Practice 09

Molecular generative model (1)

Seongok Ryu

Department of Chemistry, KAIST

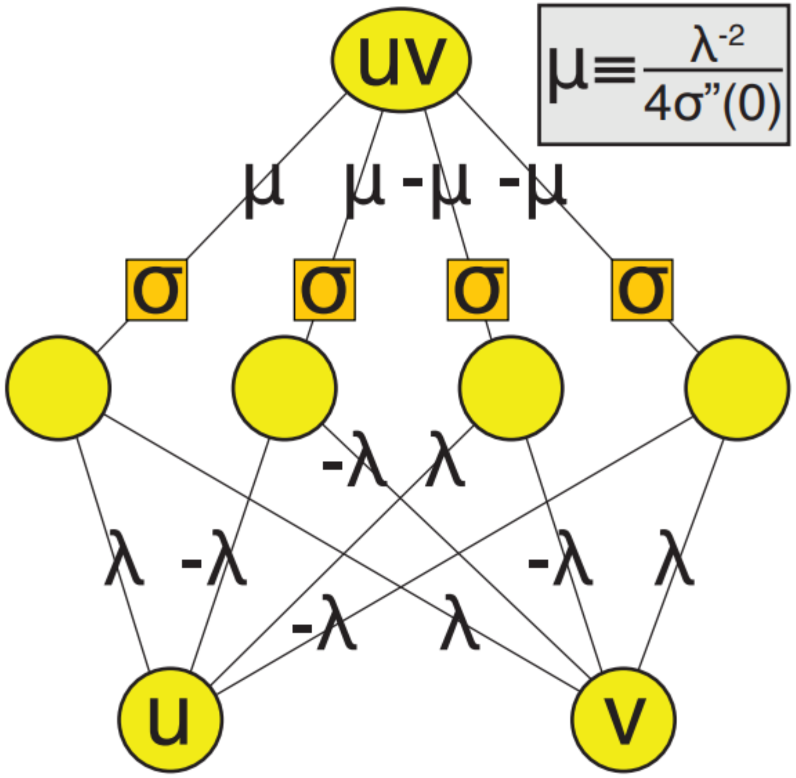
Contents

- Why does deep learning work so well?
- Autoencoder
- Variational autoencoder
- VAE for molecule

Why does deep learning work so well?

MLP can work as a “multiplication gate”, with “non-linearity”.

Continuous multiplication gate:



Output

$$m(u,v) = \mu \cdot \{ \sigma[\lambda(u+v)] + \sigma[-\lambda(u+v)] - \sigma[\lambda(u-v)] - \sigma[\lambda(-u+v)] \}$$

The nonlinear activation function can be expanded as

$$\sigma(u) \approx \sigma_0 + \sigma_1 \cdot u + \sigma_2 \cdot \frac{u^2}{2}$$

Substituting it into the output function to obtain

$$m(u,v) = \mu \cdot (\sigma_1 \lambda \cdot [(u+v) + (-u-v) - (u-v) - (-u+v)] + \sigma_2 \lambda^2 \cdot [(u+v)^2 + (-u-v)^2 - (u-v)^2 - (-u+v)^2)$$

$$= 4\mu \cdot \sigma_2 \cdot \lambda^2 \cdot uv$$

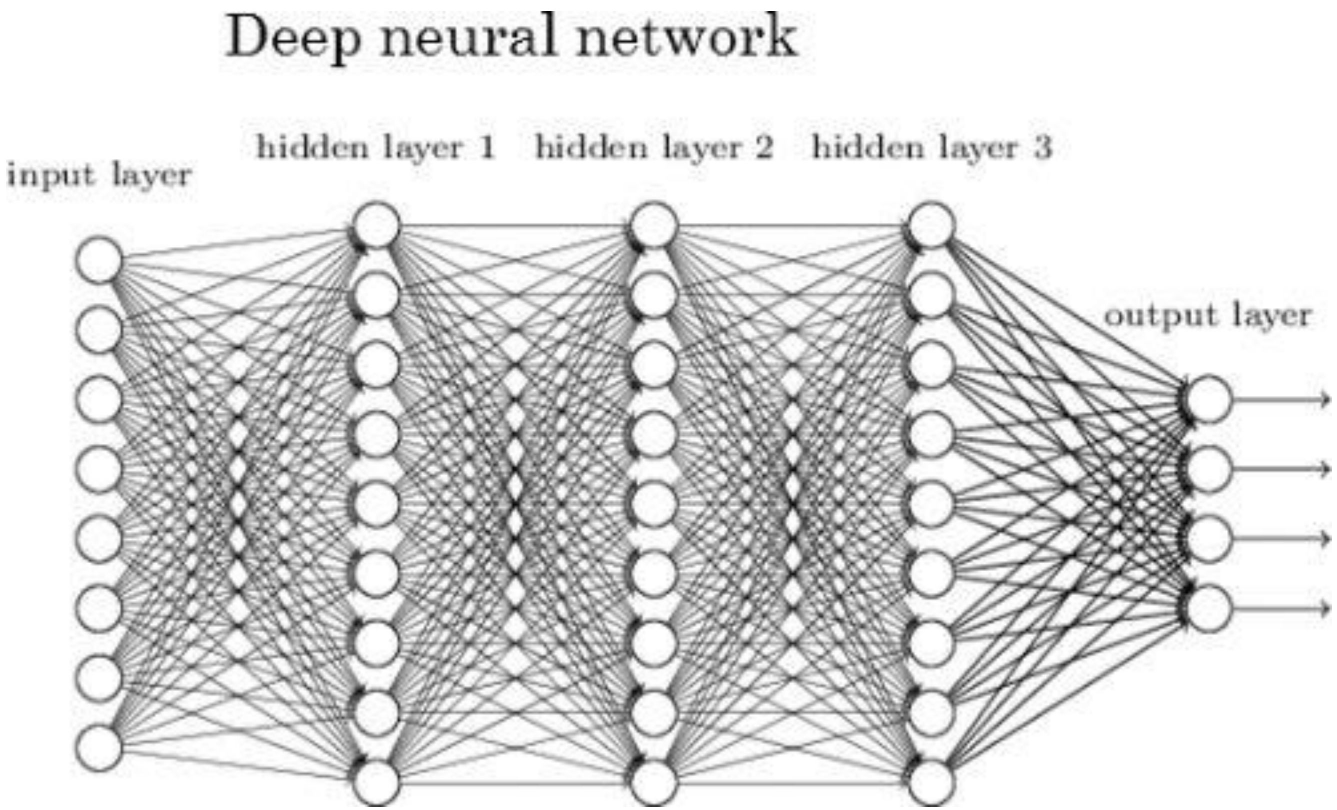
The result is not a linear combination of inputs but multiplication (or nonlinear)!

KAIST

Lin, Henry W., Max Tegmark, and David Rolnick. "Why does deep and cheap learning work so well?." *Journal of Statistical Physics* 168.6 (2017): 1223-1247.

Why does deep learning work so well?

However, why neural networks need to be deep?



KAIST

Why does deep learning work so well?

Sufficient statistics

Given $P(y|x)$, a *sufficient statistic* $T(x)$ is defined by the equation $P(y|x) = P(y|T(x))$ and has played an important role in statistics for almost a century. All the information about y contained in x is contained in the sufficient statistics.

A *minimal sufficient statistic* is some sufficient statistic T_* which is a sufficient statistic for all other sufficient statistics. This means that if $T(y)$ is sufficient, then there exists some function f such that $T_*(y) = f(T(y))$. T_* can be thought of as an information distiller, optimally compressing the data so as to retain all information relevant to determining y and discarding all irrelevant information.

KAIST

Lin, Henry W., Max Tegmark, and David Rolnick. "Why does deep and cheap learning work so well?." *Journal of Statistical Physics* 168.6 (2017): 1223-1247.

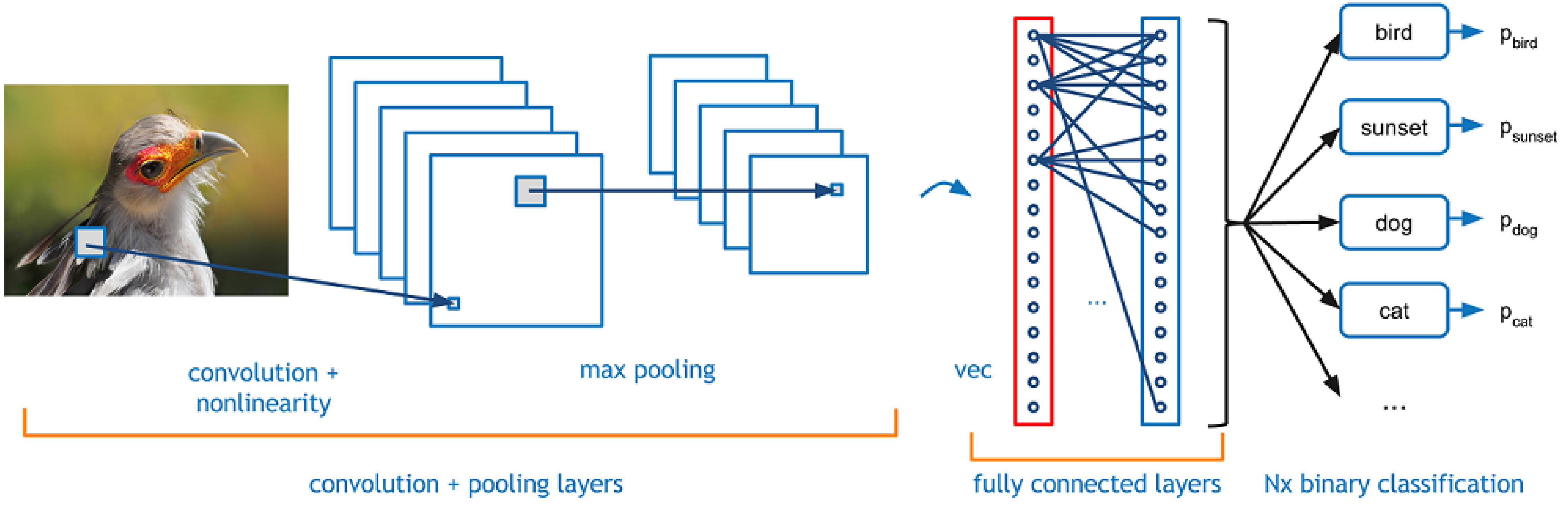
Why does deep learning work so well?

