Lecture 11

AI for Life Science – Selected Topics

Stan Z. Li

西湖大學
WESTLAKE UNIVERSITY

Outline

1. Introduction
2. Proteomics Models
3. Single-Cell Data Analysis
4. Some Future Directions

AI for Sciences

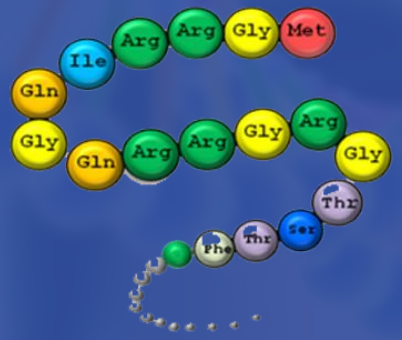
Optimization/ Decision Making

Protein Folding and Design

Drug Design and Synthesis

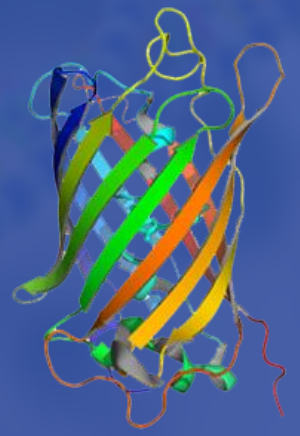
Protein Sequence – Structure – Function

Protein Sequence



Structure Prediction

Sequence Design

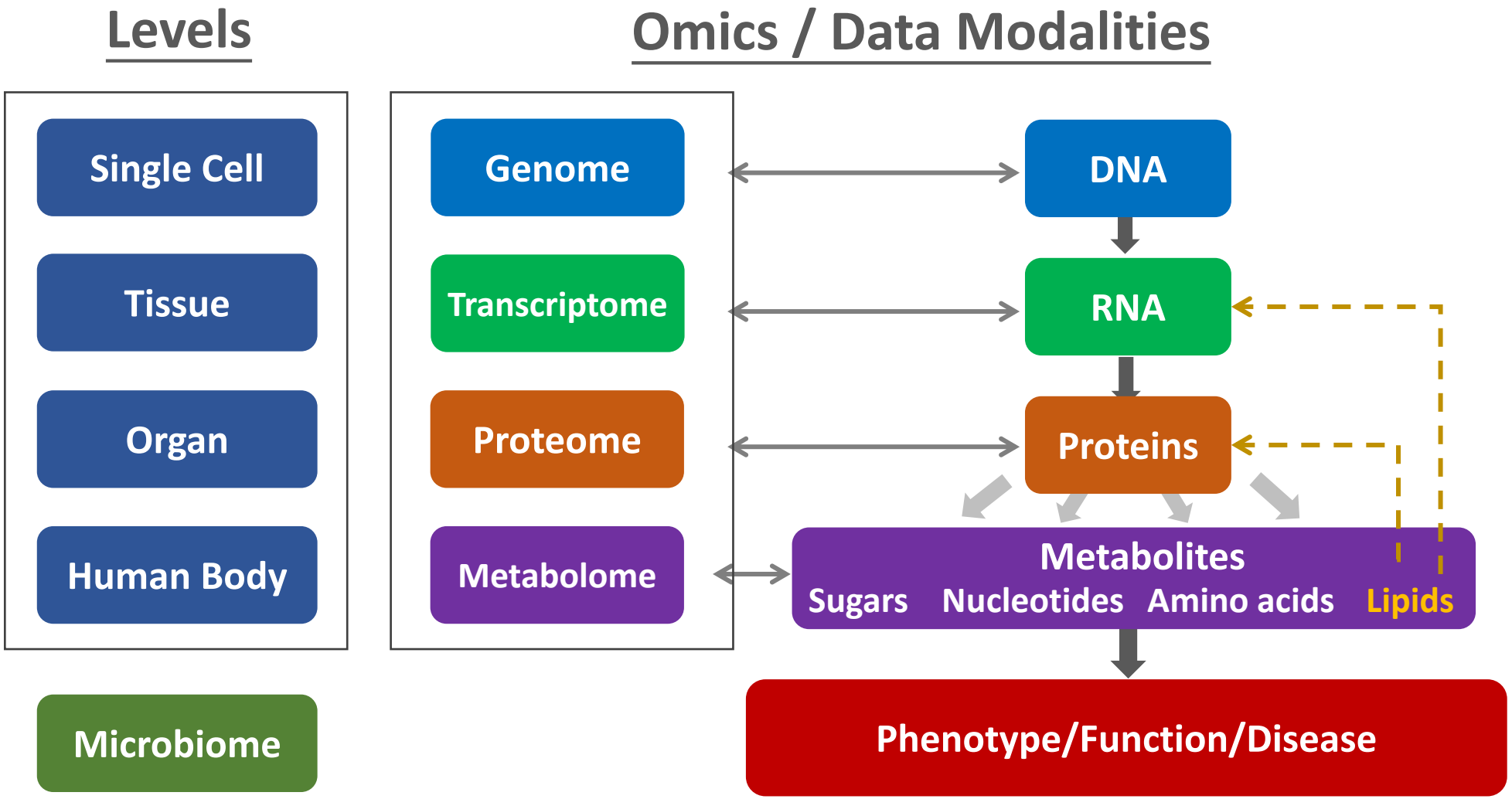


Protein Structure

Function

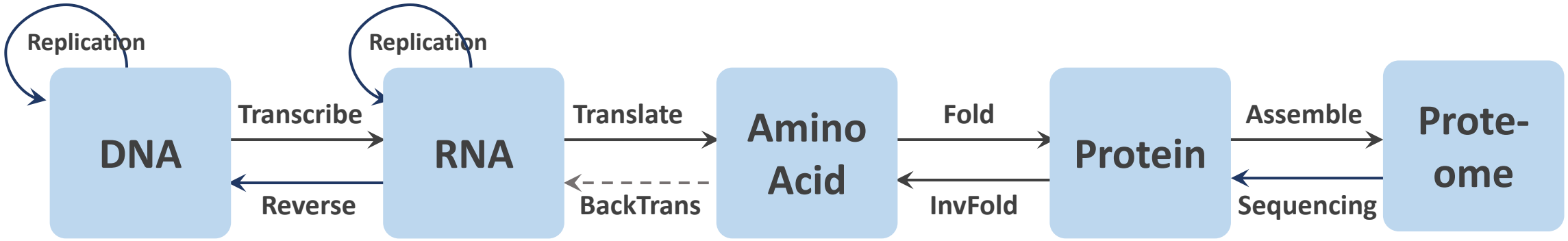
Drug Design

Omics in Life Science and Biomedicine



Source: ***Wikipedia***

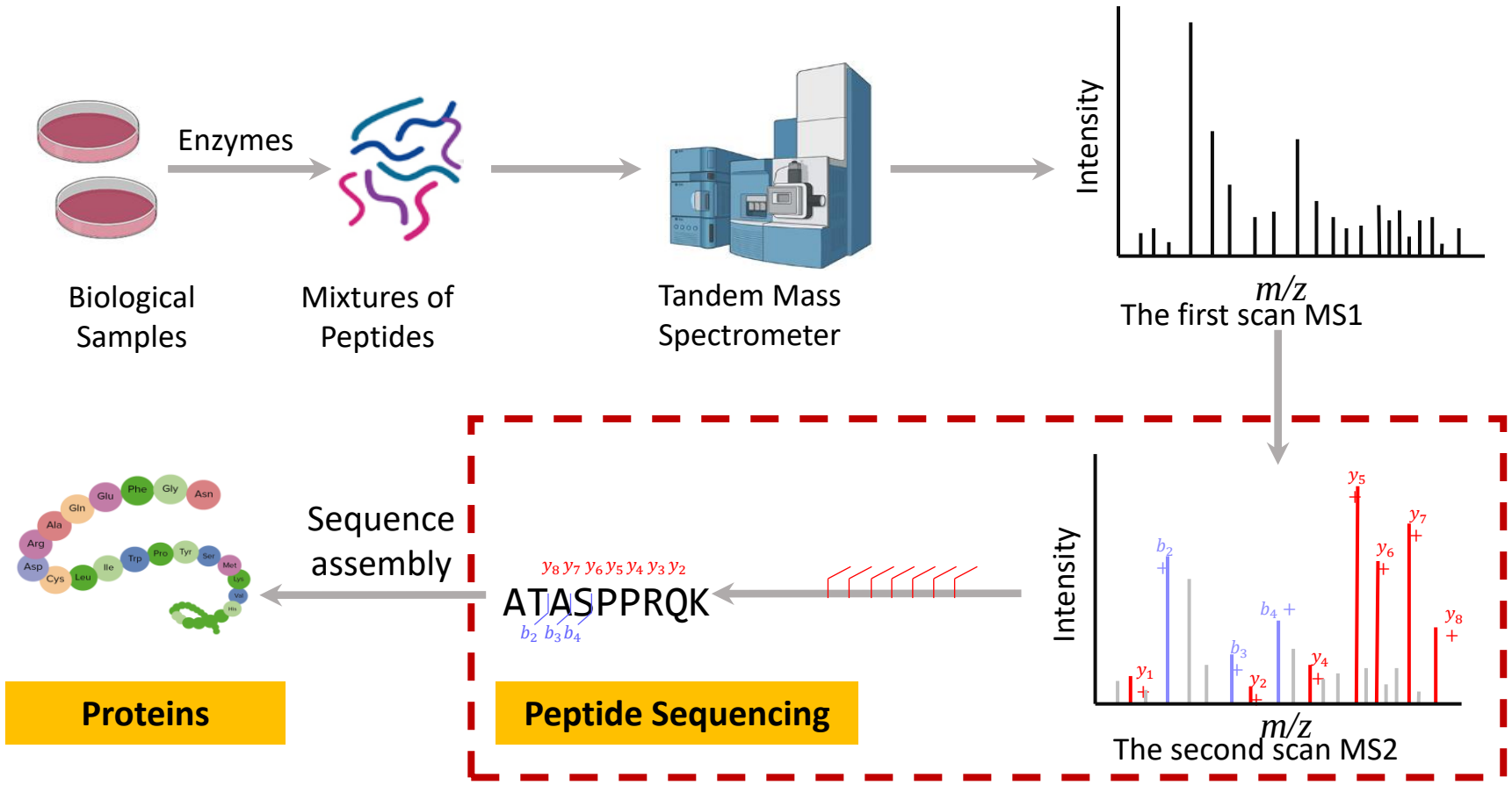
Biological Processes and Data



Outline

1. Introduction
2. Proteomics Models
