block b6

The AI Virtual Cell

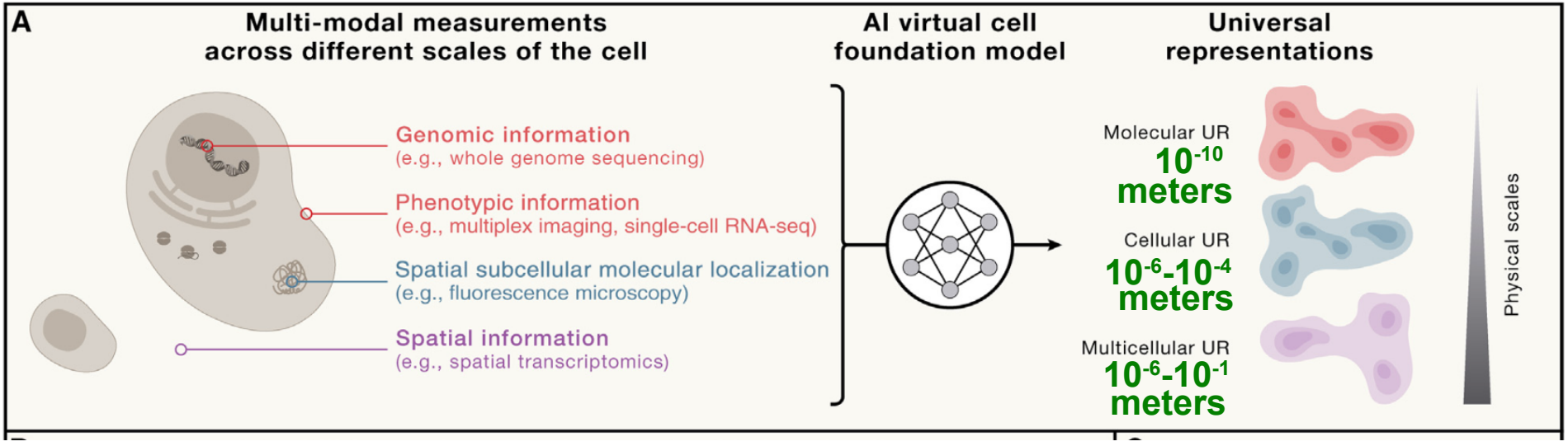
block b6

The AI Virtual Cell

"A conceptual feat at present"

Lin Tang, Nat Meth Editor, Dec 2025.

Multiple scales to unify Multiple data types

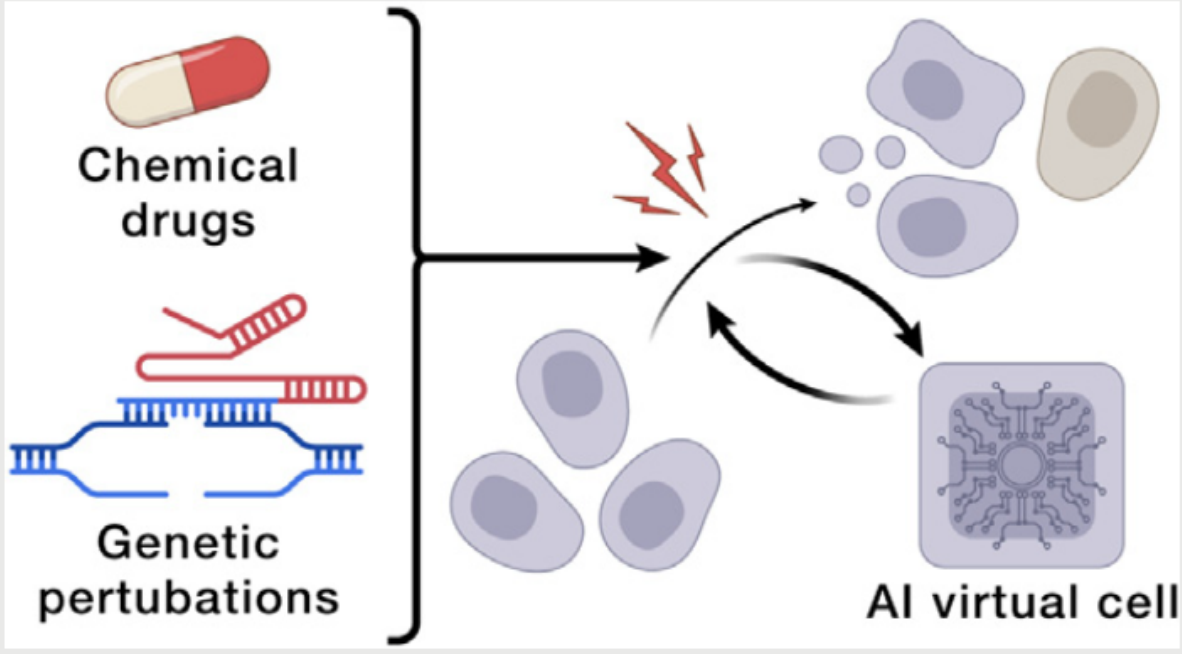


'How to build the virtual cell with AI: prioriities and opportunities", *Bunne et al, Cell Dec2024*