Sufficient statistics

Information distillation (Feature extraction)

Classification based on the feature

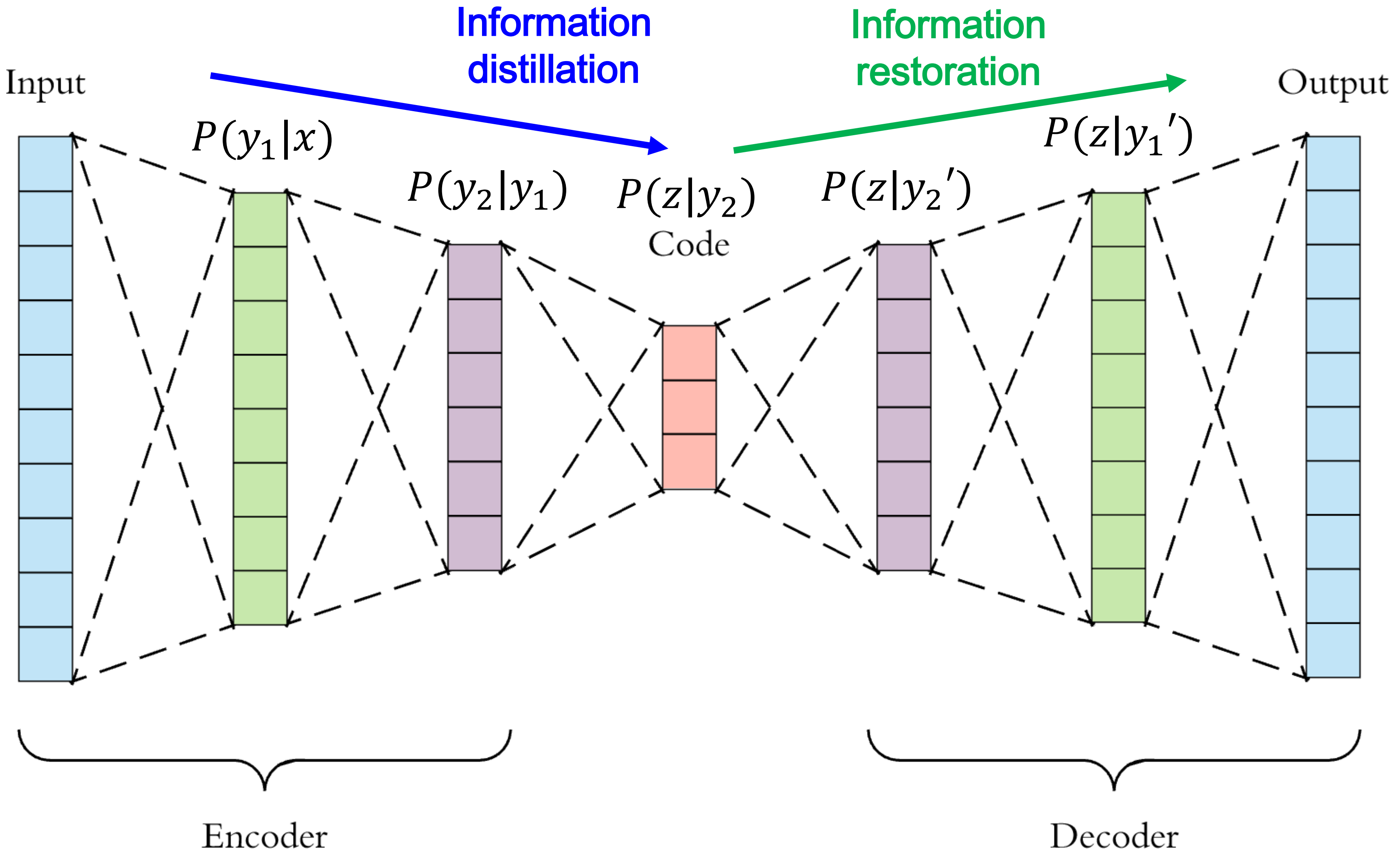
$$x \longrightarrow P(y_1|x) \longrightarrow P(y_2|y_1) \longrightarrow P(y_3|y_2) \longrightarrow P(c = bird|y_3)$$



KAIST

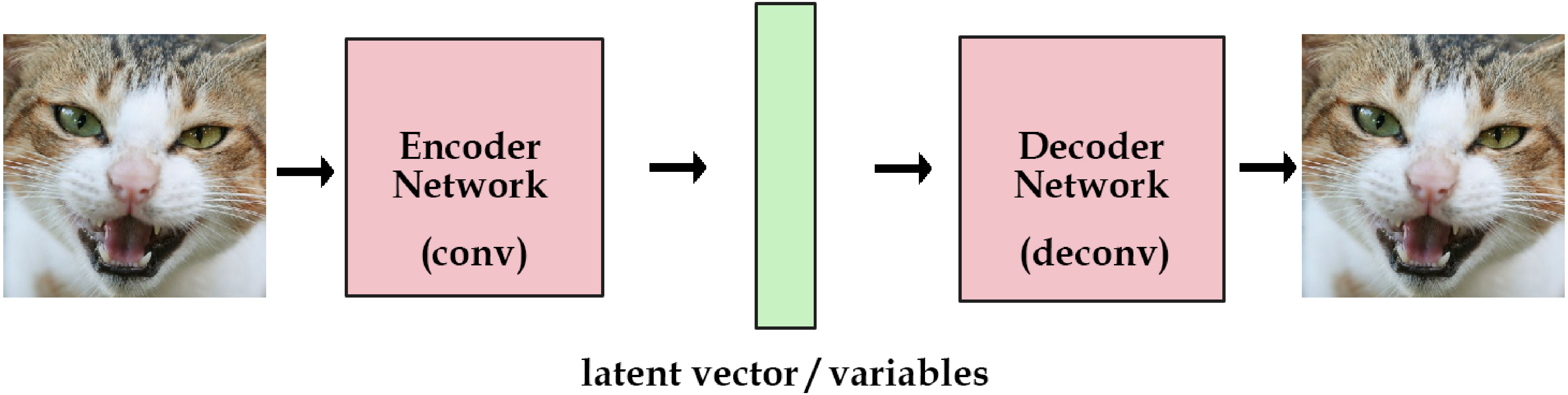
Autoencoder

Can we extract the feature without given labels?



