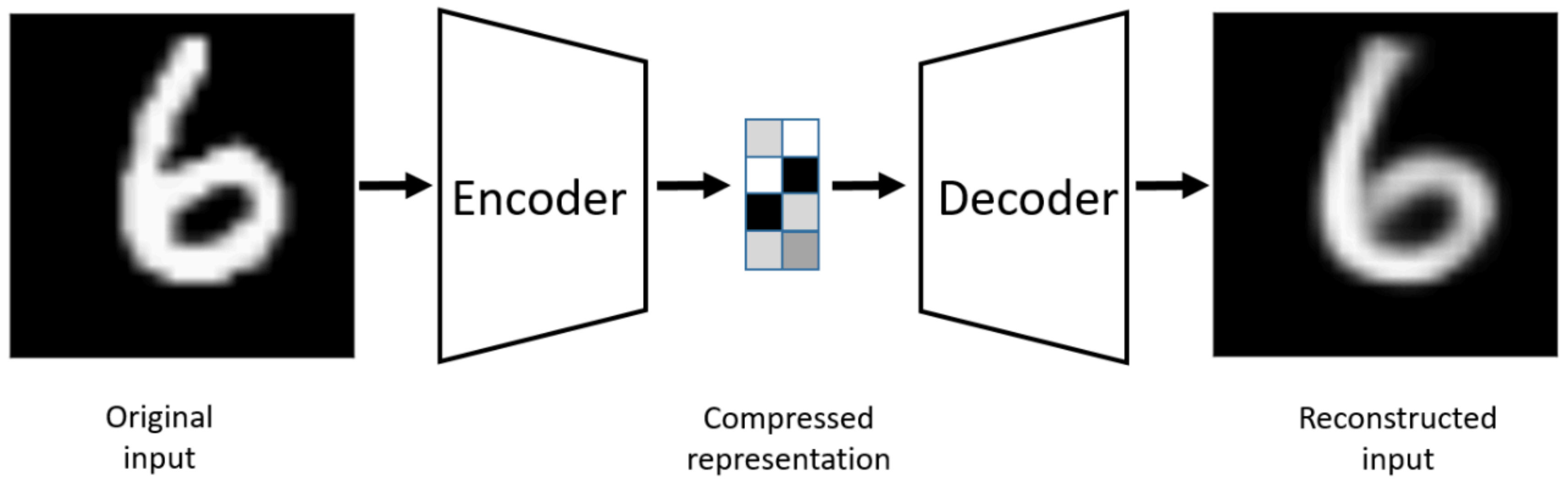
19-11–2024

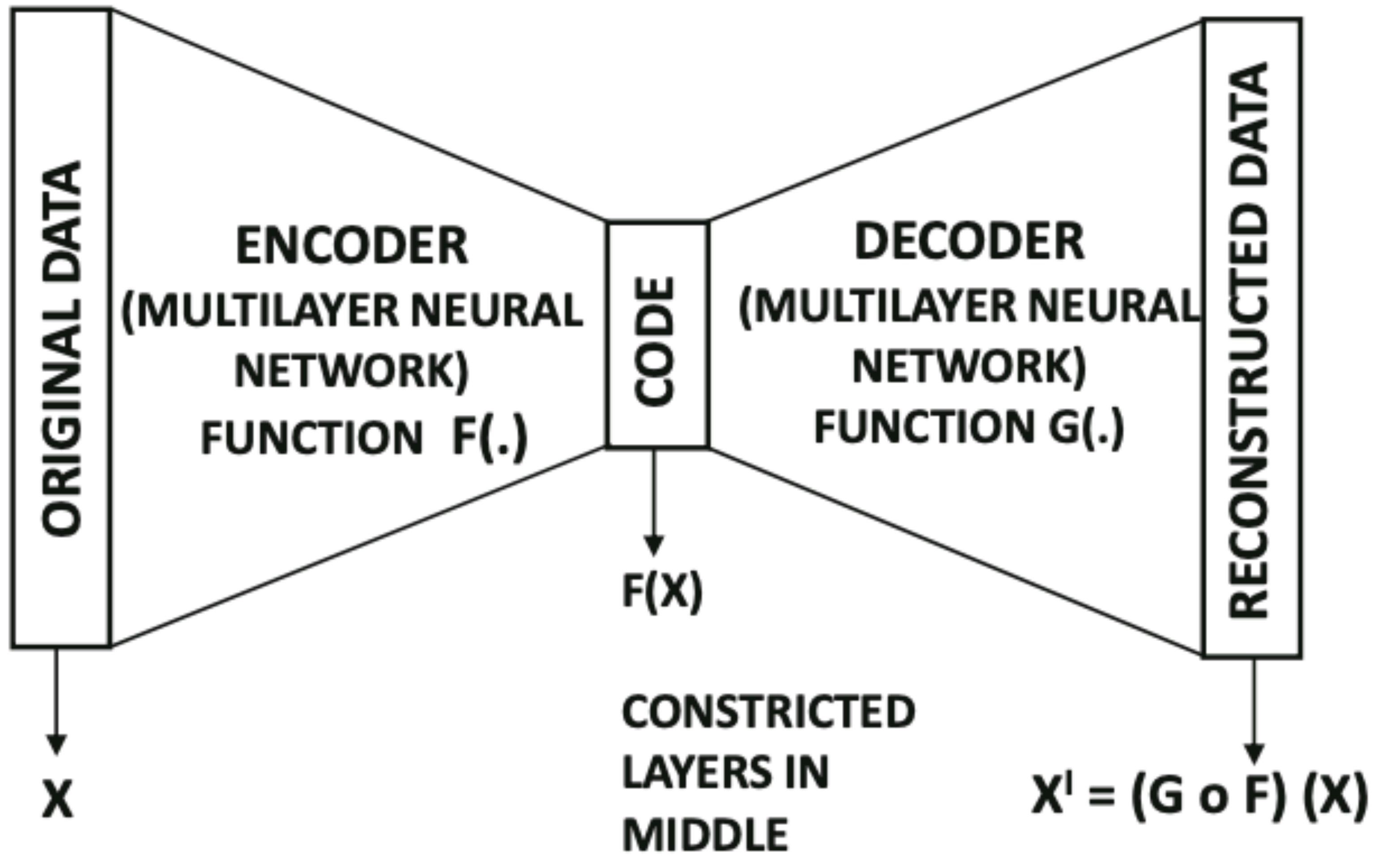
Projects

- Can be done in a group (max two students)
- Be careful about your project partner!
- If he is auditing the course then you will be in trouble!
- Define your own project
- Submit a one page project proposal- within fixed time (first four weeks)?
- Finished the work within the time-line
- Report submission
  - Submission deadline: seven days before the final exam date, is strict and you can adjust your assignment buffer days here - 24-11-2024
  - We will consider 11:59PM as our day end
- Final presentation
  - 20 min (divided into group members)
  - Five days before the final exam date - 26-11-2024 & 27-11-2024

