Variational Auto-Encoder

Yingzi Bu

Department of Pharmaceutical Sciences
University of Michigan

Aug 21, 2023

References

1. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013). Cited by 28916
2. Kingma, Durk P., et al. "Semi-supervised learning with deep generative models." Advances in neural information processing systems 27 (2014). Cited by 3102
3. Gomez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276. Cited by 2569
4. Zhavoronkov, Alex, et al. "Deep learning enables rapid identification of potent DDR1 kinase inhibitors." Nature biotechnology 37.9 (2019): 1038-1040. Cited by 769

Probability is all you need

$p(\text{The Sun will rise tomorrow}) \approx 1$

$p(\text{cure cancer} \mid \text{compounds}) = ???$

$p(\text{compounds} \mid \text{cure cancer}) = ???$

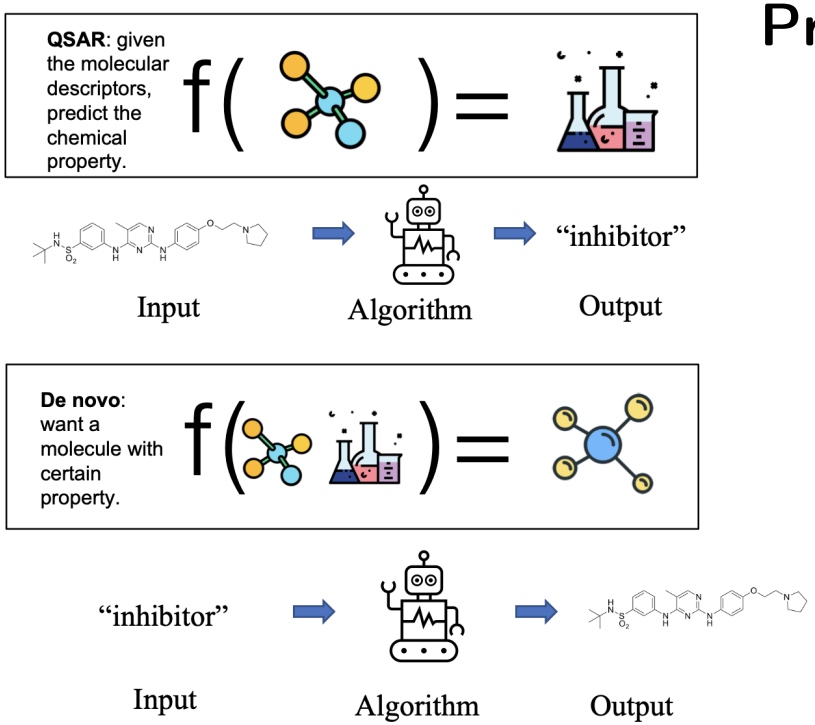
$p(\text{approximation, interesting} \mid \text{Biologist, Chemist}) \approx 1$

$p(\text{Talks in math} = \text{boring} \mid \text{Biologist, Chemist}) > 0.5$

$p(\text{Math proof} \mid \text{you want to know ML}) = 1$

Tasks: classification and generation

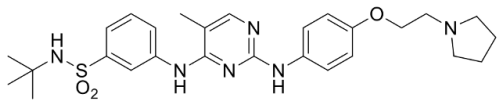
De novo design as the inverse task of molecule property prediction



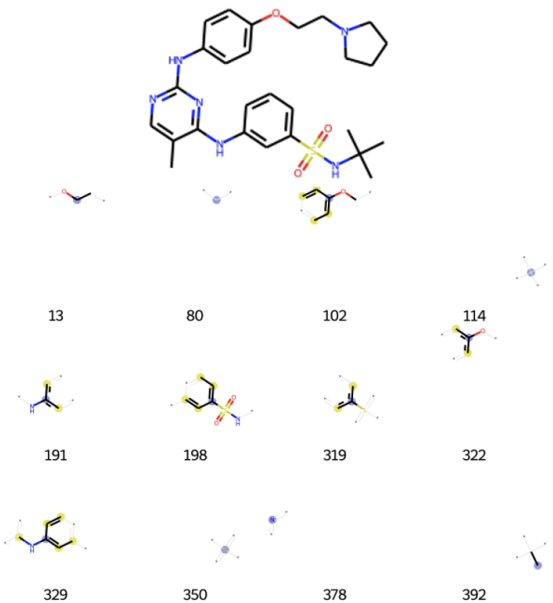
Preparation for ML

- How to represent the input?
- Algorithm. Will it be suitable for your task?
- What is your goal? A numerical way to define it?
- Data set? ML need something to learn from.

Review: Representation of drugs

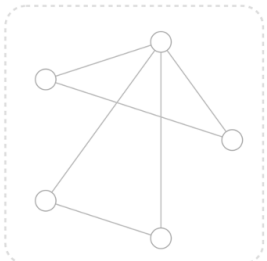
Input 

- **Fingerprint**: 010001..



https://distill.pub/2021/gnn-intro/

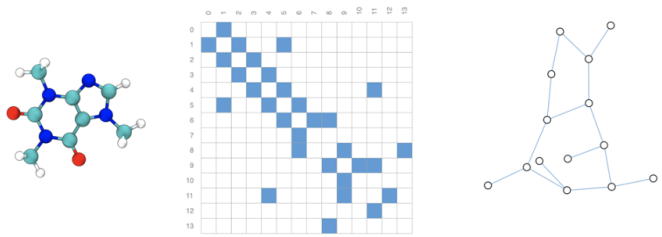
- **Graph**



V Vertex (or node) attributes
e.g., node identity, number of neighbors

E Edge (or link) attributes and directions
e.g., edge identity, edge weight

U Global (or master node) attributes
e.g., number of nodes, longest path



(Left) 3d representation of the Caffeine molecule (Center) Adjacency matrix of the bonds in the molecule (Right) Graph representation of the molecule.

- **Language (SMILES)**
  CC1=CN=C(N=C1NC2=CC(=CC=C2)S(=O)(=O)NC(C)(C)C)NC3=CC=C(C=C3)OCCN4CCCC4

Dataset, train & eval model

Professor vs student

Prof wants to help student learn sth

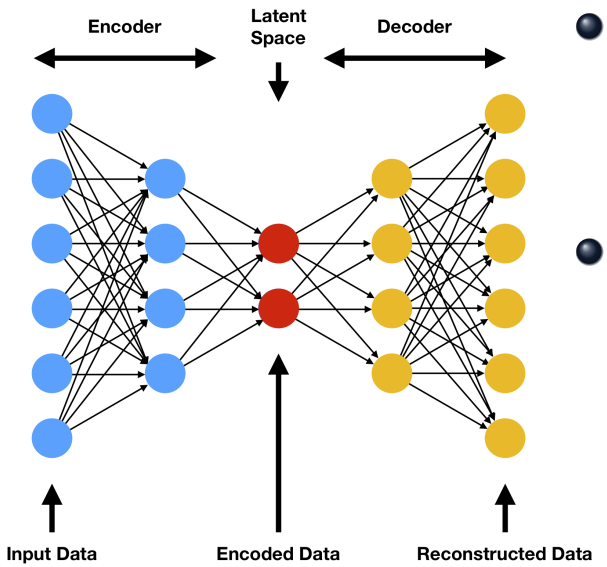
- Prof: limited Q & A collected
- 80% HW, 10% mid, 10% Final
  - Students use HW to learn
  - mid-term for better learning
  - final to test knowledge
- The more questions, the better students may learn
- Prof will give score to guide students in the right direction
  e.g. $\min ||A(s) - A_c||^2$
- Prof may not want students to know final exam questions
  Evaluate whether students really understand concept

Programmer vs model

Coder wants model to learn sth.