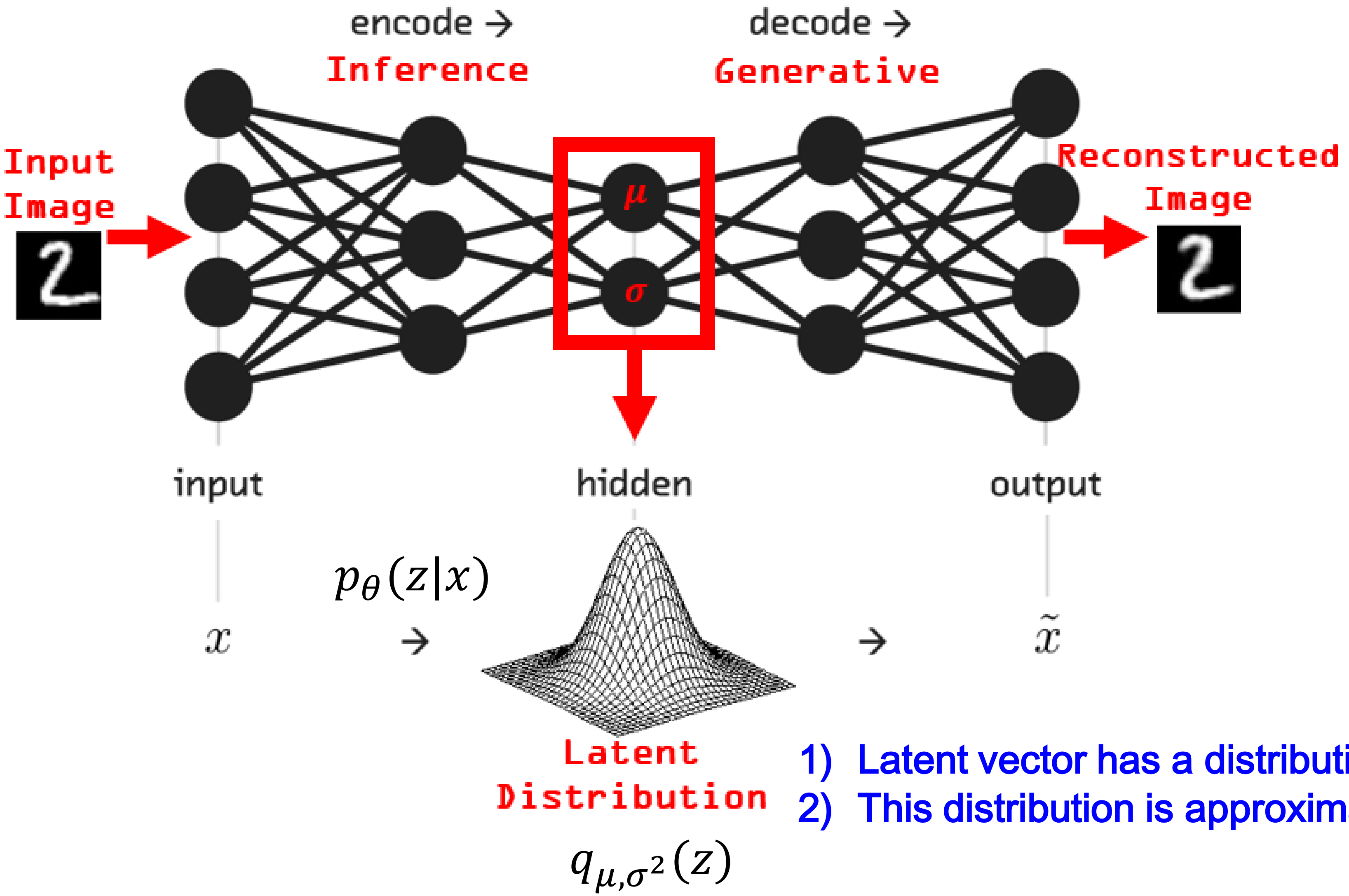
Autoencoder

Can we extract the feature without given labels?



: contains the essential information of given input

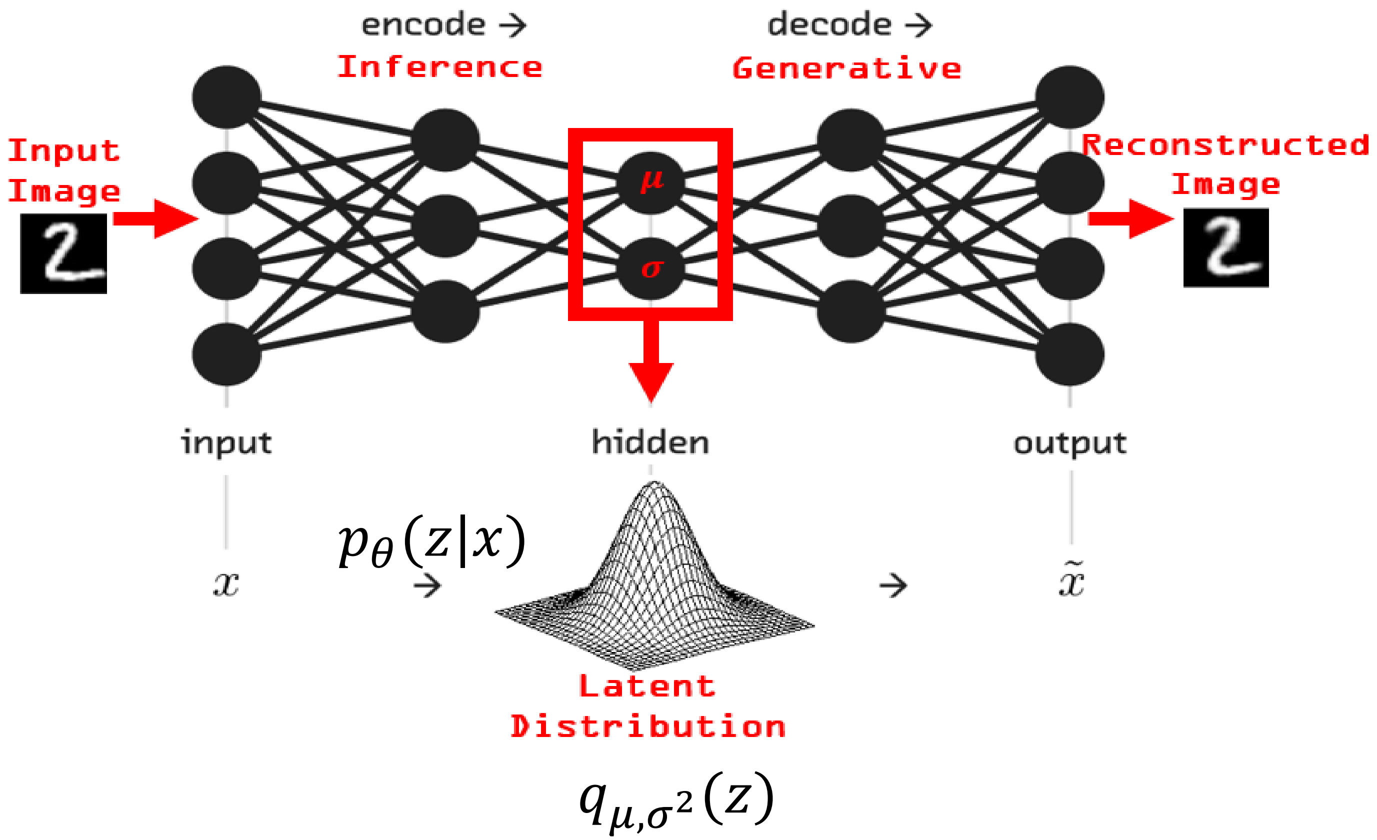
Variational autoencoder



1) Latent vector has a distribution
2) This distribution is approximated with $\mathcal{N}(\mu, \sigma^2)$

KAIST

Variational autoencoder

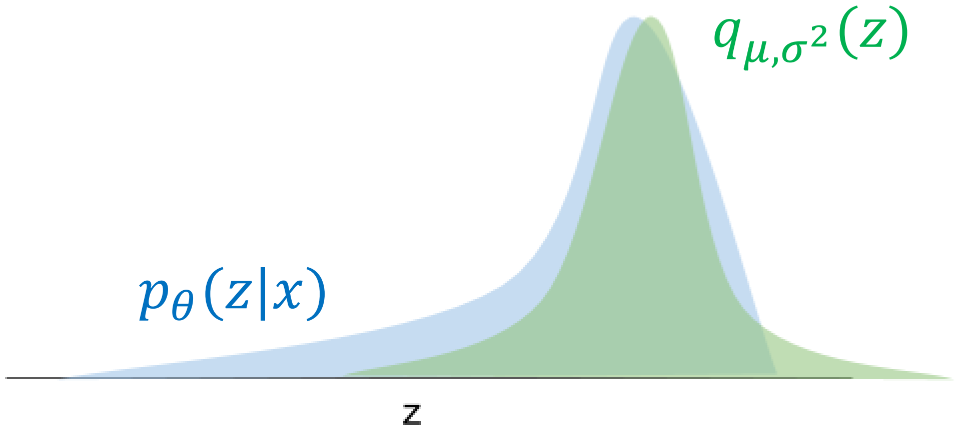


Our minimization objective

1. Reconstruction of the image
$$\mathcal{L}_{reconst} = \|x - \tilde{x}\|^2$$

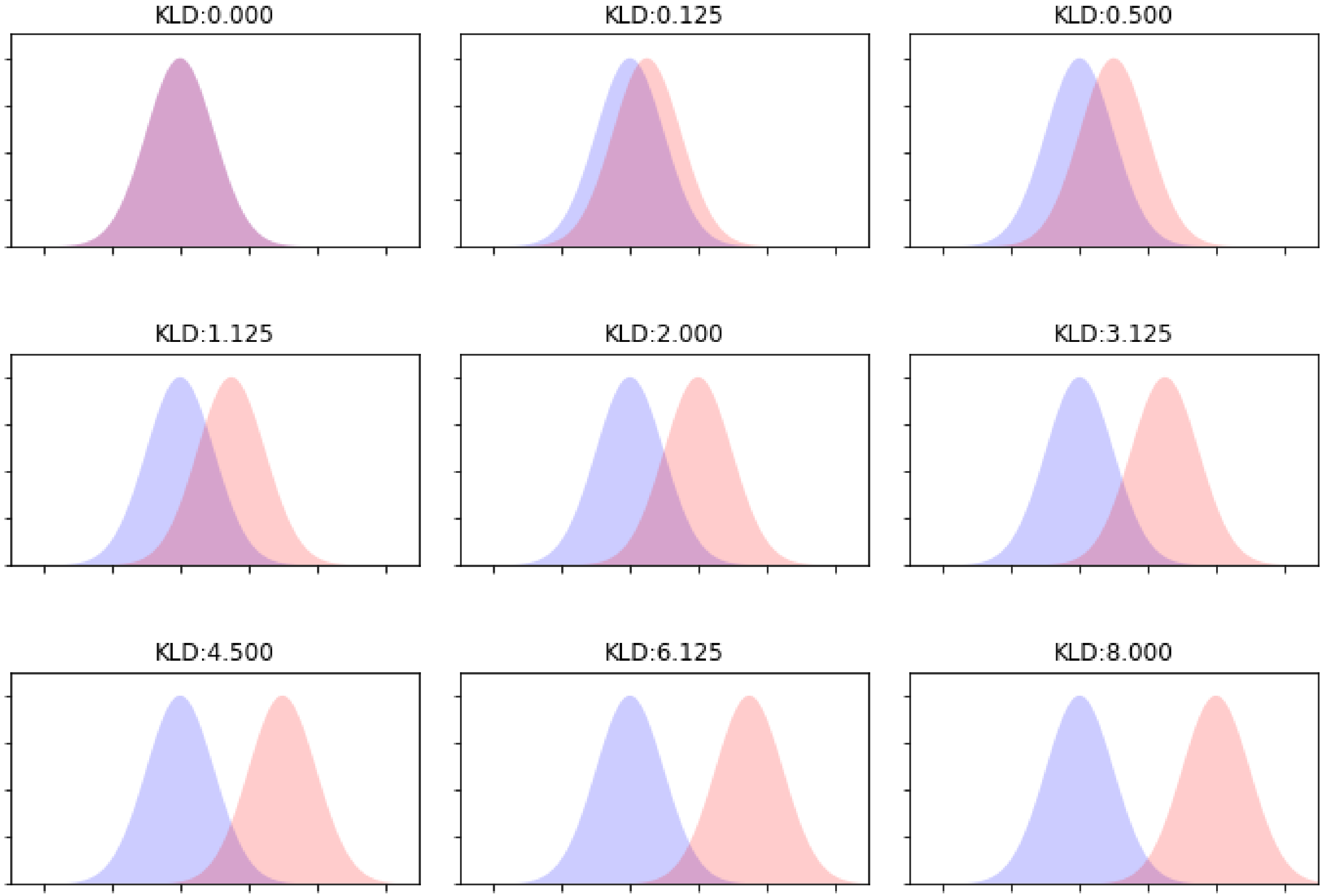
2. Variational inference of the latent distribution
$$\mathcal{L}_{VI} = \mathrm{KL}\big(q_{\mu,\sigma^2}(z)||p_\theta(z|x)\big)$$