Auto-encoder

- Dimensionality reduction:
  - Classical method (linear): PCA
  - Data matrix: $X \in \mathbb{R}^{d \times N}$
  - Projection matrix: $W \in \mathbb{R}^{d \times d_1}$
  - Lower dims. representation (LDR): $Z = W^T X$; $Z \in \mathbb{R}^{d_1 \times N}$
  - Reconstruction: $\bar{X} = [WW^T]^{-1} WZ$
  - In general - $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d_1}$; $g : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^d$
  - LDR: $Z = f(X)$
  - Reconstruction: $\bar{X} = g(Z) = g[f(X)]$
  - Non-linear dimensionality reduction

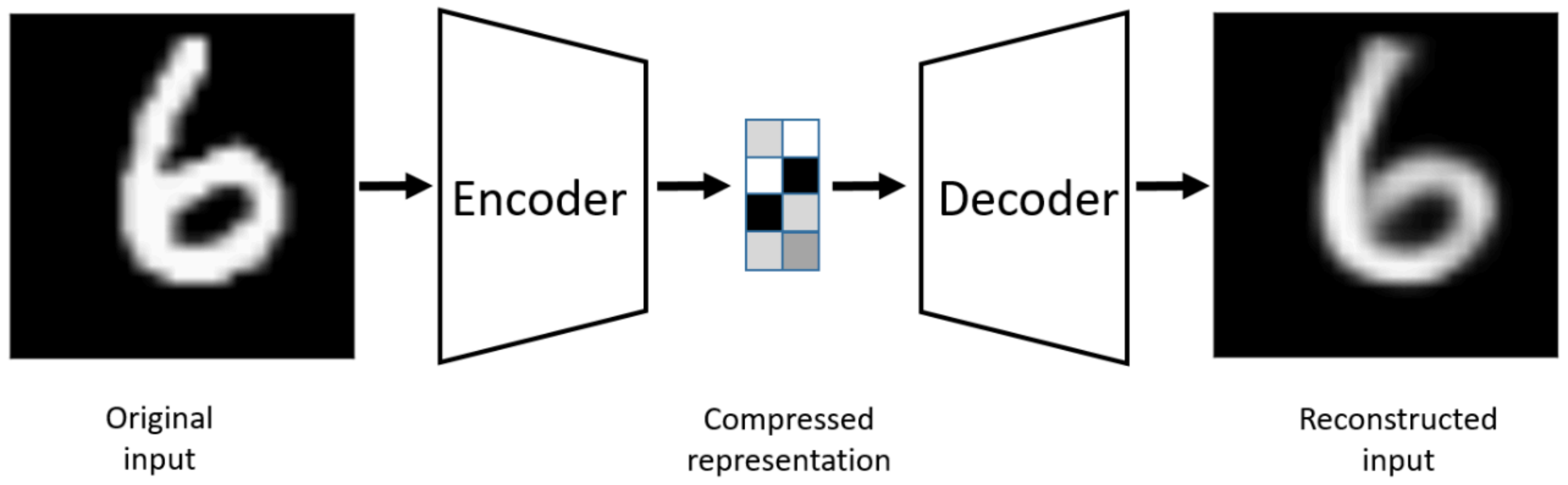




Images: Bank et al. and Chart Agarwal

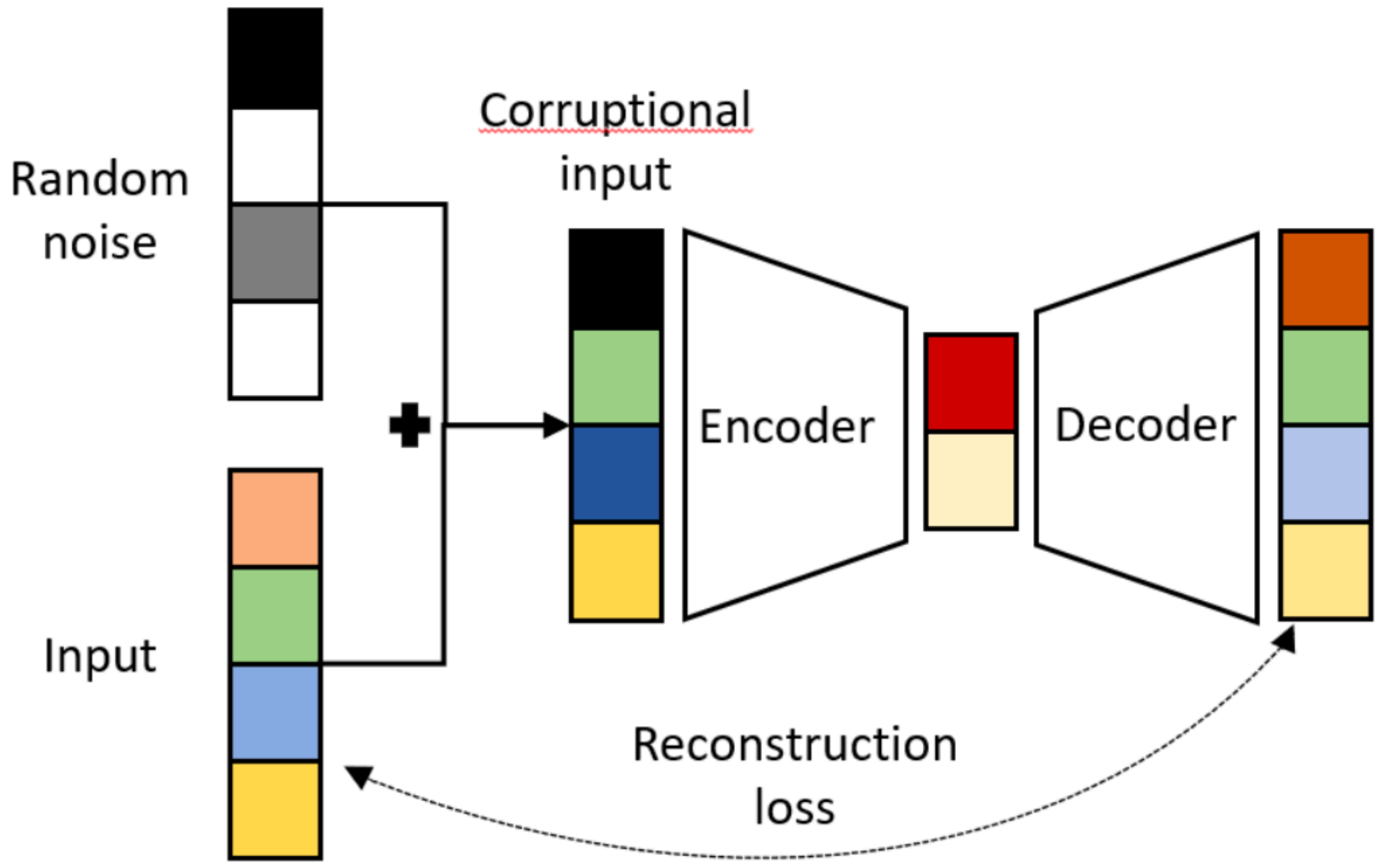
Auto-encoder (cont…)