- Coder: limited data w/o label.
- 80% train, 10% val, 10% test
  - Model use train set to learn
  - val to fine tune
  - test to evaluate performance
- The more data, the better model may learn
- Coder will define function to tell model how to perform
  e.g. $\min ||\hat{y} - y||^2$
- Coder does not want to use test set to train model
  Then the evaluation reflects model's true ability

Autoencoders



- Two sets:
  - the space of decoded messages $\mathscr{X} \subseteq \mathbb{R}^n$;
  - the space of encoded messages $\mathscr{Z} \subseteq \mathbb{R}^m$
- Two parametrized families of functions:
  - the encoder family: $E_\phi : \mathscr{X} \to \mathscr{Z}$, parametrized by ϕ
  - the decoder family: $D_\theta : \mathscr{Z} \to \mathscr{X}$, parametrized by θ

for $x \in \mathscr{X}$, latent variable $z = E_\phi(x) \in \mathscr{Z}$
for $z \in \mathscr{Z}$, decoded message $x' = D_\theta(z) \in \mathscr{X}$

Training Autoencoders

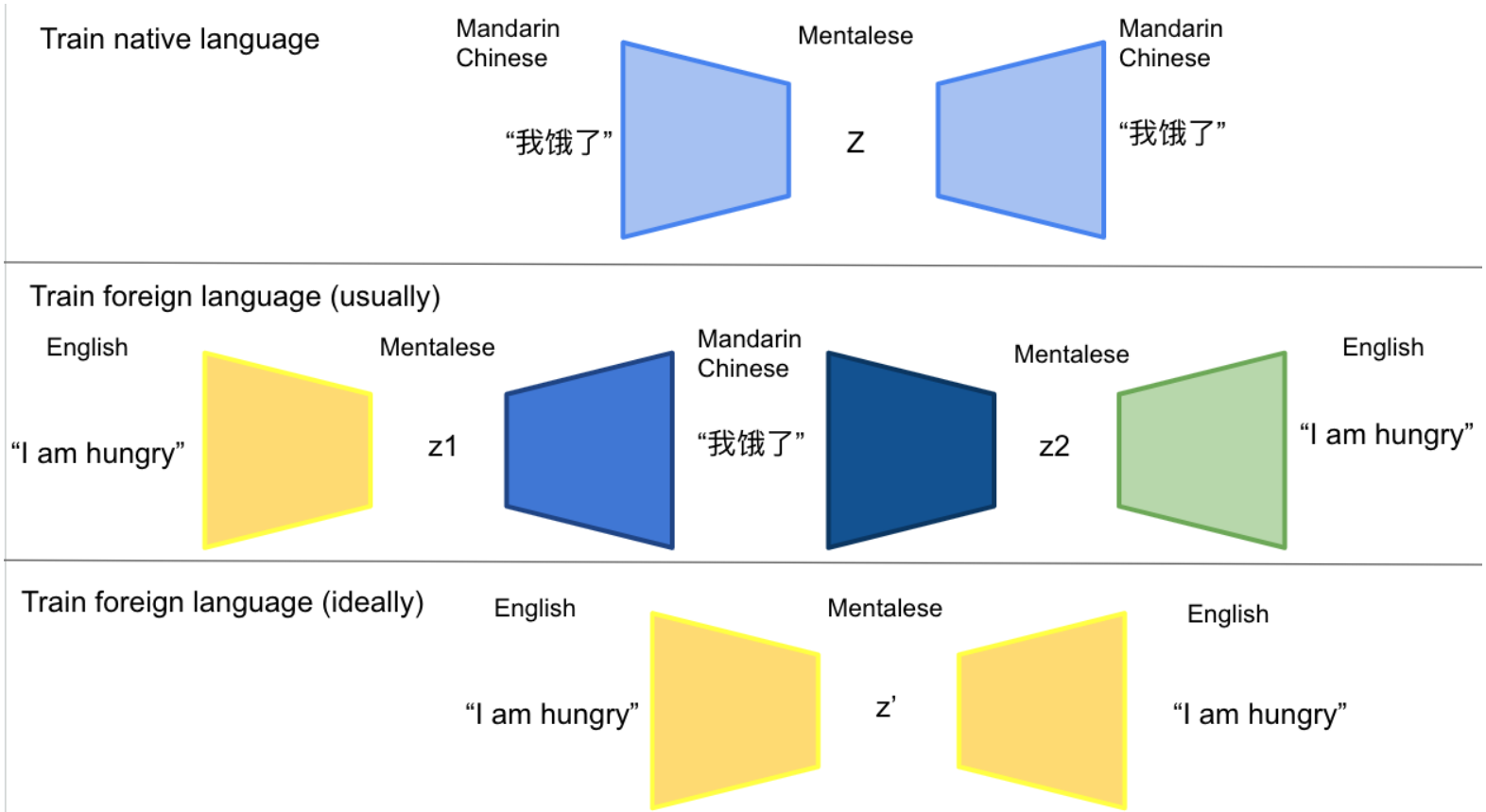
A reference probability distribution μ_{ref} over $\mathscr{X}$,
a function $d : \mathscr{X} \times \mathscr{X} \to [0, \infty)$ measures how much x' differs from x

$$L(\theta, \phi) := \mathbb{E}_{x \sim \mu_{\text{ref}}}[d(x, D_\theta(E_\phi(x))]$$

The least-squares optimization:

$$\min_{\theta,\phi} L(\theta, \phi), \text{ where } L(\theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} ||x_i - D_\theta(E_\phi(x_i))||_2^2$$

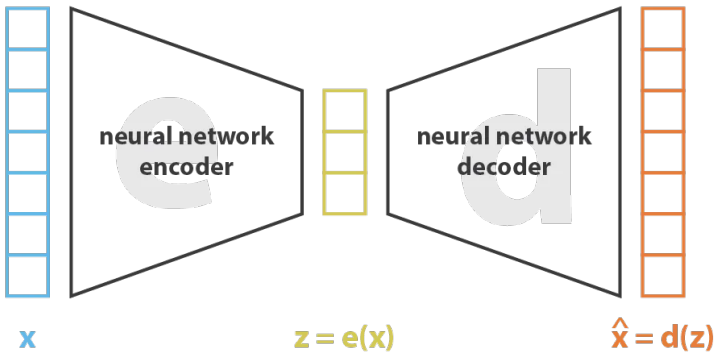
We used AE to learn (e.g. languages)!



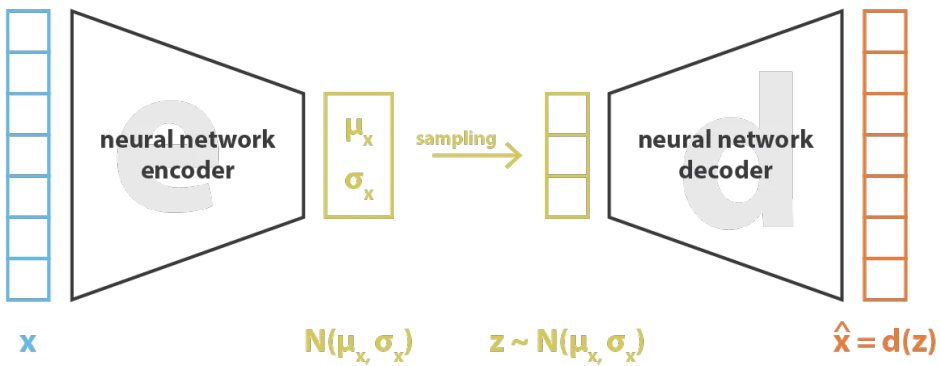
- Native language model: $x \in$ Mandarin, $z \in$ Mentalese, goal: $\min ||x - D(z)||^2 = \min ||x - D(E(x))||^2$, when speak, sample $z \in$ Mentalese, decode mother tongue $x = D(z)$.
- foreign language model: Trained one AE to do English - Chinese, another AE Chinese - English
- Trained ideally: $p(z'|\text{hungry}) \approx p(z_1|\text{hungry}) \approx p(z_2|\text{'e'}) \approx p(z|\text{'e'})$. Then when speak foreign language, sample $z \in$ Mentalese, decode foreign language $x = D(z)$.

