

Variational Auto-encoders

Shuwen Yue

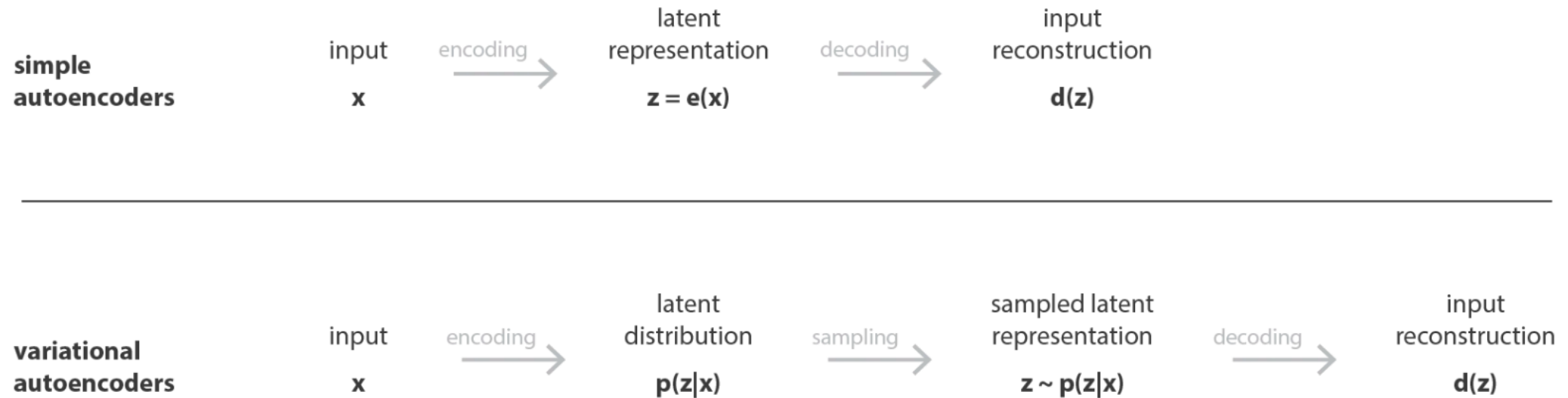
i-CoMSE ML for Molecular Science

Auto-encoders vs PCA

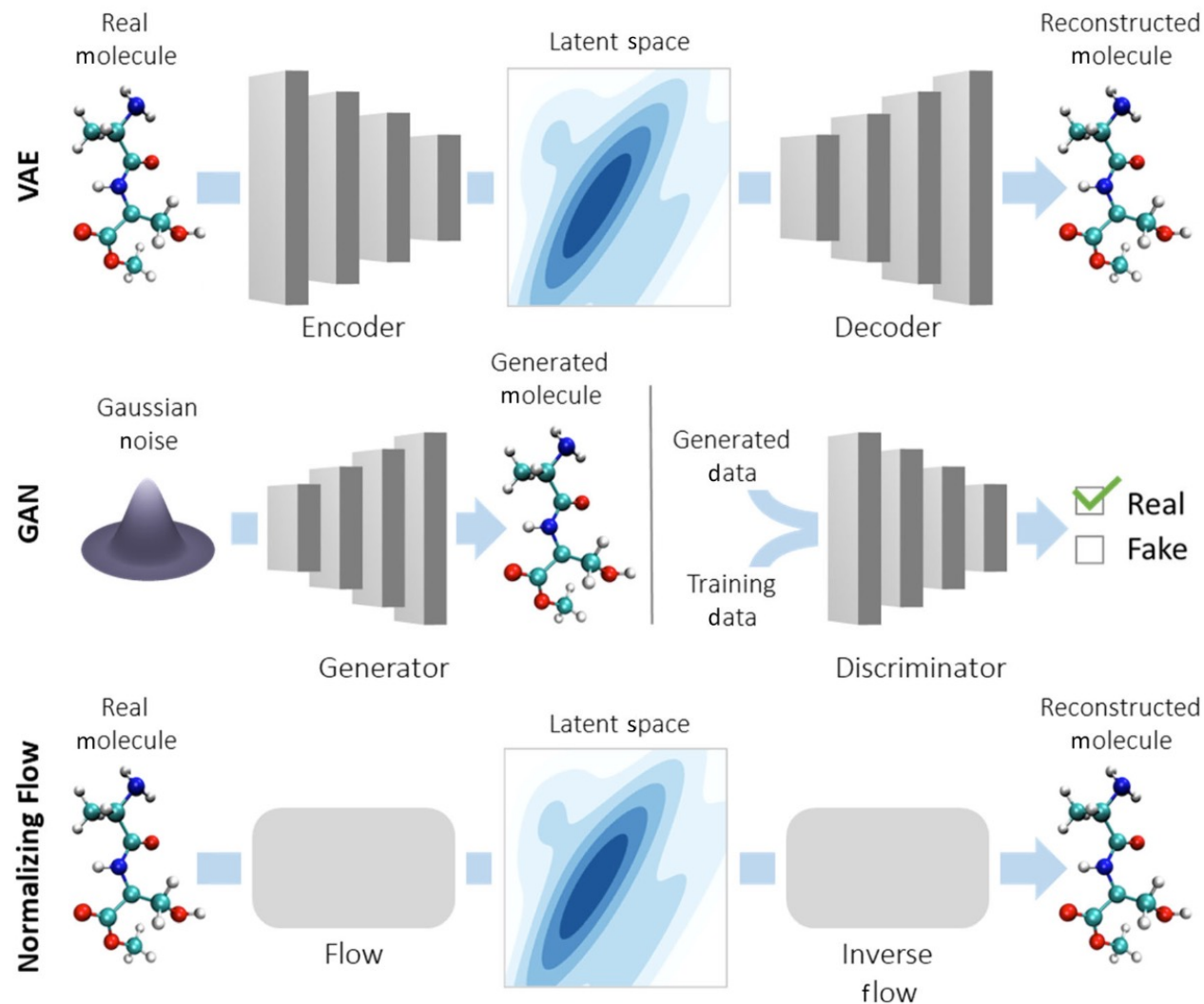
- PCA is a linear transformation, auto-encoders can describe complicated non-linear processes
- PCA features projects in orthogonal basis. Auto-encoders features optimize for reconstruction, could have correlated features
- PCA is cheaper to compute than autoencoders
- Auto-encoders have a large number of parameters, prone to overfitting

Autoencoder vs variational autoencoder

VAE encodes data as probability distribution instead of a single point



Types of Generative models



Generation tries to recover correct molecule reconstruction AND regularization from learned molecular embedding

Generates molecules from Gaussian noise, where a discriminator learns to identify molecules as real or fake. Two networks competing against each other.

Model learns a series of invertible transformations between a prior distribution and molecular data. Can calculate exact data likelihood.

Example of generative models in Molecular Science

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

Rafael Gómez-Bombarelli,^{†,‡,§} Jennifer N. Wei,^{‡,§} David Duvenaud,^{¶,§} José Miguel Hernández-Lobato,^{§,‡} Benjamín Sánchez-Lengeling,[‡] Dennis Sheberla,[‡] Jorge Aguilera-Iparraguirre,[†] Timothy D. Hirzel,[†] Ryan P. Adams,^{¶,||} and Alán Aspuru-Guzik^{*,†,‡,§}

[†]Kyulux North America Inc., 10 Post Office Square, Suite 800, Boston, Massachusetts 02109, United States

