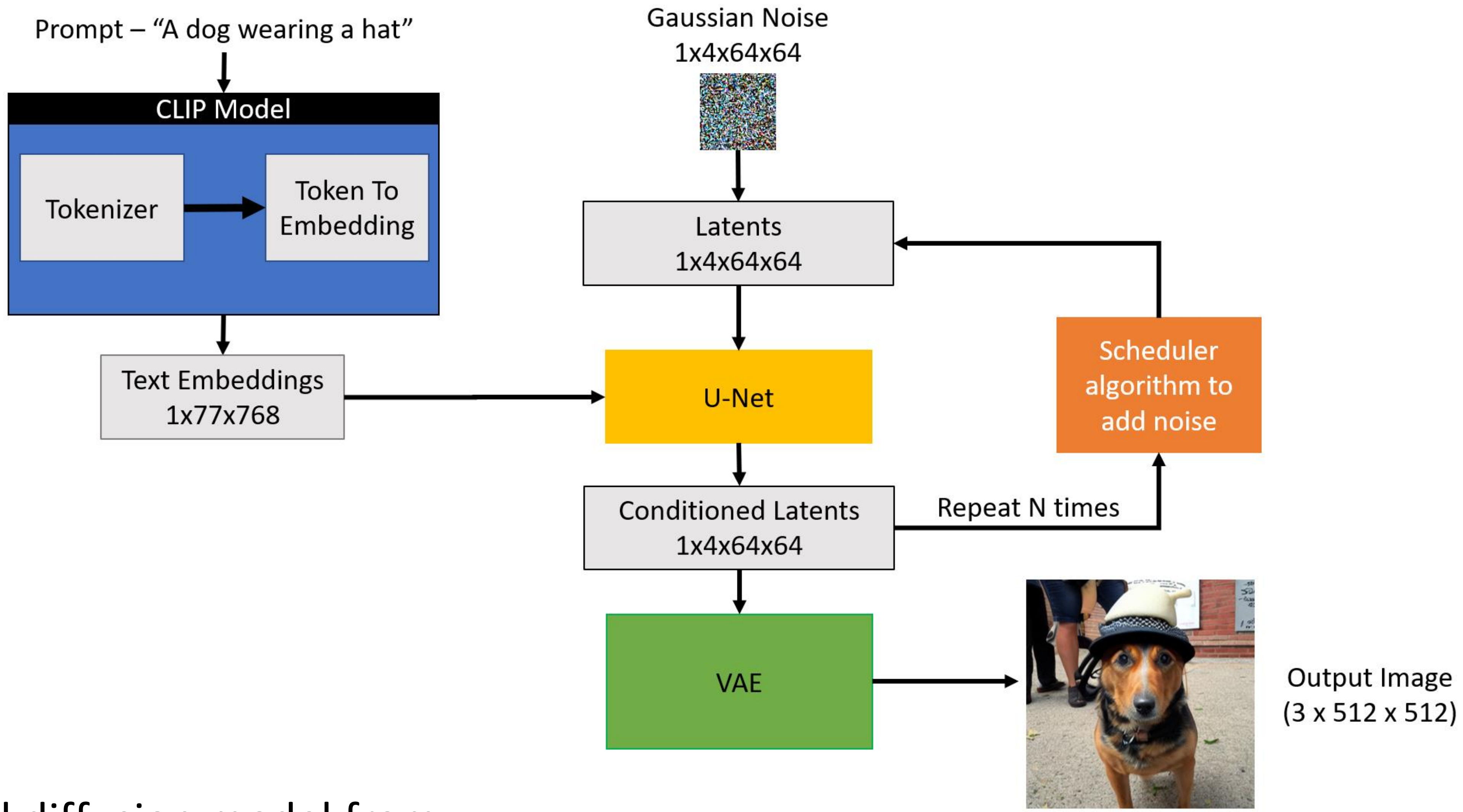
Generative models and Variational Autoencoders