Sample? How?

Variational Autoencoder (VAE)



$$\text{loss} = ||x - \hat{x}||^2 = ||x - d(z)||^2 = ||x - d(e(x))||^2$$



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

Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

Evidence lower bound (ELBO)

p^* true distribution of x, p_θ estimated distribution of x
Objective: $\max \mathbb{E}_{x\sim p^*(\cdot)}[\log p_\theta(x)]$
Alternative objective: $\max ELBO = \max_{\theta,\phi} \mathbb{E}_{z\sim q_\phi(\cdot|x)} log \frac{p_\theta(x,z)}{q_\phi(z|x)}$

$$
\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{x\sim p^*(\cdot)}[\log p_\theta(x)] \\
&= KL(q_\phi(z|x)||p_\theta(z|x)) + \mathbb{E}_{q_\phi(z|x)}[-\log q_\phi(z|x) + \log p_\theta(x,z)] \\
&\geq ELBO = -KL(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]
\end{aligned}
$$

$$
\tilde{\mathscr{L}}_{\text{VAE}}(x;\theta,\phi) = -KL(q_\phi(z|x)||p_\theta(z)) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x|z^{(l)})
$$

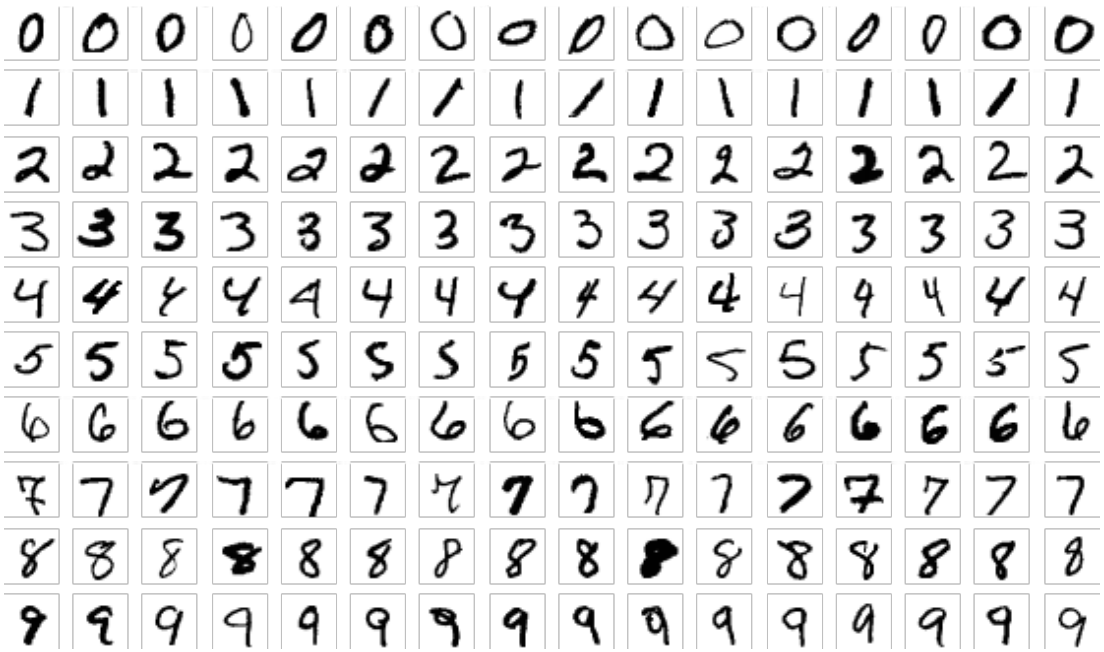
Loss function

$$
\begin{aligned}
\min Loss(x;\theta,\phi) &= \min -\tilde{\mathscr{L}}_{\text{VAE}}(x;\theta,\phi) \\
&= \min KL(q_\phi(z|x)||p_\theta(z)) - \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x|z^{(l)}) \\
&= \min\left[KL(\mathscr{N}(\mu,\boldsymbol{\Sigma})||\mathscr{N}(0,\boldsymbol{I})) - \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x|z^{(l)})\right] \\
&= \min\left[(-\frac{1}{2}\sum_{j=1}^{J}(1+\log\sigma_j^2-\mu_j^2-\sigma_j^2)) - \sum_{l=1}^{L}\log p_\theta(x|z^{(l)})\right] \\
&= \min(\text{KLD} + \text{BCE / MSE})
\end{aligned}
$$

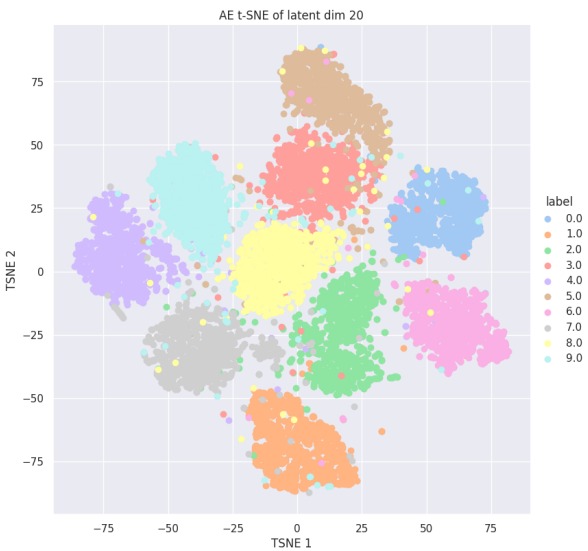
```
def loss_function(recon_x, x, mu, logvar):
    BCE = F.binary_cross_entropy(recon_x.cuda(),
                                 x.view(-1, 784).cuda(),
                                 reduction='sum')
    KLD = -0.5 * torch.sum(1 + logvar - mu.pow(2) - logvar.exp())
    return BCE + KLD
```

Example: MNIST data set

P(Application | AE, VAE math is correct) = ?



AE vs VAE classification, 100 epochs on MNIST data set



Regression task

Toy exmaple:
Using MACCS fingerprint of JAK dataset to predict pIC_{50}, regression task

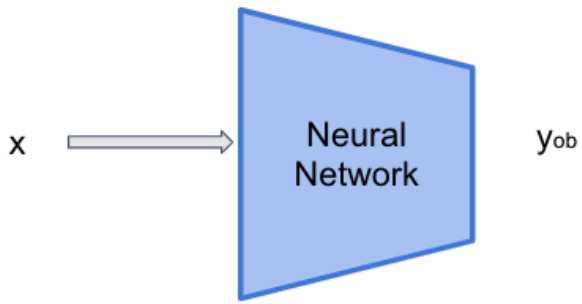


Figure 1: simple DNN

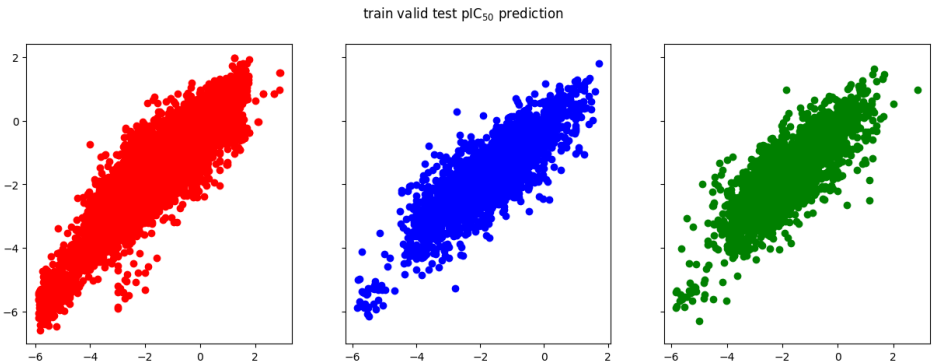


Figure 2: JAK pIC_{50} prediction

Question for Audience: Why need reconstruction of latent space?