[‡]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States

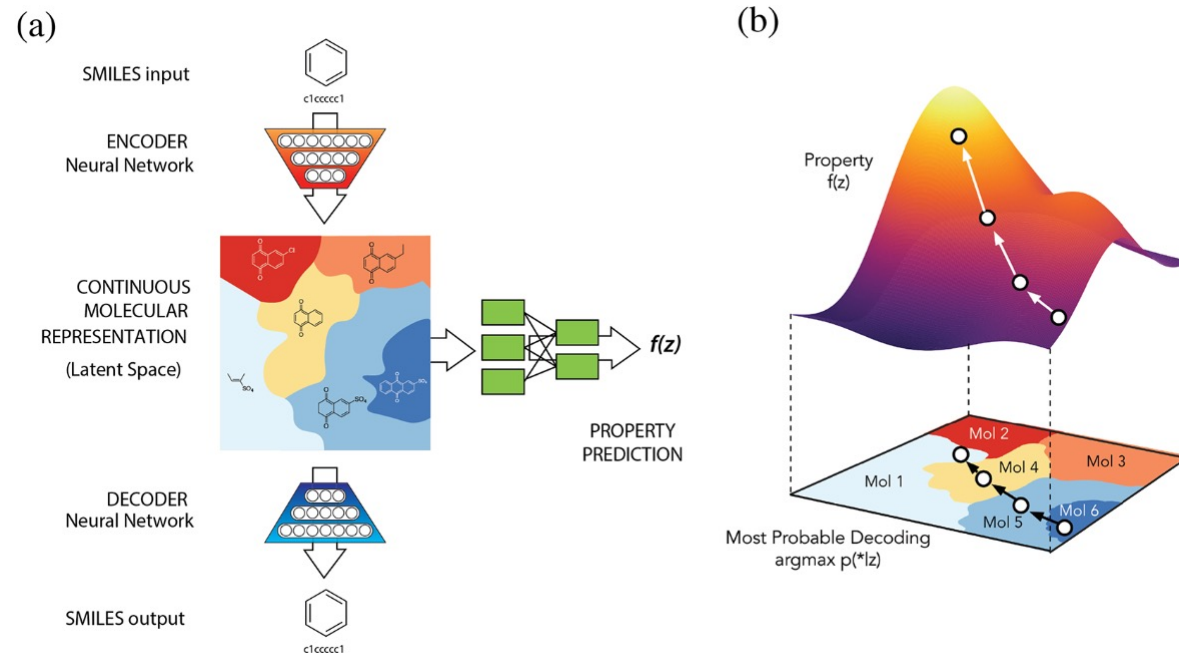
[¶]Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3H5, Canada

[§]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.

[¶]Google Brain, Mountain View, California, United States

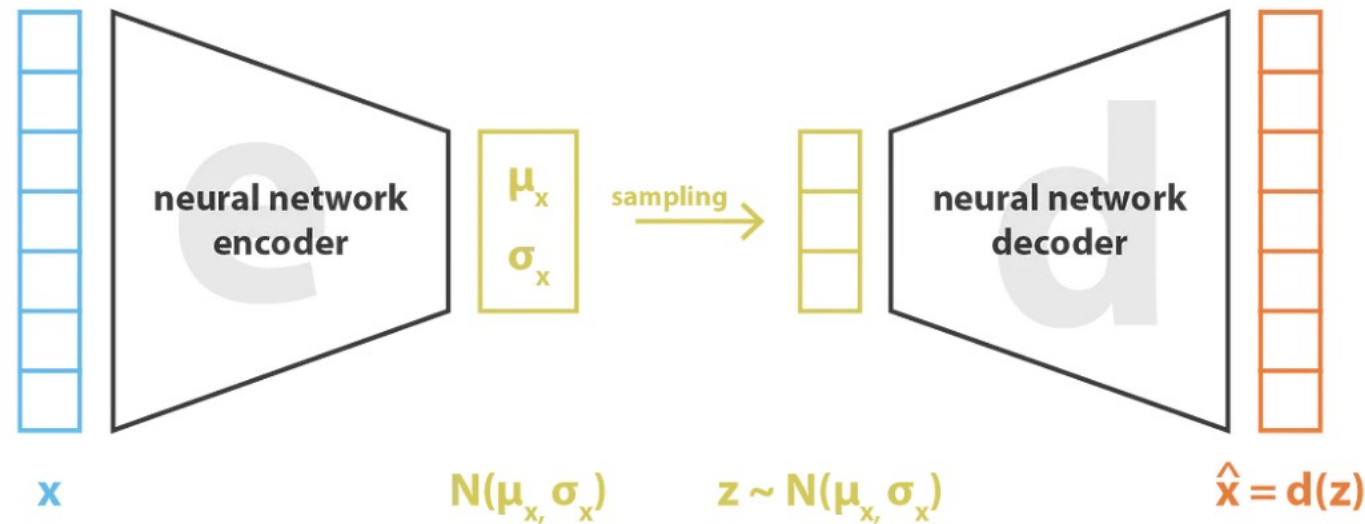
^{||}Princeton University, Princeton, New Jersey, United States