- Issues in Auto-encoder - $f : \mathbb{R}^d \to \mathbb{R}^{d_1}; g : \mathbb{R}^{d_1} \to \mathbb{R}^d$
  - Overfitting
    - What about the size of the latent dims. (d_1) ?
    - $d < d_1$
    - $d = d_1$
    - $d > d_1$
    - $d_1 = 1$
  - Bias-variance tradeoff
    - We want the architecture of the auto-encoder to be able to reconstruct the input well
    - We want the low representation to generalise to a meaningful one
  - Solution:
    - Sparse Auto-encoder; Denoising Auto-encoder; …



Auto-encoder (cont…)

- Denoising Auto-encoder:
  - Add small noise to the input data
  - What about the random noise ?
    - $\tilde{X} = X + \epsilon; \epsilon \sim \mathcal{N}(0, \sigma^2)$
    - $C_\sigma(\tilde{X} | X) = \mathcal{N}(X, \sigma^2 I)$
    - $C_p(\tilde{X} | X) = \beta \odot X; \beta \sim Ber(p)$

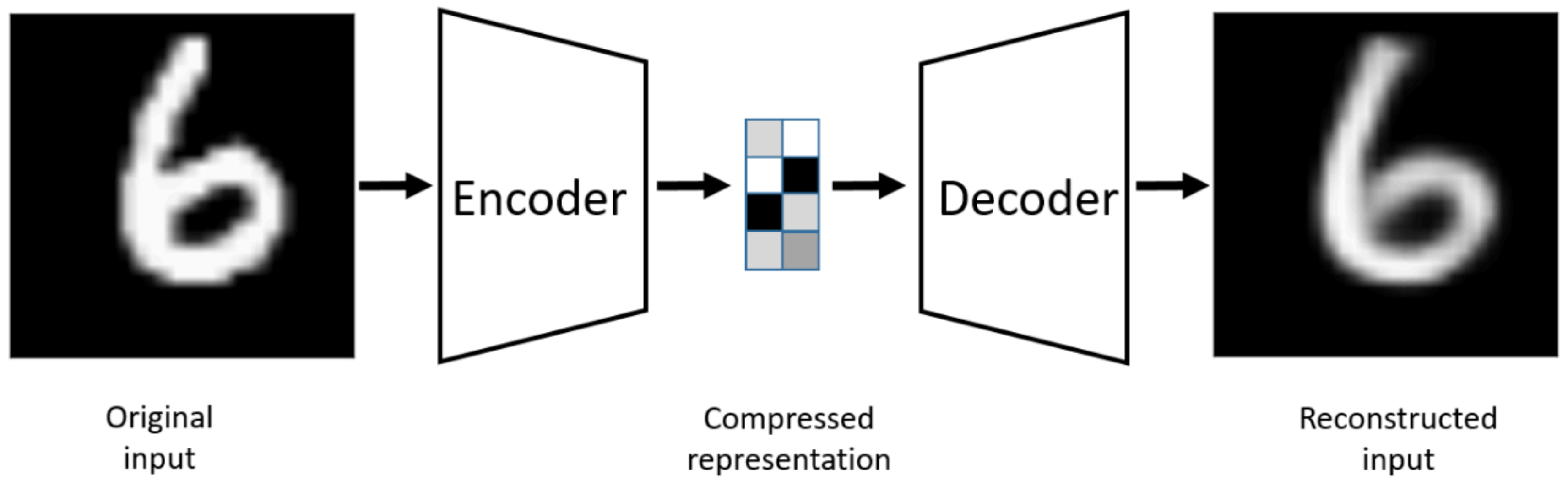


Auto-encoder (cont…)

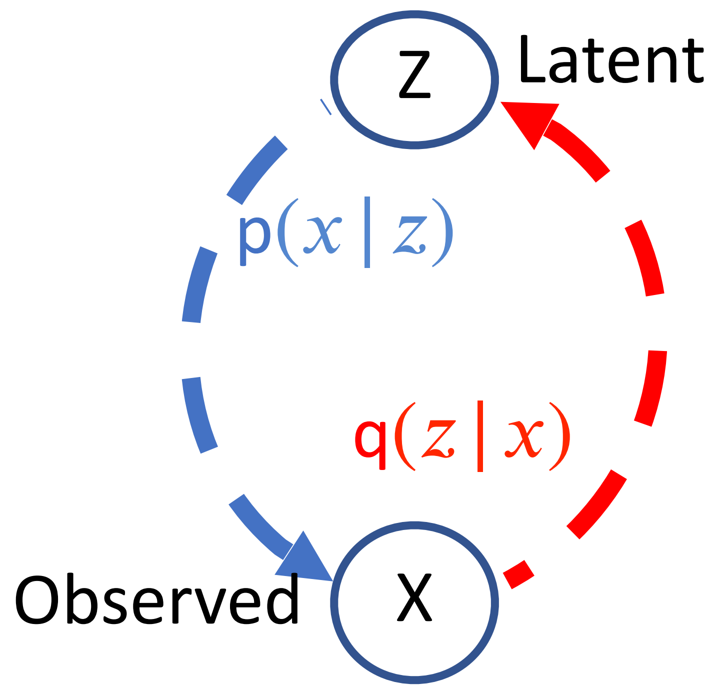
- Demo:
  - https://cs.stanford.edu/~karpathy/convnetjs/demo/autoencoder.html