Shuwen Yue

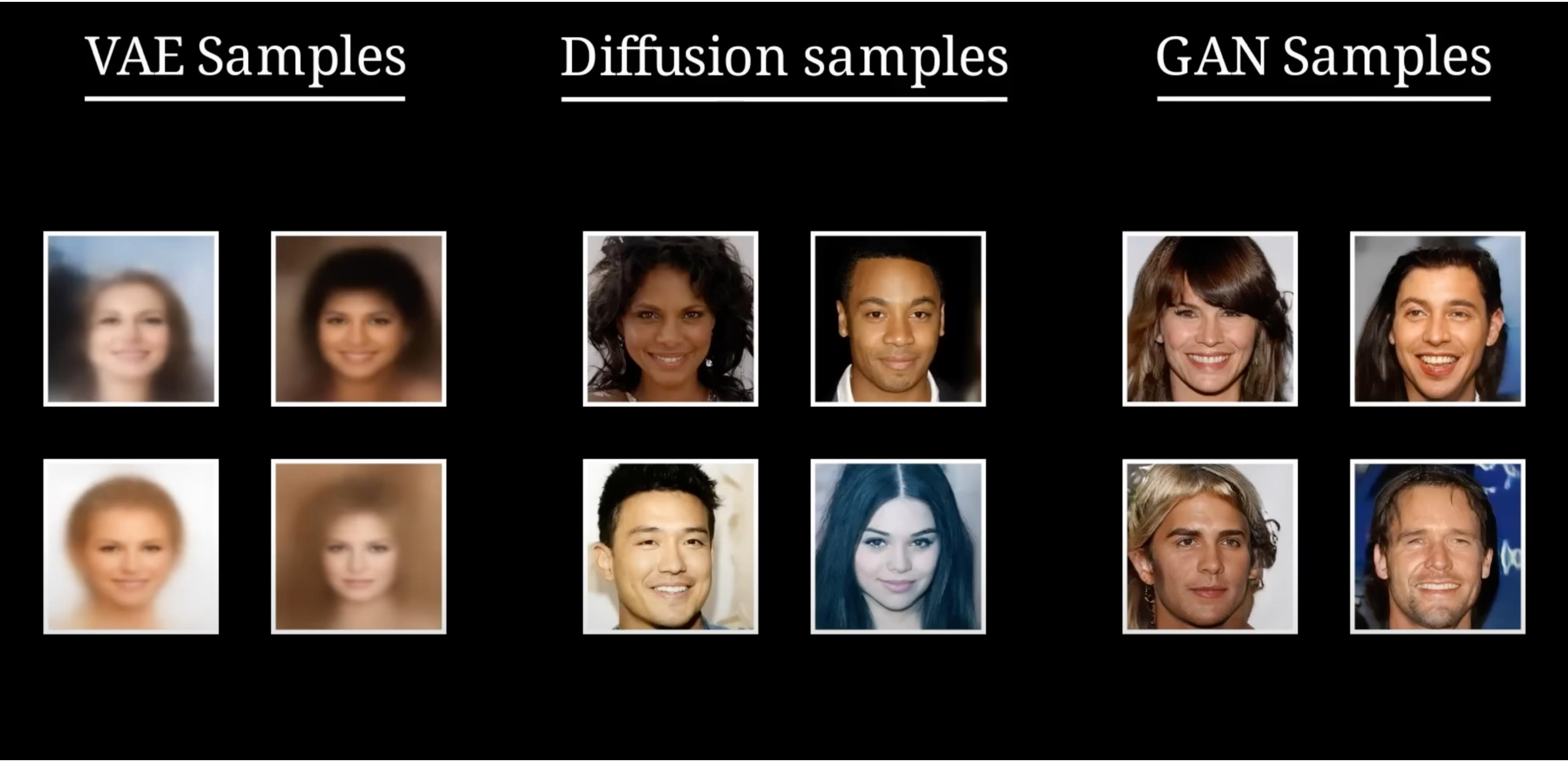
Assistant Professor, Cornell University

May 1, 2025

Generative models for general use

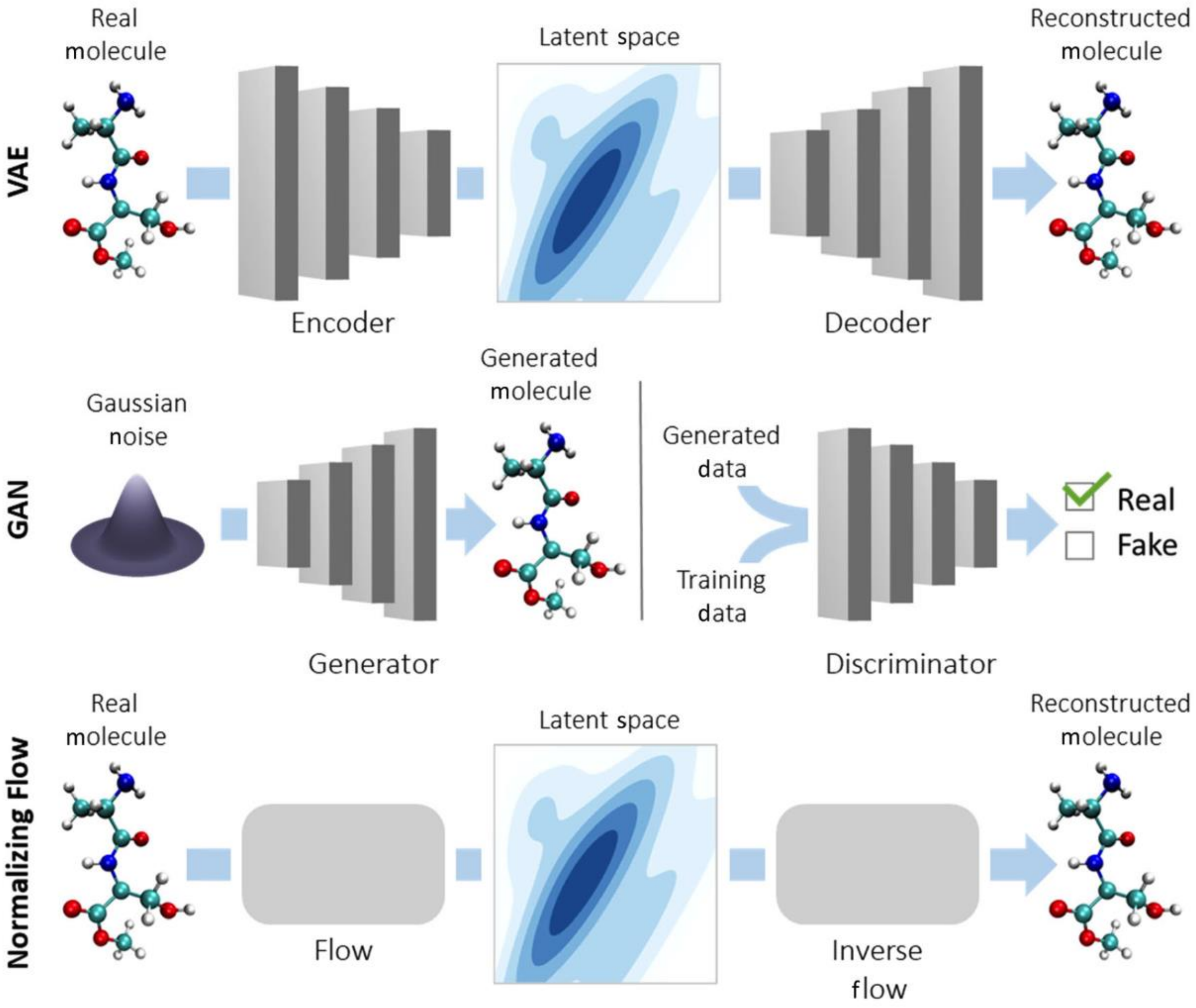


Stabilized diffusion model from Hugging Face



https://www.youtube.com/watch?v=qJeaCHQ1k2w

Types of Generative models for molecules



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.