Code for binary classification
Code for regression

AE vs VAE generation, 100 epochs on MNIST data set

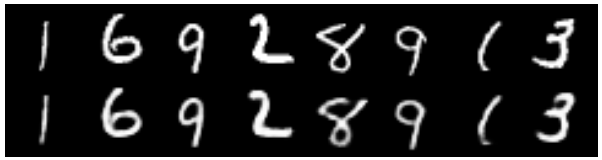


Figure 3: AE $x' \sim p_\theta(*|x)$

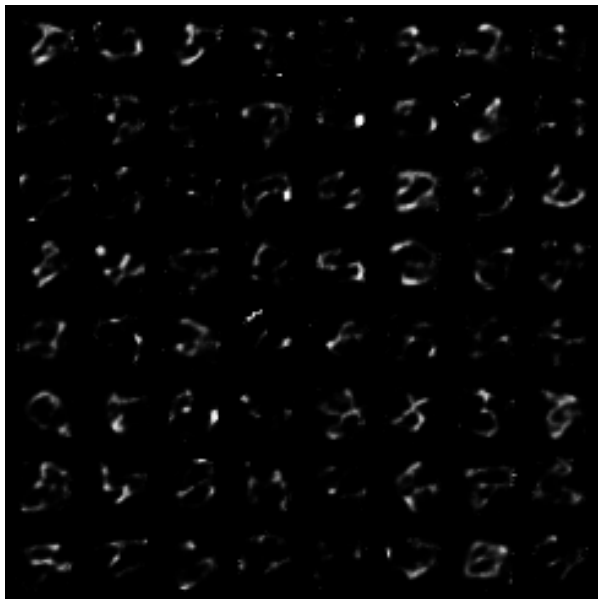


Figure 4: AE $x' \sim p_\theta(*|z)$, $z \sim \mathcal{N}(0, \boldsymbol{I})$

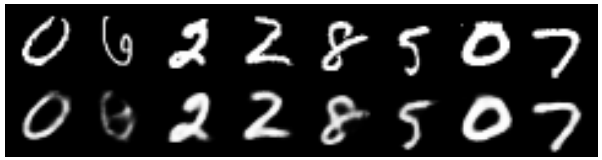


Figure 5: VAE $x' \sim p_\theta(*|x)$

Figure 6: VAE $x' \sim p_\theta(*|z)$, $z \sim \mathcal{N}(0, \boldsymbol{I})$

ML question: AE latent space does not follow normal distribution anyway.

Frame Title

i went to the store to buy some groceries .
i store to buy some groceries .
i were to buy any groceries .
horses are to buy any groceries .
horses are to buy any animal .
horses the favorite any animal .
horses the favorite favorite animal .
horses are my favorite animal .

Table 1: Sentences produced by greedily decoding from points between two sentence encodings with a conventional autoencoder. The intermediate sentences are not plausible English.

Bowman, Samuel R., et al. "Generating sentences from a continuous space." arXiv preprint arXiv:1511.06349 (2015). Cited by 2433

Application in Drugs, prediction

P (application in drugs | works fine in CV) = ?
MNIST data set: figure <-> label | data set (ours): drug <-> label
1. Classification using latent space

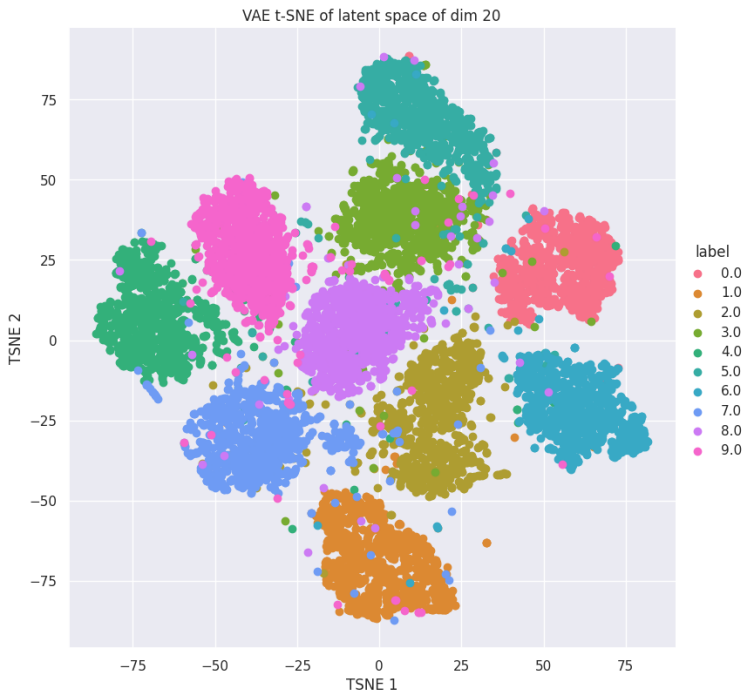


Figure 7: t-SNE of VAE latent space, MNIST dataset

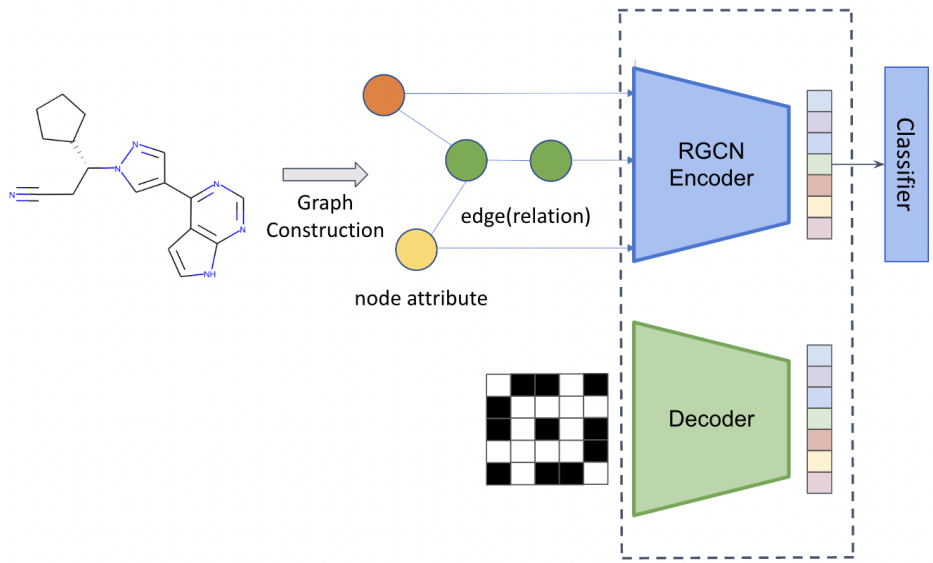


Figure 8: GraphVAE, used in JAK inhibitor binary classification

Application in generation

sample VAE from latent space $z \sim \mathcal{N}(0, I)$, then reconstruct $x \sim p_\theta(\cdot|z)$, image will be obtained by converting $x \in \mathbb{R}^{\mathbf{784}}$ to a 28×28 image matrix.

- Question for Audience: What if we train this on compound/molecule data set?
- Improvements?

conditional generative models

What if I want compounds that have certain properties?