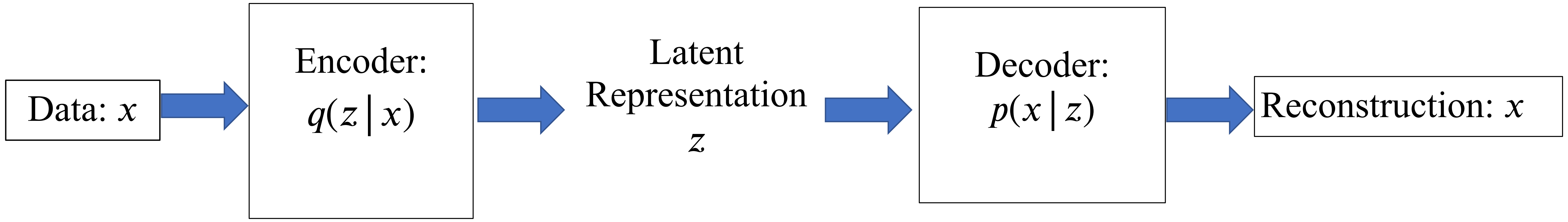
Auto-encoder (cont…)

- Take-home on Auto-encoder
  - Non-linear dimensionality reduction
  - Useful for unsupervised feature learning



Variational auto-encoder





$$
\begin{aligned}
log p_\theta(x) &= E_z[p_\theta(x \mid z)] - KL[q_\phi(z \mid x) \| p_\theta(z)] + \boxed{\underbrace{KL[q_\phi(z \mid x) \| p_\theta(z \mid x)]}_{\geq 0}} \\
&= E_z[p_\theta(x \mid z)] - KL[q_\phi(z \mid x) \| p_\theta(z)]
\end{aligned}
$$

Variational auto-encoder

- Latent space

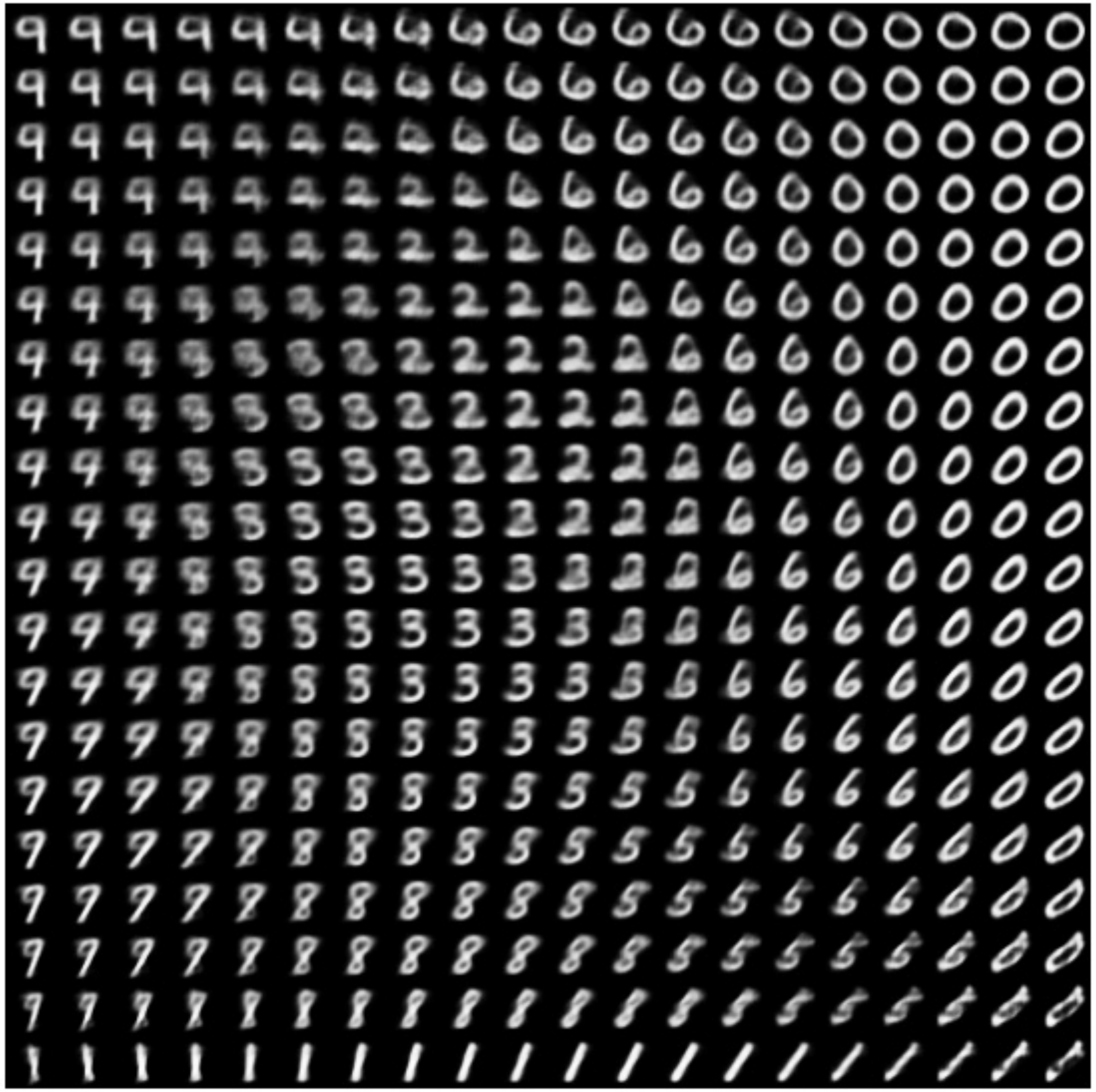
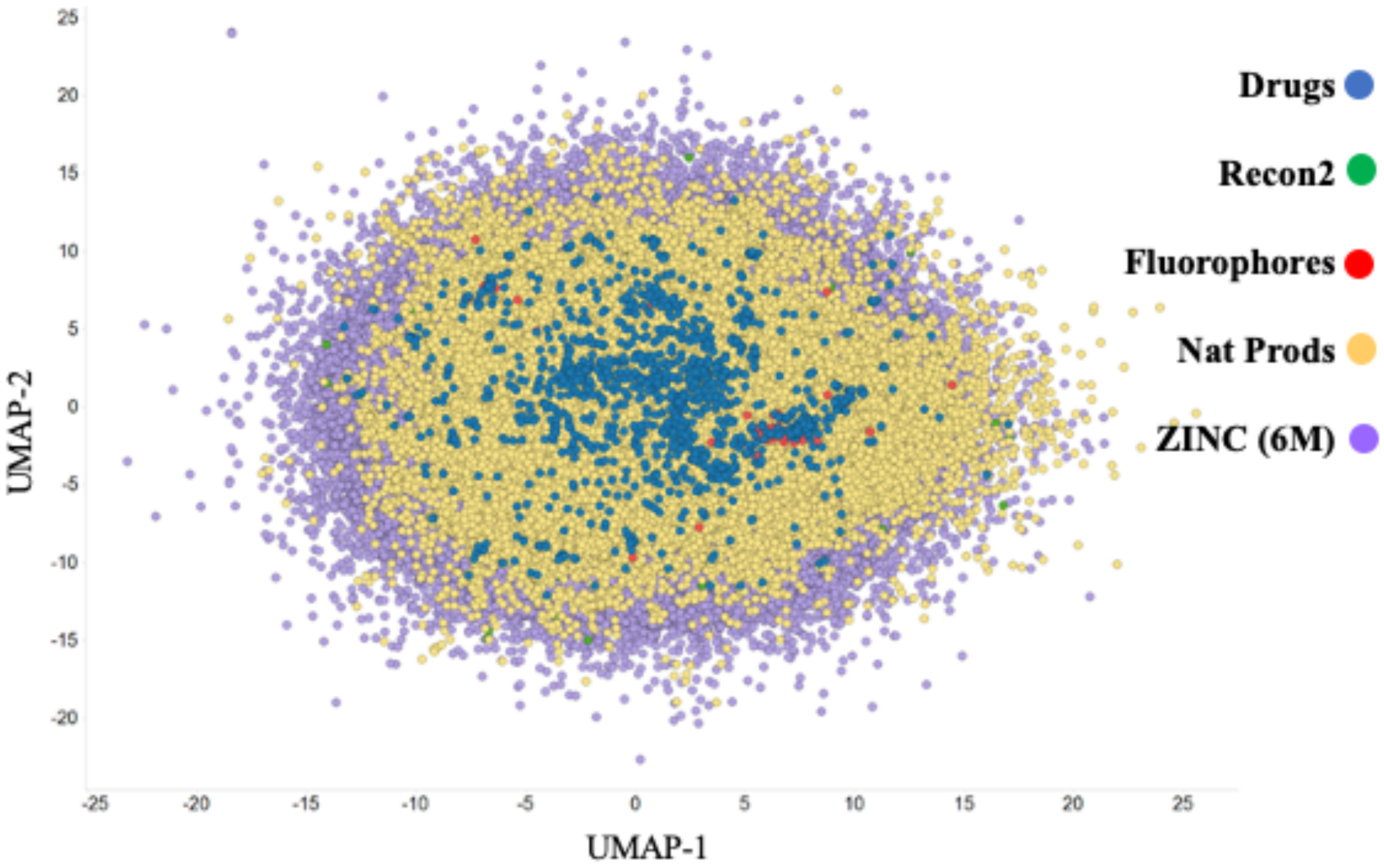