Bilodeau et al. WIREs Computational Molecular Science. (2022)

Latent space optimization for target properties

ACS central science

ACS AuthorChoice

Research Article

Cite This: ACS Cent. Sci. 2018, 4, 268−276

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

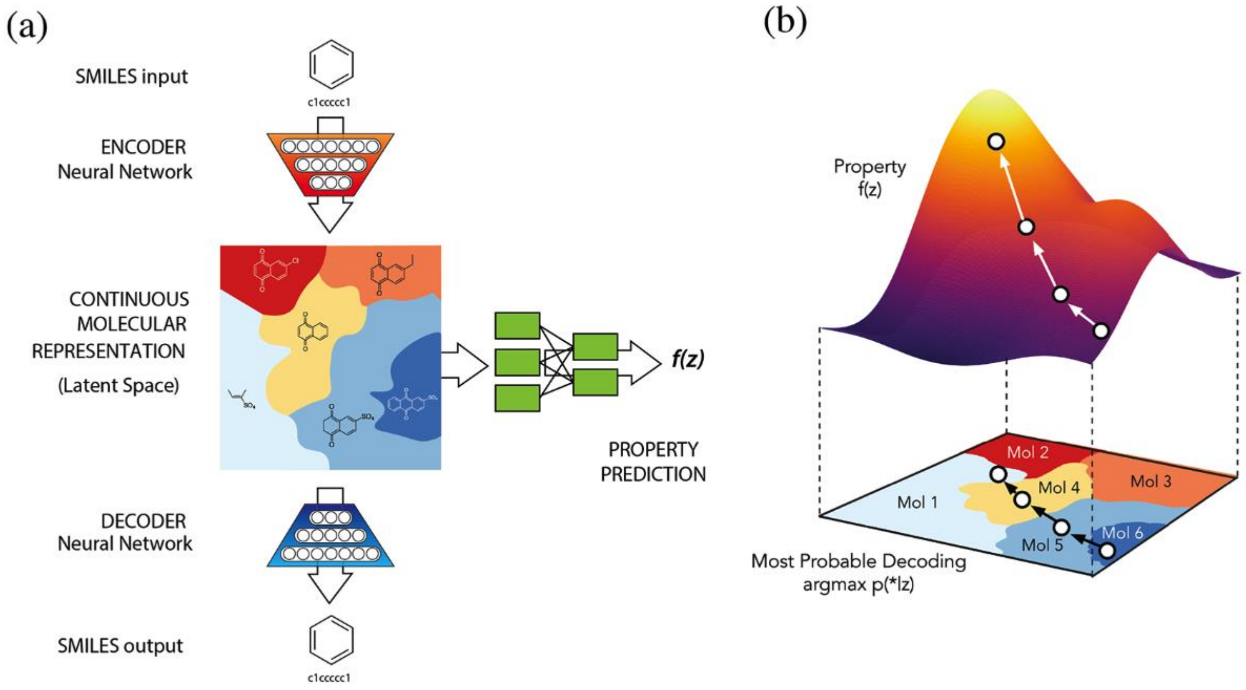
[¶]Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3H5, Canada