^{*}Biologically-Inspired Solar Energy Program, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario M5S 1M1, Canada



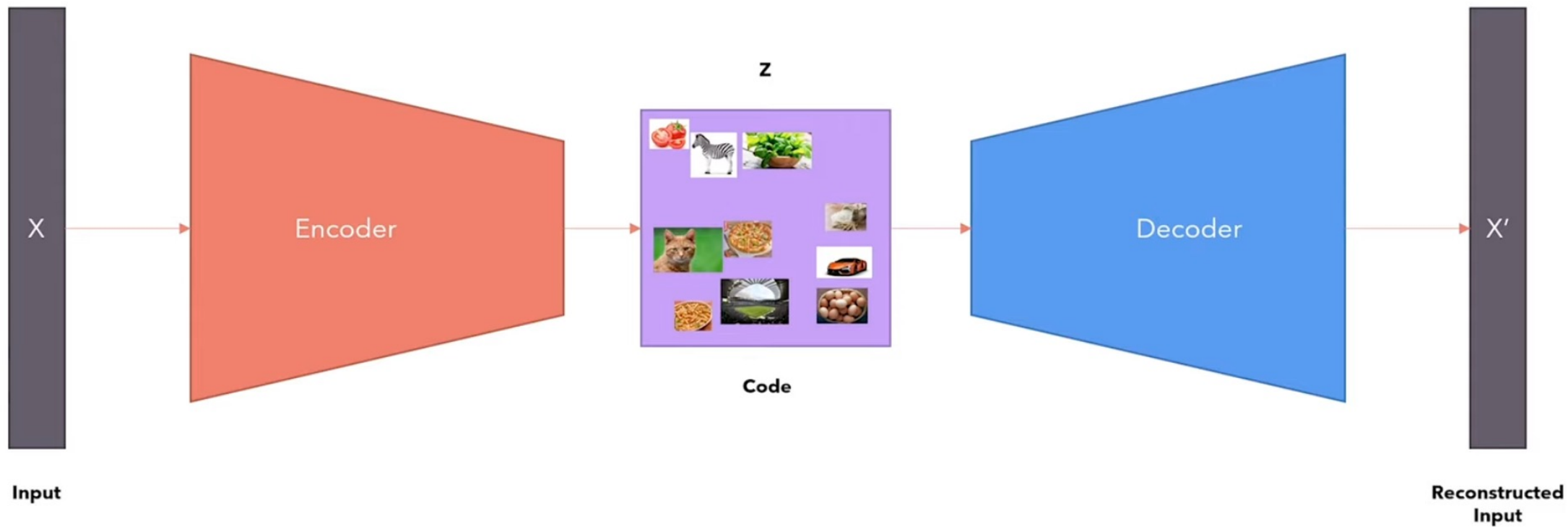
Autoencoder vs variational autoencoder

Regularization in the form of the Kullback-Leibler divergence -> this induces better organization in the latent space