Image source: https://www.tensorflow.org/

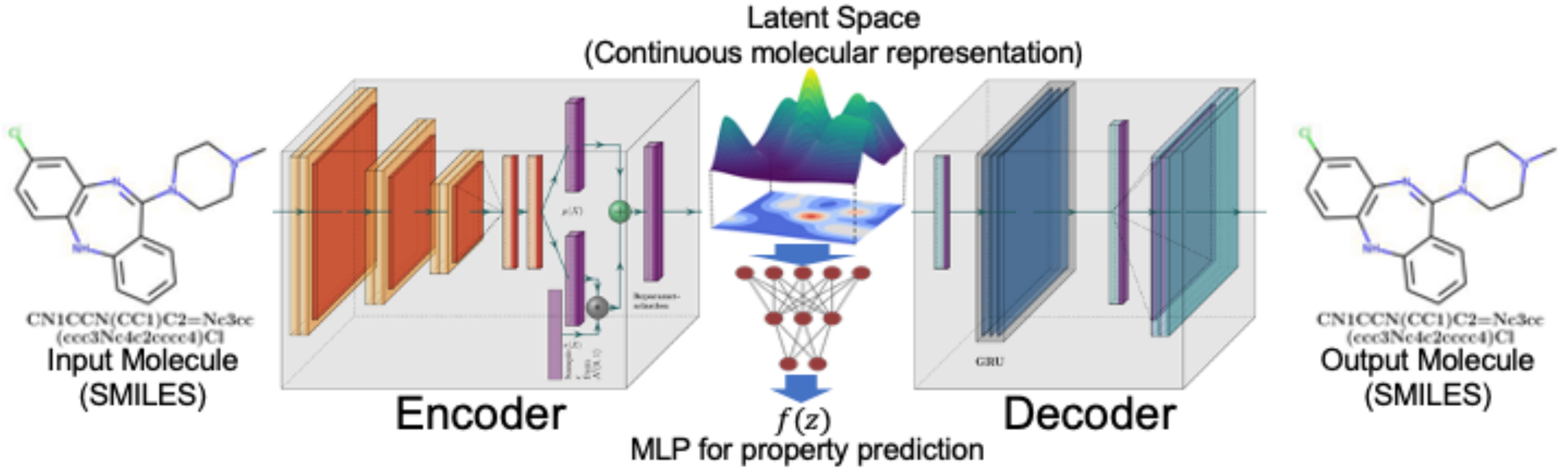
Molecule representation in continuous space



[1]**Samanta** et al. VAE-Sim: A Novel Molecular Similarity Measure Based on a Variational Autoencoder. *Molecules* 2020, *25*, 3446