Universal Representations (URs)
Universal Embeddings

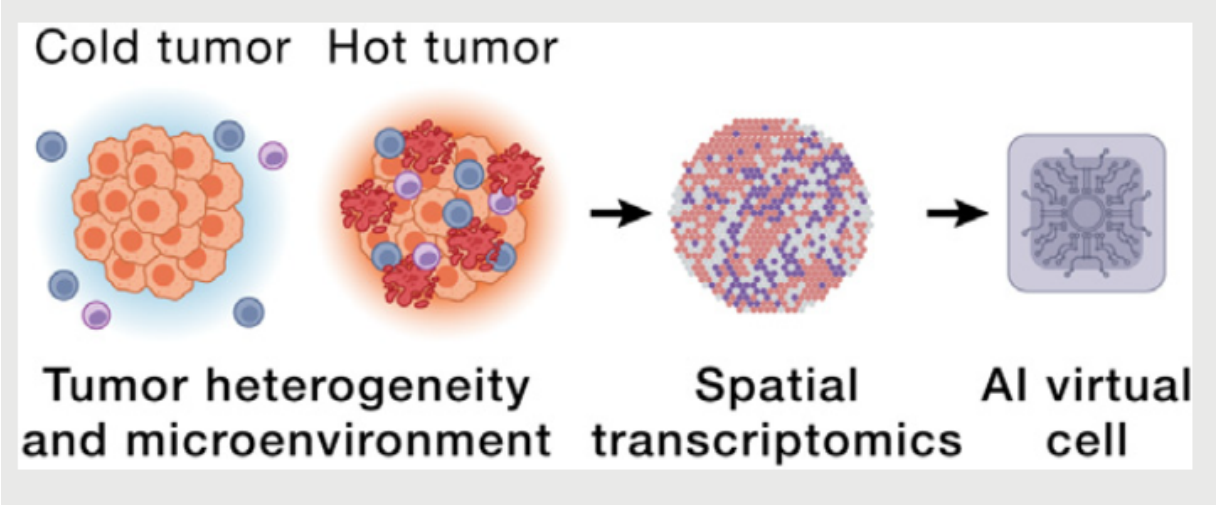
Desired Capabilities

drug-discovery and cell-based Therapeutics



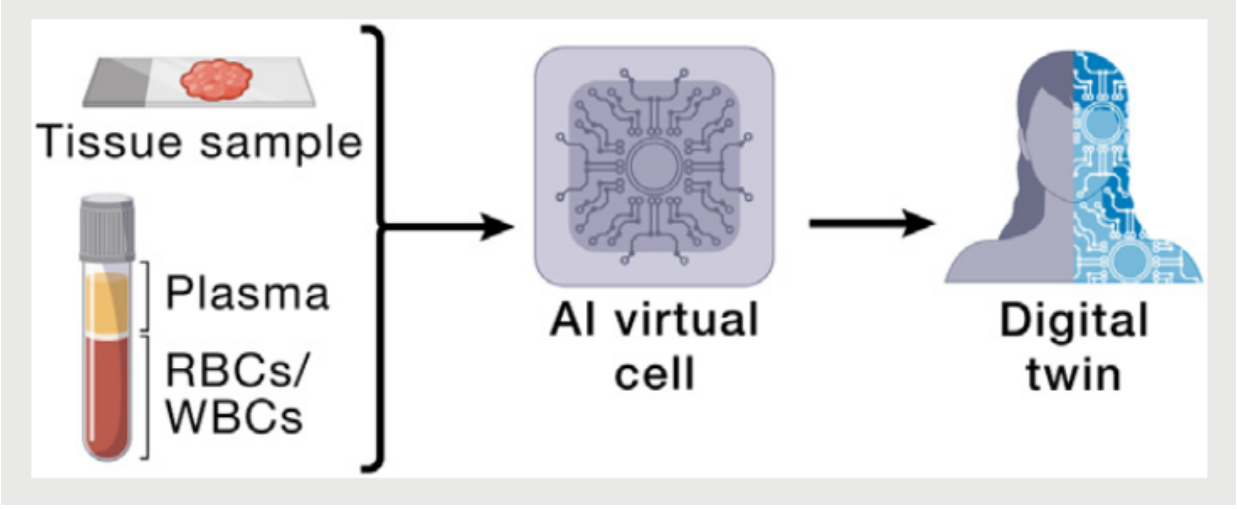
Desired Capabilities

identifying pan-cancer markers



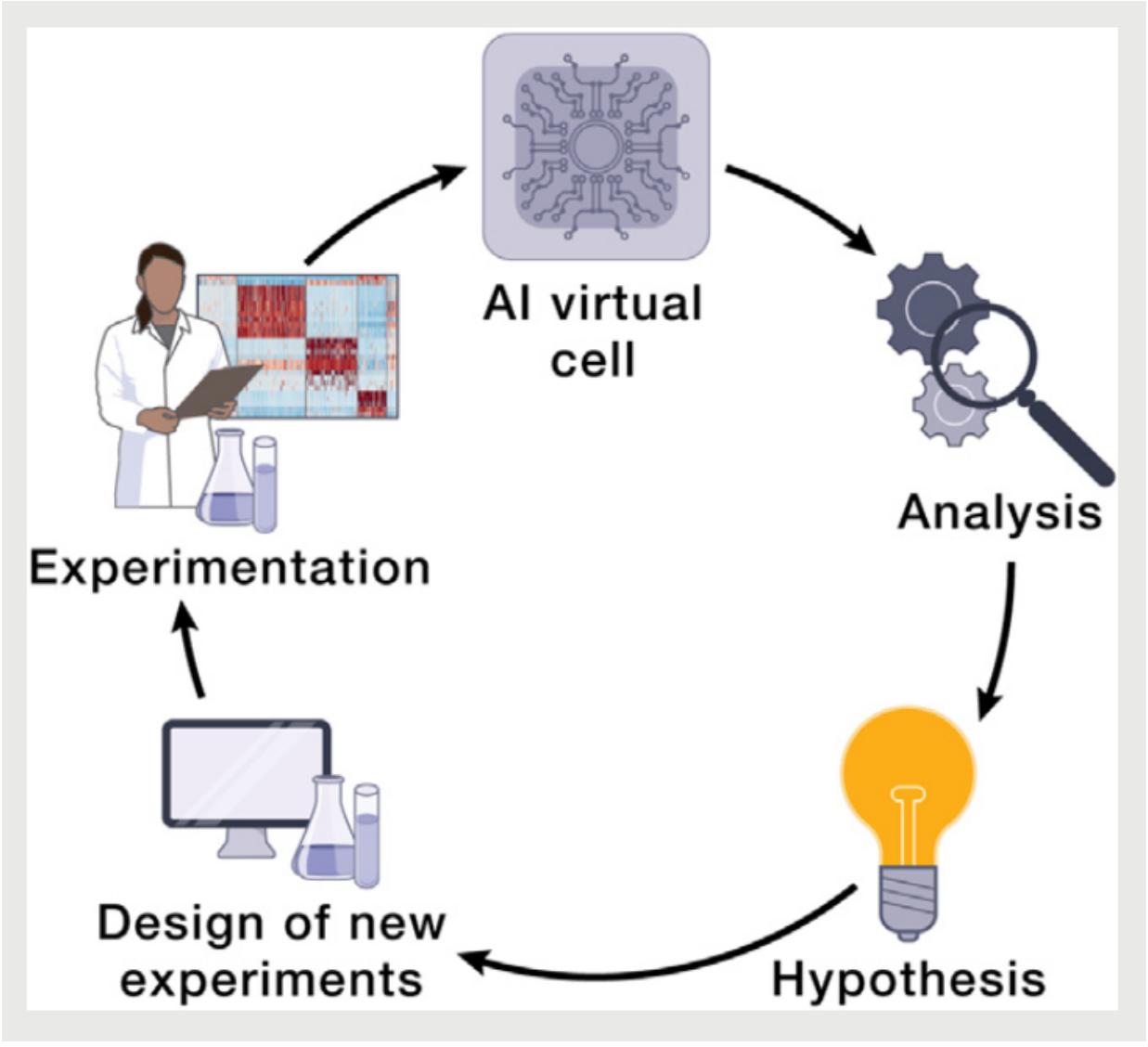
Desired Capabilities

cell models for individual pacients

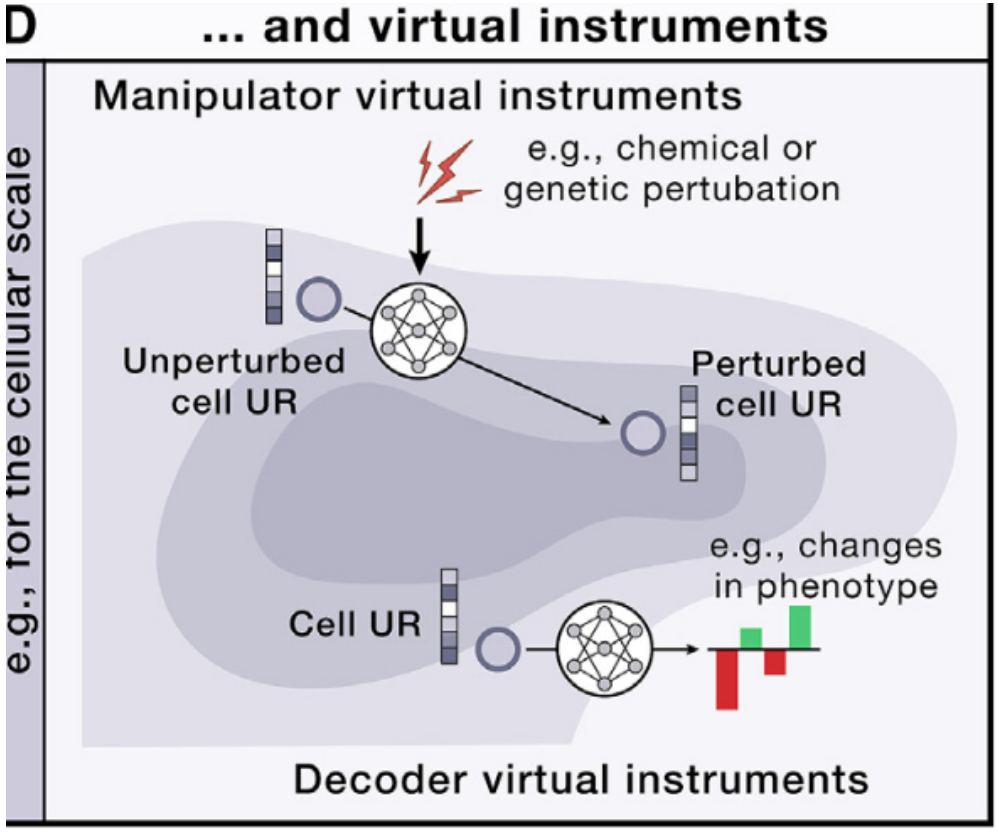


Desired Capabilities

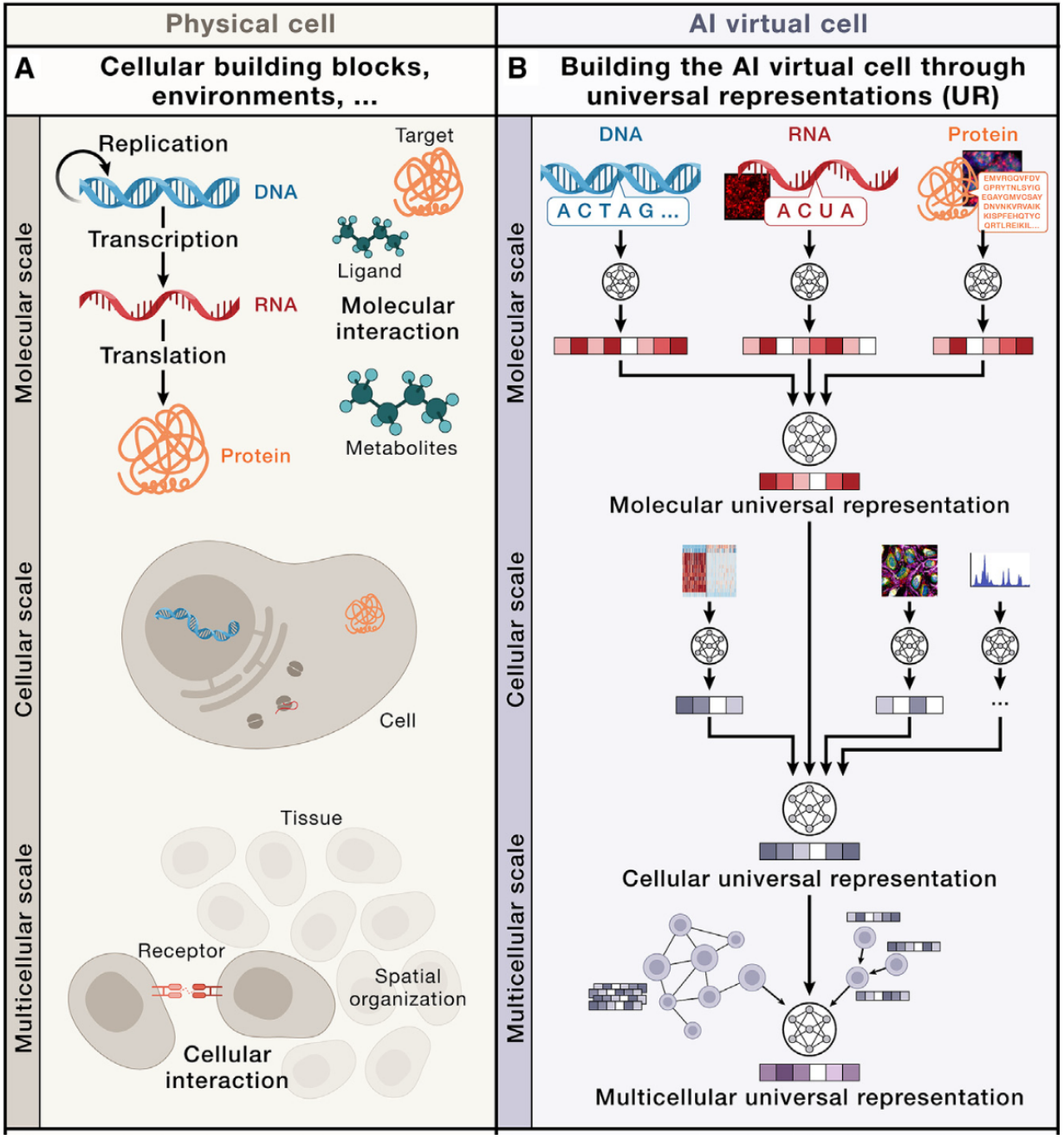
HYPOTHESIS GENERATING VA



Desired Capabilities



Methods



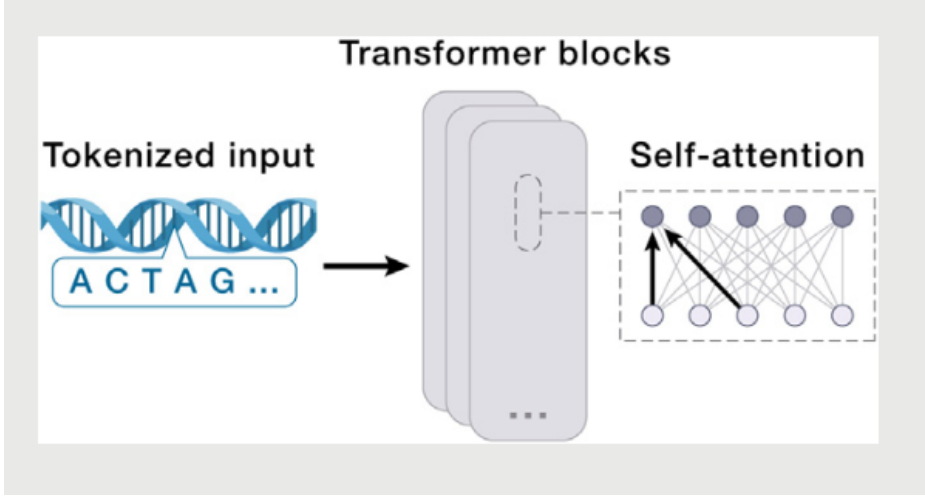
whole genome sequencing
cryo-EM
x-Ray crystals