[§]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

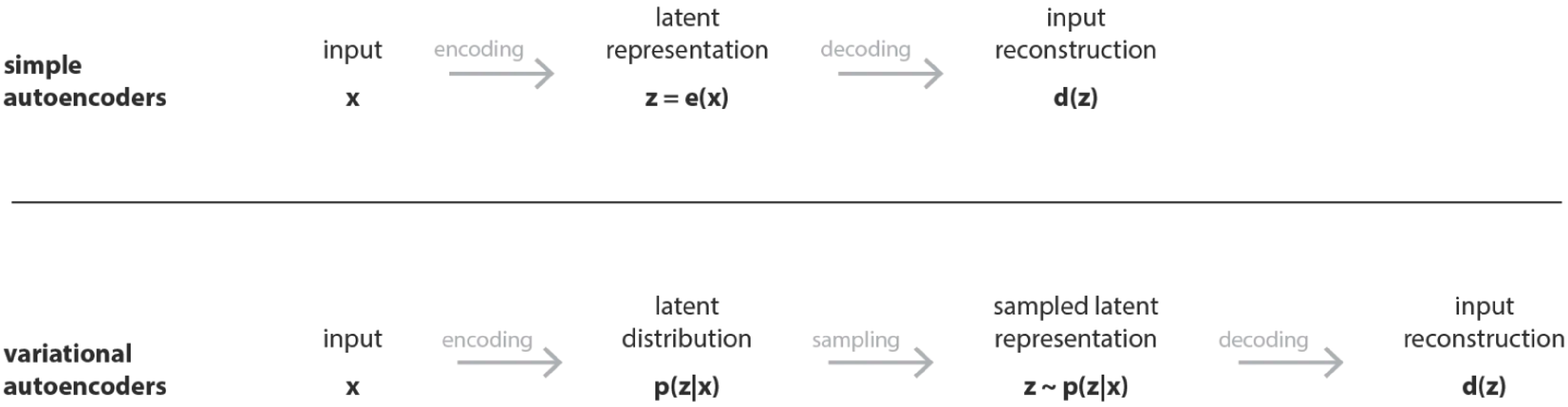


Auto-encoders vs PCA

- 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

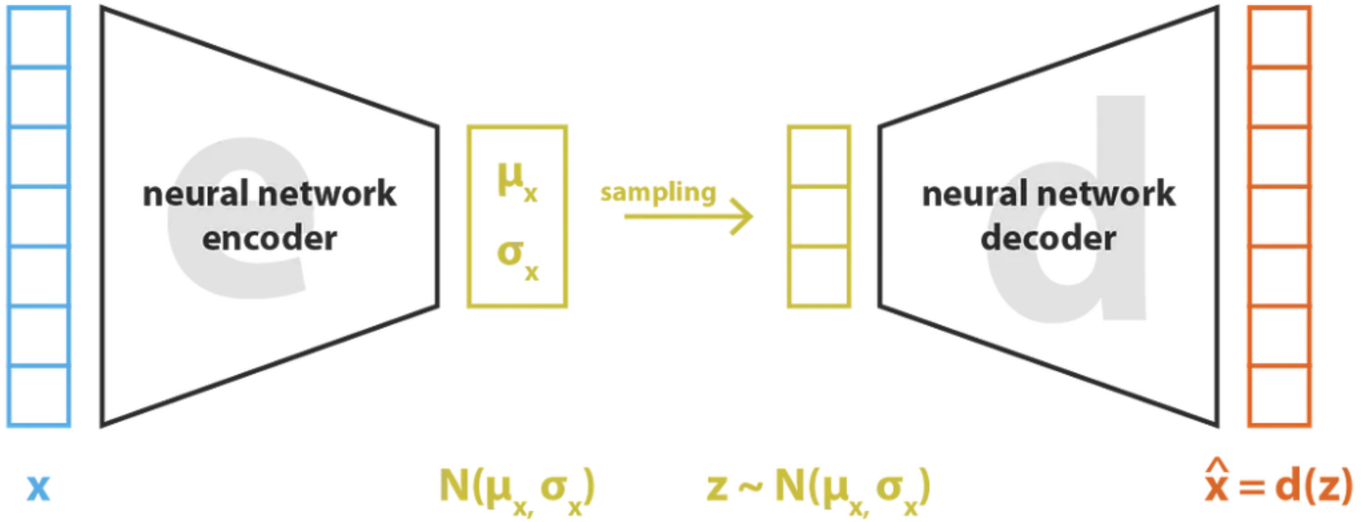
VAE encodes data as probability distribution instead of a single point



https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Autoencoder vs variational autoencoder