   - Proteomics Empowered by Protein Language Models
   - De Novo Protein Sequencing
3. Single-Cell Data Analysis
4. Some Future Directions

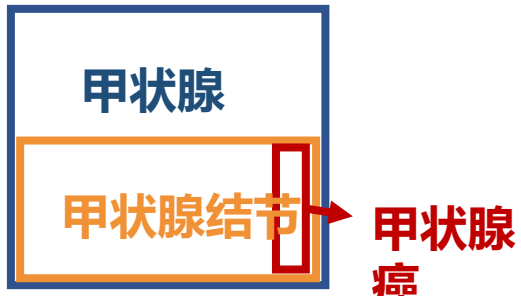
Proteomics Data Acquisition by MS Peptide Sequencing

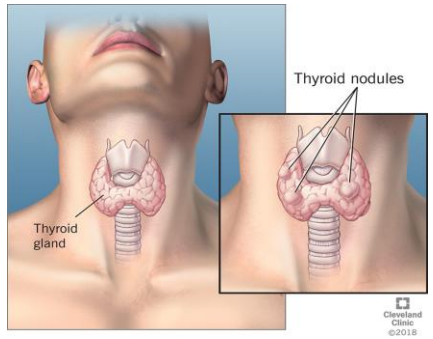


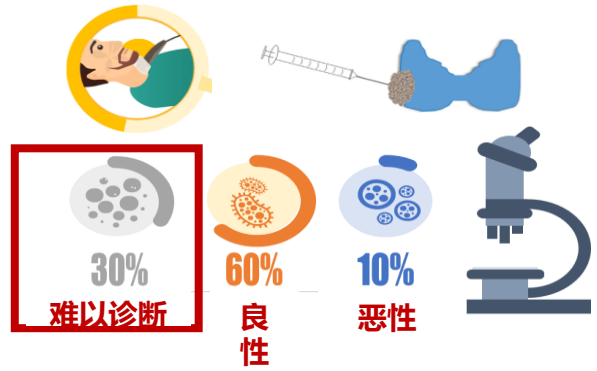
Challenges in Protein Sequencing

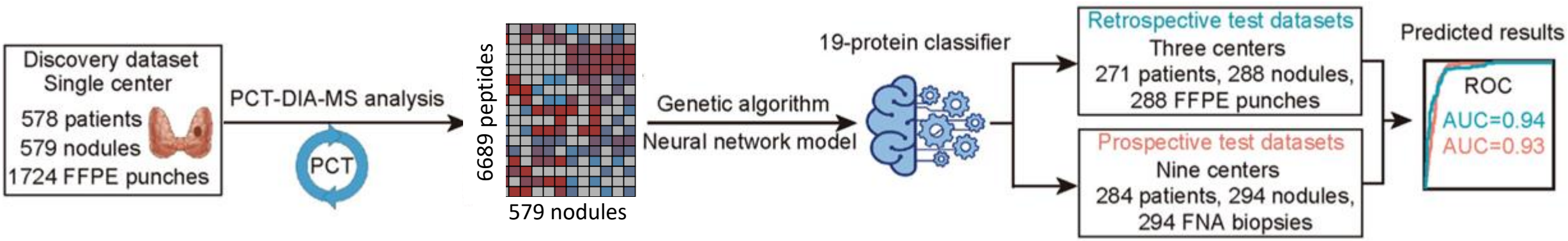
1. Individual proteomics-based tasks are limited by small data and foundations are needed to
2. Current database search methods are unable to identify new proteins (dark matters in proteomics).
3. Current de novo sequencing methods perform poorly in identifying post-translational modifications (PTMs)
4. Current spectrum prediction methods are limited by the difference of fragmentation types or instrument settings.