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}, y) = \mathcal{N}(\mu_\phi(\boldsymbol{x}), \text{diag}(\sigma^2_\phi(\boldsymbol{x})))$$
$$p_\theta(\boldsymbol{z}) = \mathcal{N}(0, \boldsymbol{I})$$
$$\boldsymbol{x} \sim p_\theta(\cdot|\boldsymbol{z}, y)$$

Question for audience: What will the architecture look like?

Math proof is left for your practice.

MNIST application: digit generation

Figure 9: sample conditional VAE from latent space $z \sim \mathcal{N}(0, I)$, given y, then reconstruct $x \sim p_\theta(\cdot|z, y)$

Improvements?

Frame Title

Stacked generative semi-supervised model

Only a subset of the observations have corresponding class labels.
How to make use of observations without labels?

Latent-feature discriminative model (M1):
VAE $p(z) = \mathcal{N}(z|0, I), p_\theta(x|z) = f(x; z, \theta)$

Generative semi-supervised model (M2)
VAE $p(y) = \text{Cat}(y|\pi), p(z) = \mathcal{N}(z|0, I), p_\theta(x|y, z) = f(x; y, z, \theta)$

Stacked generative semi-supervised model (M1 + M2)
z_1 from M1, learn model M2 using embeddings from z_1 instead of x.
Two layers of stochastic variables:
$p_\theta(x, y, z_1, z_2) = p(y)p(z_2)p_\theta(z_1|y, z_2)p_\theta(x|z_1)$

M1: $p(\mathbf{z_1}|x) = \mathcal{N}(z|\mu, \text{diag}(\sigma^2)), p(\mathbf{z_1}) = \mathcal{N}(z|0, I)$, decoder: $x \sim p_\theta(x|\mathbf{z_1})$
M2: $p(\mathbf{z_2}|\mathbf{z_1}, y) = \mathcal{N}(\mathbf{z_2}|\mu_\mathbf{1}, \text{diag}(\sigma^\mathbf{2}_\mathbf{1})), p(\mathbf{z_2}) = \mathcal{N}(\mathbf{z_2}|0, I)$, decoder: $\mathbf{z_1} \sim p_\theta(\mathbf{z_1}|\mathbf{z_2}, y)$

Kingma, Durk P., et al. "Semi-supervised learning with deep generative models." Advances in neural information processing systems 27 (2014).

Stacked generative semi-supervised model: M1 loss

M1 loss:

$p(z|x) = \mathcal{N}(z|\mu, \text{diag}(\sigma^2)), p(z) = \mathcal{N}(z|0, I)$, decoder: $x \sim p_\theta(x|z)$

$\log p_\theta(x) \geq \mathbb{E}_{q_\theta(z|x)}[\log p_\theta(x|z) - KL(q_\phi(z|x)||p_\theta(z))] = -\mathscr{J}(x)$

Stacked generative semi-supervised model: M2 Loss

M2 loss:
$p(z_2|z_1,y) = \mathcal{N}(z_2|\mu_1, \mathrm{diag}(\sigma_1^2)), p(z_2) = \mathcal{N}(z_2|0,I)$,
decoder: $z_1 \sim p_\theta(z_1|z_2,y)$, ($x$ means z_1, z means z_2) for eqs. below:

$$\begin{aligned}
\log p_\theta(x,y) &\geq \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(x|y,z) + \log p_\theta(y) + \log p(z) - \log q_\phi(z|x,y)] \\
&= -\mathscr{L}(x,y) \text{ data with labels} \\
\log p_\theta(x) &\geq \mathbb{E}_{q_\phi(y,z|x)}[\log p_\theta(x|y,z) + \log p_\theta(y) + \log p(z) - \log q_\phi(y,z|x)] \\
&= \sum_y q_\phi(y|x)(-\mathscr{L}(x,y)) + \mathscr{H}(q_\phi(y|x)) = -\mathscr{U}(x) \text{ data no labels} \\
\mathscr{J} &= \sum_{(x,y)\sim\tilde{p}_l} \mathscr{L}(x,y) + \sum_{x\sim\tilde{p}_u} \mathscr{U}(x) \\
\mathscr{J}^\alpha &= \mathscr{J} + \alpha \cdot \mathbb{E}_{\tilde{p}_l(x,y)}[-\log q_\phi(y|x)]
\end{aligned}$$

Example in drug discovery using VAE

Abstract

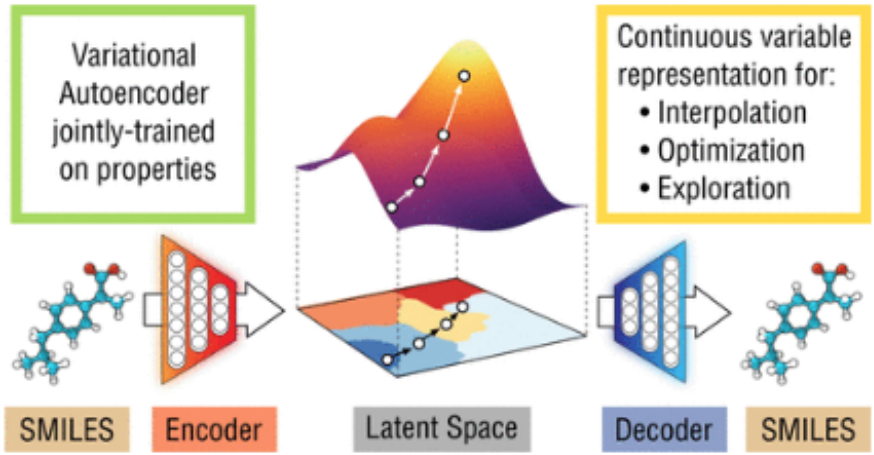
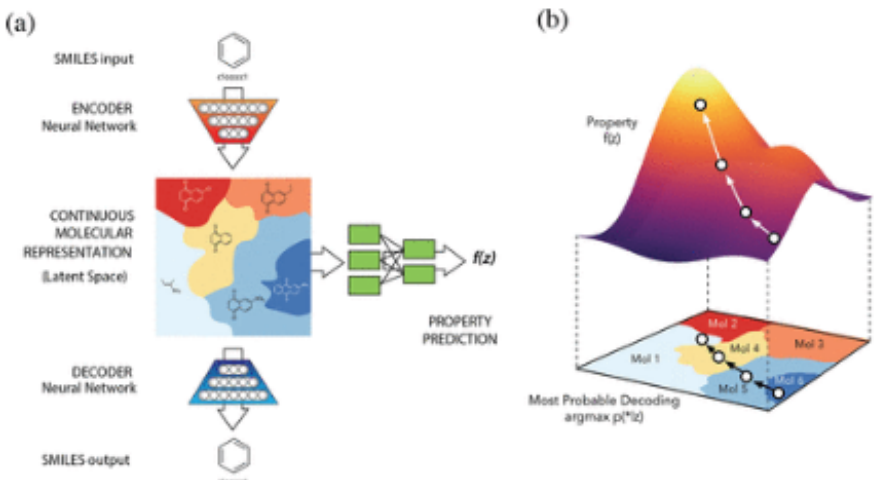
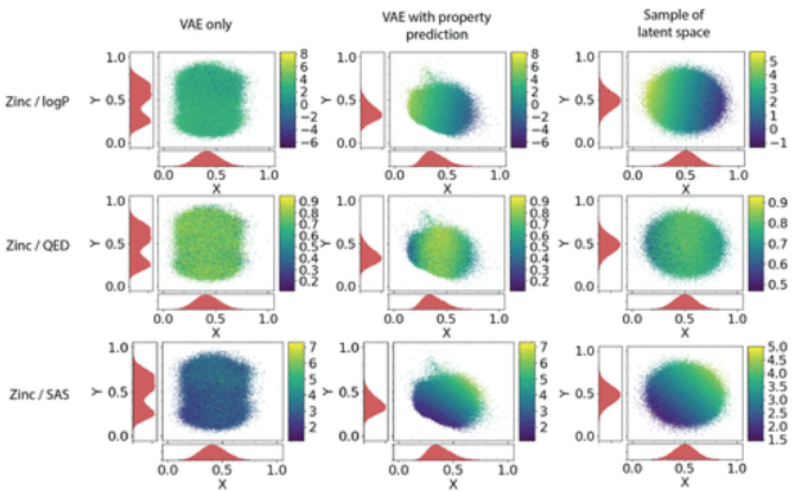