[2]Yash Khemchandani, Steve O'Hagan, **Soumitra Samanta**, Neil Swainston, Timothy J. Roberts, Danushka Bollegala and Douglas B. Kell, **DeepGraphMolGen, a multi-objective, computational strategy for generating molecules with desirable properties: a graph convolution and reinforcement learning approach**, In Journal of Cheminformatics, 12, 53, September 2020

Generative model for new molecule generation: Variational Autoencoder



SMILES: Simplified molecular-input line-entry system

[1]Samanta et al. VAE-Sim: A Novel Molecular Similarity Measure Based on a Variational Autoencoder. Molecules 2020, 25, 3446

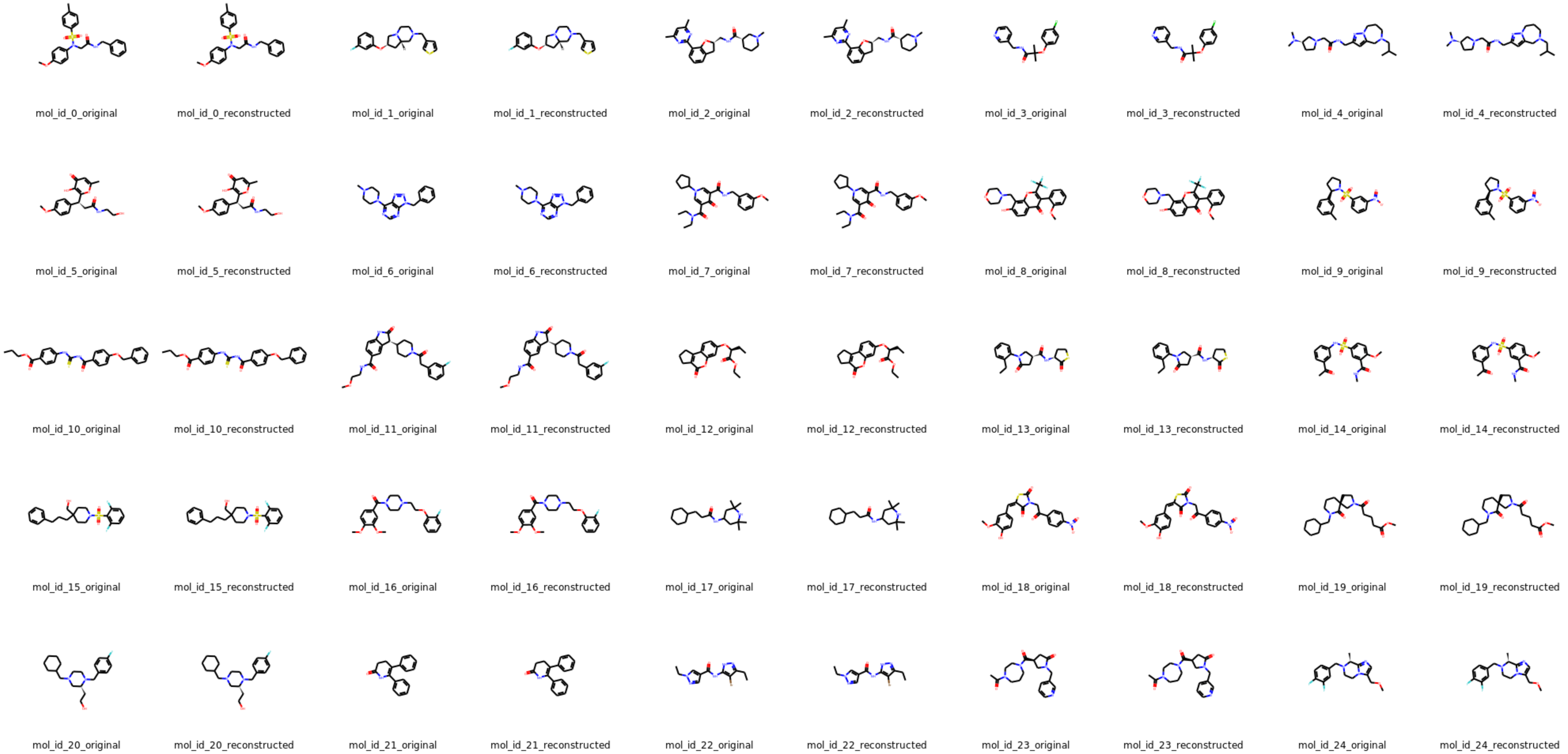
[2]Yash Khemchandani, Steve O'Hagan, Soumitra Samanta, Neil Swainston, Timothy J. Roberts, Danushka Bollegala and Douglas B. Kell, DeepGraphMolGen, a multi-objective, computational strategy for generating molecules with desirable properties: a graph convolution and reinforcement learning approach, In Journal of Cheminformatics, 12, 53, September 2020

Results

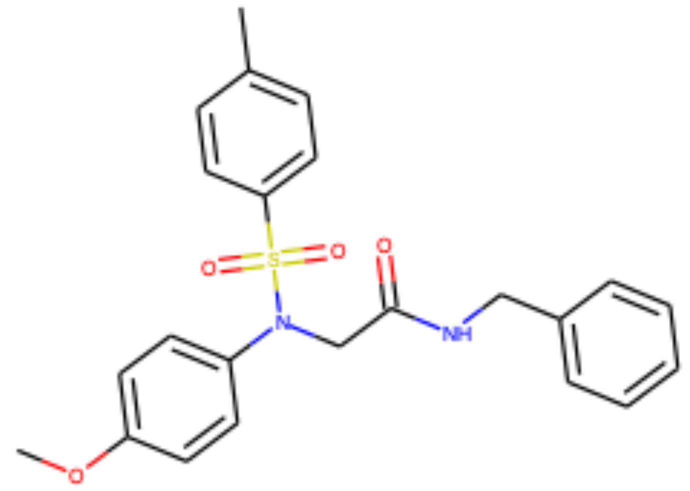
- ZINC15: 2D Drug-Like, clean and in-stock (6,202,415 substances)
- Data partition: randomly partition the dataset into train-50% (3,101,207), validation-20% (1,240,483) and test-30% (1,860,725)

Data partition	#Samples	#Valid reconstructed samples	Accuracy(%)
Train	3,101,207	2,984,669	96.24
Validation	1,240,483	1,180,189	95.13
Test	1,860,725	1,771,064	95.18