Case: AI-based Thyroid Nodule Diagnosis





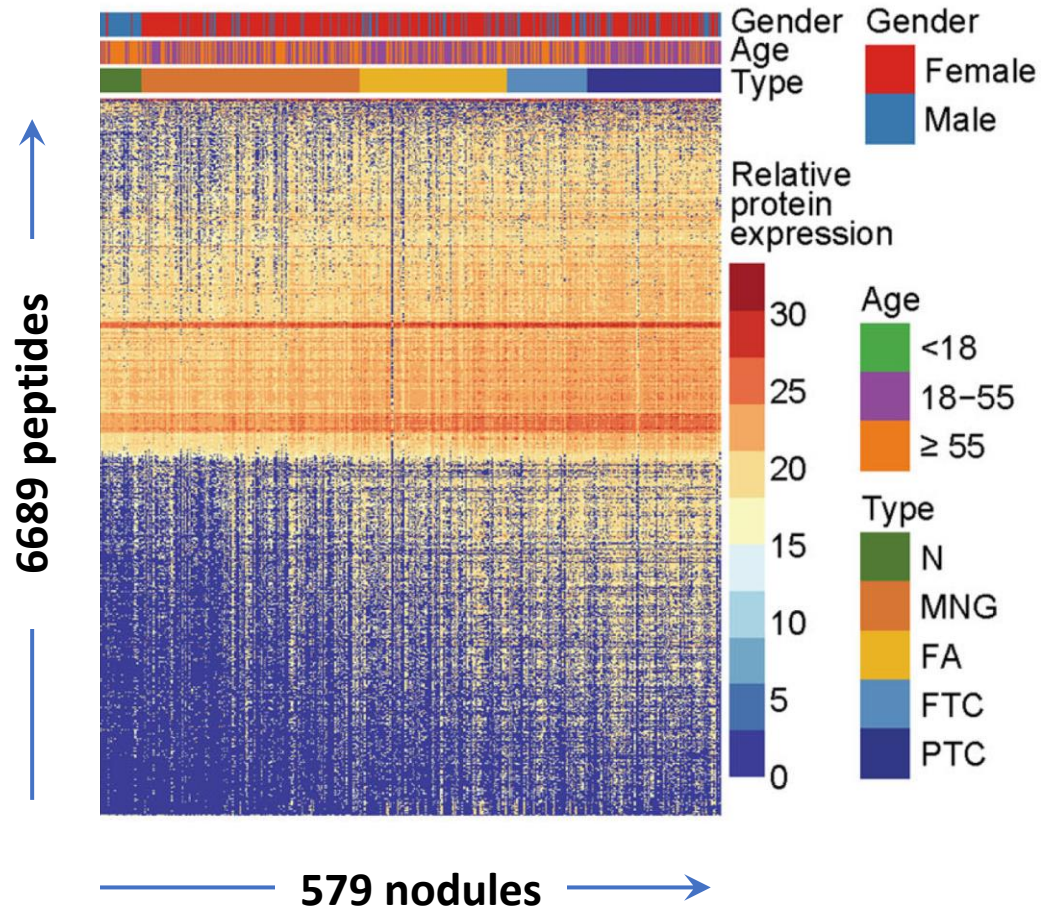


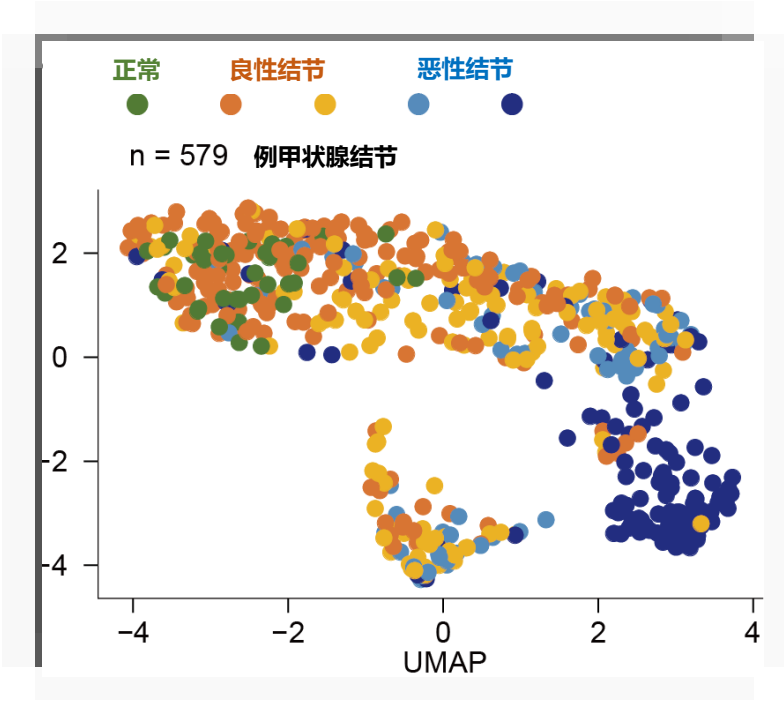


In Collaboration with Guomics Lab

AI Analysis of Proteomic Abundance Matrix

MS Data from Protein Sequencing

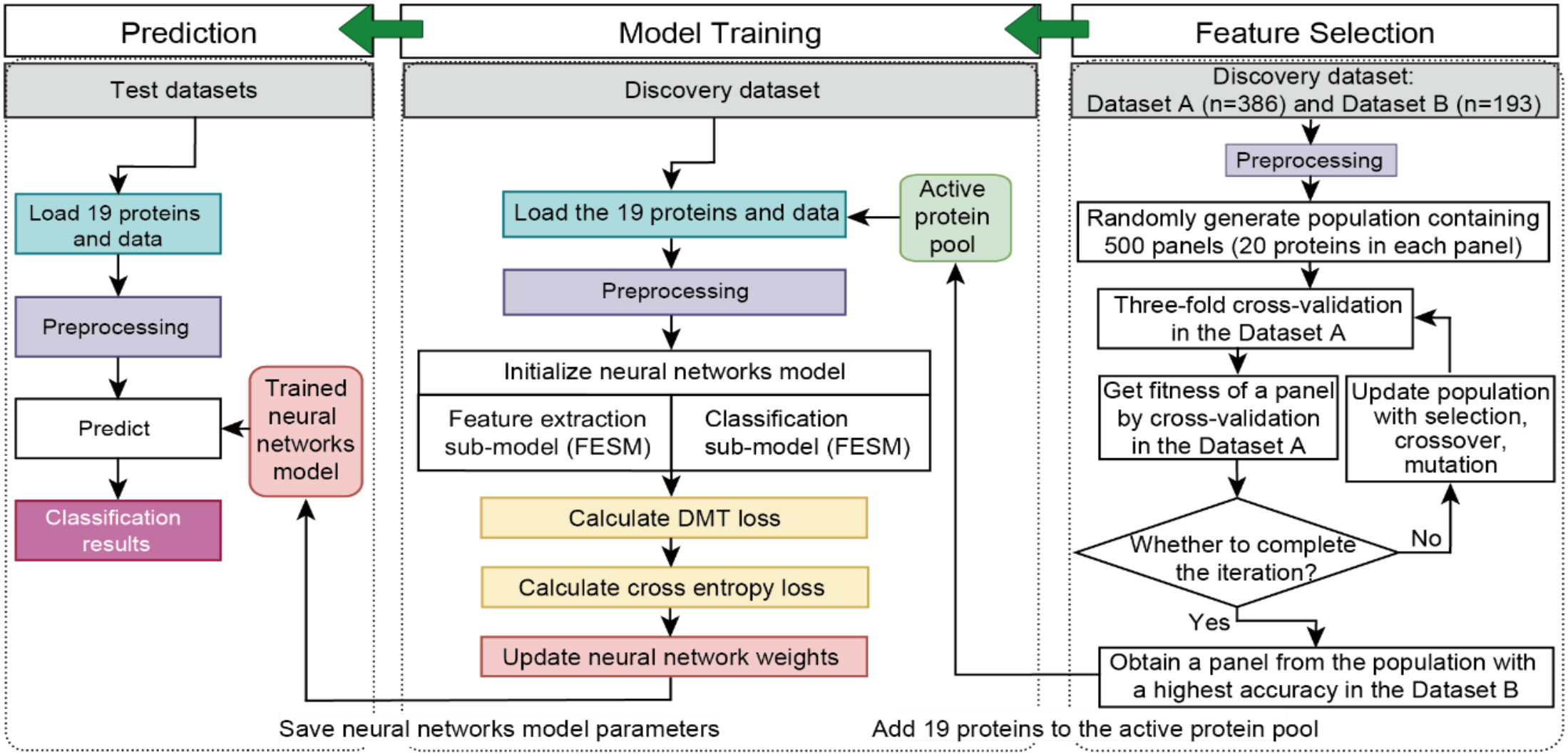




从蛋白质表达水平上，恶性结节与良性结节有一定区分度，但未完全区分开。

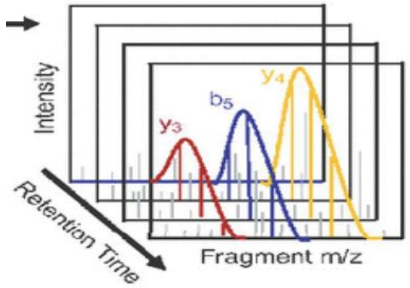
Sun, et al. “Artificial intelligence defines protein-based classification of thyroid nodules”, Cell Discovery 2022

AI Modeling


Prediction
Model Training
Feature Selection
Test datasets
Discovery dataset
Discovery dataset:
Dataset A (n=386) and Dataset B (n=193)
Load 19 proteins and data
Preprocessing
Predict
Classification results
Trained neural networks model
Load the 19 proteins and data
Active protein pool
Preprocessing
Initialize neural networks model
Feature extraction sub-model (FESM)
Classification sub-model (FESM)
Calculate DMT loss
Calculate cross entropy loss
Update neural network weights
Preprocessing
Randomly generate population containing 500 panels (20 proteins in each panel)
Three-fold cross-validation in the Dataset A
Get fitness of a panel by cross-validation in the Dataset A
Update population with selection, crossover, mutation
Whether to complete the iteration?
No
Yes
Obtain a panel from the population with a highest accuracy in the Dataset B
Save neural networks model parameters
Add 19 proteins to the active protein pool