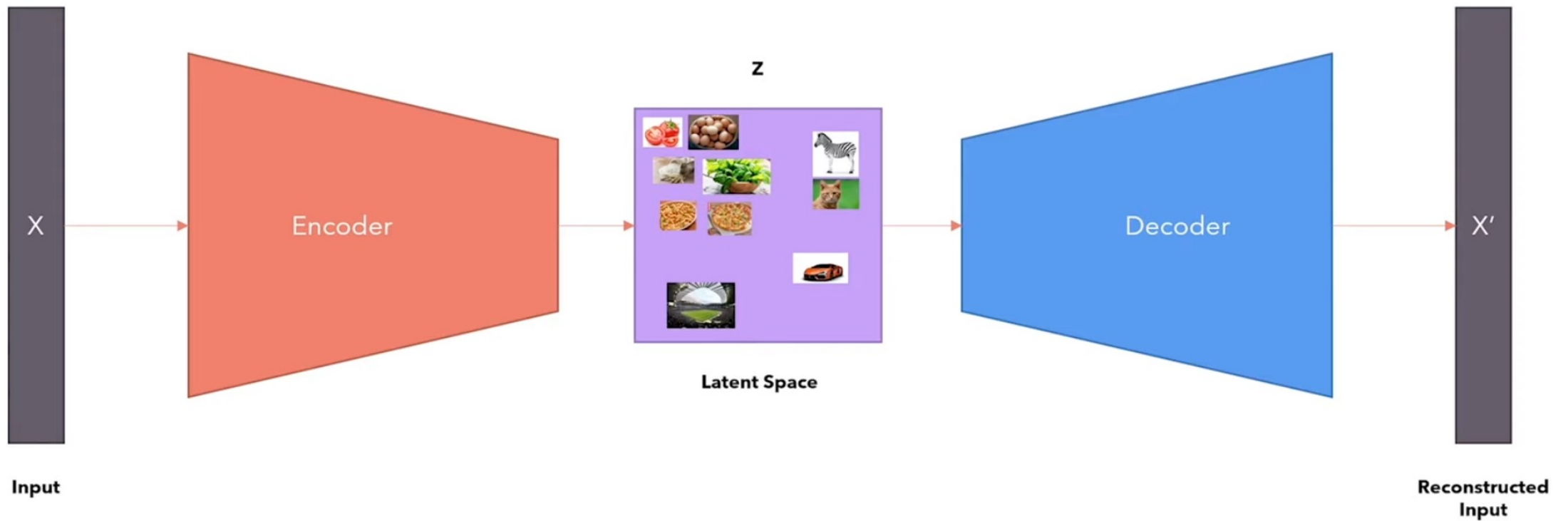


$$\text{loss} = \|x - \hat{x}\|^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)] = \|x - d(z)\|^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)]$$