3. Total loss
$$\mathcal{L}_{total} = \mathcal{L}_{reconstr} + \mathcal{L}_{VI}$$



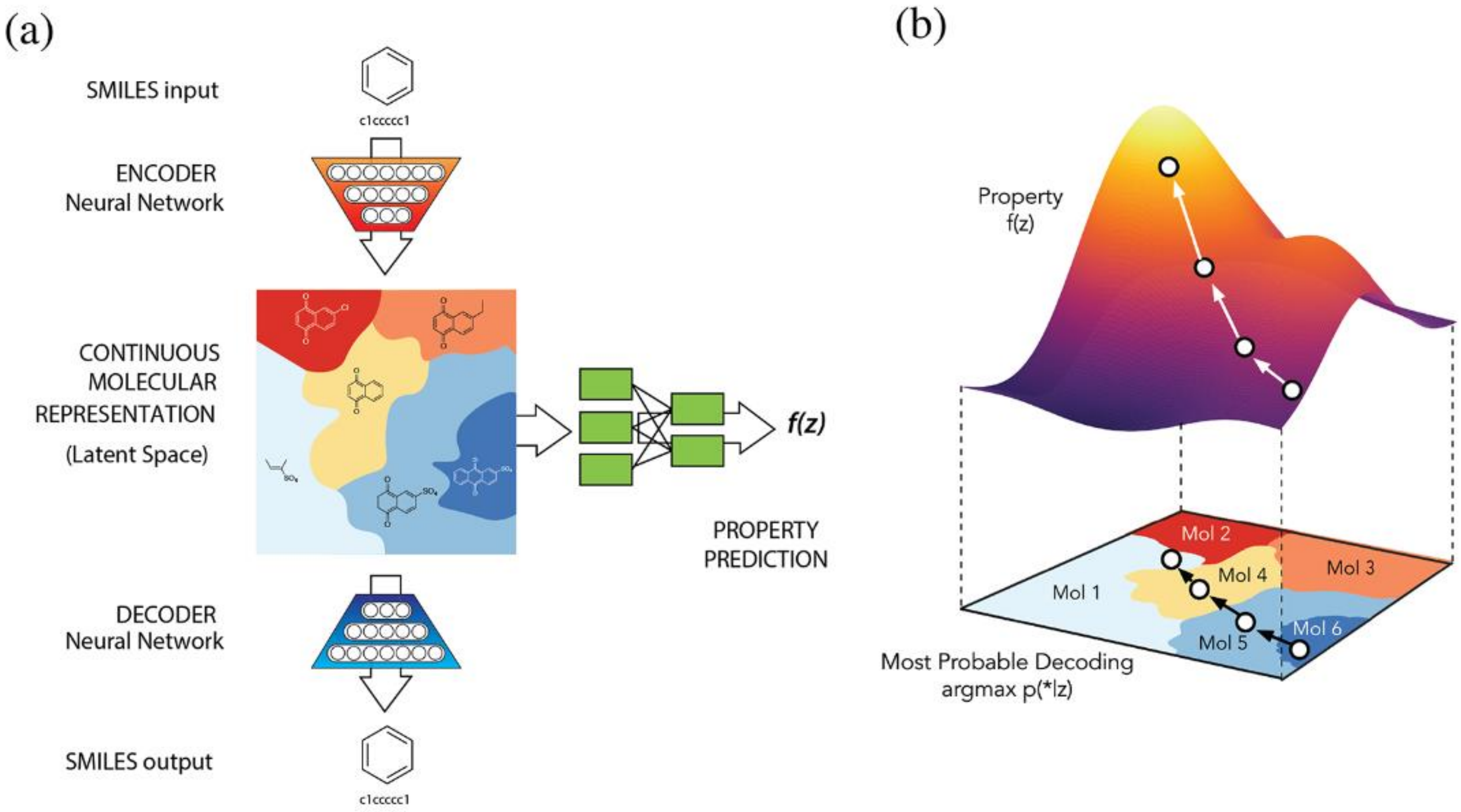
Variational autoencoder

$\mathrm{KL}\left(q_{\mu,\sigma^2}(z)||p_\theta(z|x)\right)$



Variational autoencoder

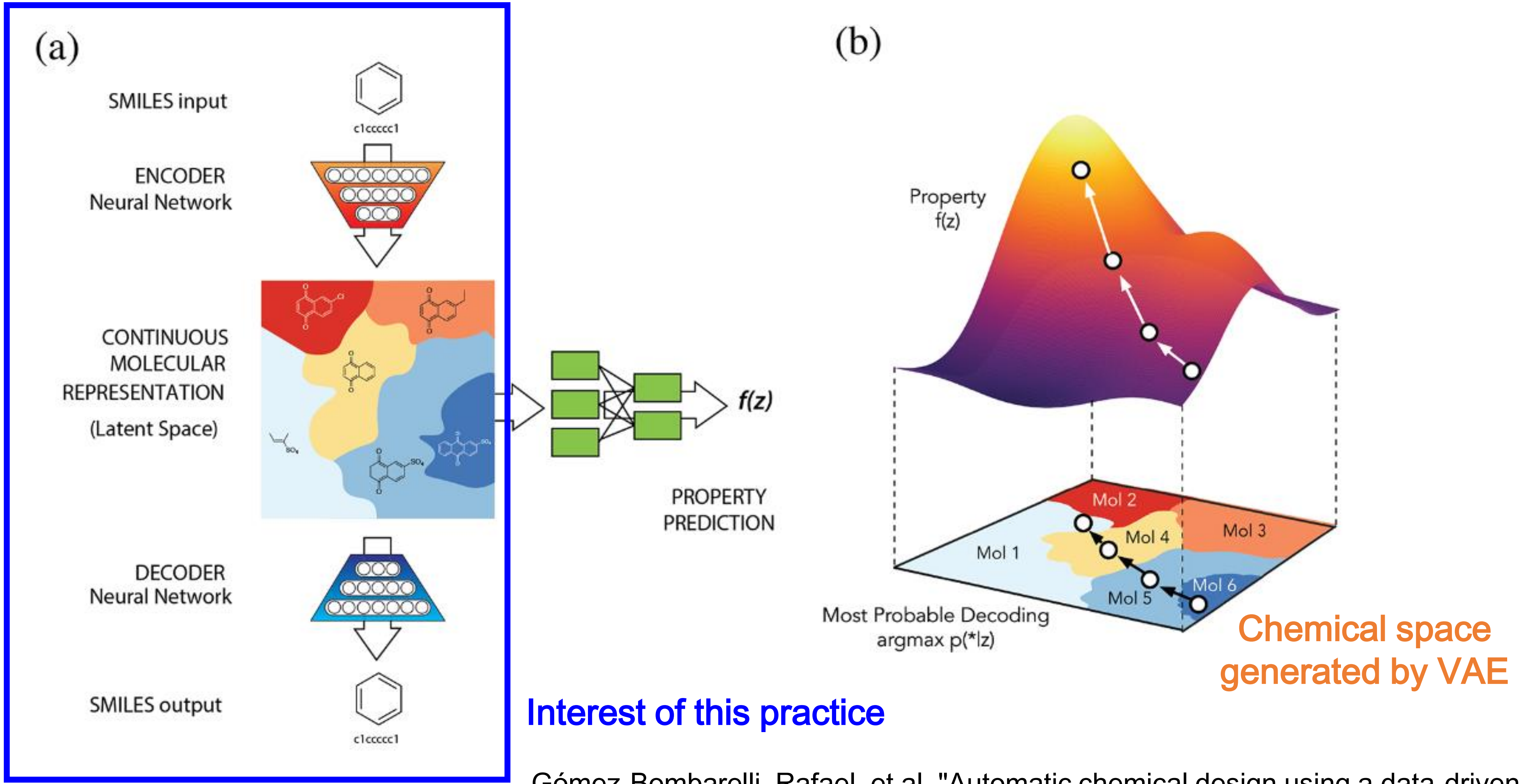
ChemicalVAE



Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *ACS central science* 4.2 (2018): 268-276.