PLM & PPI-Empowered Proteomics Model

Proteomics Data

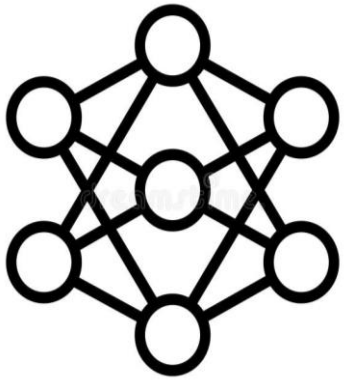


PLM & PPI

Protein Language Model
Protein-Protein Interaction

Diffusion-based Sample Generator

Proteomics Model



Downstream Tasks

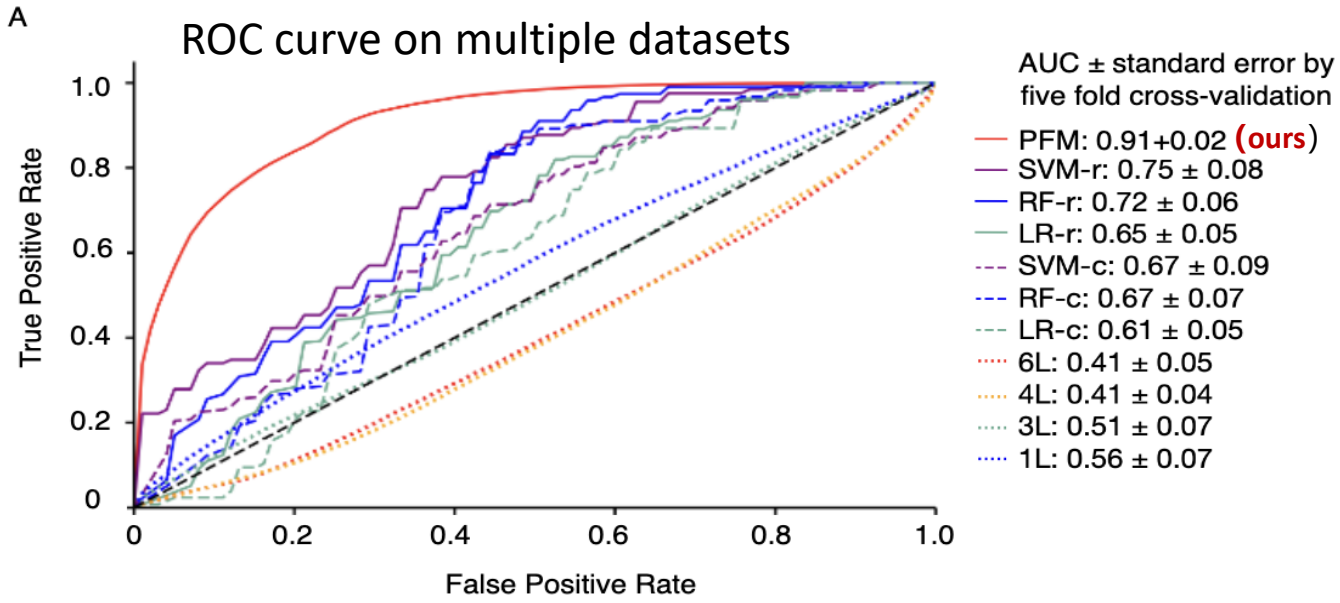
- Biomarker
- Diagnosis
- Other Tasks

Zang Z, et al. Boosting Unsupervised Contrastive Learning Using Diffusion-Based Data Augmentation From Scratch. ICML 2024

Performances on Cancer Diagnosis

Test on 3 Datasets

1. Mixed Cancer Datasets
2. Gastric Cancer Dataset
3. Thyroid Cancer Dataset



Gastric Cancer Dataset

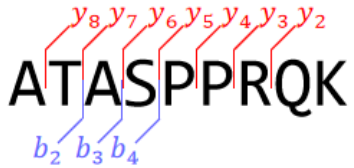
Methods	ACC (%)	AUC (%)	F1 (%)	P (%)	R(%)
SVM	75.2	77.5	75.0	75.3	75.1
lasso	75.1	77.4	74.9	75.2	75.0
DT	54.8	55.2	54.7	54.9	54.8
mlp	75.5	77.7	75.3	75.4	75.4
rf	75.1	77.4	74.9	75.2	75.0
lr	75.3	77.6	75.1	75.3	75.2
nn	76.8	79.1	76.5	76.7	76.6
LSTM-one-hot	77.6	79.8	77.2	77.4	77.3
LSTM-w2v	77.5	79.7	77.1	77.3	77.2
Ours	80.7	85.0	79.1	79.3	89.1

Thyroid Cancer Dataset

Methods	ACC (%)	AUC (%)	F1 (%)	P (%)	R(%)
SVM	88.5	91.2	88.0	88.3	88.2
lasso	88.4	91.2	87.9	88.2	88.1
DT	60.1	59.0	60.0	60.2	60.0
mlp	88.7	91.4	88.4	88.6	88.5
rf	88.4	91.2	87.9	88.2	88.1
lr	88.6	91.3	88.2	88.4	88.3
nn	90.5	93.2	90.2	90.4	90.3
LSTM-one-hot	90.6	90.1	90.2	90.4	89.3
LSTM-w2v	89.2	92.0	91.1	91.3	90.0
Ours	92.7	95.8	92.1	92.3	92.1