scRNA-seq
fluorescence microscopy
scATAC-seq
CITE-seq

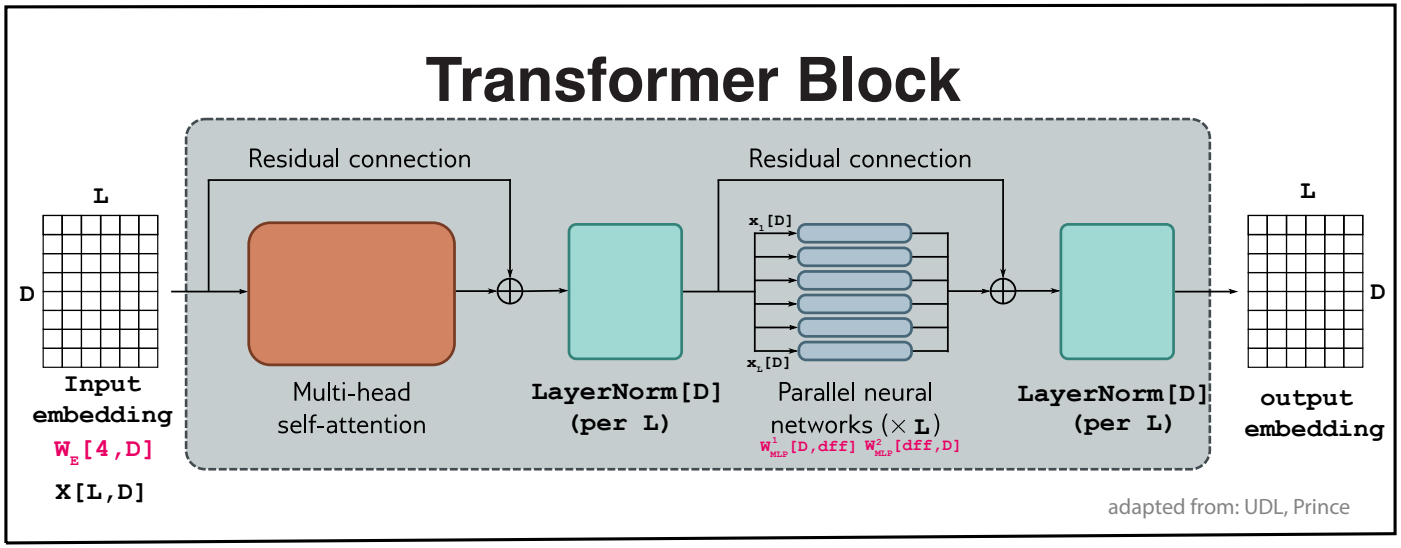
spatial transcriptomics

Methods

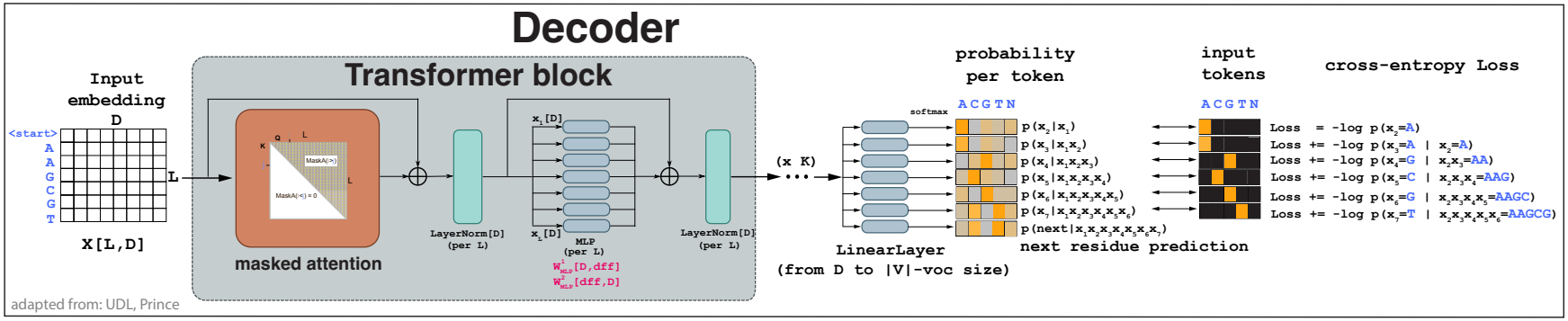
Transformers

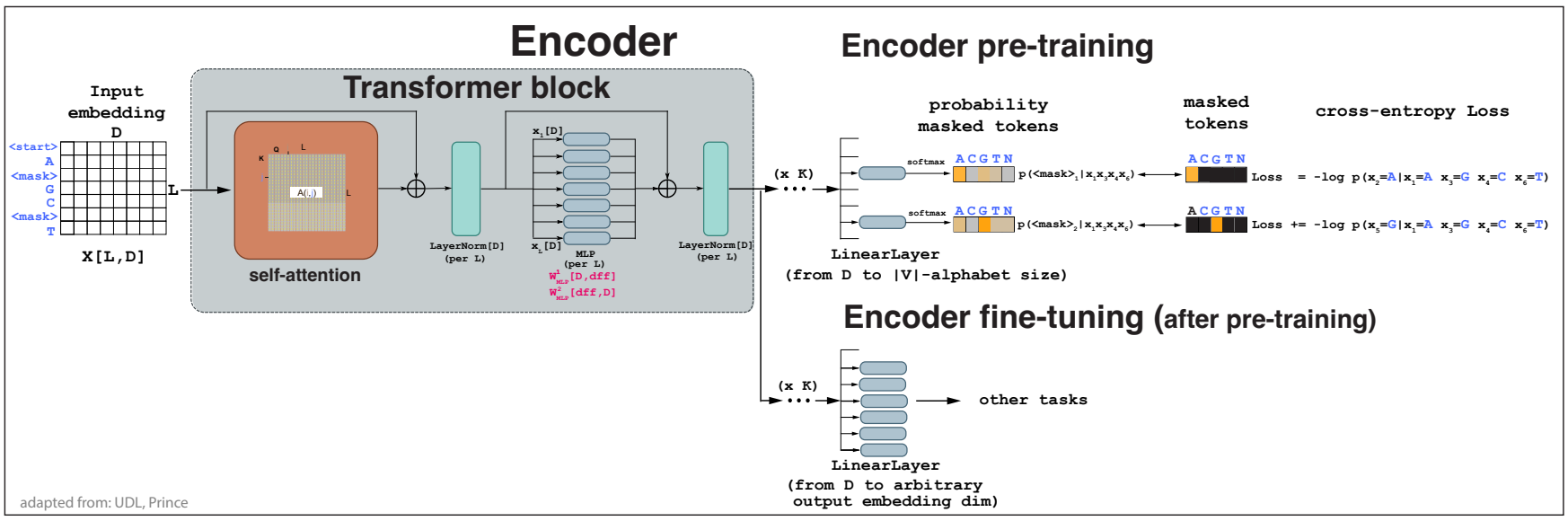


Methods

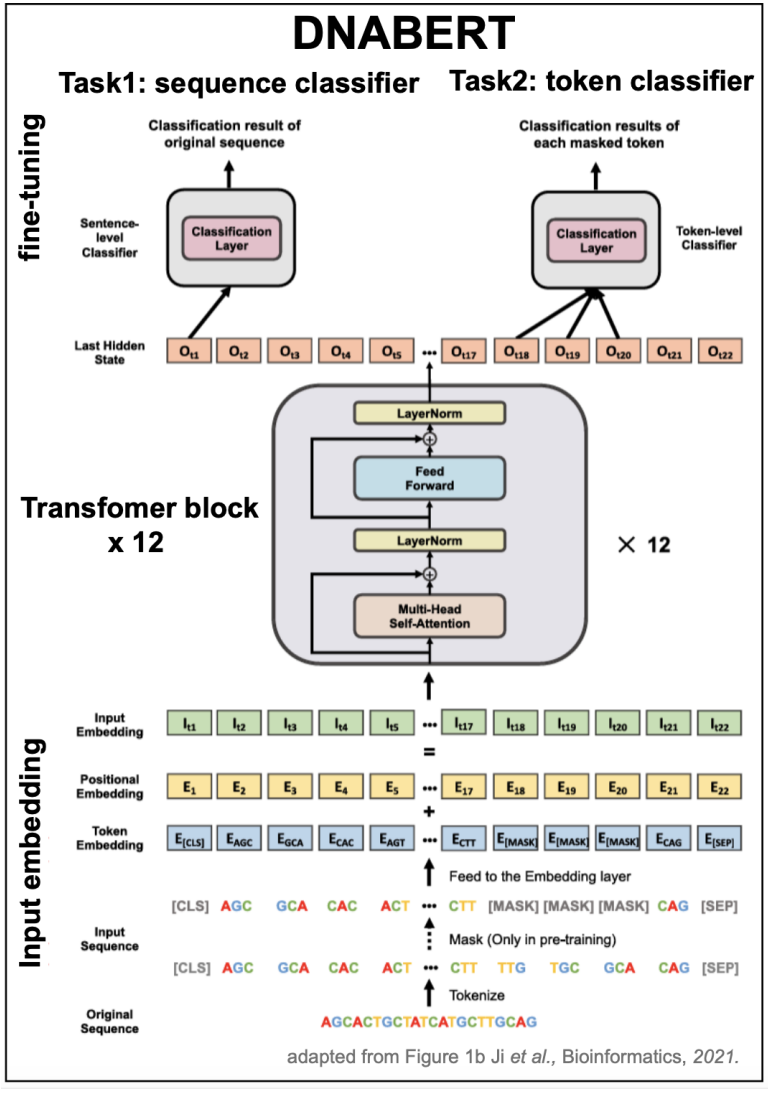

Transformer Block
Residual connection
Residual connection
L
D
Input embedding
W_E[4,D]
X[L,D]
Multi-head self-attention
LayerNorm[D] (per L)
x_1[D]
x_L[D]
Parallel neural networks (× L)
W^1_MLP[D,dff] W^2_MLP[dff,D]
LayerNorm[D] (per L)
L
D
output embedding
adapted from: UDL, Prince

Methods


Decoder
Input embedding D
<start>
A
A
G
C
G
T
L
X[L,D]
Transformer block
masked attention
LayerNorm[D] (per L)
MLP (per L)
LayerNorm[D] (per L)
(x K)
probability per token
softmax
A C G T N
LinearLayer
(from D to |V|-voc size)
next residue prediction
input tokens
cross-entropy Loss
Loss = -log p(x2=A)
Loss += -log p(x3=A | x2=A)
Loss += -log p(x4=G | x2x3=AA)
Loss += -log p(x5=C | x2x3x4=AAG)
Loss += -log p(x6=G | x2x3x4x5=AAGC)
Loss += -log p(x7=T | x2x3x4x5x6=AAGCG)
adapted from: UDL, Prince

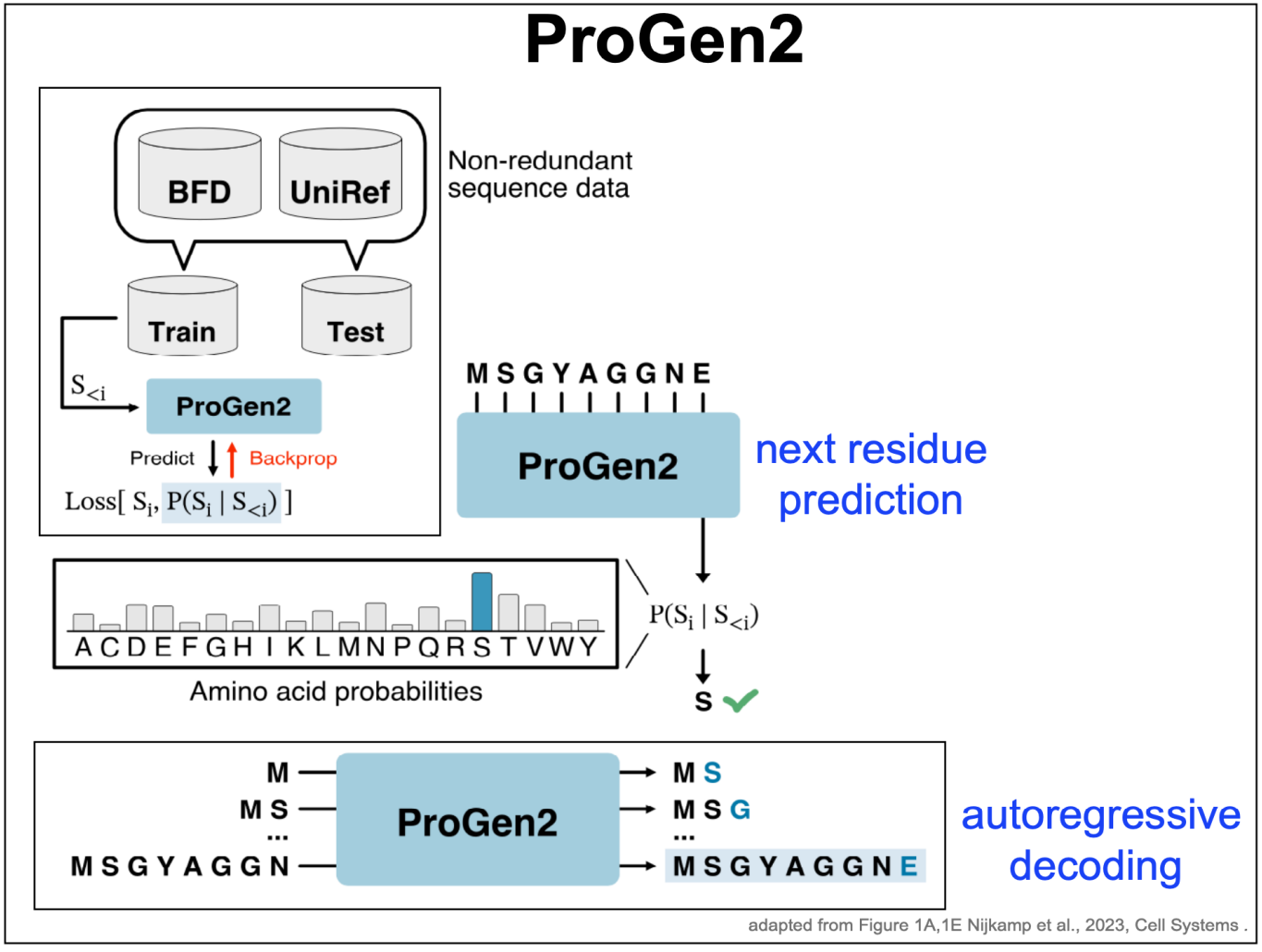

Encoder
Input embedding D
<start>
A
<mask>
G
C
<mask>
T
L
X[L,D]
Transformer block
self-attention
LayerNorm[D] (per L)
MLP (per L)
LayerNorm[D] (per L)
(x K)
Encoder pre-training
probability masked tokens
masked tokens
cross-entropy Loss
softmax
A C G T N
LinearLayer
(from D to |V|-alphabet size)
Loss = -log p(x2=A|x1=A x3=G x4=C x6=T)
Loss += -log p(x5=G|x1=A x3=G x4=C x6=T)
Encoder fine-tuning (after pre-training)
other tasks
LinearLayer
(from D to arbitrary output embedding dim)
adapted from: UDL, Prince