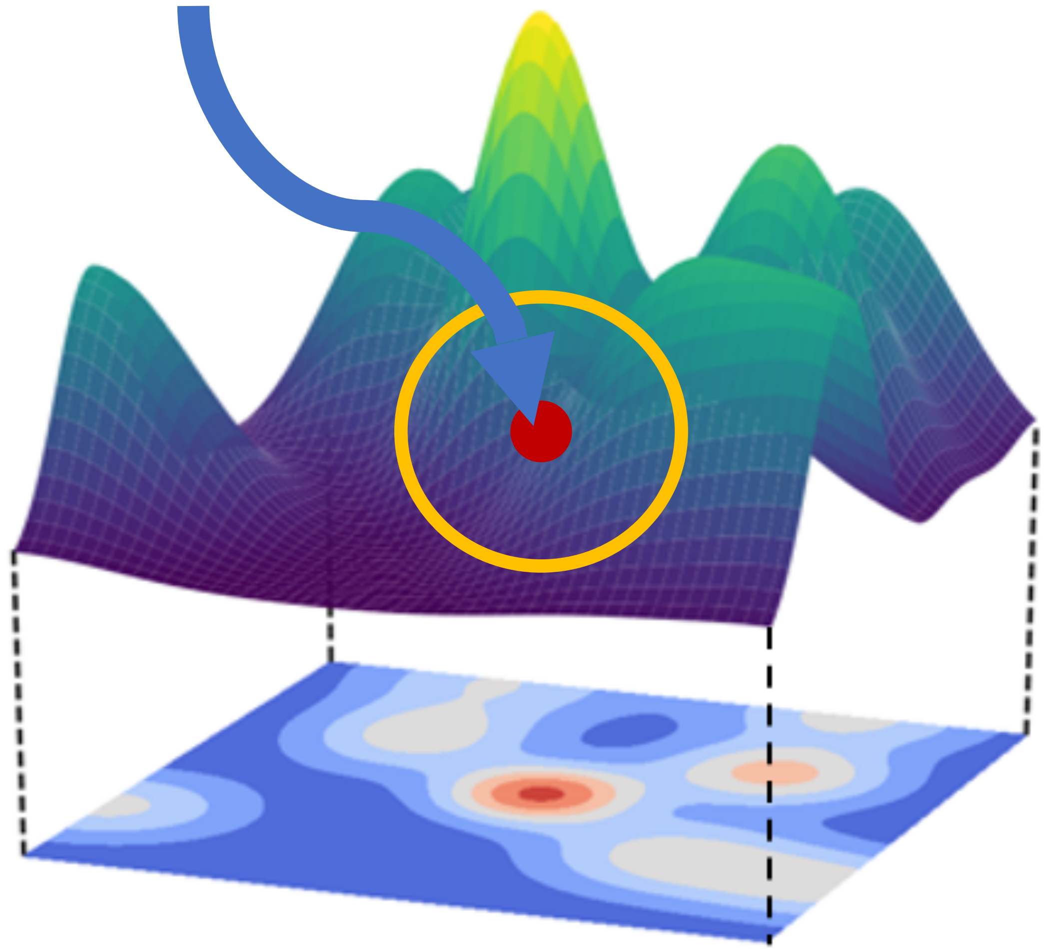
Reconstructed molecules on test dataset



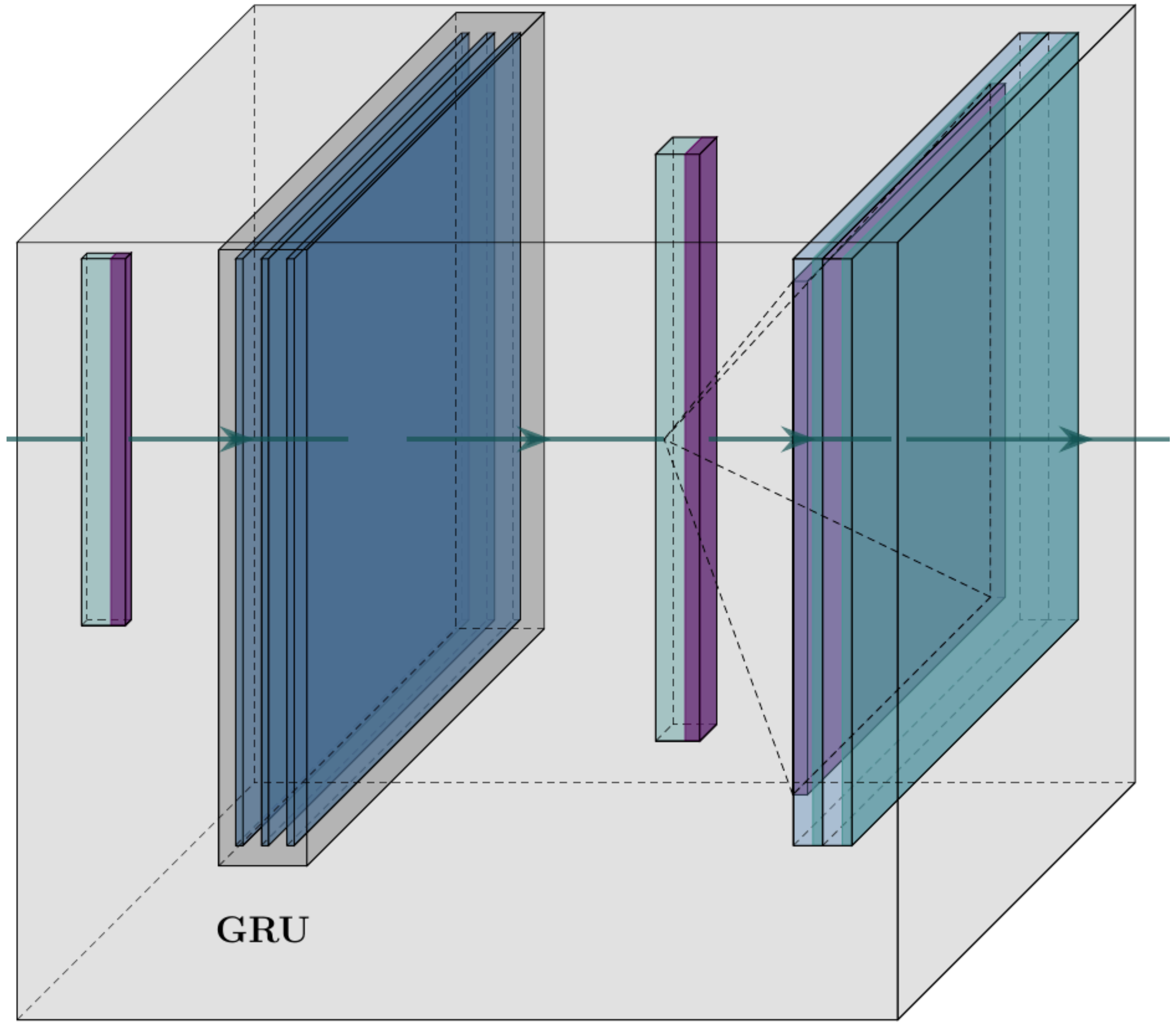
Sampling molecules from the latent space