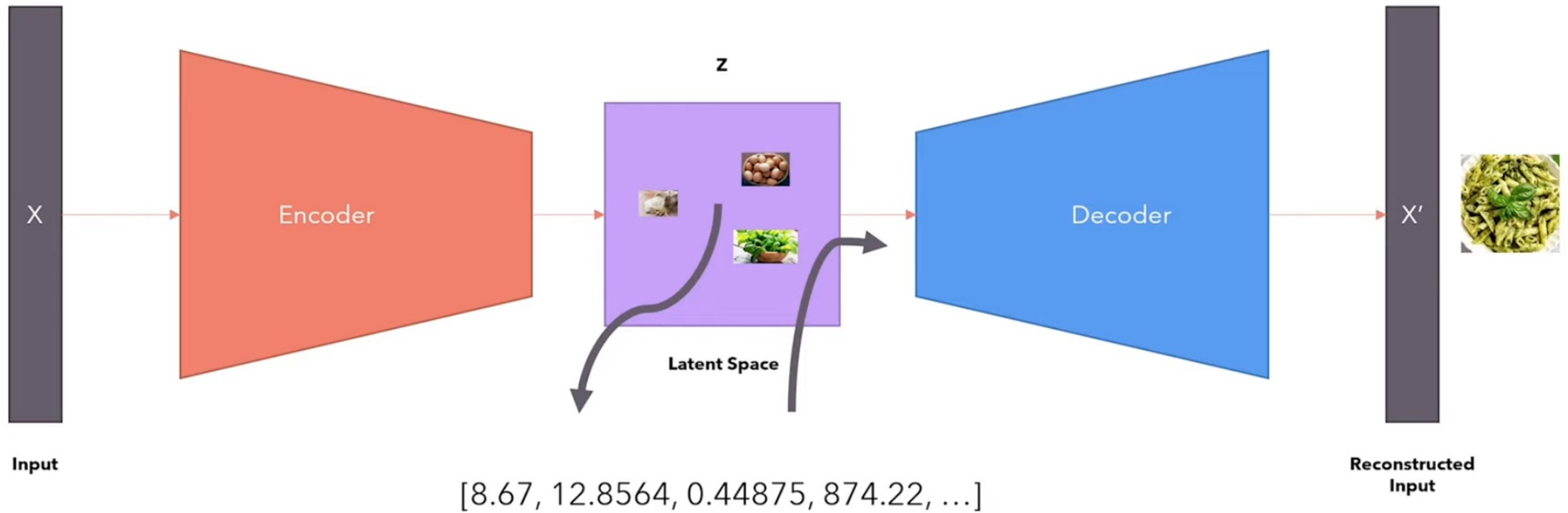
Autoencoder



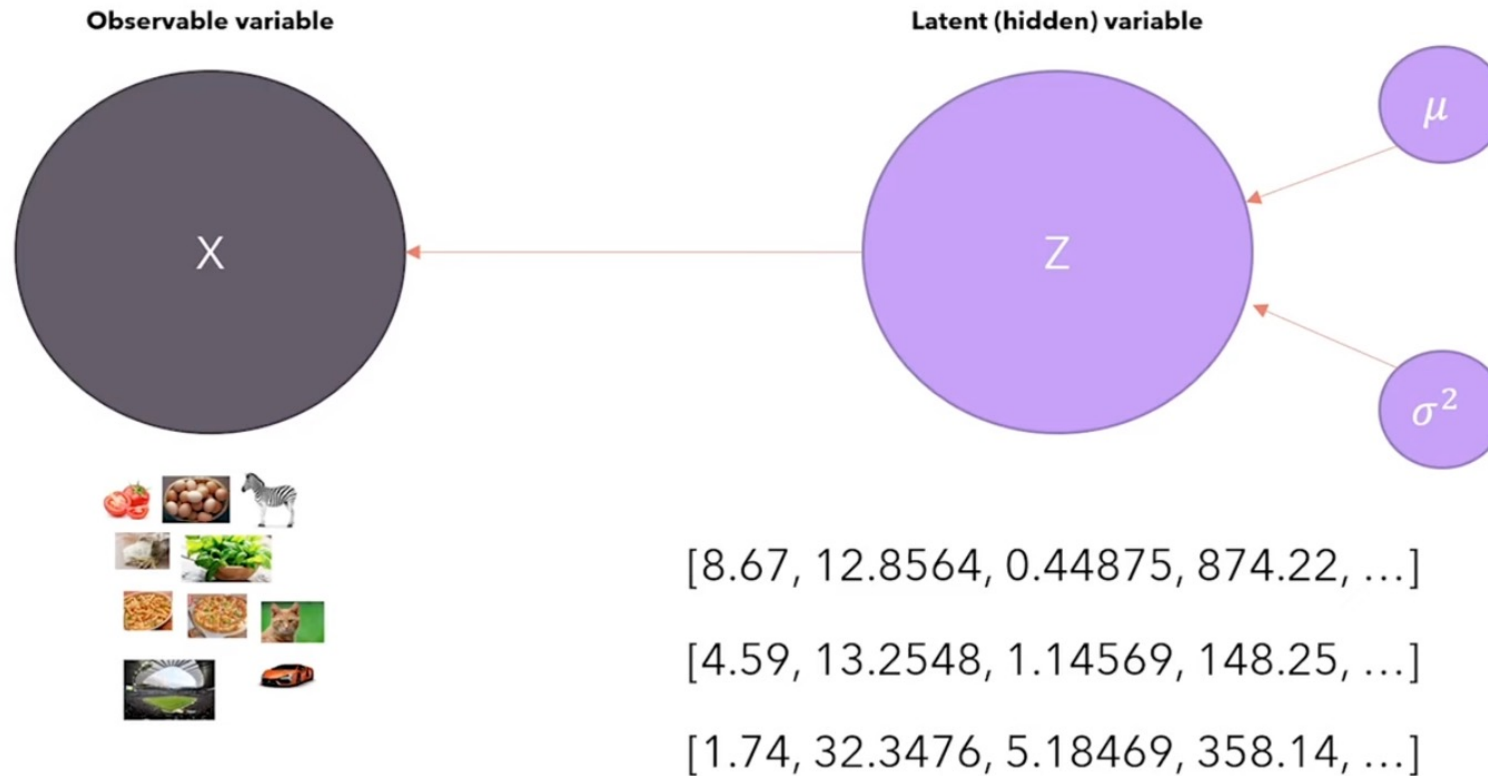
Variational Autoencoder



Sampling the latent space



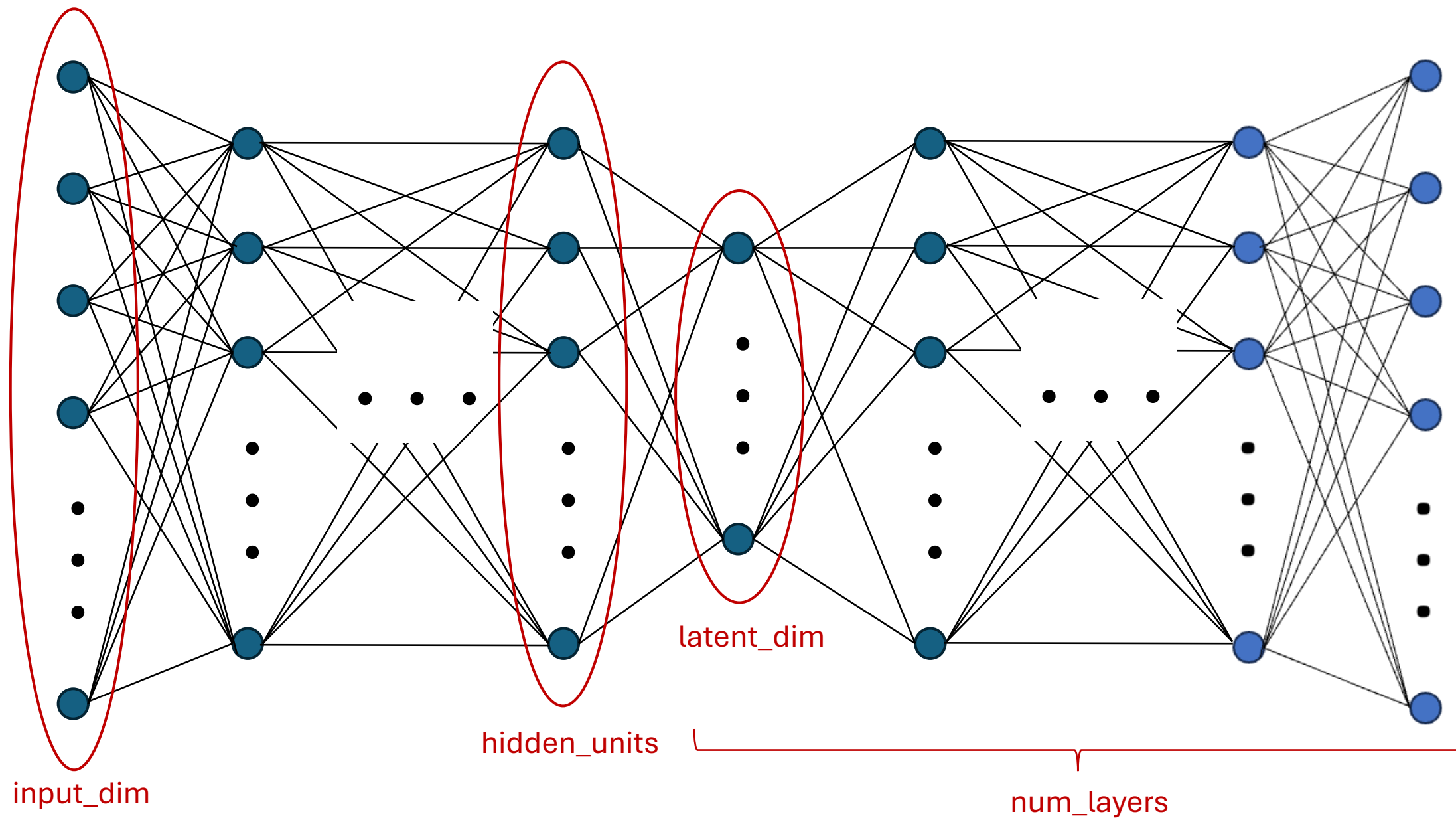
Distributions in the latent space



Math

VAE notebook

VAE Schematic



Considerations when generating molecules

- Lack of benchmarks for generated molecule validity, synthesizability, uniqueness, diversity
- Synthesizability scores
- Representation considerations
- Generative model vs genetic algorithm vs enumeration vs high throughput screening...etc