Variational autoencoder

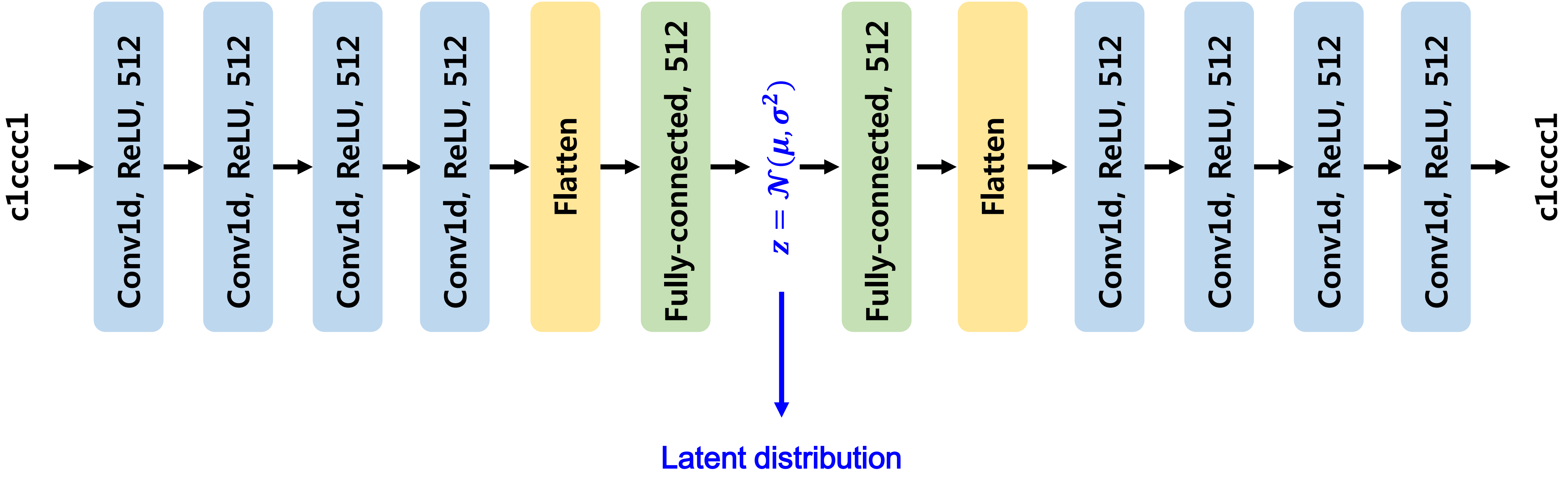
ChemicalVAE



Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *ACS central science* 4.2 (2018): 268-276.

Variational autoencoder

ChemicalVAE



Variational autoencoder

ChemicalVAE

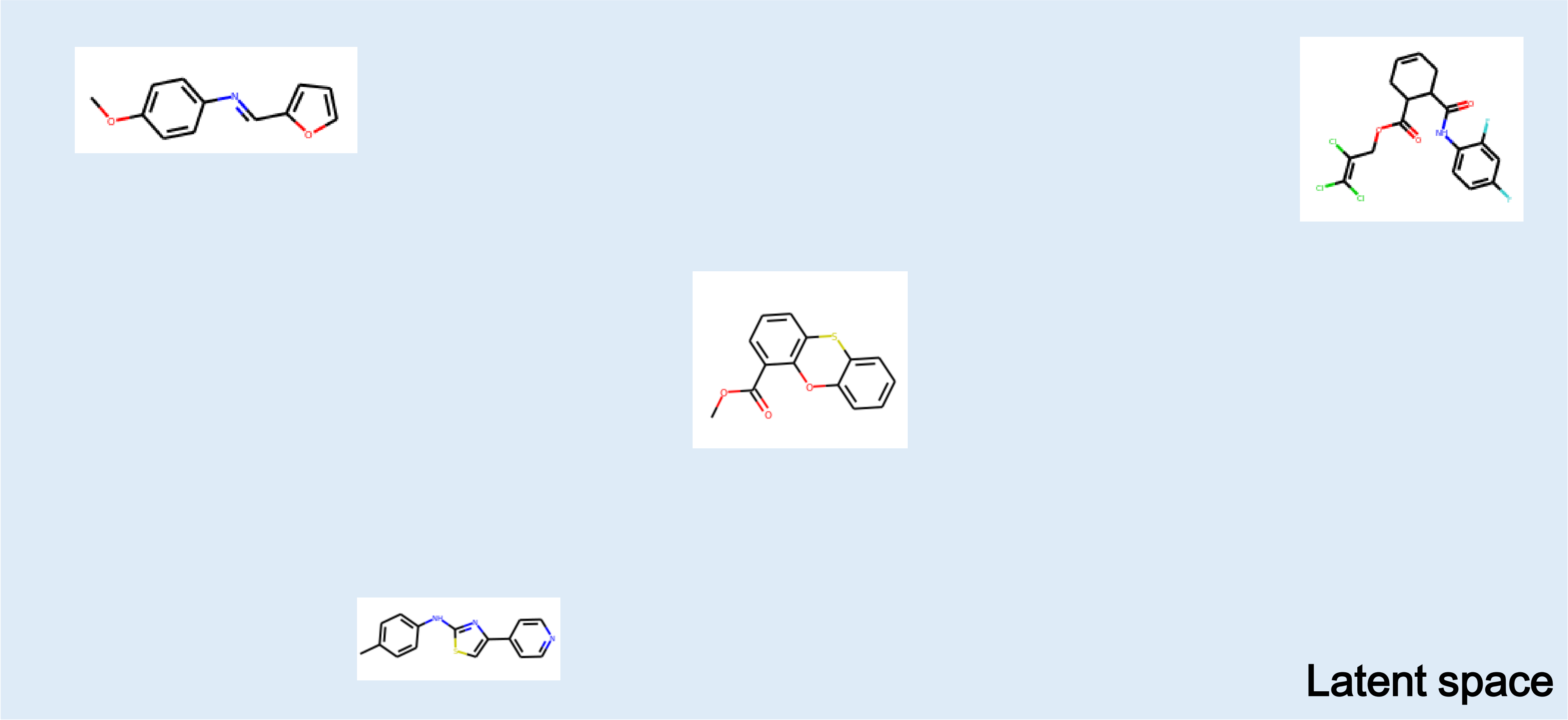
Input SMILES	Reconstructed SMILES
CC(C)(C)c1ccccc1OCC(=O)Nc1cccc(C(=O)[O-])c1	CC(C)(C)c1ccccc1OCC(=O)Nc1cccc(C(=O)[O-])c1
COc1ccc(-c2nnc(SCC(=O)OC3CCCC3)n2C)cc1	COc1ccc(-c2nnc(SCC(=O)OC3CCCC3)n2C)cc1
C=C(C)C(=O)OCC**S**C	C=C(C)C(=O)OCC**C**C
C**N**c1oc(-c2ccc(Cl)cc2)nc1S(=O)(=O)c1ccc(C)cc1	C**O**c1oc(-c2ccc(Cl)cc2)nc1S(=O)(=O)c1ccc(C)cc1
CC(=O)N(c1ccccc1C)S(=O)(=O)c1ccc(Cl)cc1	CC(=O)N(c1ccccc1C)S(=O)(=O)c1ccc(Cl)cc1
O=C(COC(=O)c1cccc(S(=O)(=O)N2CCCCCC2)c1)Nc1ccc(F)cc1F	O=C(COC(=O)c1cccc(S(=O)(=O)N2CCCCCC2)c1)Nc1ccc(F)cc1F
CCOc1ccc(N2C3=[N+](CCCCC3)CC2(O)c2ccc([N+](=O)[O-])cc2)cc1	CCOc1ccc(N2C3=[N+](CCCCC3)CC2(O)c2ccc([N+](=O)[O-])cc2)cc1
CC(C)(C)c1csc(NC(=O)COc2ccc(F)cc2)n1	CC(C)(C)c1csc(NC(=O)COc2ccc(F)cc2)n1
O=C1CN(C(=O)C=Cc2ccc3c(c2)OCO3)c2ccccc2**N**1	O=C1CN(C(=O)C=Cc2ccc3c(c2)OCO3)c2ccccc21
COC(=O)c1ccccc1NC(=O)CSc1ccc(C)cc1C	COC(=O)c1ccccc1NC(=O)CSc1ccc(C)cc1C