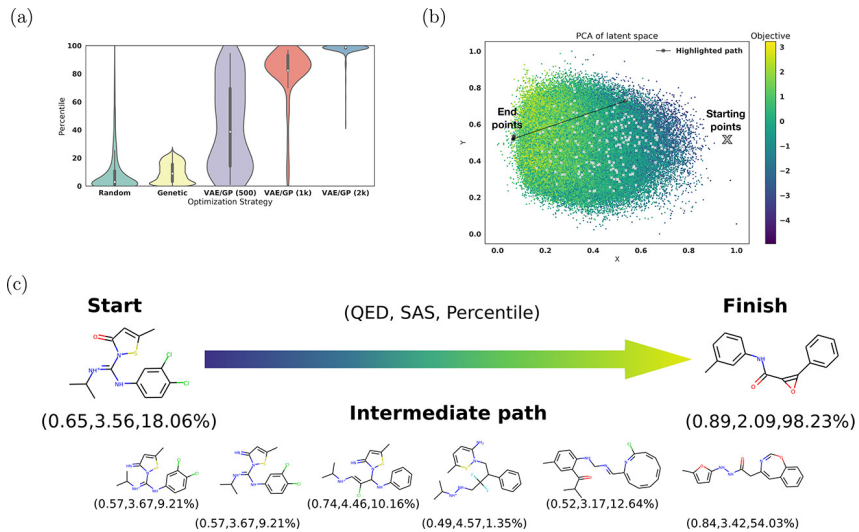
Figure 1



Gomez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276. Cited by 2569

Example in drug discovery using VAE

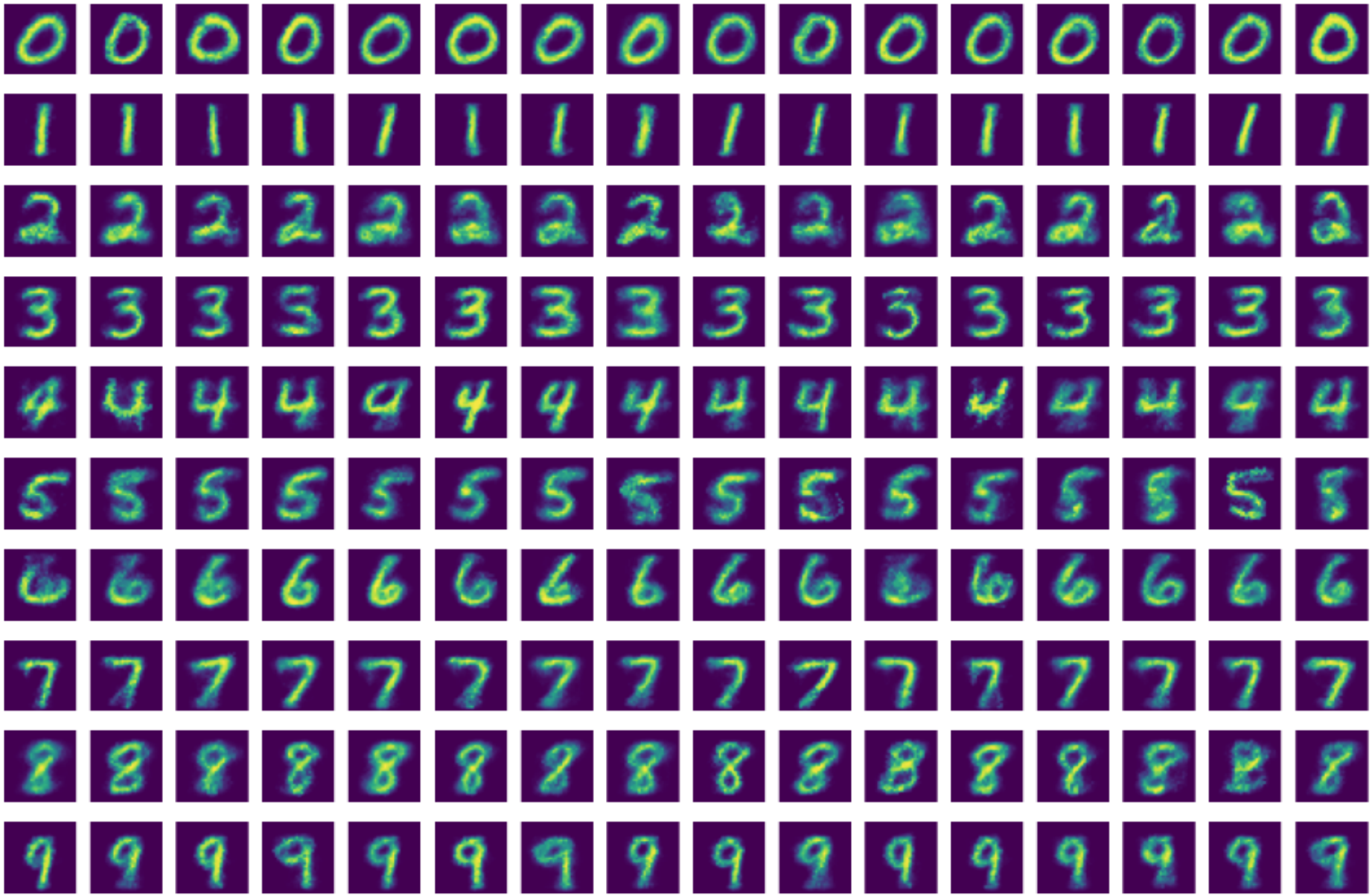




Question for audience: What is the difference between this VAE and simply using M2?

Gomez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276. Cited by 2569

Stacked Deep Generative Model

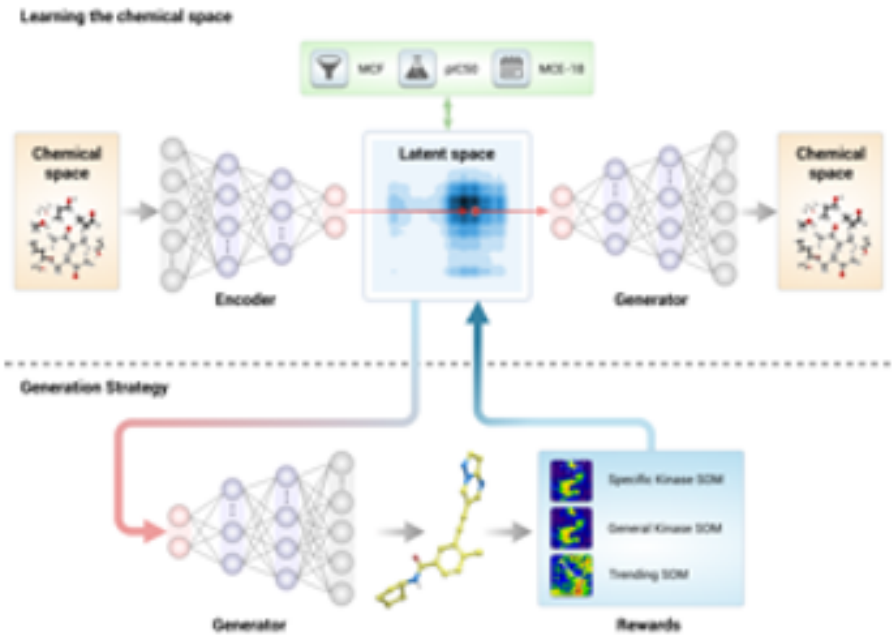


https://colab.research.google.com/drive/1v-UEbcQBygEwAoRscoKxY3zLP9QXpveK#scrollTo=_ZB-myHkWm1T&uniqifier=1