Analog Comparison: Proteomics vs Structural Biology

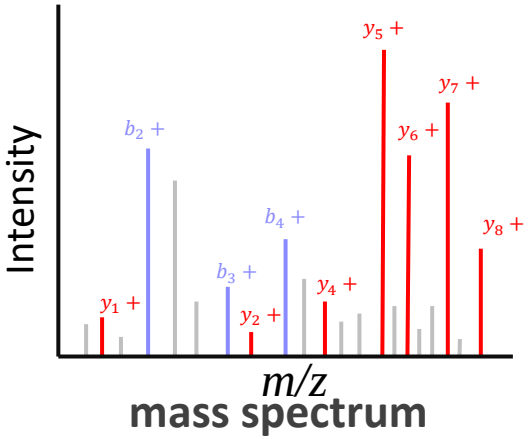
Proteomics



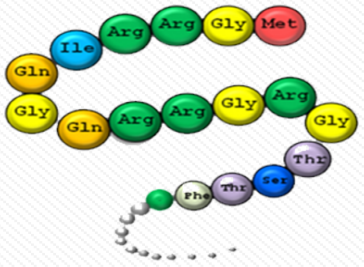
Peptide Sequence

Spectrum Prediction

Peptide Sequencing



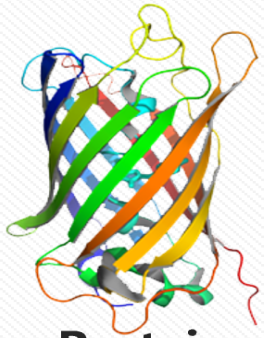
Structural Biology



Protein Sequence

Structure Prediction

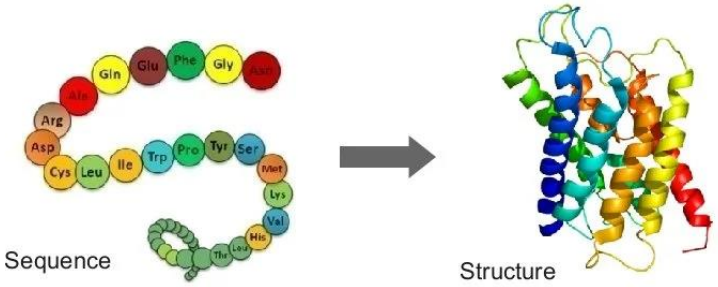
Sequence Design

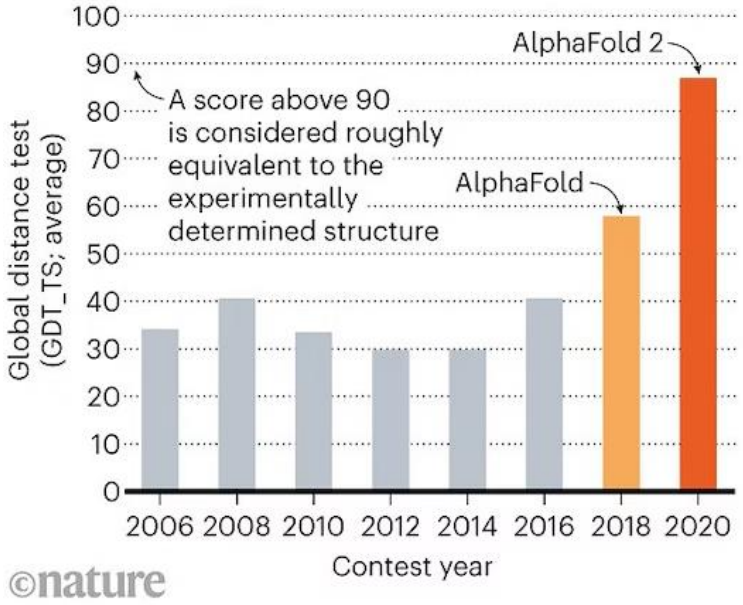


Protein Structure

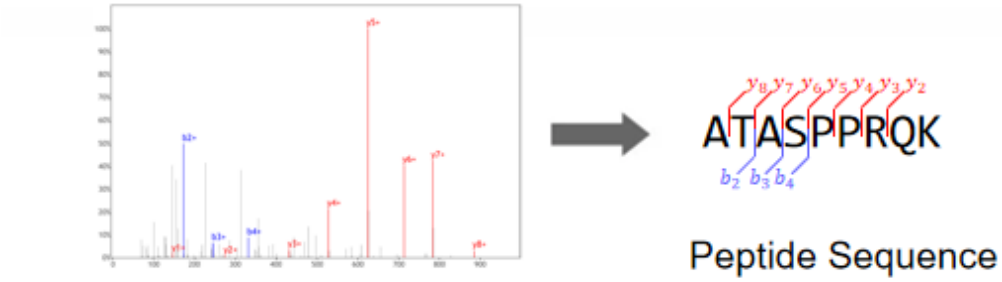
Foundation Model "AlphaFold2" for Proteomics?

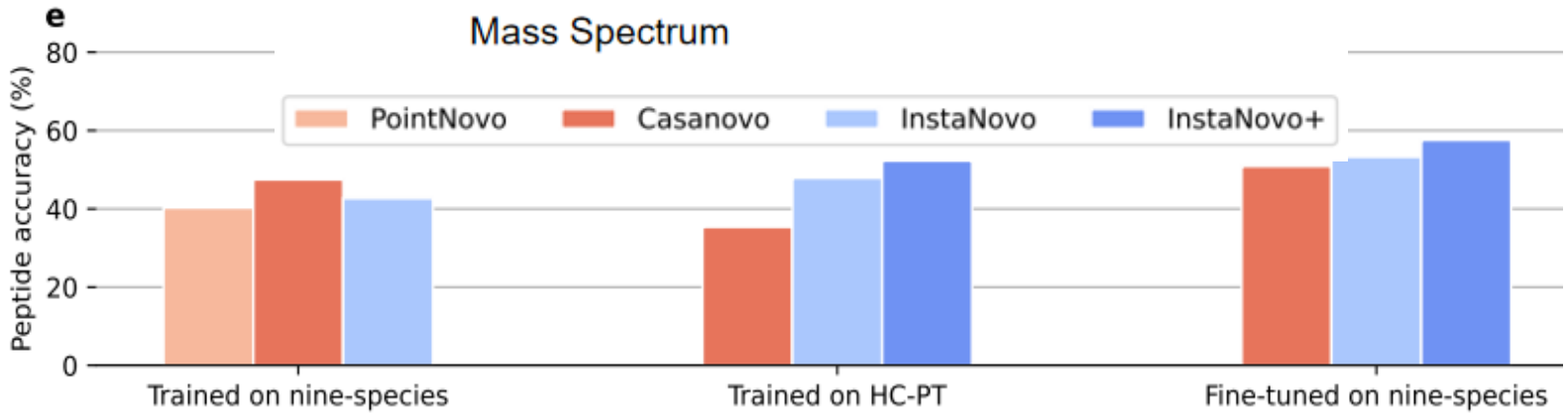
Protein Structure Prediction





Peptide Sequencing

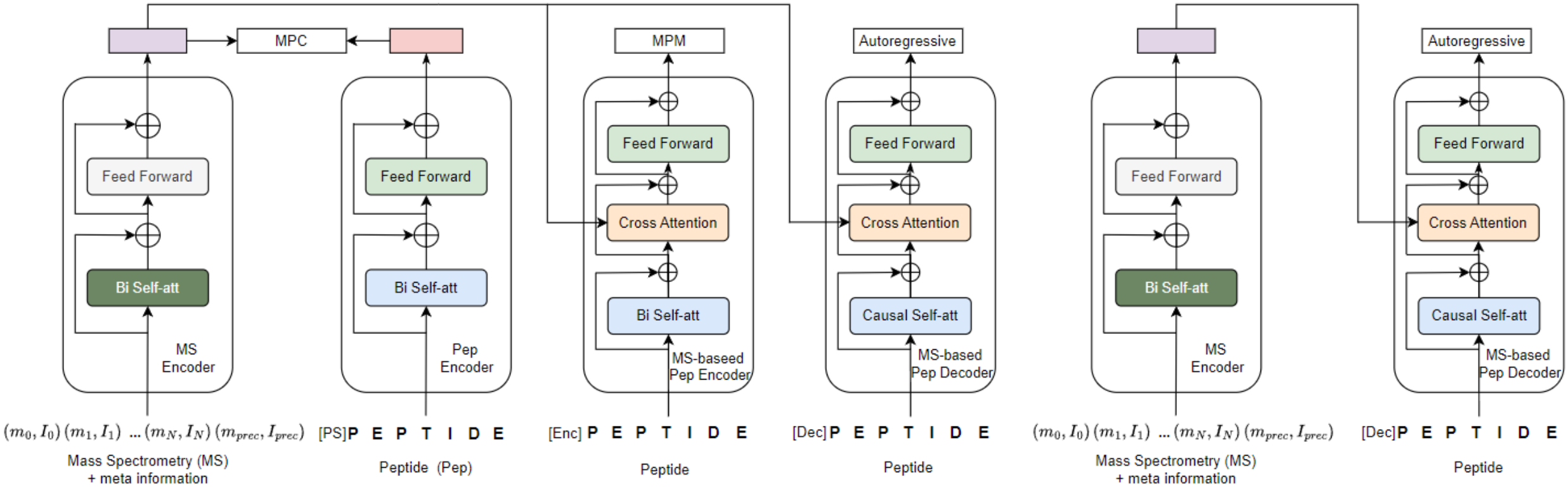




The SOTA accuracy for peptide sequencing is about 40%-60%. Achieving 80% accuracy would be a turning point for proteomics, effectively similar to AlphaFold2 for structural biology.

[1] Highly accurate protein structure prediction with AlphaFold (Nature 2021)
[2] De Novo Peptide Sequencing with InstaNovo (Biorxiv 2023)

π-UniMass – Foundation Model for De Novo Peptide Sequencing



MPC: **M**ass Spectrum and **P**eptide **C**ontrastive Loss

MPM: **M**ass Spectrum and **P**eptide **M**atching Loss

Autoregressive: **Autoregressive** Loss

De novo: **De Novo** Peptide Sequencing Loss

π-UniMass: Performance in De Novo Peptide Sequencing

Model	Training Data Num.	Human	Mouse	Yeast	Honeybee
CasaNovo [1]	30 M	0.446	0.483	0.599	0.493
InstaNovo [2]	20 M	0.431	0.436	0.581	0.477
π-UniMass	20 M	0.535	0.543	0.633	0.559

(Metric: peptide-level precision)

[1] De novo peptide sequencing with InstaNovo (Biorxiv 2023)
[2] De Novo Mass Spectrometry Peptide Sequencing with a Transformer Model (ICML 2022)

π-UniMass: Performances in Spectrum Prediction

Models / Datasets	OG	OL	O2	OC
Prosit [1]	82.35%	81.05%	86.53%	82.17%
pDeep 3 [2]	84.06%	86.70%	91.46%	85.80%
π-UniMass	**93.68%**	**94.17%**	**96.19%**	**90.52%**

(Metric: The proportion of Pearson correlation coefficients greater than 0.9)

[1] Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning (Nature Methods, 2019)