Recunstruction accuracy (mol) = 70%, accuracy(character) > 99%

KAIST

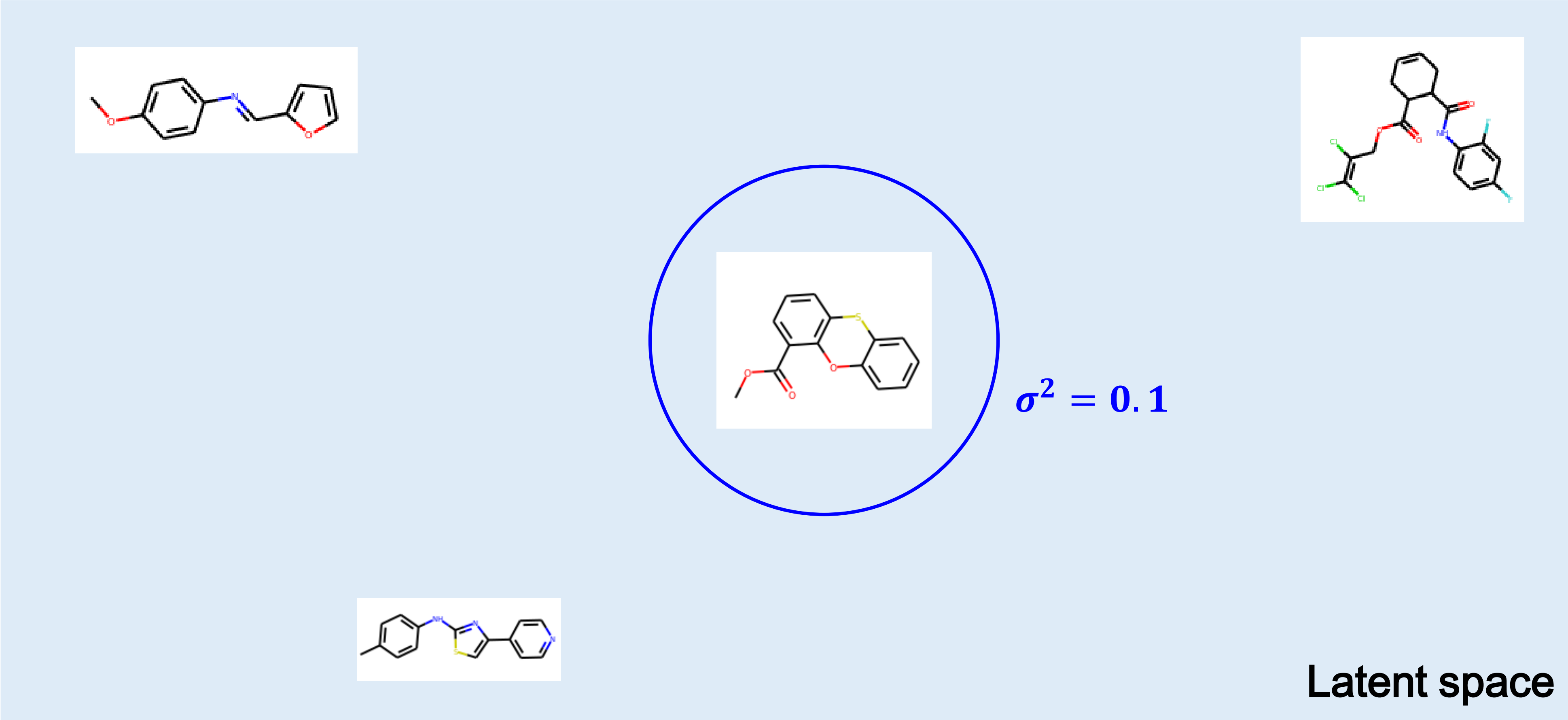
Variational autoencoder

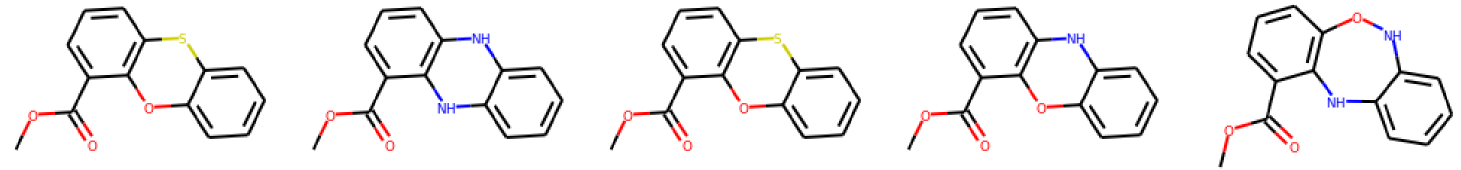
ChemicalVAE



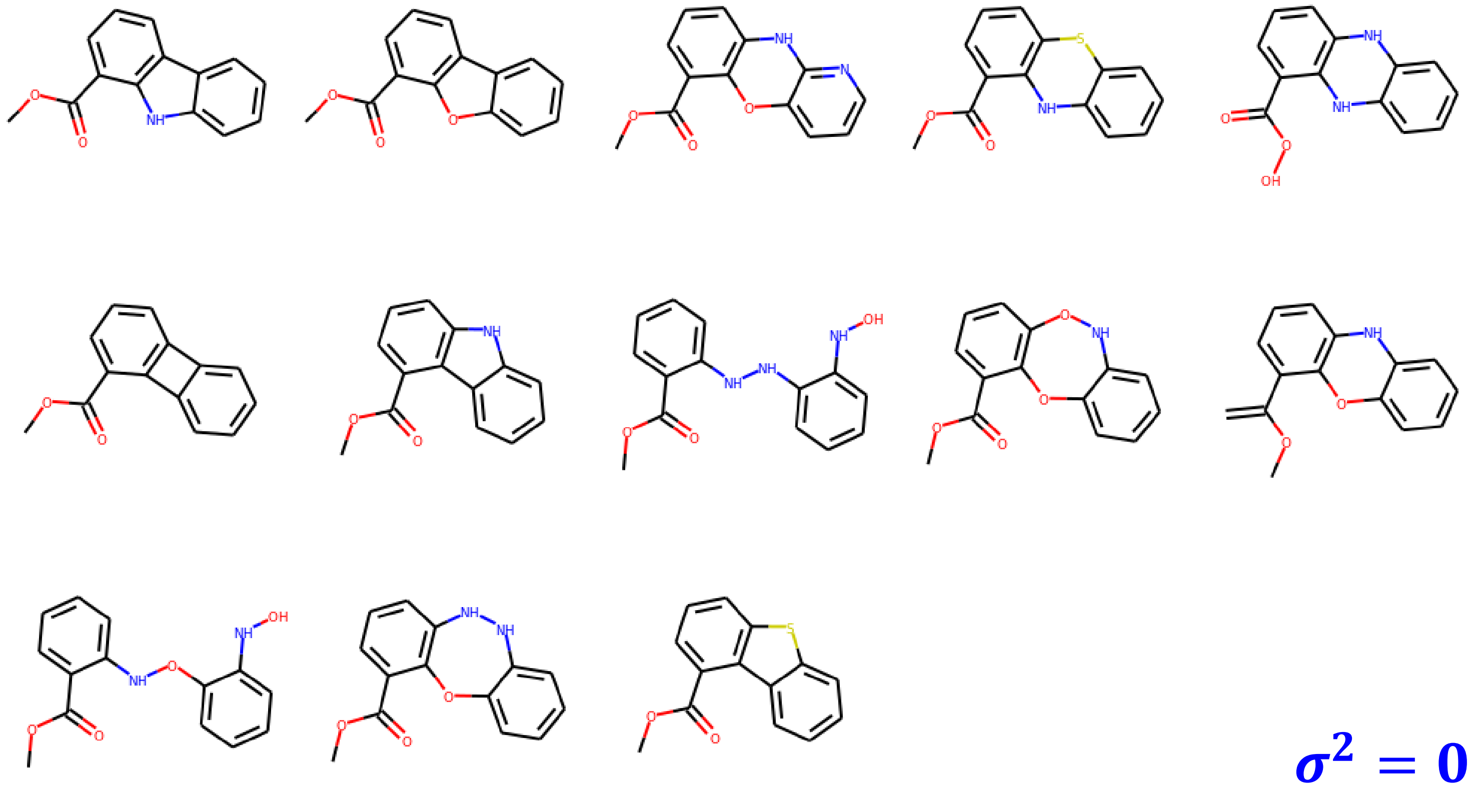
Variational autoencoder

Searching latent space

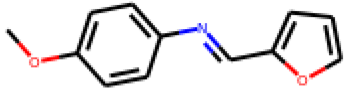




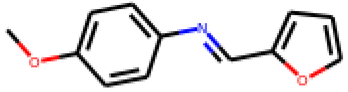
Query

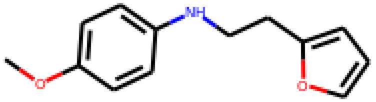


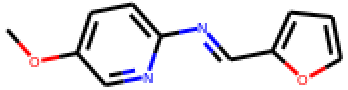
$\sigma^2 = 0.1$

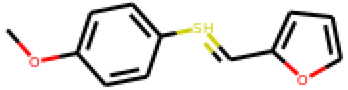


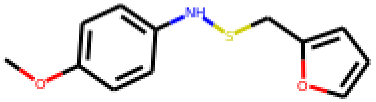
Query







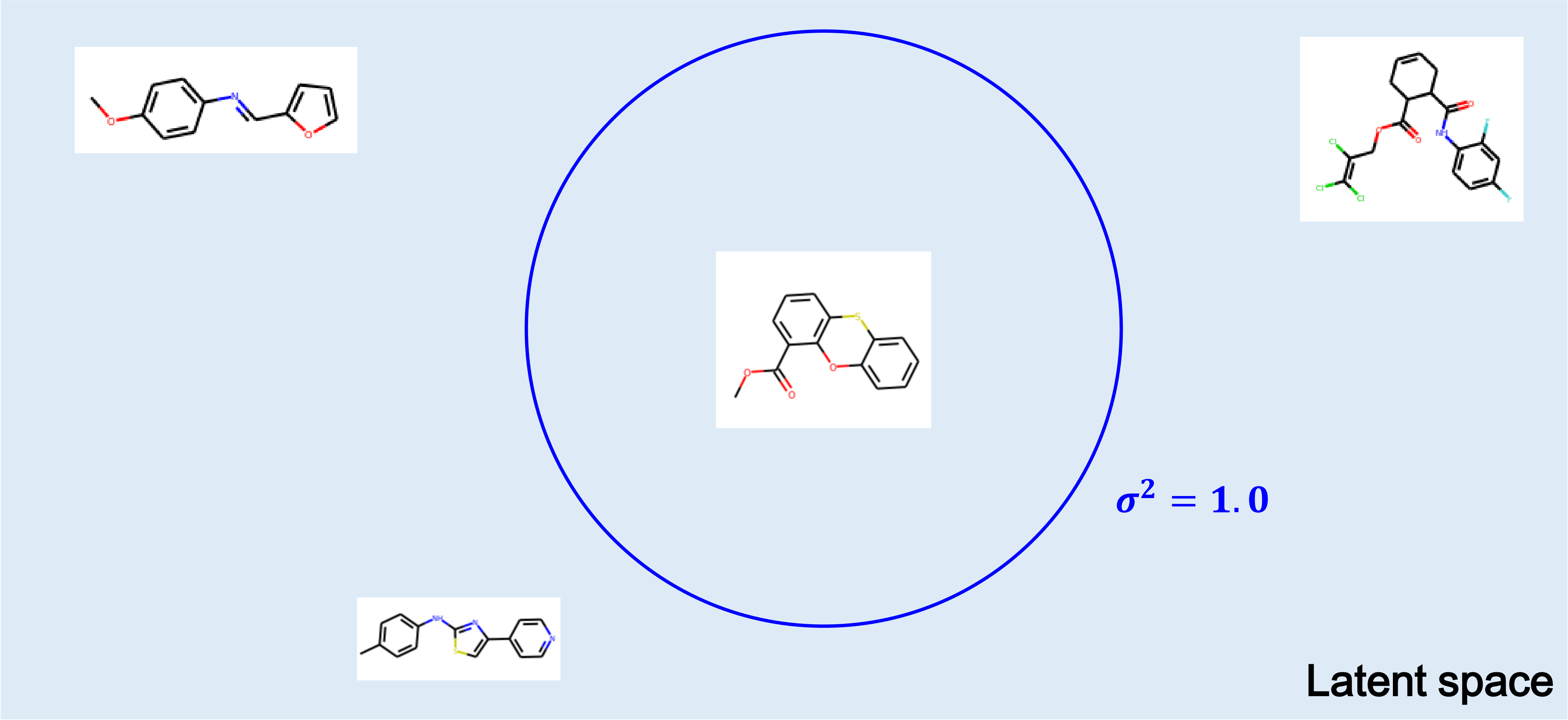


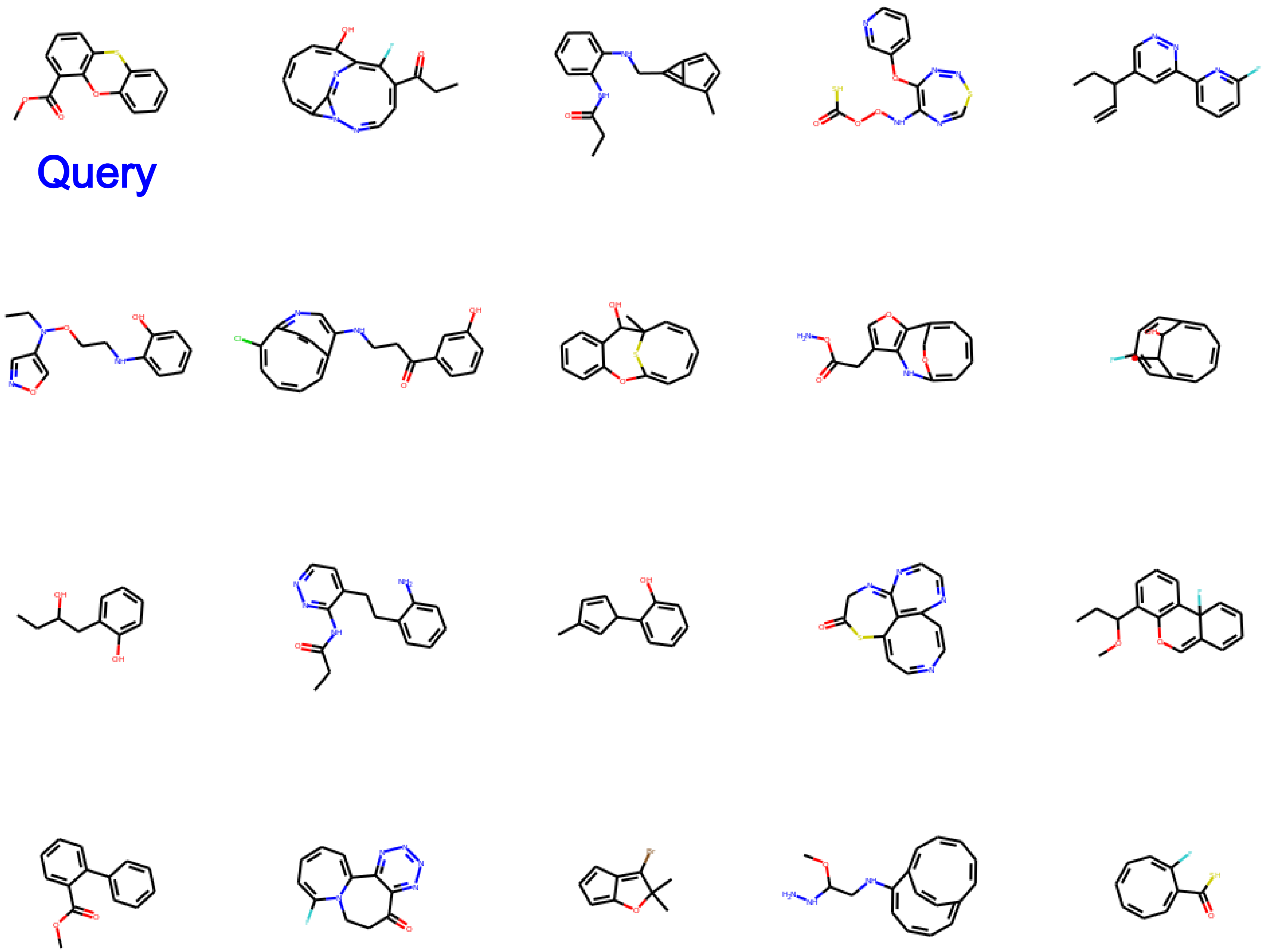


$$\sigma^2 = 0.1$$

Variational autoencoder

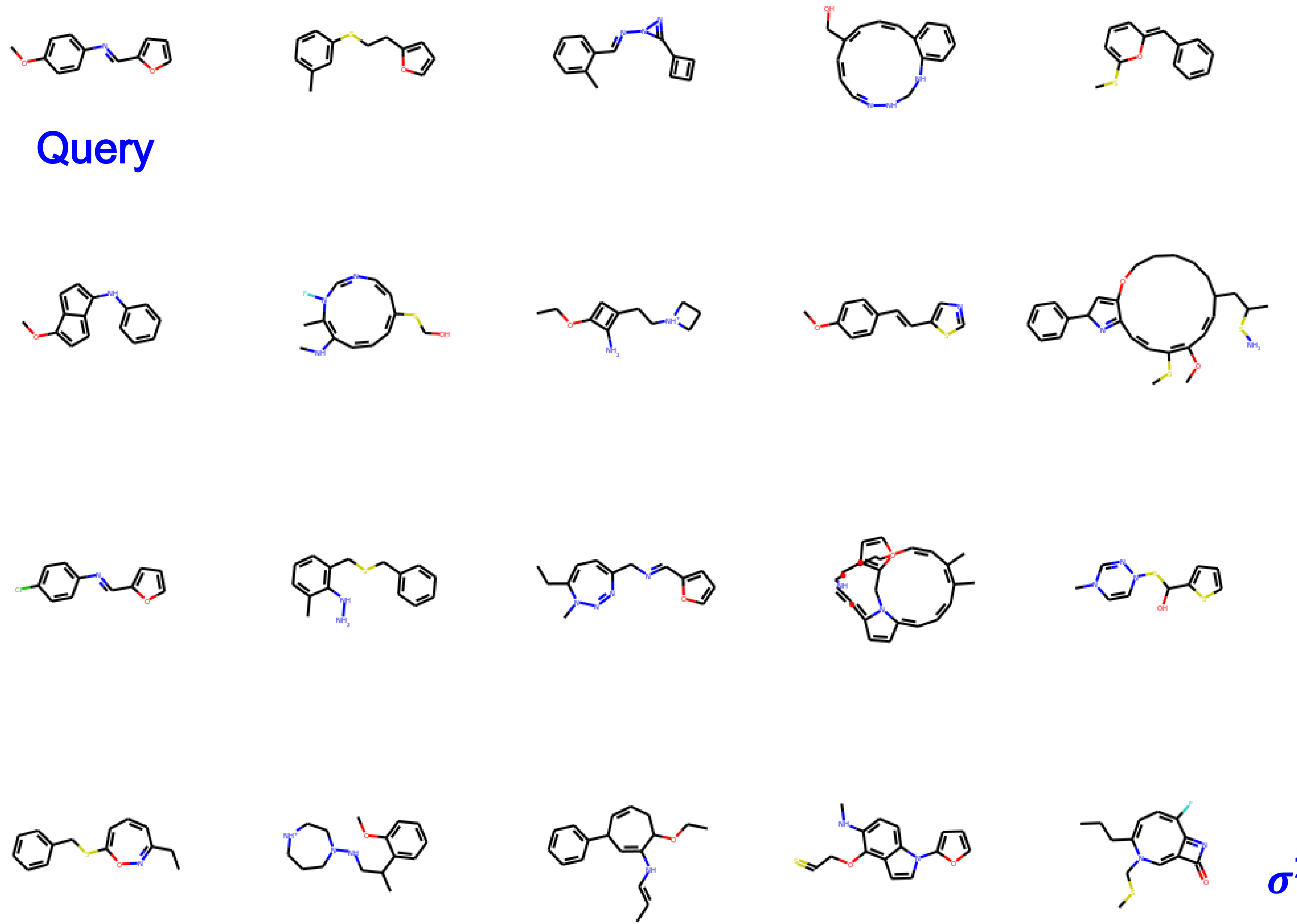
Searching latent space





Query

$\sigma^2 = 1.0$