DDR1 inhibitor generation

BRIEF COMMUNICATION
https://doi.org/10.1038/s41587-019-0224-x
nature biotechnology

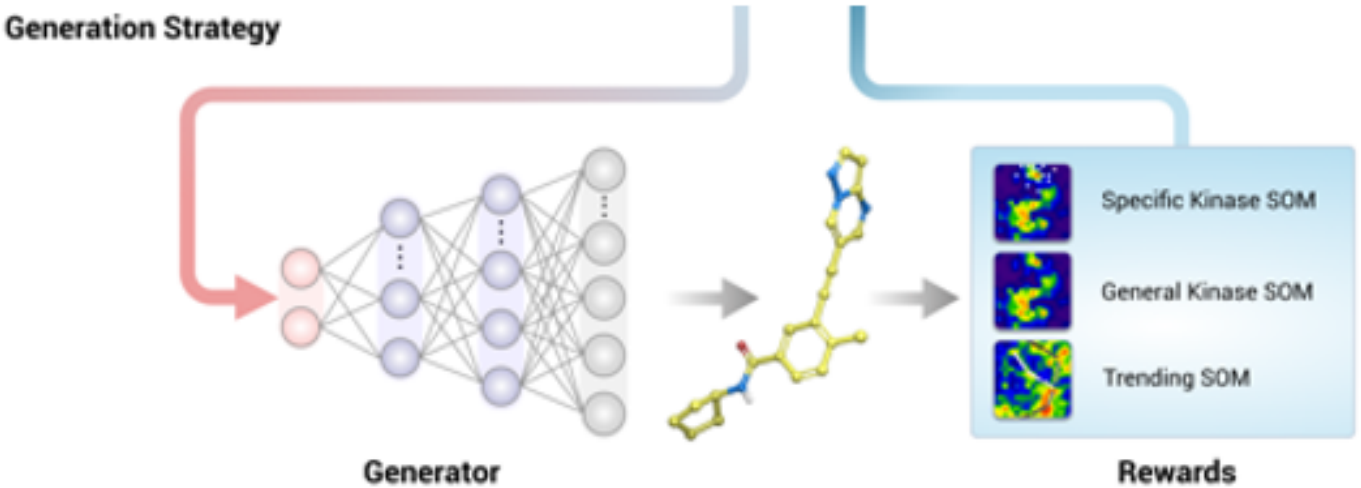
Deep learning enables rapid identification of potent DDR1 kinase inhibitors



Dataset:
a large set of molecules derived from a ZINC data set

Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A. and Volkov, Y., 2019. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology*, *37*(9), pp.1038-1040.

DDR1 inhibitor generation



Datasets:

known DDR1 kinase inhibitors

common kinase inhibitors (positive set)
molecules that act on non-kinase targets (negative set)

patent data for biologically active molecules that have been claimed by pharmaceutical companies

$$\max_{\psi} \mathbb{E}_{\mathbf{z}\sim p_{\psi}(\mathbf{z})} R(\mathbf{z}), \quad R(\mathbf{z}) = \mathbb{E}_{\mathbf{x}\sim p_{\theta}(\mathbf{x}|\mathbf{z})}[R_{\text{general}}(\mathbf{x}) + R_{\text{specific}}(\mathbf{x}) + R_{\text{trending}}(\mathbf{x})]$$

Datasets (JAK task):

known JAK inhibitors and noninhibitors

Rewards:
Calculated using CoGT instead of simply using SOM

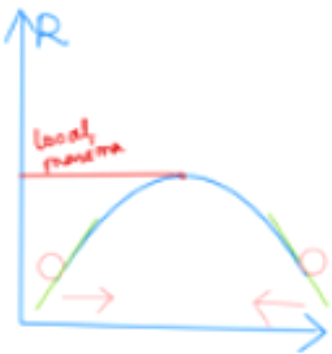
$$R(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{x}\sim p_{\theta}(\boldsymbol{x}|\boldsymbol{z})} \sum_{i} w_i R_{\text{JAK}_i}(\boldsymbol{x})$$

DDR1 inhibitor generation

Reinforcement learning

REINFORCE algorithm:

1. sample $\{\tau^i\}$ from $\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)$ (run the policy)
2. $\nabla_\theta J(\theta) \approx \sum_i (\sum_t \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t^i|\boldsymbol{s}_t^i))(\sum_t r(\boldsymbol{s}_t^i, \boldsymbol{a}_t^i))$
3. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$



$$\nabla_\psi \mathbb{E}_{\mathbf{z}\sim p_\psi(\mathbf{z})} R(\mathbf{z}) = \mathbb{E}_{\mathbf{z}\sim p_\psi(\mathbf{z})} \nabla_\psi \log p_\psi(\mathbf{z}) \cdot \mathrm{R}(\mathbf{z})$$

$$\nabla_\psi \mathbb{E}_{\mathbf{z}\sim p_\psi(\mathbf{z})} R(\mathbf{z}) \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_\psi \log p_\psi(\mathbf{z}_i) \left[R(\mathbf{z}_i) - \frac{1}{N} \sum_{j=1}^{N} R(\mathbf{z}_j) \right]$$

$$\theta^* = \arg\max_\theta \underbrace{\mathbb{E}_{\tau\sim p_\theta(\tau)} \left[\sum_t r(\boldsymbol{s}_t, \boldsymbol{a}_t) \right]}_{J(\theta)}$$

$$J(\theta) = \mathbb{E}_{\tau\sim p_\theta(\tau)} \left[\sum_t r(\boldsymbol{s}_t, \boldsymbol{a}_t) \right] \approx \frac{1}{N} \sum_i \sum_t r(\boldsymbol{s}_{i,t}, \boldsymbol{a}_{i,t})$$

$$J(\theta) = \mathbb{E}_{\tau\sim p_\theta(\tau)} [\underbrace{r(\tau)}_{\sum_{t=1}^{T} r(\boldsymbol{s}_t, \boldsymbol{a}_t)}] = \int p_\theta(\tau) r(\tau) d\tau \quad \text{Direct policy differentiation}$$

$$\nabla_\theta J(\theta) = \int \nabla_\theta p_\theta(\tau) r(\tau) d\tau$$

$$\begin{aligned}
\nabla_\theta J(\theta) &= \mathbb{E}_{\tau\sim p_\theta(\tau)} [\nabla_\theta \log p_\theta(\tau) r(\tau)] \\
&= \mathbb{E}_{\tau\sim p_\theta(\tau)} \left[\nabla_\theta \left(\log p(\boldsymbol{s}_1) + \sum_{t=1}^{T} \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t) + \log p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t) \right) r(\tau) \right] \\
&= \mathbb{E}_{\tau\sim p_\theta(\tau)} \left[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t) \right) \left(\sum_{t=1}^{T} r(\boldsymbol{s}_t, \boldsymbol{a}_t) \right) \right] \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\boldsymbol{a}_{i,t}|\boldsymbol{s}_{i,t}) \right) \left(\sum_{t=1}^{T} r(\boldsymbol{s}_{i,t}, \boldsymbol{a}_{i,t}) \right)
\end{aligned}$$

A convenient identity: $p_\theta(\tau) \nabla_\theta \log p_\theta(\tau) = \nabla_\theta p_\theta(\tau)$

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \log p_\theta(\tau) [r(\tau) - b]$$

$$b = \frac{1}{N} \sum_{i=1}^{N} r(\tau) \quad \text{but... are we allowed to do that??}$$

$$\begin{aligned}
\mathbb{E}[\nabla_\theta \log p_\theta(\tau) b] &= \int p_\theta(\tau) \nabla_\theta \log p_\theta(\tau) b d\tau = \int \nabla_\theta p_\theta(\tau) b d\tau \\
&= b \nabla_\theta \int p_\theta(\tau) d\tau = b \nabla_\theta 1 = 0
\end{aligned}$$

Technical problems

Problem I encountered: I cannot generate JAK1 specific inhibitors.