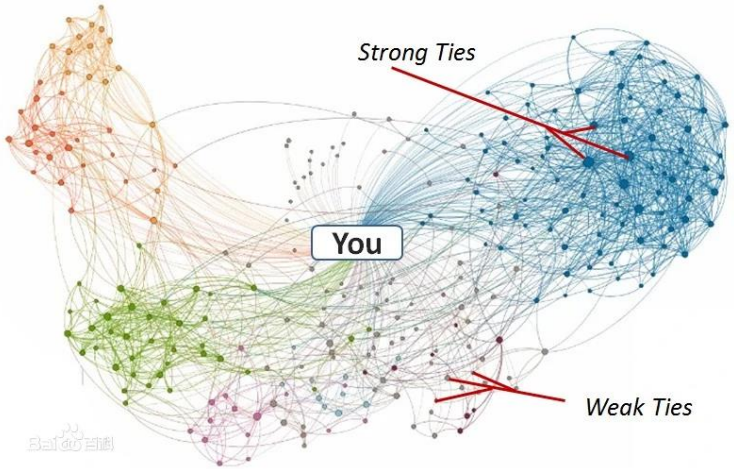
[2] pDeep3: Toward More Accurate Spectrum Prediction with Fast Few-Shot Learning (Anal. Chem., 2021)

Outline

1. Introduction
2. Proteomics Models
3. Single-Cell Data Analysis
   - High-Dimensional Data Analysis
   - Developmental/Evolutional Data Visualization
4. Some Future Directions

High-Dimensionality of Data

- DNA Sequences
- RNA Sequences
- Protein Sequences
- Images, Videos, Text, Audio



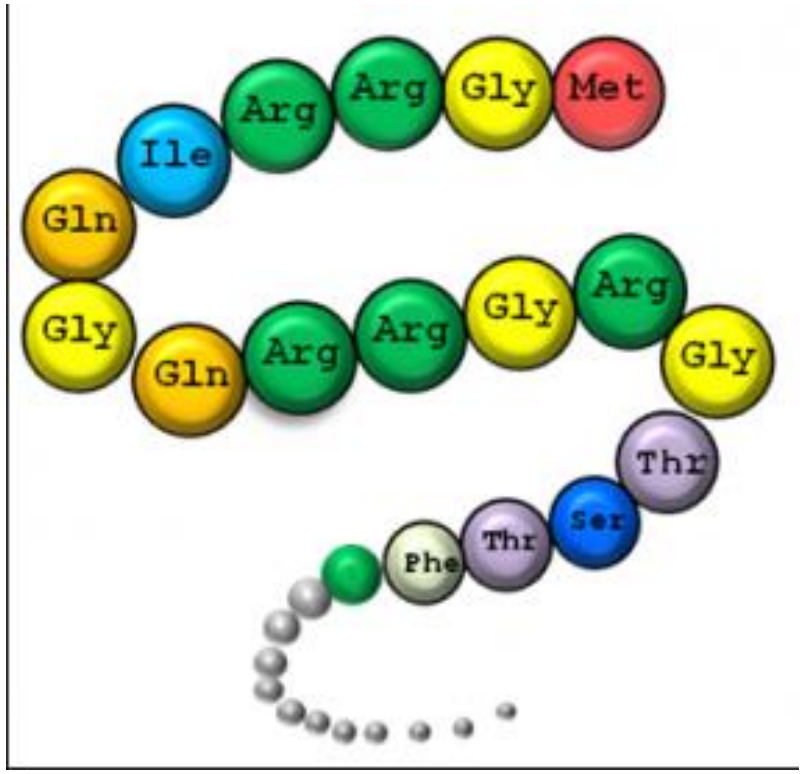
Face Image Data

- Image size 100x100 = 10^4 pixels
- RGB image size $3x10^4$ pixels
- Dimensionality = $3x10^4$
- Pixel values in {0,…,255}
- #Possibility = $256^{30,000} \cong$ infinity
- Only a tiny portion is of faces
- Face pattern lives in low dim subspace (Face Manifold)

Protein Data

- 20 amido acids
- Length L
- Total number 20^L
- Stable natural protein << 20^L

Forming "**Protein Manifold**"



Manifold Assumption

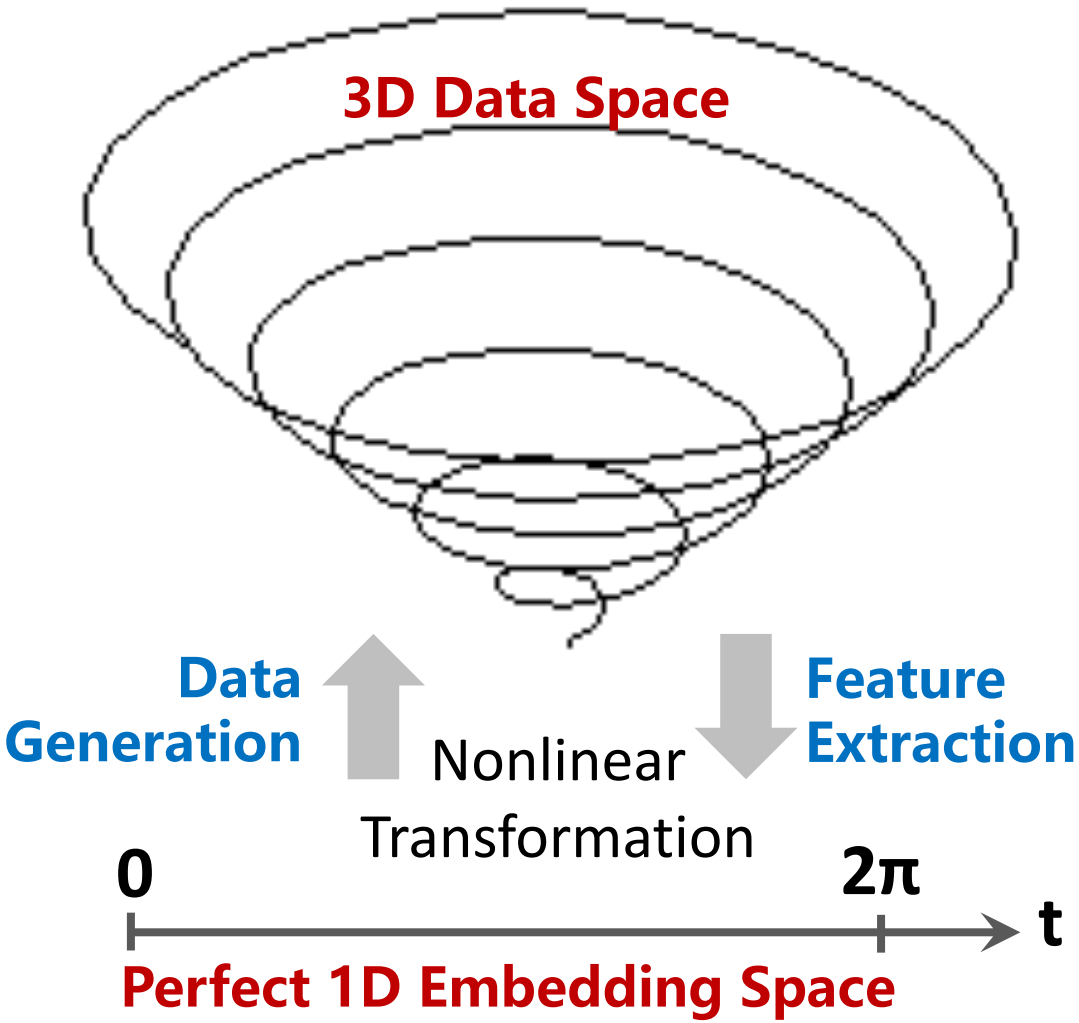
High-Dimensional Data: Images, Web pages, Gene sequences, ….