Methods



adapted from Figure 1b Ji *et al.*, Bioinformatics, *2021.*

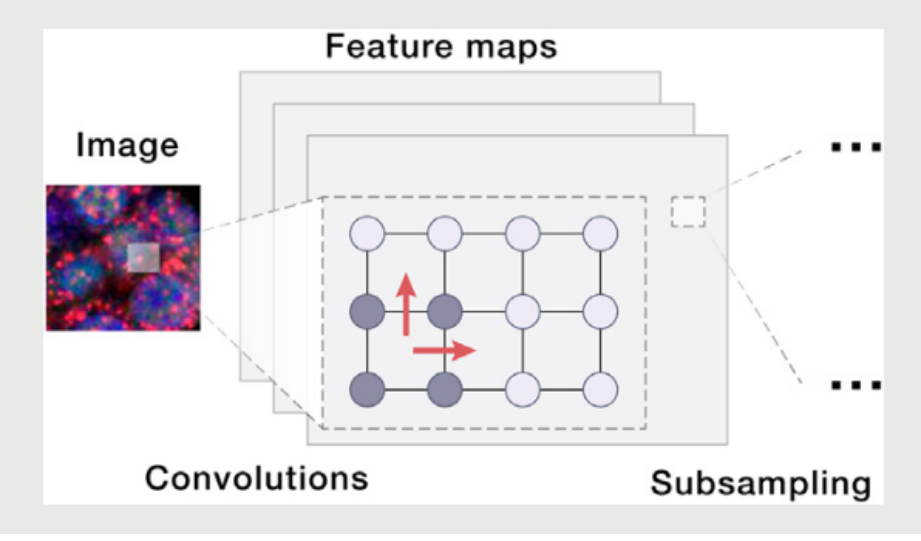
Methods



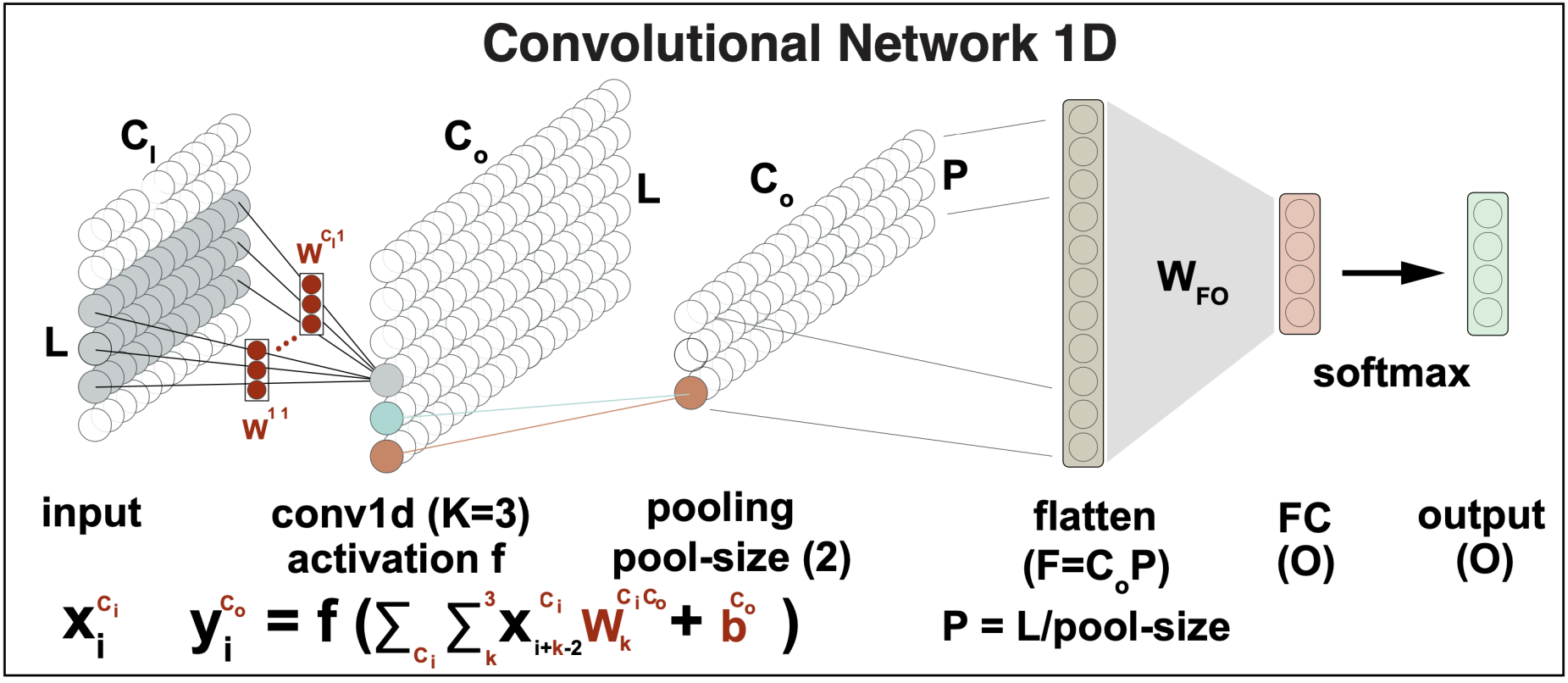
adapted from Figure 1A,1E Nijkamp et al., 2023, Cell Systems .

Methods

CNNs



Methods


Convolutional Network 1D
C_I
C_o
L
C_o
P
L
$W^{C_I 1}$
$W^{1 1}$
W_{FO}
softmax
input
conv1d (K=3)
activation f
pooling
pool-size (2)
flatten
(F=C_oP)
FC
(O)
output
(O)
$x_i^{C_i}$
$y_i^{C_o} = f\left(\sum_{C_i}\sum_{k}^{3} x_{i+k-2}^{C_i} W_k^{C_i C_o} + b^{C_o}\right)$
P = L/pool-size

Methods

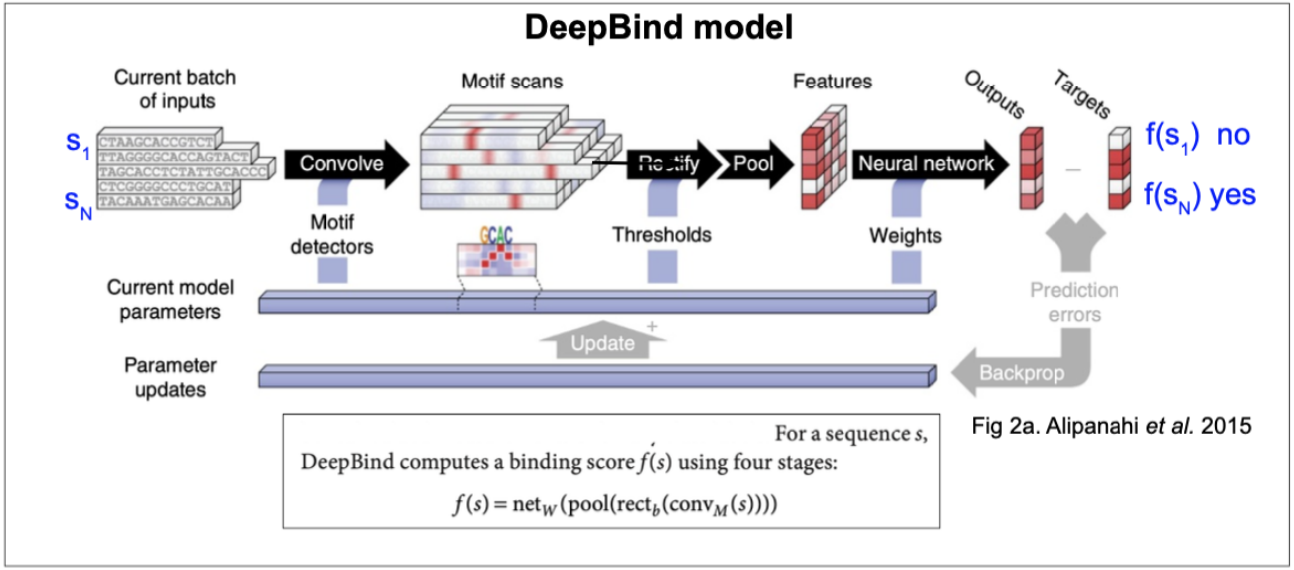
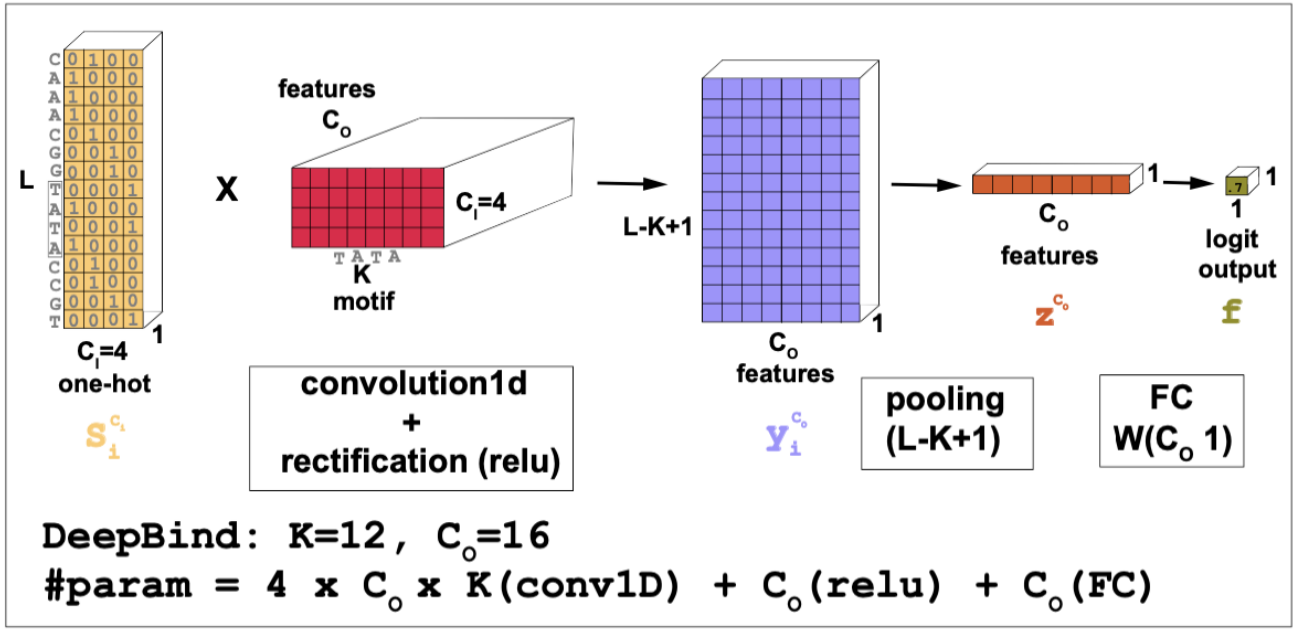
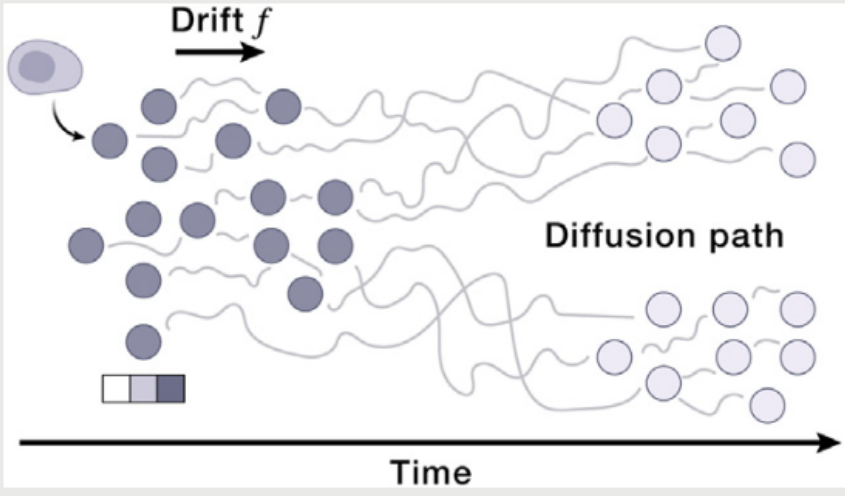