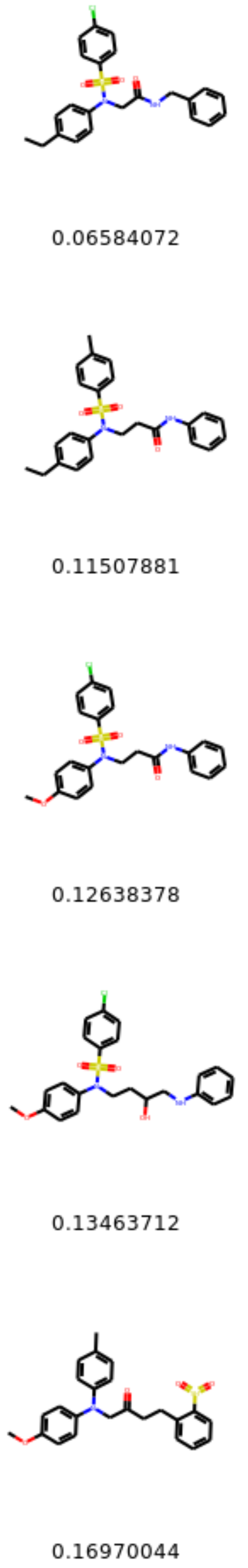


Oc1ccc(N(CC(=O)NCc2ccccc2)S(=O)(=O)c2ccc(C)cc2)c



Sample from latent space

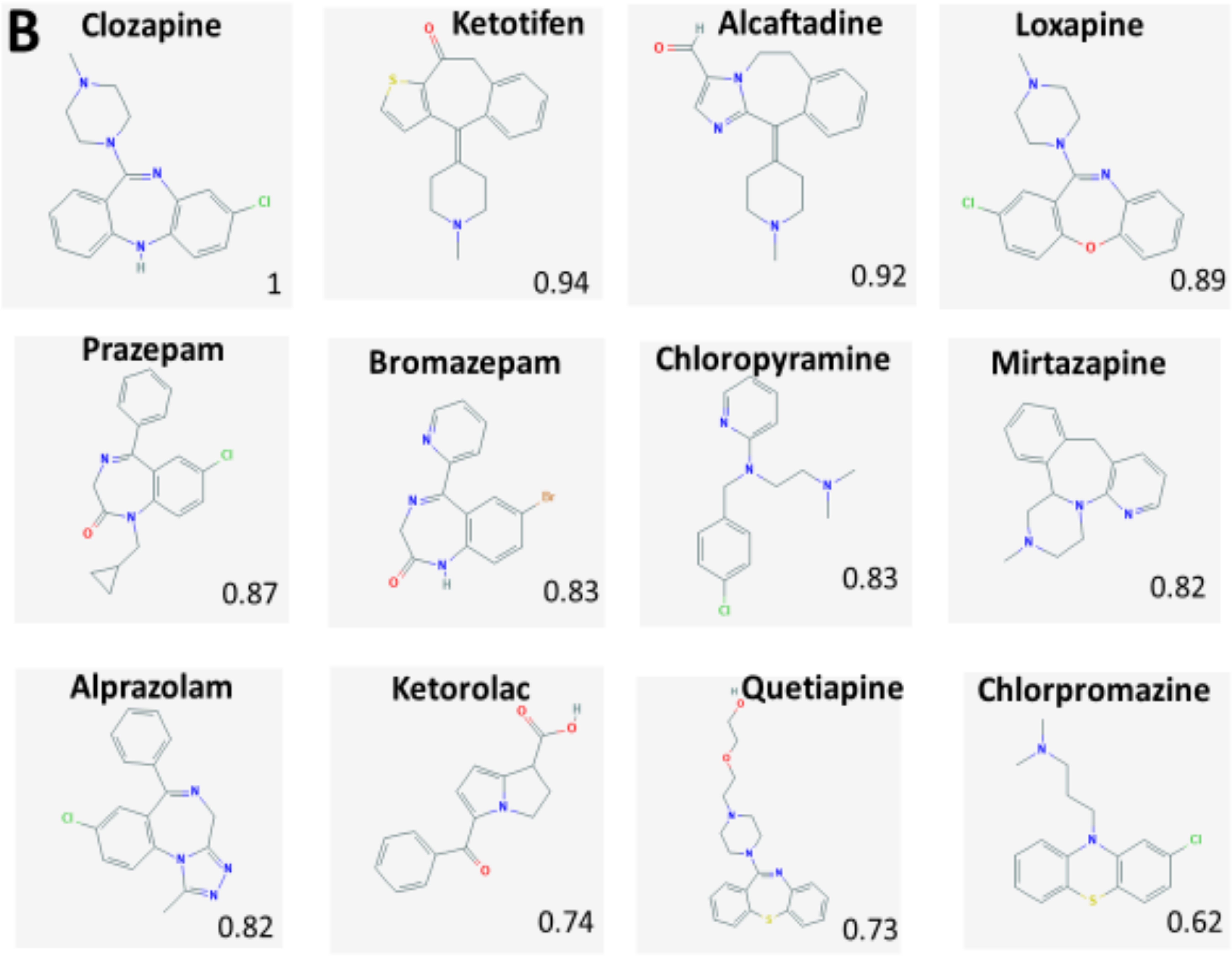




Output

Nearest neighbour sampling for a test data

Nearest neighbour search



Latent dimension evaluation