Dimension Reduction into Coordinate System of a Lower Dim

- **For representation learning (feature extraction)**
- **For data visualization – in 2D or 3D**

Manifold Assumption: an interesting pattern in high dimensional data resides on a low dimensional manifold

Manifold in Hi-D Data Space: 1D Curve in 3D Space



Conical Helix:

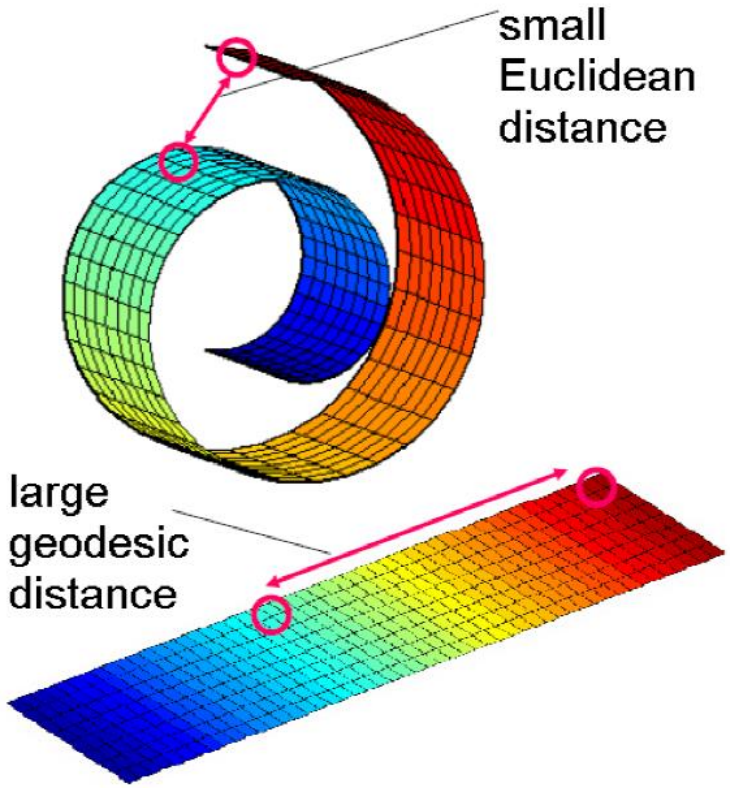
$x=t*\cos(6t)$, $y=t*\sin(6t)$, $z=t$

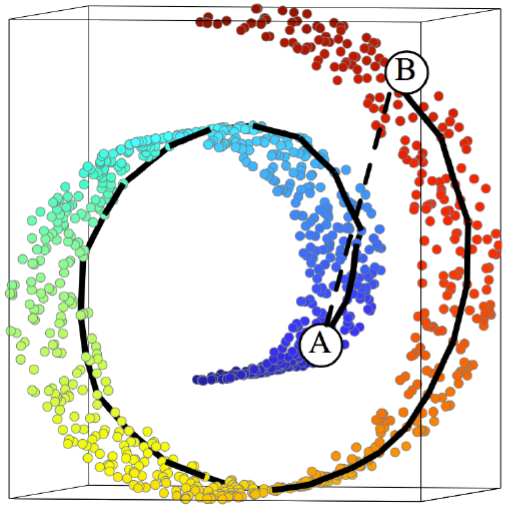
$0 \le t \le 2\pi$

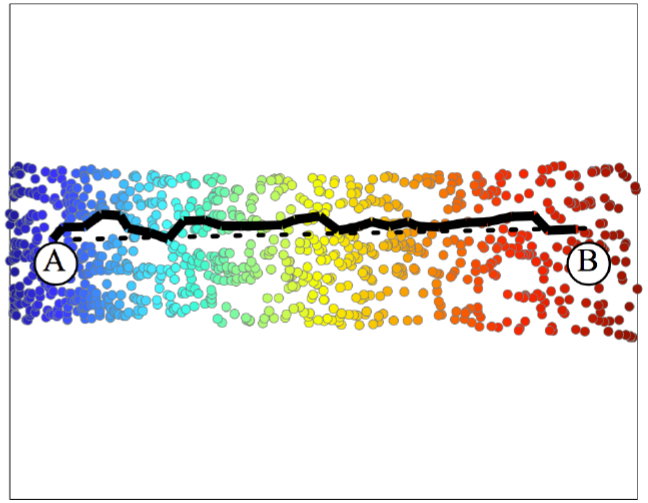
1D line segment

Latent variable t

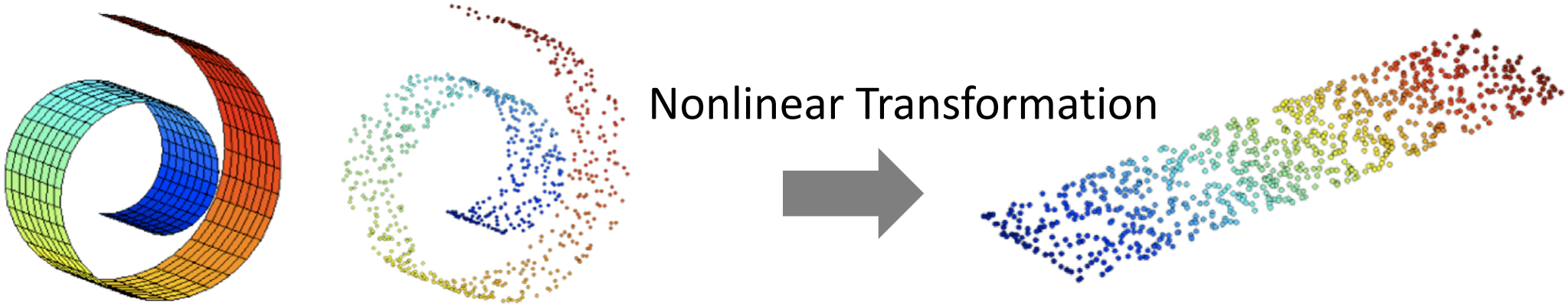
Geodesic Distance on Manifolds







Flattening of Curved Manifolds



Swiss Roll:

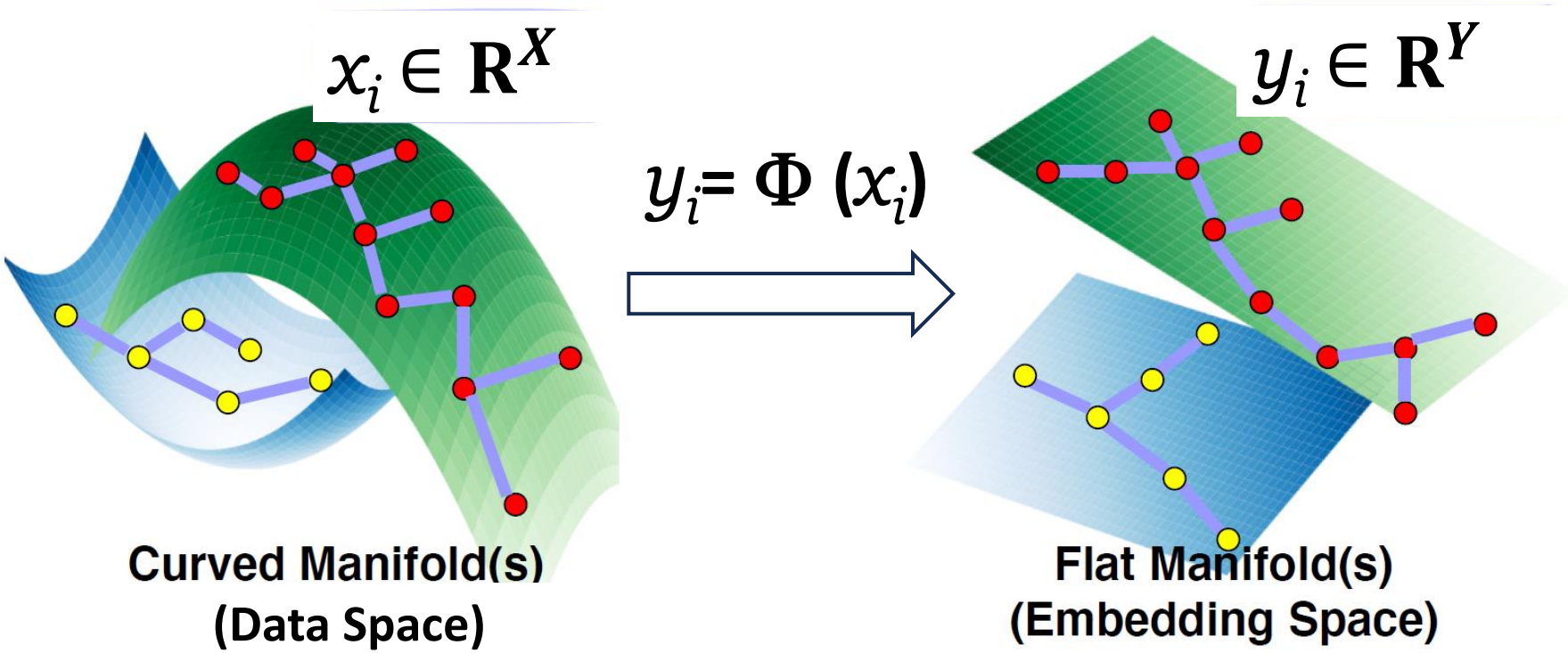
$x=\varphi\cos(\varphi)$, $y=\varphi\sin(\varphi)$, $z=\psi$