Fig 2a. Alipanahi *et al.* 2015



Methods

Denoising Diffusion Probabilistic models



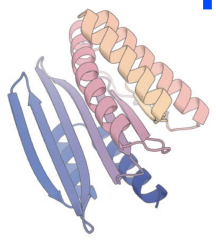
DIFFUSION Training

$t \sim \text{Uniform}[1...0]$

Neural Network

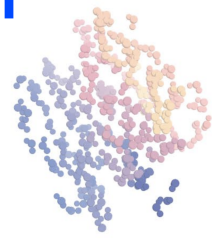
forward diffusion

backward diffusion $\epsilon_\Theta(X_t)$



data X_1

$t=1$



X_t

$N(0,I)$

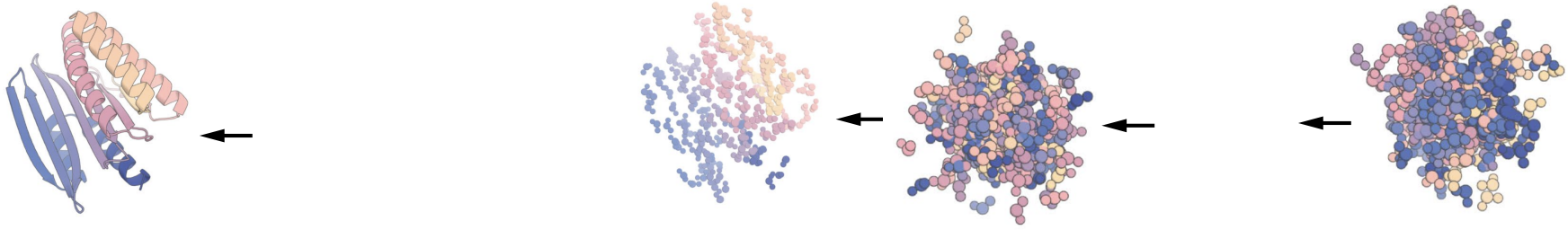
$$X_t = \sqrt{\alpha_t}\, X_1 + \sqrt{1-\alpha_t}\, \epsilon$$

$$\text{Loss} = \| \epsilon - \epsilon_\Theta(X_t) \|^2$$

Methods

DIFFUSION Sampling

backward diffusion



X_1 t=1

$X_{t+1=0.2}$

$X_{t=0.1}$

X_0

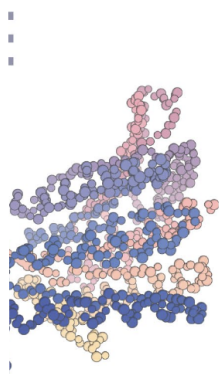
substracts noise

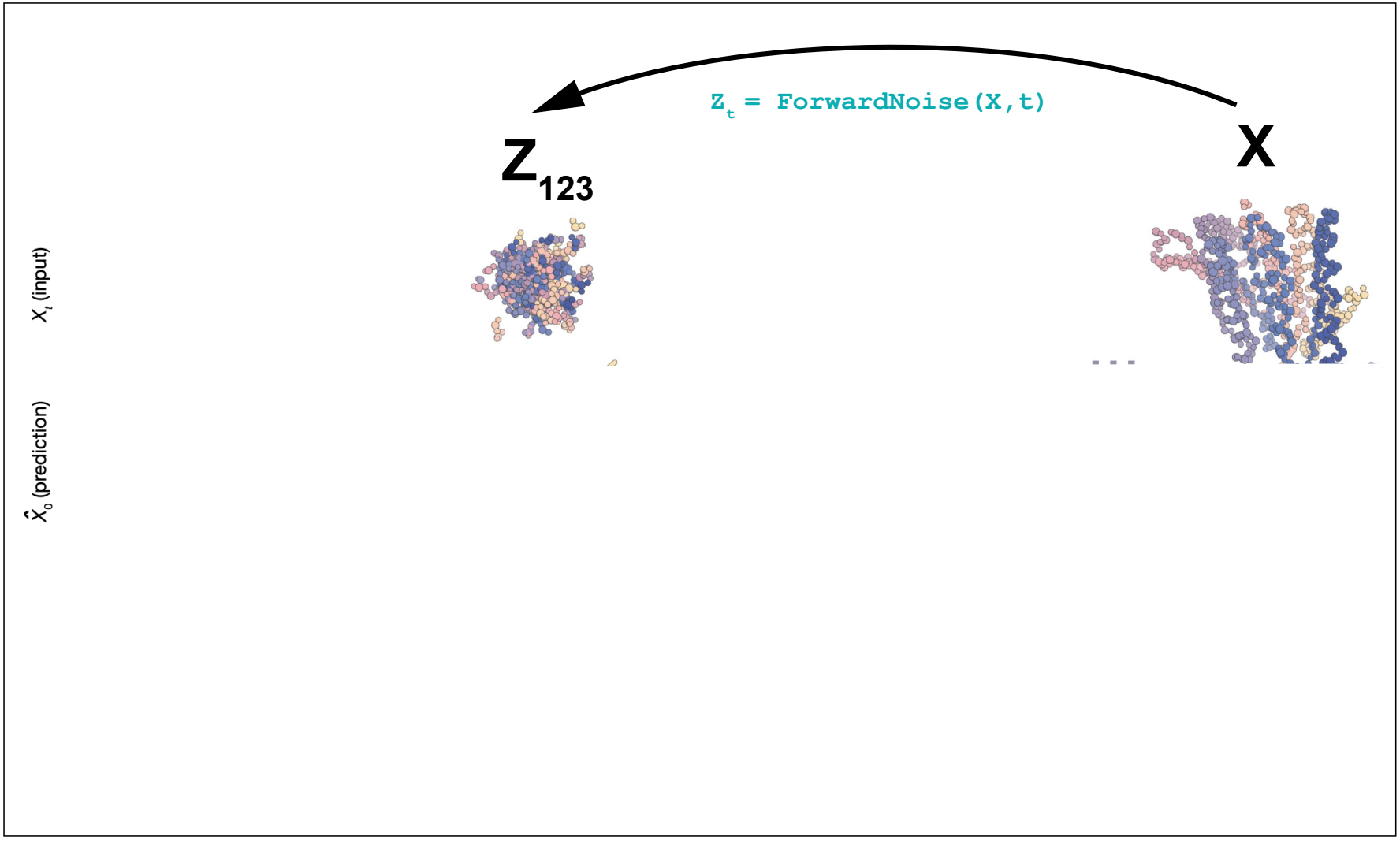
Neural Network

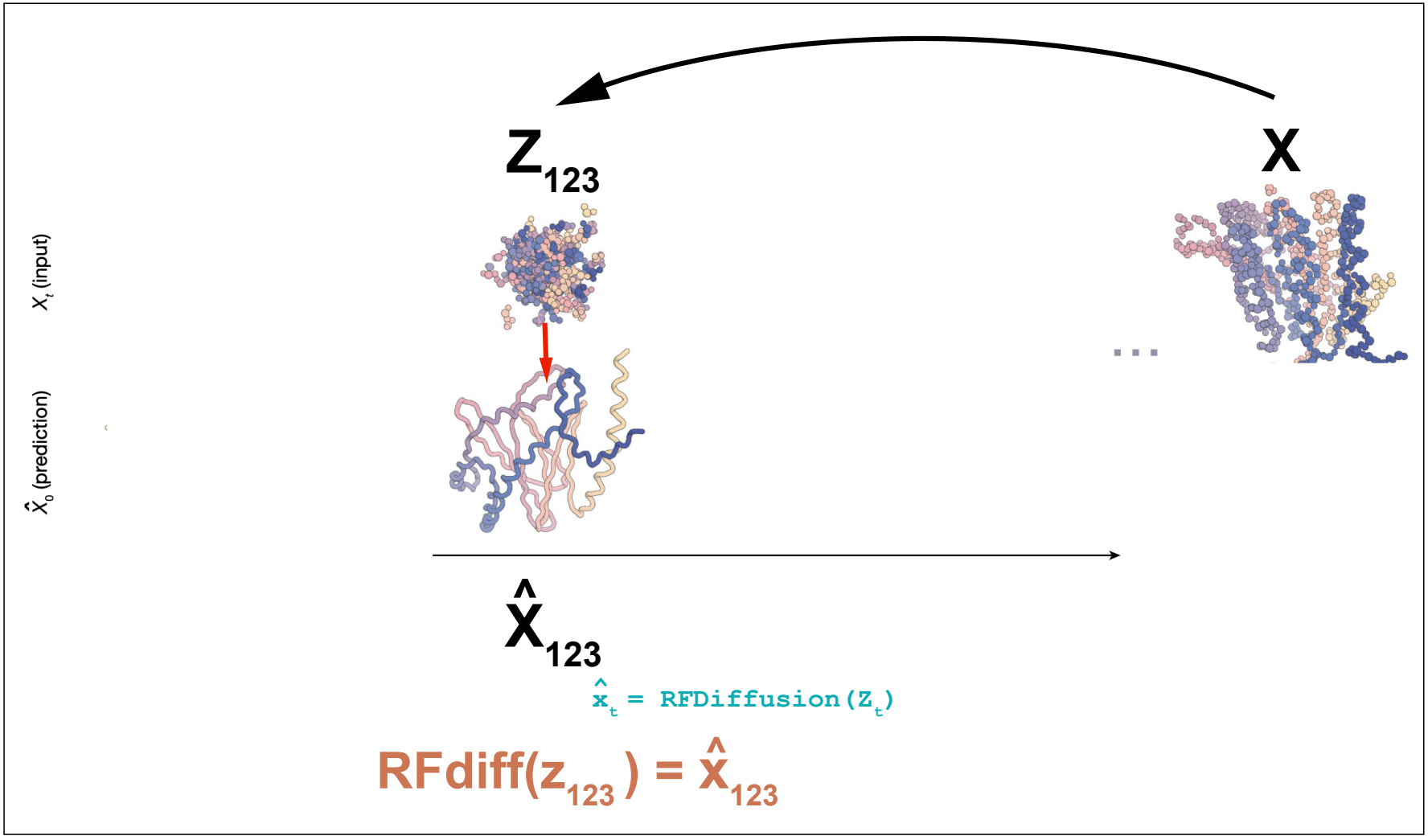
$$X_{t+1} = \frac{1}{\sqrt{1-\beta_t}} X_t - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}} \epsilon_\Theta^t(X_t)$$

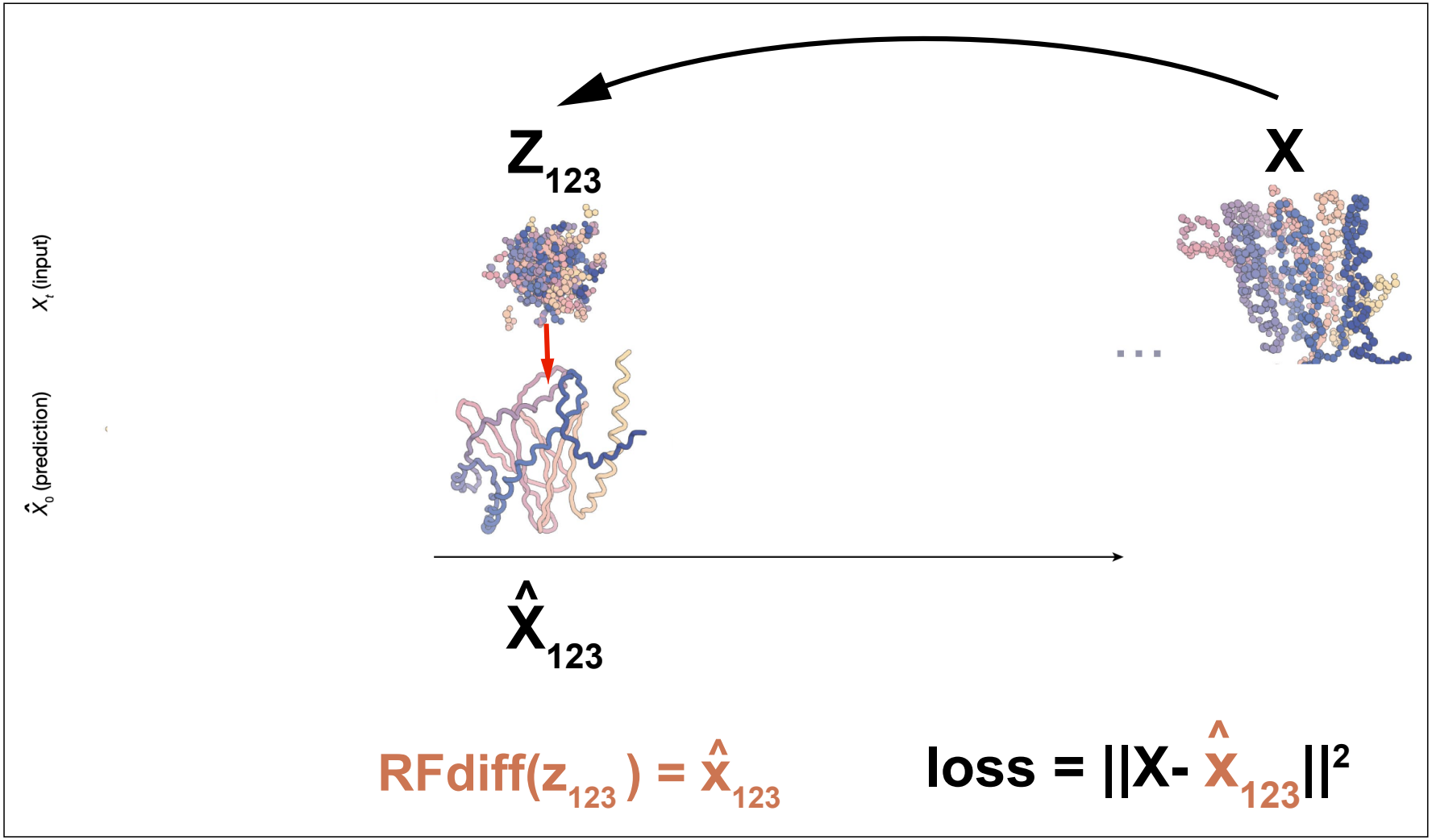
$\hat{X}_0$ (prediction)