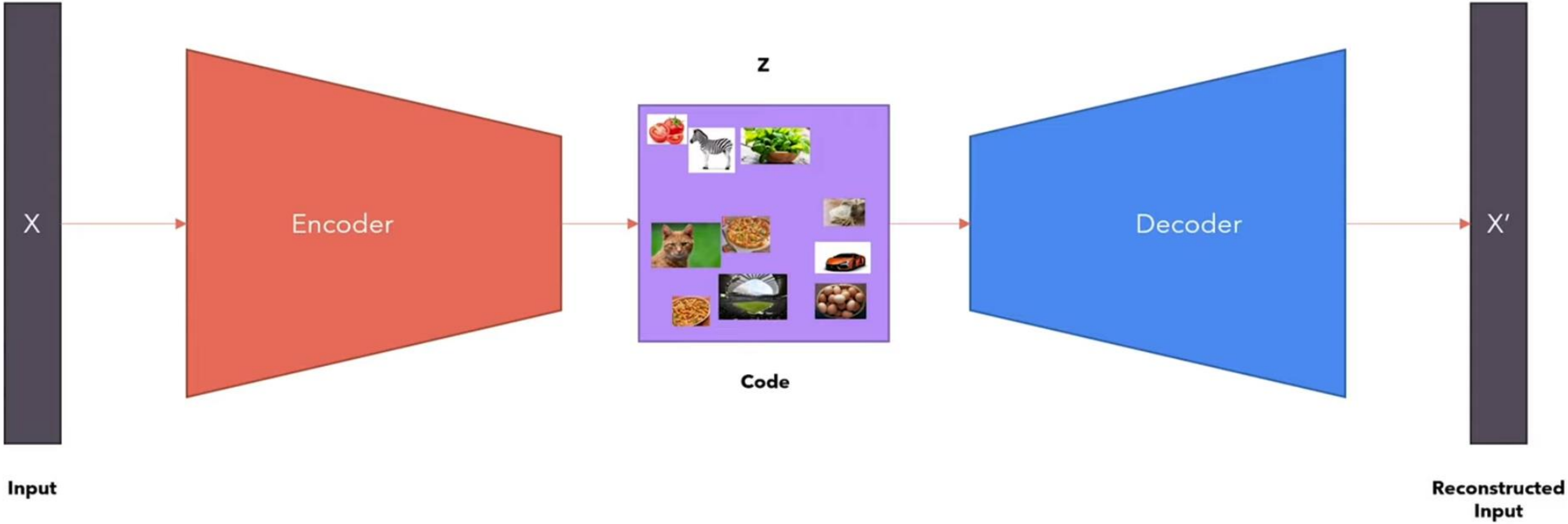
Regularization in the form of the Kullback-Leibler divergence -> this induces better organization in the latent space



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

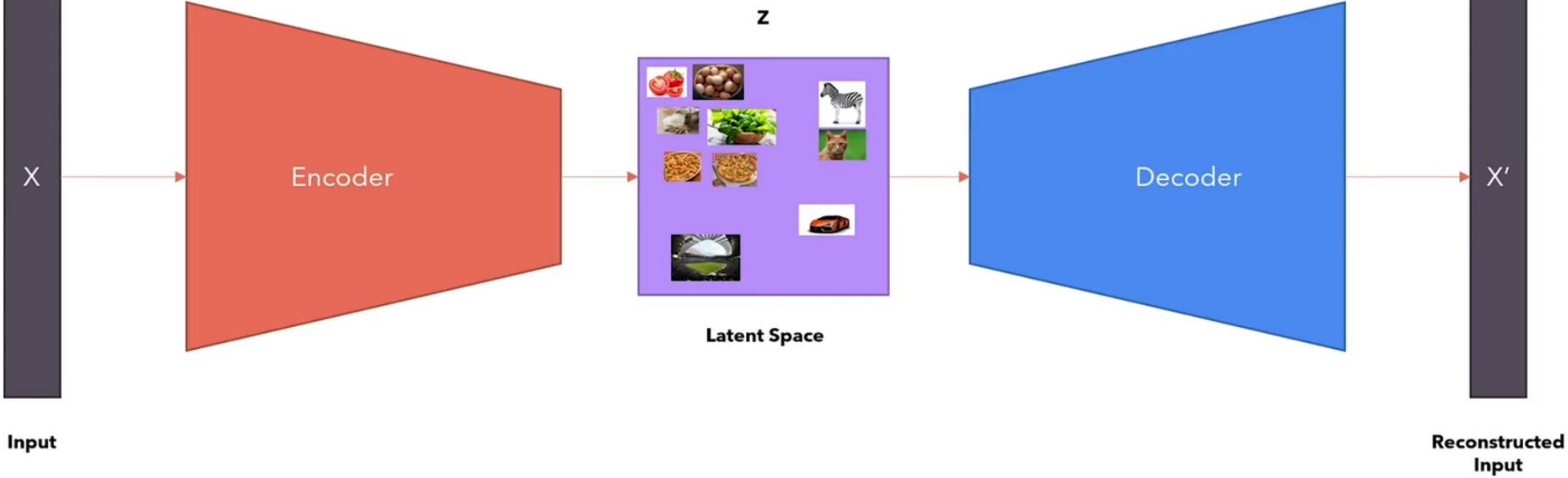
https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Autoencoder