$\sigma^2 = 1.0$

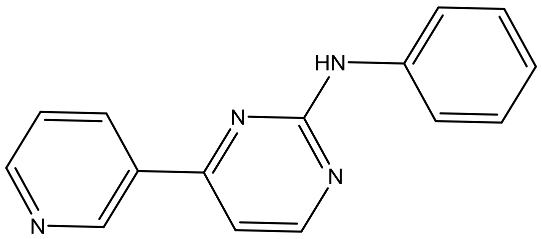
Questions

1. What is the meaning of latent vectors?
   → It may contains the “structural information”, whose original representation is SMILES.

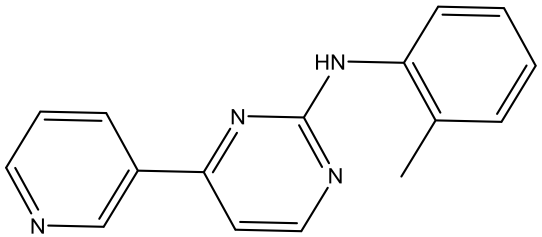
Questions

1. What is the meaning of latent vectors?
   → It may contains the “structural information”, whose original representation is SMILES.

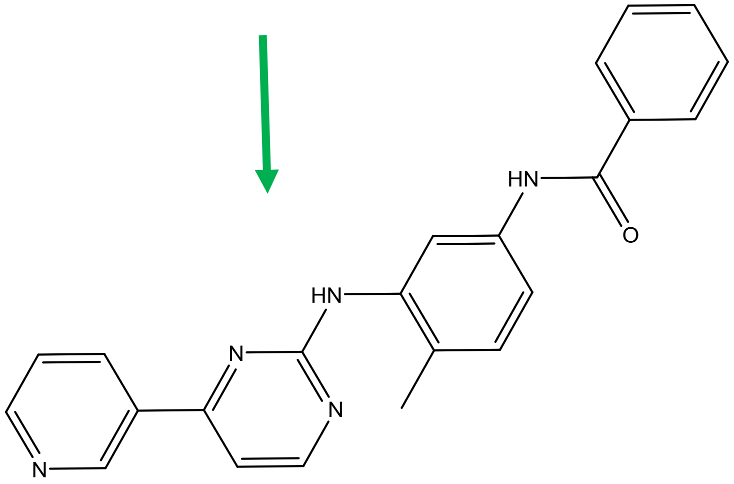
2. Is using SMILES relevant?



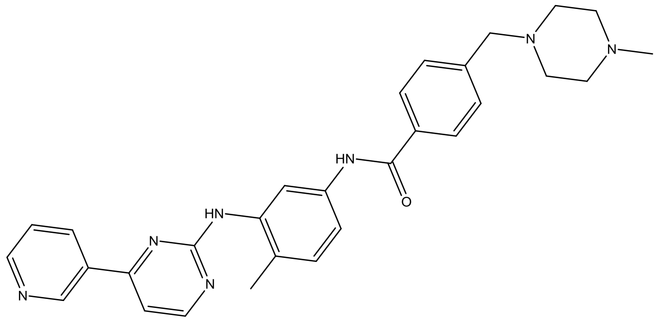
C1(C2=NC(NC3=CC=CC=C3)=NC=C2)=CN=CC=C1



CC(C=CC=C1)=C1NC2=NC=CC(C3=CN=CC=C3)=N2



CC(C=CC(NC(C1=CC=CC=C1)=O)=C2)=C2NC3=NC=CC(C4=CN=CC=C4)=N3



Gleevec (Imatinib)

CC(C=CC(NC(C1=CC=C(CN2CCN(C)CC2)C=C1)=O)
=C3)=C3NC4=NC=CC(C5=CN=CC=C5)=N4

Questions

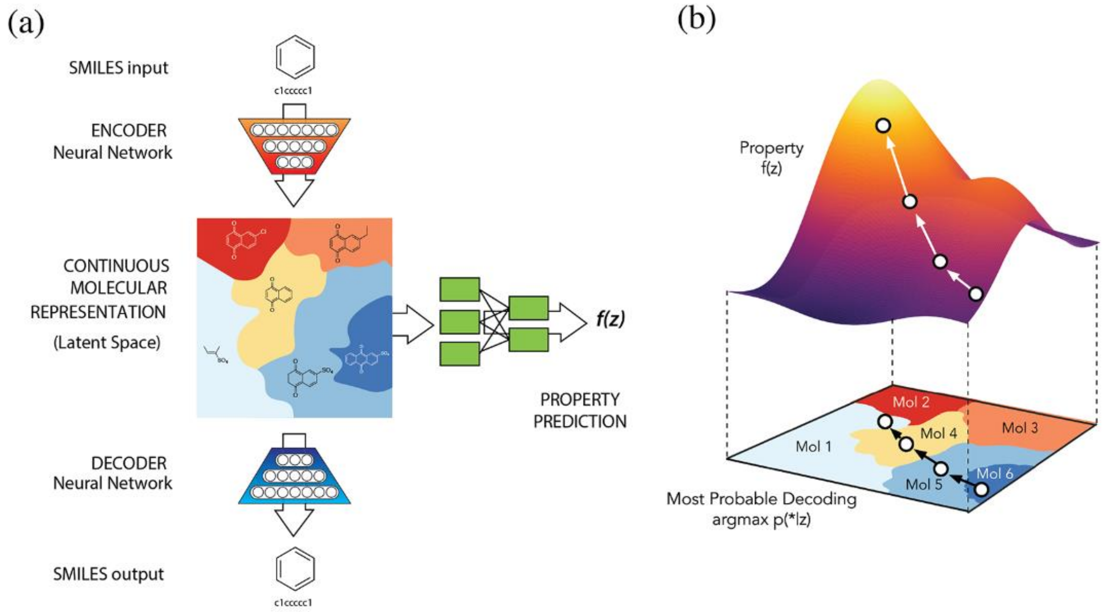