X_t (input)

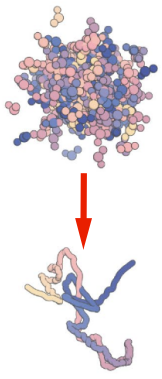



Z_t = ForwardNoise(X,t)
Z_{123}
X
X_t (input)
$\hat{X}_0$ (prediction)


Z_{123}
X
X_t (input)
$\hat{X}_0$ (prediction)
...
$\hat{X}_{123}$
$\hat{x}_t = \text{RFDiffusion}(Z_t)$
$\text{RFdiff}(z_{123}) = \hat{x}_{123}$


Z_{123}
X
X_t (input)
$\hat{X}_0$ (prediction)
$\hat{X}_{123}$
RFdiff(z_{123}) = $\hat{x}_{123}$
loss = ||X- $\hat{x}_{123}$||²

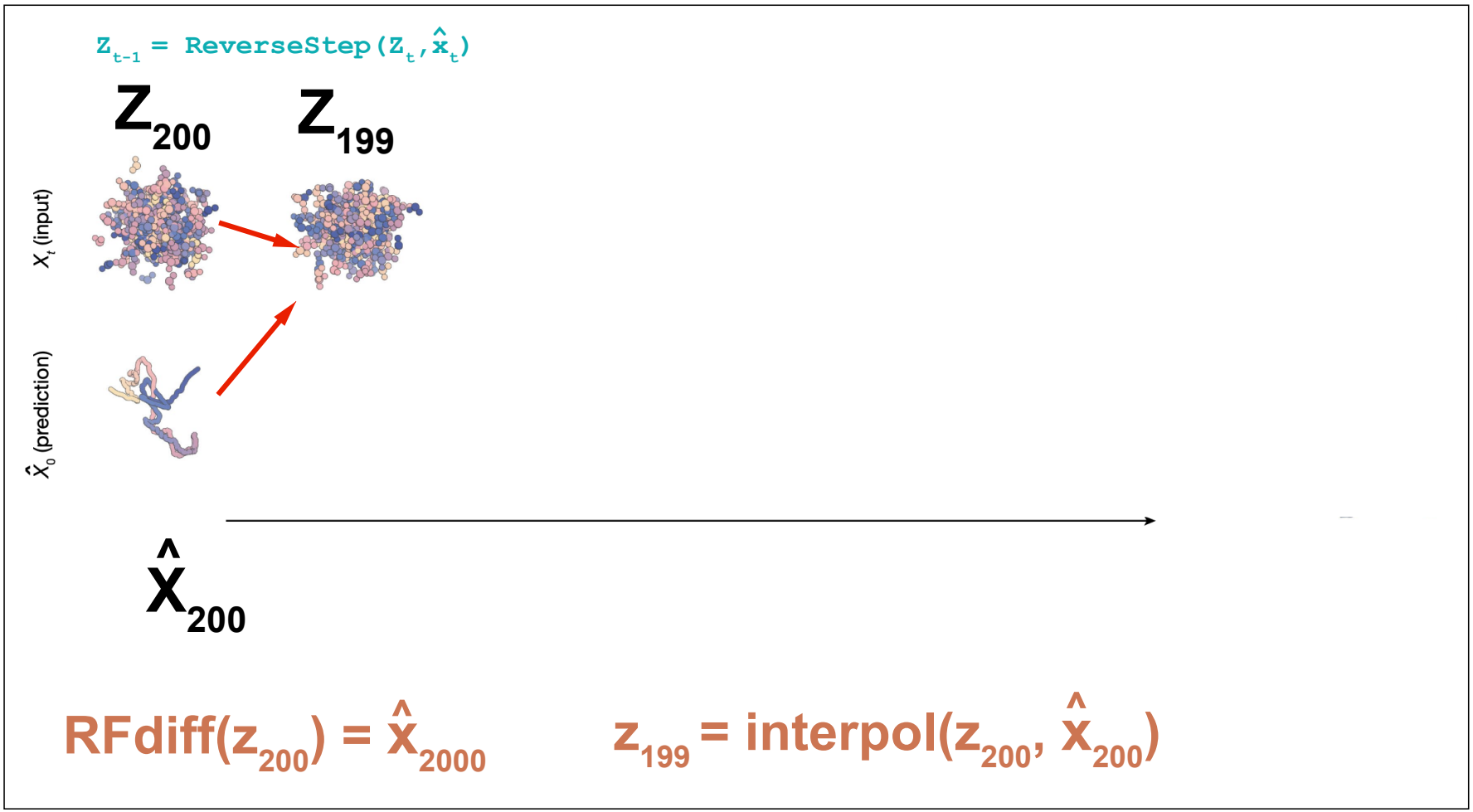
Z_{200}



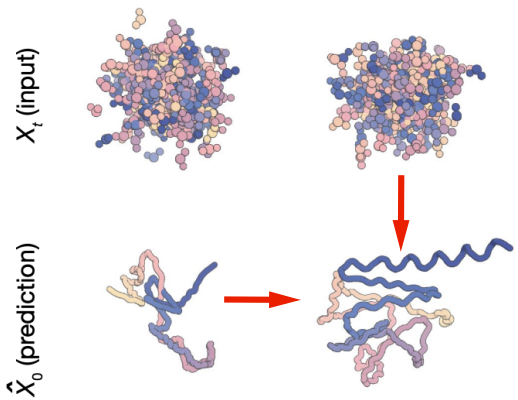
$\hat{X}_{200}$

$\hat{x}_t = \texttt{RFDiffusion}(Z_t)$

$\text{RFdiff}(z_{200}) = \hat{x}_{2000}$


$Z_{t-1} = \text{ReverseStep}(Z_t, \hat{x}_t)$
Z_{200}
Z_{199}
X_t (input)
$\hat{X}_0$ (prediction)
$\hat{X}_{200}$
$\text{RFdiff}(z_{200}) = \hat{x}_{2000}$
$z_{199} = \text{interpol}(z_{200}, \hat{x}_{200})$

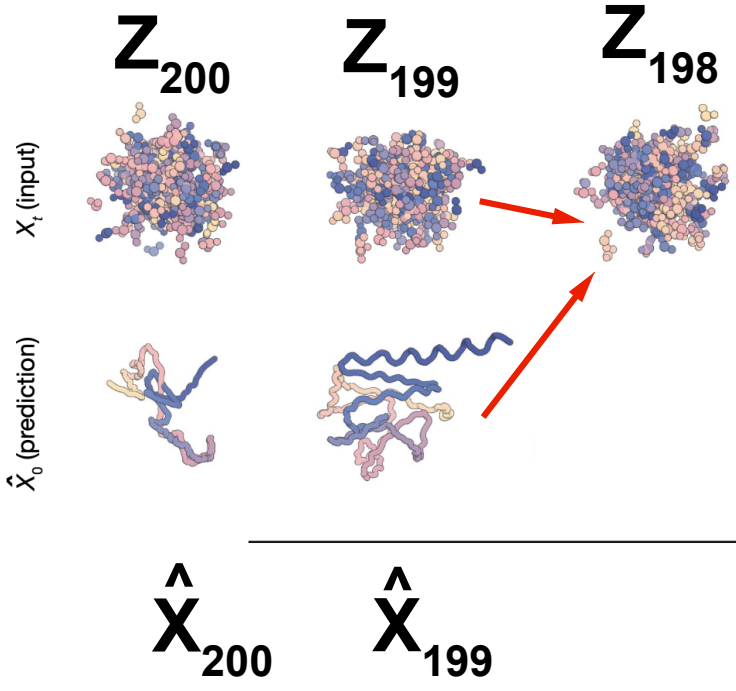
Z_{200} Z_{199}



$\hat{X}_{200}$ $\hat{x}_t = \text{RFDiffusion}(Z_t, \hat{x}_{t+1})$

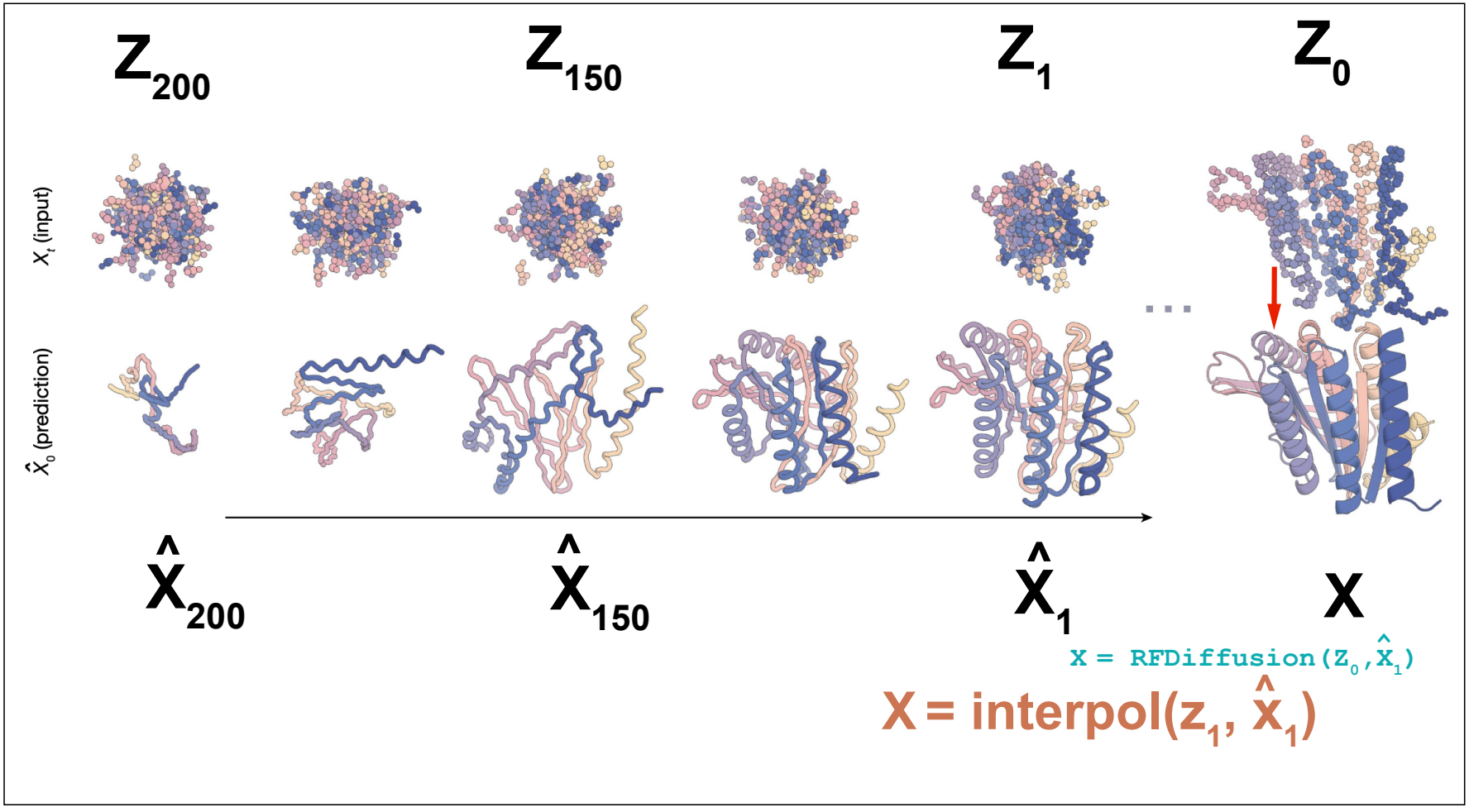
$\text{RFdiff}(z_{199}, \hat{x}_{200}) = \hat{x}_{199}$