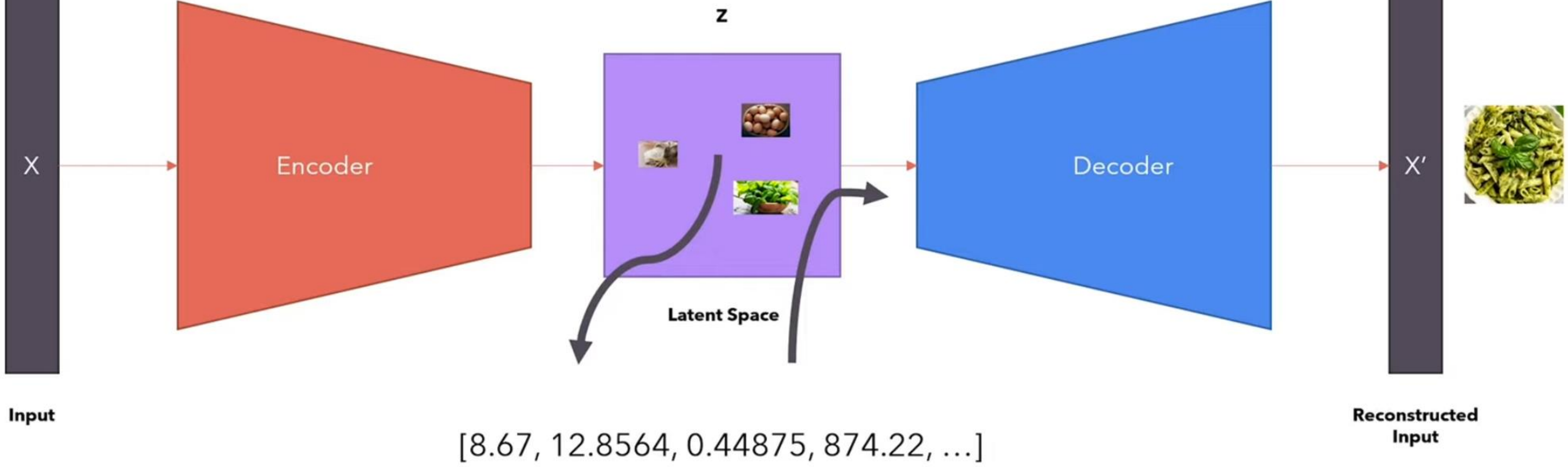
https://www.youtube.com/watch?v=iwEzwTTalbg

Variational Autoencoder



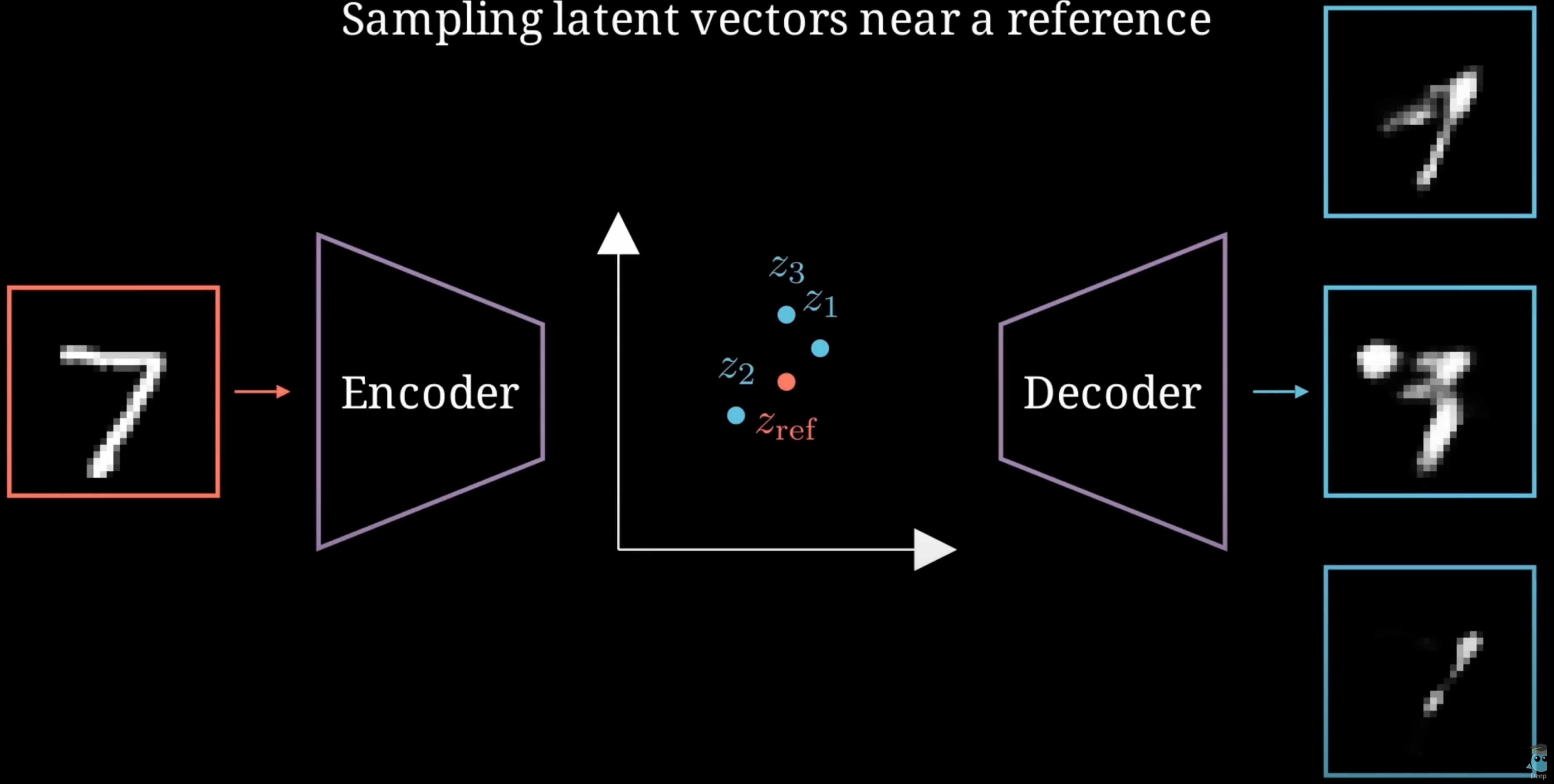
https://www.youtube.com/watch?v=iwEzwTTalbg