1. What is the meaning of latent vectors?
   → It may contains the "structural information", whose original representation is SMILES.

2. Is using SMILES relevant?
   → Graph generative model might be better

3. How can we generate the molecules with desired properties?
   → Conditional VAE, training VAE jointly with a property controler

Next time

Conditional VAE, training VAE jointly with a property controler



Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *ACS central science* 4.2 (2018): 268-276.

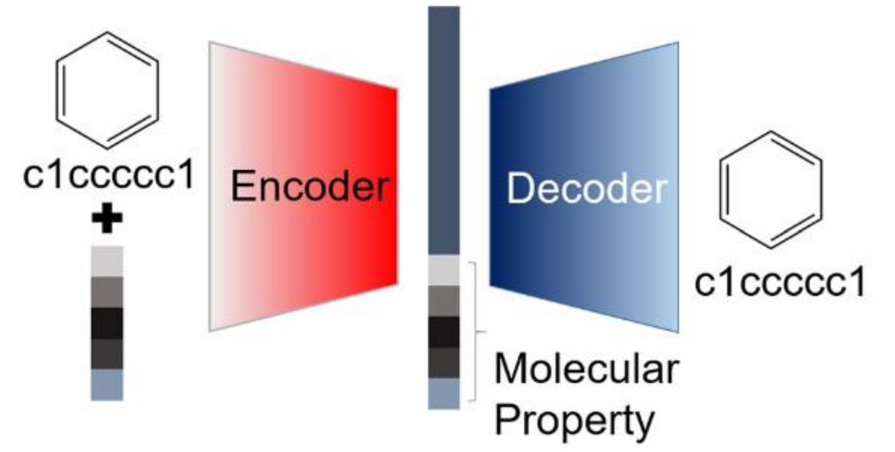


Fig. 1 Schematic representation of conditional variational autoencoder for molecular design

Lim, Jaechang, et al. "Molecular generative model based on conditional variational autoencoder for de novo molecular design." *arXiv preprint arXiv:1806.05805* (2018).