$Z_{t-1} = \text{ReverseStep}(Z_t, \hat{x}_t)$



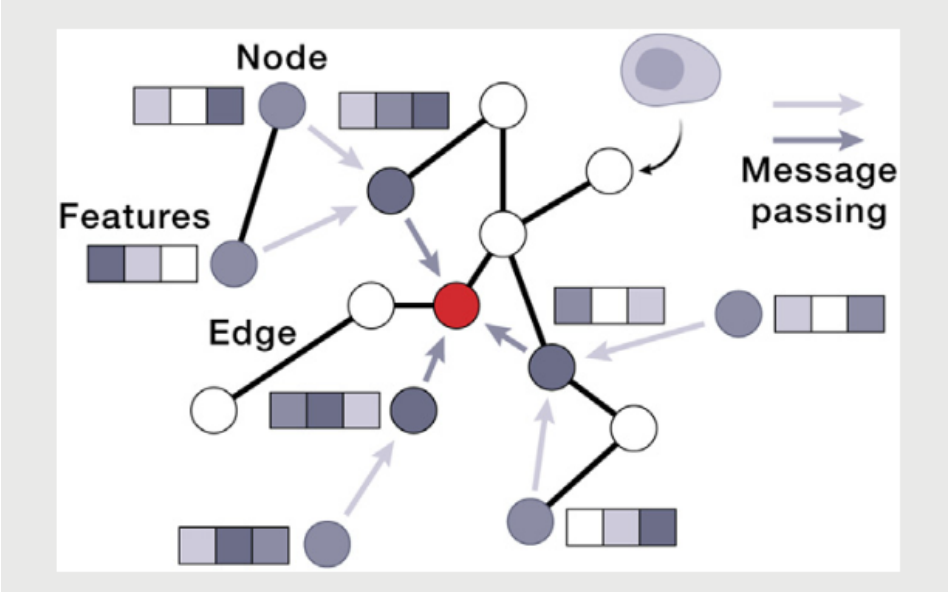
$\text{RFdiff}(z_{199}, \hat{x}_{200}) = \hat{x}_{199}$

$z_{198} = \text{interpol}(z_{199}, \hat{x}_{199})$


z_{200}
z_{150}
z_1
z_0
x_t (input)
$\hat{x}_0$ (prediction)
$\hat{x}_{200}$
$\hat{x}_{150}$
$\hat{x}_1$
x
$x = \text{RFDiffusion}(z_0, \hat{x}_1)$
$x = \text{interpol}(z_1, \hat{x}_1)$