PharmSci thought: Maybe indeed JAK1 specific cannot be achieved.

ML thought: Is it because of the algorithm?

Question for audience: How can I modify the architecture?

KL divergence

For VAE, min Loss = min (BCE + KLD)

min Loss equals min BCE + min KLD?

what will happen if KLD = $KL(q_\phi(z|x)||p_\theta(z))$= 0 ?

Four questions

My answer

- Aim: Obtain potent drugs with higher probability. Use p_θ to approximate the true p* of drugs with desired properties.
  $\max_\theta p_\theta(\text{drug}|\text{target})$, $\theta^* = \arg\max p_\theta$.
- Now: $\theta_{tr} \in \{\text{tradition strategy}\} \approx$ approximation of p* using human experience.
  Sample $x \in$ compound space $[10^{23}, 10^{60}]$, then determine

$$p(\text{x}|\text{target}) = \begin{cases} 1 & \text{if x binds to target} \\ 0 & \text{otherwise} \end{cases}$$

- Innovation: To find θ^*, chemists use their brains, I use computer.
- Usage: $x \sim p_{\theta^*}$ would have desired properties with higher probabilities.

Other frequently asked questions

Q: What is your innovation? No new algorithm, no innovation.
A: PCR is innovation, using PCR to solve problems is not?

Q: I do not believe ML. Why does it work?
A: It makes sense after some math proof.

Q: You model does not predict well. Thus ML does not work.
A: Data are limited / local optimum. However, I do not question ML methodology.

Thank you

Derive ELBO

KL divergence: $KL(q||p) = \int q \log \frac{q}{p} = \mathbb{E}_q[\log \frac{q}{p}] \geq 0$, =0 holds iff $q = p$.

We want $p_\theta(x) \approx p^*(x)$, which is equivalent of $\min KL(p^*(x)||p_\theta(x))$

$$\begin{aligned}\log p_\theta(x) &= \mathbb{E}_{x\sim p^*(\cdot)}[\log p_\theta(x)] = \mathbb{E}_{x\sim p^*(\cdot)}[\log \frac{p^*(x)}{p^*(x)} p_\theta(x)] \\ &= \mathbb{E}_{x\sim p^*(\cdot)} \log p^*(x) + \mathbb{E}_{x\sim p^*(\cdot)}[\log \frac{p_\theta(x)}{p^*(x)}] \\ &= -H(p^*) - KL(p^*(x)||p_\theta(x))\end{aligned}$$

$\max \log p_\theta(x)$ is $\min KL(p^*(x)||p_\theta(x))$ as $H(p^*) = const.$

Then we need to find a way of calculating $\log p_\theta(x)$

Appendix

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz = \int p_\theta(x,z)dz$$
$$= \int \frac{p_\theta(x,z)}{q_\phi(z|x)} q_\phi(z|x)dz = \mathbb{E}_{z\sim q_\phi(z|x)} \frac{p_\theta(x,z)}{q_\phi(z|x)}$$
$$\log p_\theta(x) = \log \mathbb{E}_{z\sim q_\phi(z|x)} \frac{p_\theta(x,z)}{q_\phi(z|x)} \geq \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$
$$ELBO := \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$

You can also see that:

$$ELBO = \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \frac{p_\theta(z|x)p_\theta(x)}{q_\phi(z|x)}\right]$$
$$= \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \frac{p_\theta(z|x)}{q_\phi(z|x)}\right] + \mathbb{E}_{z\sim q_\phi(z|x)} [\log p_\theta(x)]$$
$$= -KL(q_\phi(z|x)||p_\theta(z|x)) + \log p_\theta(x) \leq \log p_\theta(x)$$

Appendix

$\min KL(p*(x)||p_\theta(x)) \rightarrow \max \log p_\theta(x) \rightarrow \max ELBO$. How to $\max ELBO$?

$$
\begin{aligned}
ELBO &= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}\right] \\
&= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(z)}{q_\phi(z|x)}\right] + \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\
&= -KL(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\
&\approx -KL(q_\phi(z|x)||p_\theta(z)) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x|z^{(l)})
\end{aligned}
$$

Thus $\max ELBO = \min -ELBO = \min KL(q_\phi(z|x)||p_\theta(z)) + ||x - \hat{x}||^2$

How to calculate $KL(q_\phi(z|x)||p_\theta(z))$?

Appendix One dimension: $z \in \mathbb{R}$

$$
\begin{aligned}
q_\phi(z|x) = \mathcal{N}(\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} e^{\left[-\frac{(z-\mu)^2}{2\sigma^2}\right]} \\
\mathbb{E}_{z\sim q_\phi(\cdot|x)}(z) &= \mu, \mathbb{E}_{z\sim q_\phi(\cdot|x)}(z^2) = \mu^2 + \sigma^2 \\
p_\theta(z) = \mathcal{N}(0,1) &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] \\
KL(q_\phi(z|x)||p_\theta(z)) &= \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\log \frac{q_\phi(z|x)}{p_\theta(z)}\right] \\
&= \mathbb{E}_{z\sim q_\phi(\cdot|x)} \log \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{\left[-\frac{(z-\mu)^2}{2\sigma^2}\right]}}{\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]} \\
&= \mathbb{E}_{z\sim q_\phi(\cdot|x)}[-\log\sigma - \frac{(z-\mu)^2}{2\sigma^2} + \frac{z^2}{2}] \\
&= -\frac{1}{2}[\log\sigma^2 + 1 - \mu^2 - \sigma^2]
\end{aligned}
$$

Appendix Multi dimension: $\boldsymbol{z} \in \mathbb{R}^J$

$$\begin{aligned}
\int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{z}) d\boldsymbol{z} &= \int \mathcal{N}(\boldsymbol{z}; \mu, \sigma^2) \log \mathcal{N}(\boldsymbol{z}; 0, \boldsymbol{I}) d\boldsymbol{z} \\
&= -\frac{J}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^{J} (\mu_j^2 + \sigma_j^2) \\
\int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log q_\phi(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} &= \int \mathcal{N}(\boldsymbol{z}; \mu, \sigma^2) \log \mathcal{N}(\boldsymbol{z}; \mu, \sigma^2) d\boldsymbol{z} \\
&= -\frac{J}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^{J} (1 + \log \sigma_j^2)
\end{aligned}$$

Appendix Multi dimension: $\boldsymbol{z} \in \mathbb{R}^J$

$$
\begin{aligned}
KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})) &= \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\cdot|\boldsymbol{x})}\left[\log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z})}\right] \\
&= \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\cdot|\boldsymbol{x})}\left[\log q_\phi(\boldsymbol{z}|\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\cdot|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{z})\right] \\
&= \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log q_\phi(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} - \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{z}) d\boldsymbol{z} \\
&= -\frac{1}{2}\sum_{j=1}^{J}(1 + \log \sigma_j^2 - \mu_j^2 - \sigma_j^2)
\end{aligned}
$$