Sampling the latent space

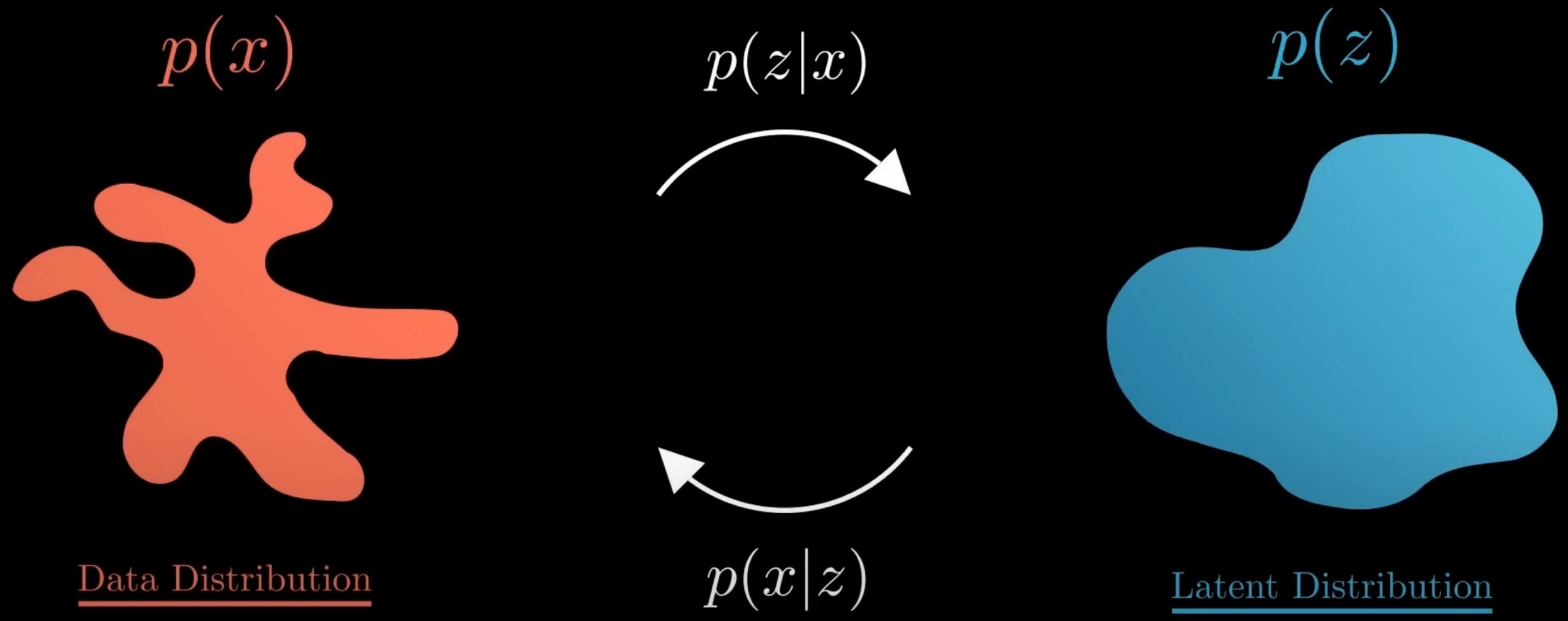


https://www.youtube.com/watch?v=iwEzwTTalbg

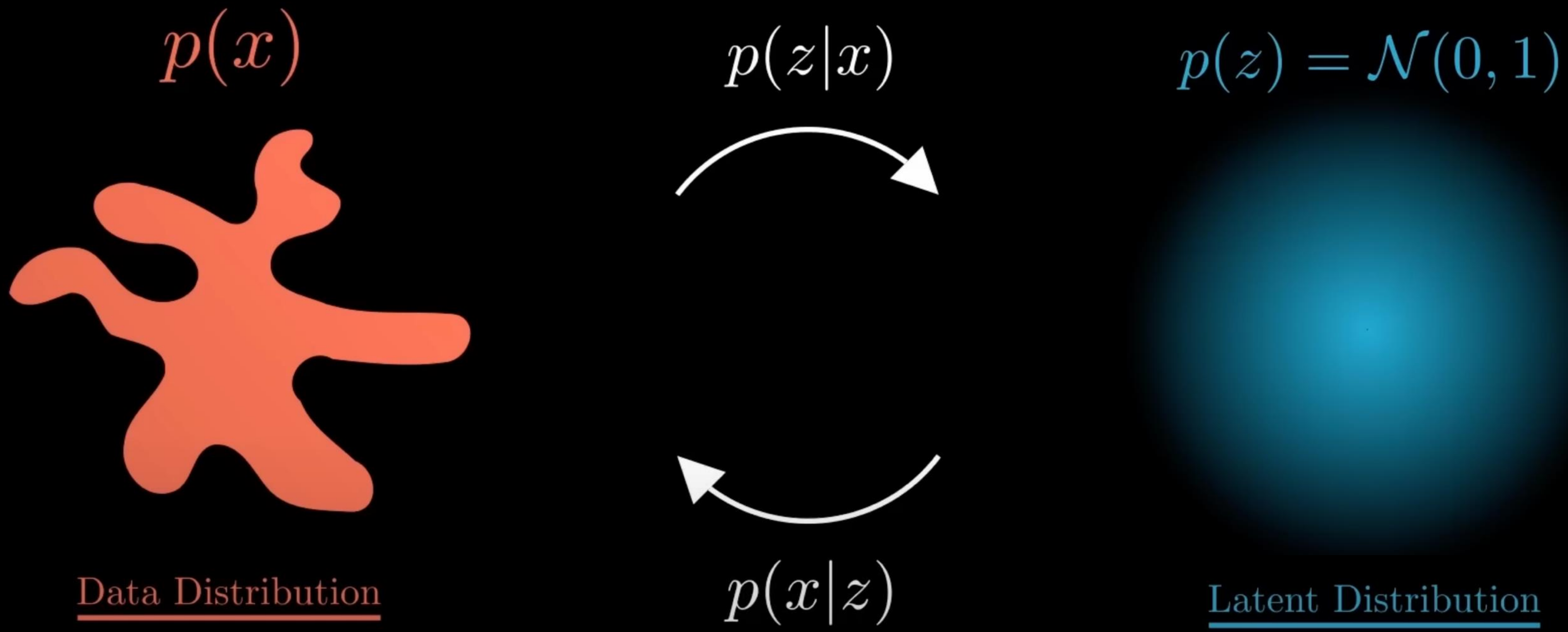
Sampling latent vectors near a reference



https://www.youtube.com/watch?v=qJeaCHQ1k2w



https://www.youtube.com/watch?v=qJeaCHQ1k2w



https://www.youtube.com/watch?v=qJeaCHQ1k2w

Loss function

L2/MSE

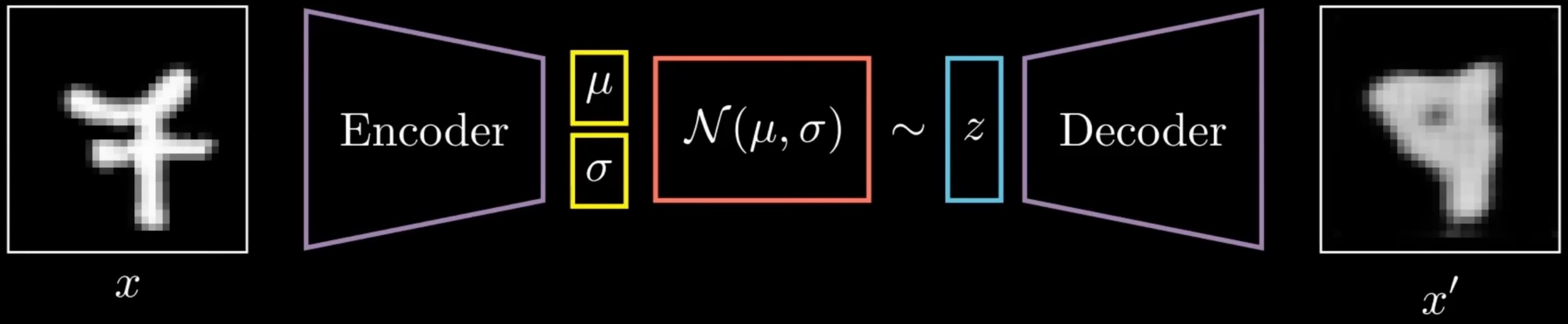
MSE between input molecule and regenerated molecule from latent space

$$\mathcal{L}(x) = \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] - \mathrm{KL}(q(z|x) \mid p(z))$$

Kullback-Leibler divergence

“regularization”, generalization, how organized the latent space is

https://www.youtube.com/watch?v=qJeaCHQ1k2w



$$\mathcal{L} = \boxed{\mathcal{L}_{KL}(\mathcal{N}(\mu, \sigma) \mid \mathcal{N}(0, 1))} + \mathcal{L}_2(x, x')$$

$$\mathcal{L}_{KL} = -\frac{1}{2}(1 + \log(\sigma^2) - \mu^2 - \sigma^2)$$

https://www.youtube.com/watch?v=qJeaCHQ1k2w