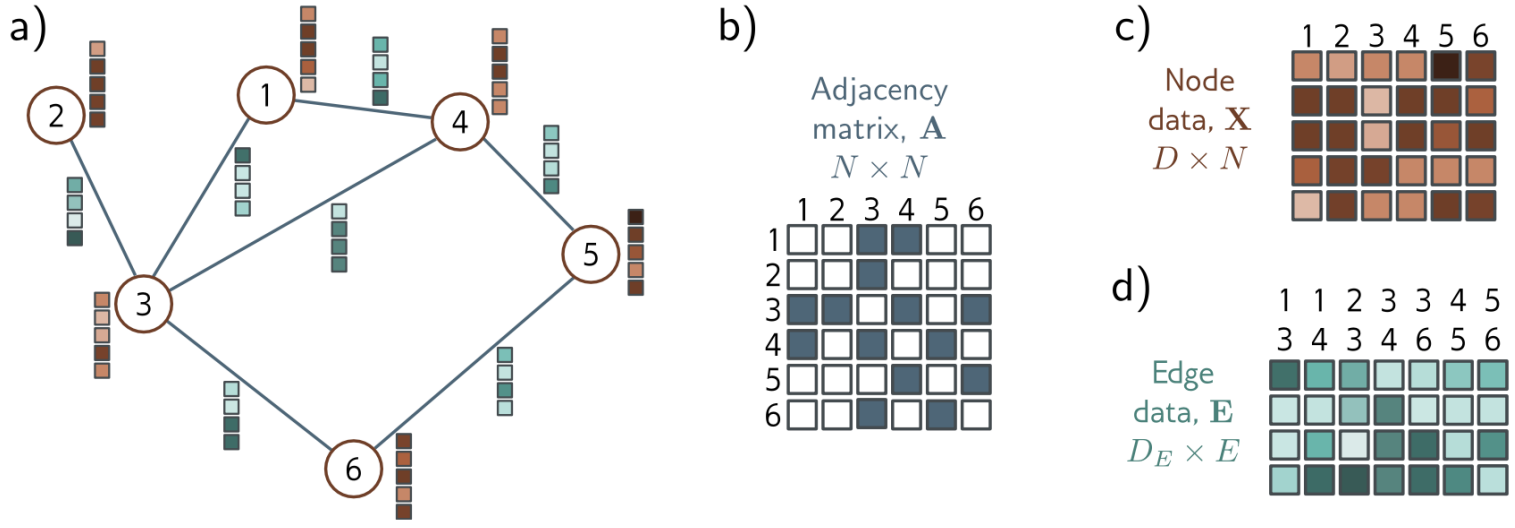
Methods

Graphical Neural Networks



Methods

Graphical Neural Networks



gene/protein networks

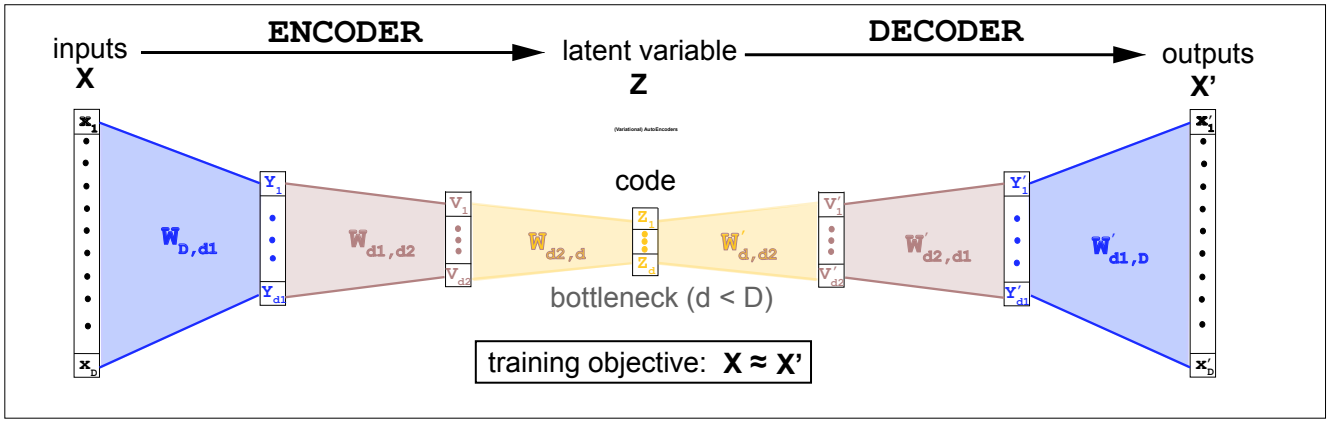
drug/drug interaction networks

cell/cell interaction networks

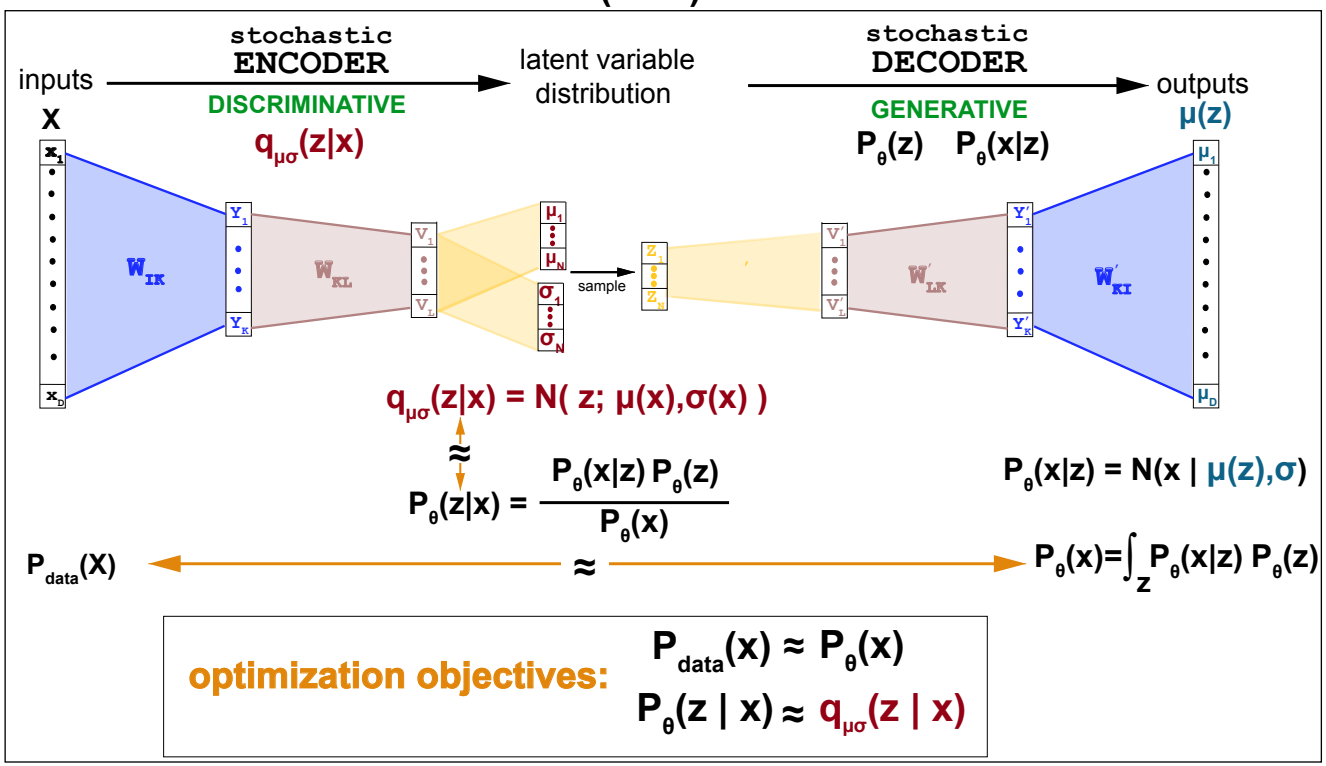
Graphical Convolutional Networks

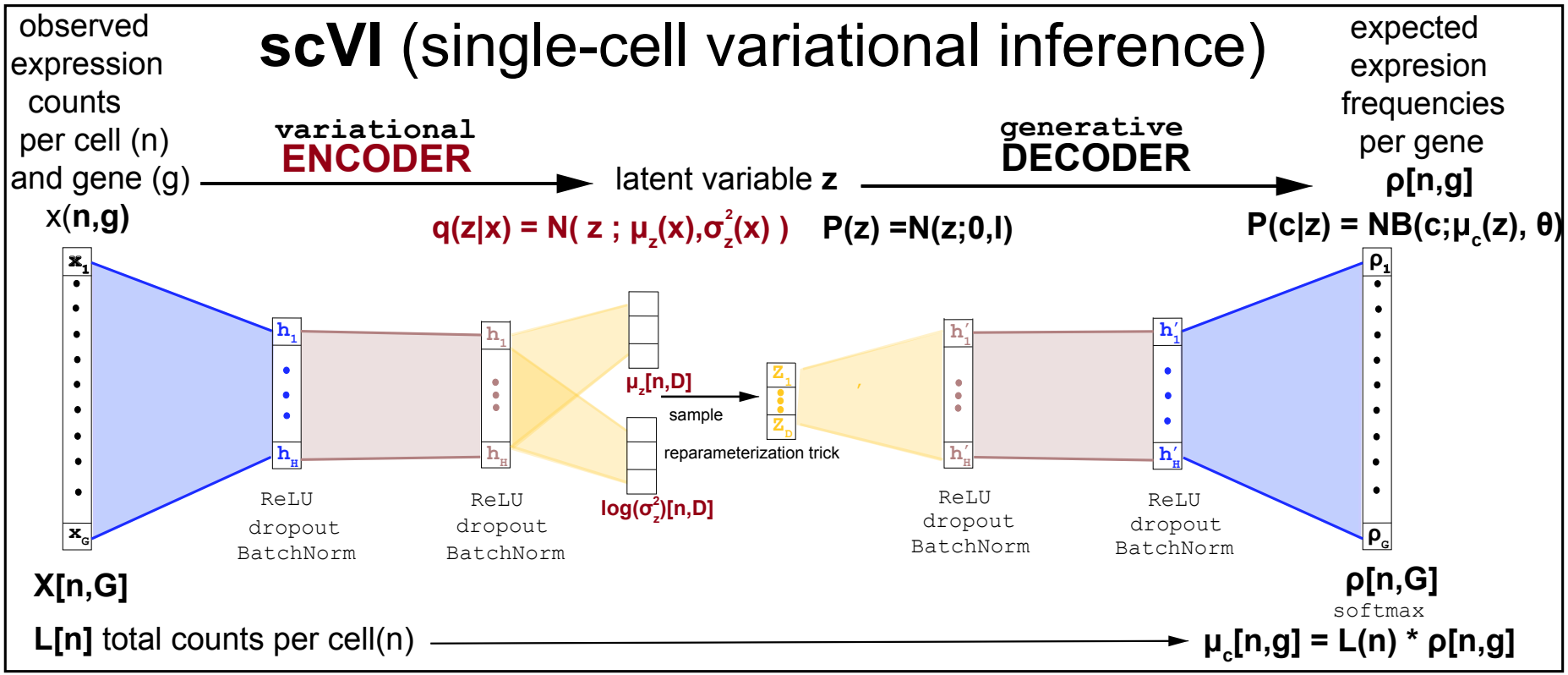
Methods

AUTOENCODER

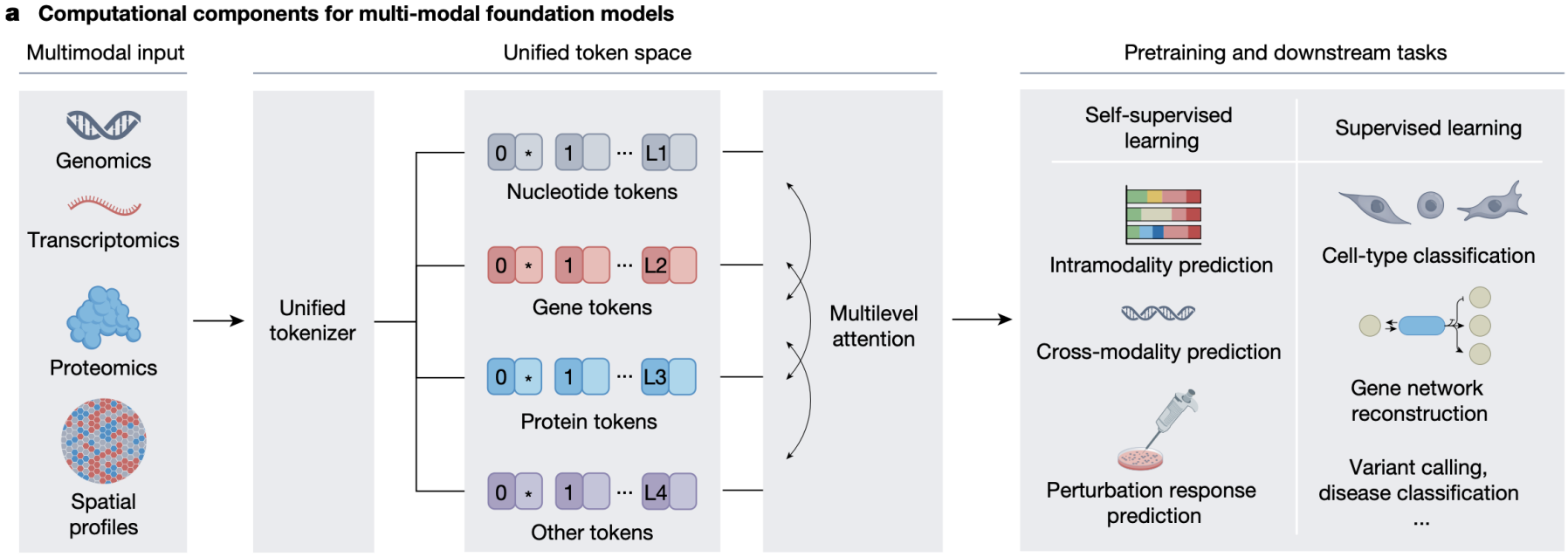


VARIATIONAL AUTOENCODER (VAE)



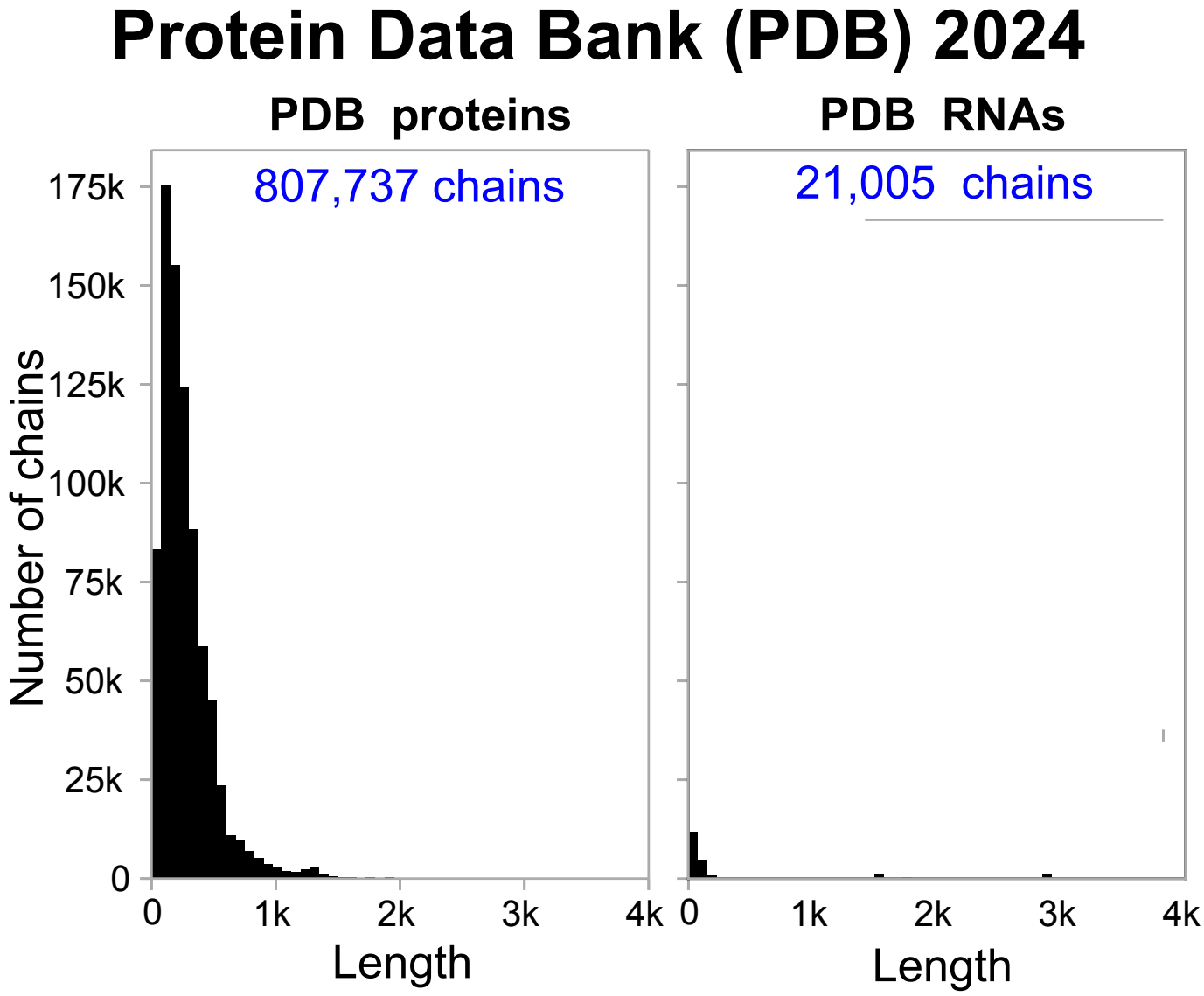

scVI (single-cell variational inference)
observed expression counts per cell (n) and gene (g) x(n,g)
variational ENCODER
latent variable z
generative DECODER
expected expresion frequencies per gene ρ[n,g]
q(z|x) = N(z ; μz(x),σ2z(x))
P(z) =N(z;0,I)
P(c|z) = NB(c;μc(z), θ)
ReLU dropout BatchNorm
ReLU dropout BatchNorm
μz[n,D]
sample
reparameterization trick
log(σ2z)[n,D]
ReLU dropout BatchNorm
ReLU dropout BatchNorm
X[n,G]
ρ[n,G]
softmax
L[n] total counts per cell(n)
μc[n,g] = L(n) * ρ[n,g]

Multimodal Foundational Models (MFMs)



'Towards multimodal foundation models in molecular cell biology", *Cui et al, Nat Perspectives, 2023-25*

Challenges - Always more data



Challenges - Always more data

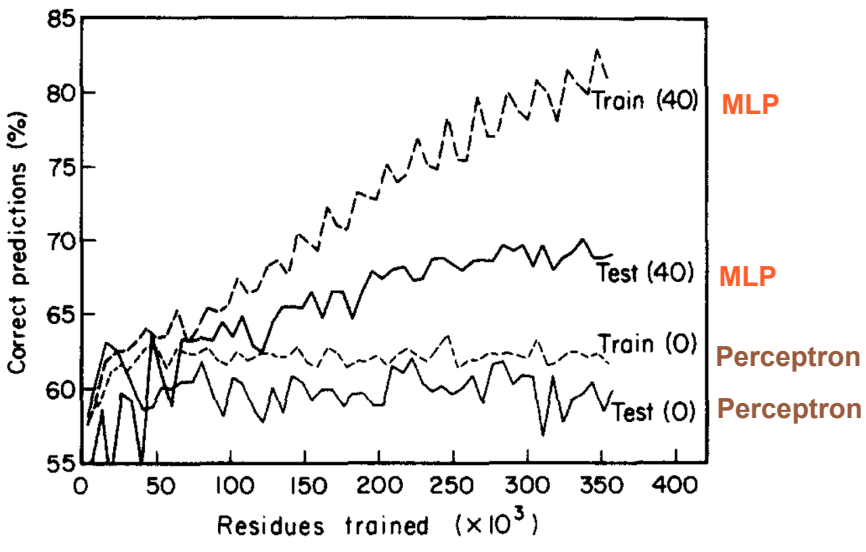
"Calling all data", Nath Meth Editorial, July 2025.

- Data generation w/o hypothesis hard to fund
- Data generation w/o hypothesis hard to publish
- Data storage costs money
- New experimental technologies to explore more of the data space needed

Challenges - Always more computational resources

Challenges - Always less data leakage:

Rigurous evaluation

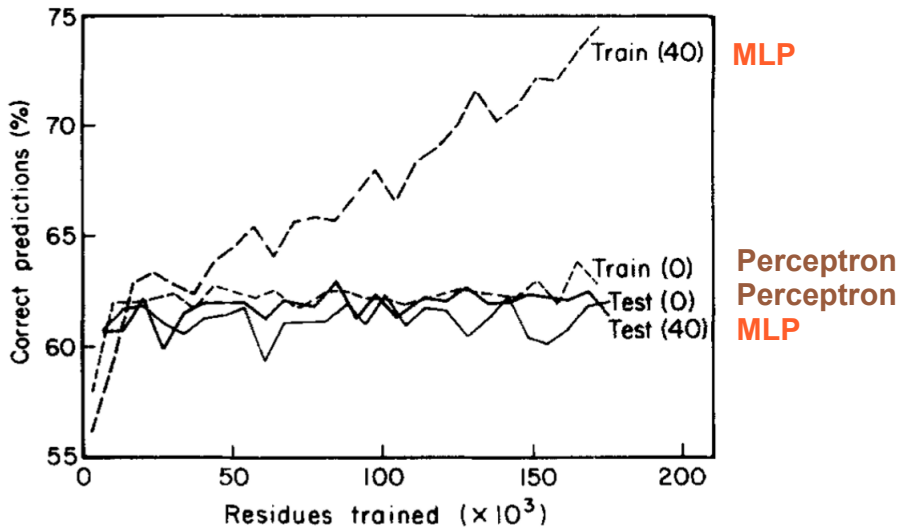


Qian-Sejnowski FIgure 14

Train set = Globin proteins

Test set = homologous Globin proteins

train/test overlap



Qian-Sejnowski FIgure 8

Train set = Globin proteins

Test set = non-homologous Globin proteins

train/test *less* overlap

Challenges - Always less hallucination:

To generate samples based on data

$$P_\theta \approx P_{data}$$

then

$$[P_\theta \sim] \, x_{pred} \approx x_{pdata} \, [\sim P_{data}]$$

Challenges - Always more ethical

1. Models should be open source
2. Data should be publicly accessible
3. Data privacy

Challenges - Always more Bayesian

Challenges - WHAT TO LEARN (teach)?

Thank You