$1.5\pi \le \varphi \le 4.5\pi$, $0 \le \psi \le 10$

Manifold: 2D rectangle

generated by two latent variables φ, ψ

Euclidean Embedding: Transforming Curved Surfaces into Planes

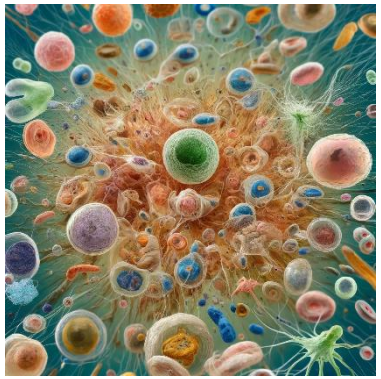


Why Hyperbolic Embedding for Single-Cell Analysis

Characteristic of sc-Data

- Tree-like hierarchical structure
- High heterogeneity

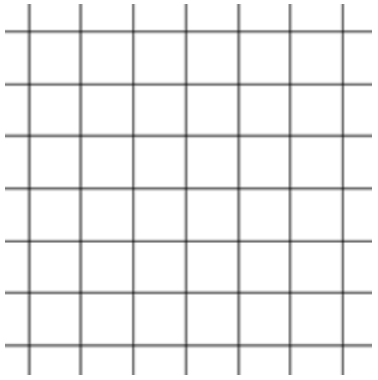
Cellular Differentiation



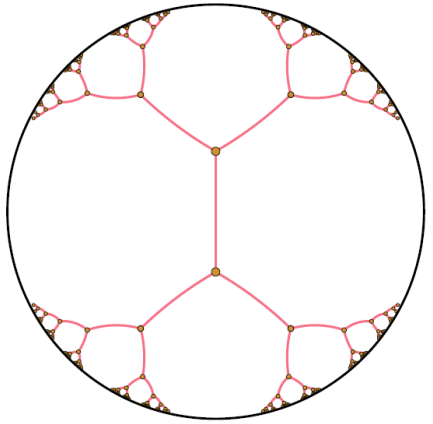
High Heterogeneity

Why Hyperbolic Embedding

- Embedding trees distortion-free
- Exponential volume capacity



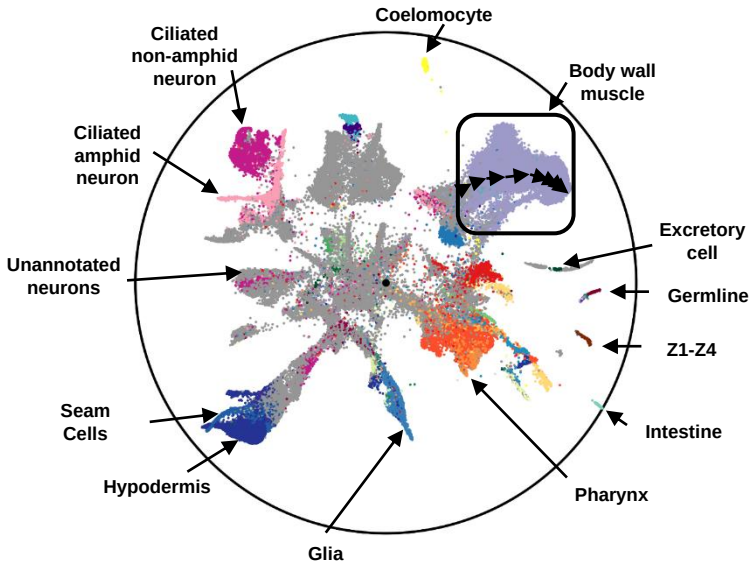
Euclidean Grid

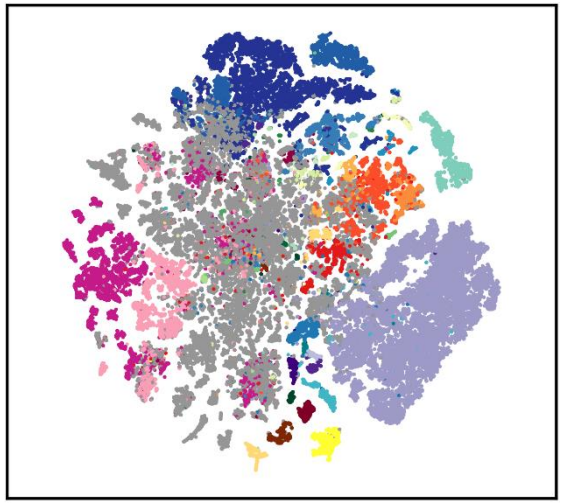


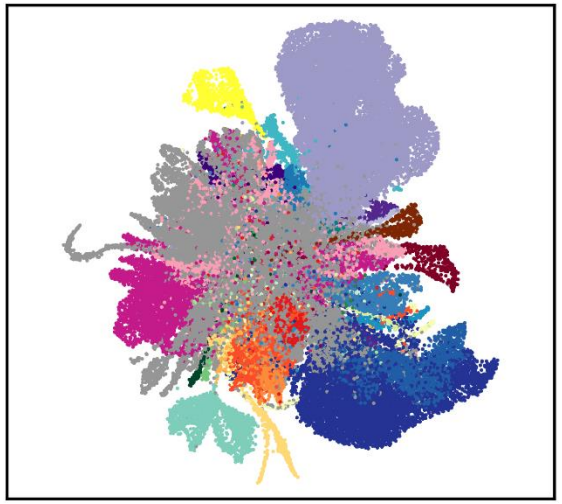
Tree Nodes on Hyperbolic Grid

Visualization for C. Elegans Embryonic Data

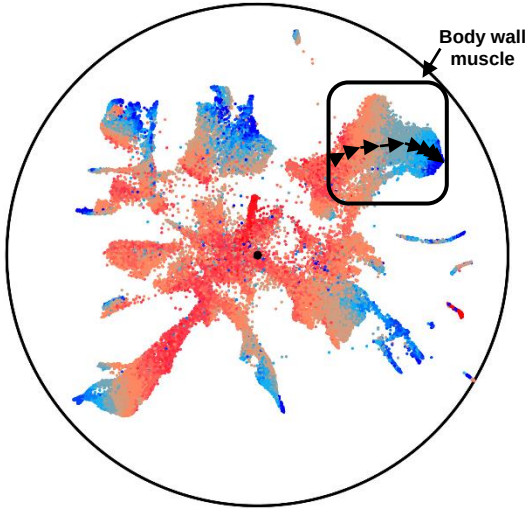
Cell Type

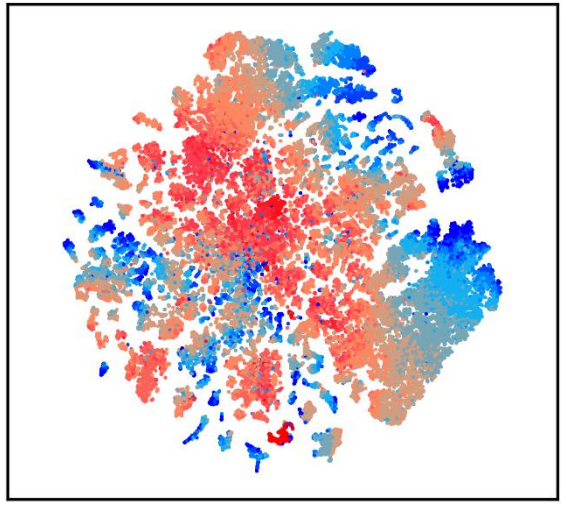


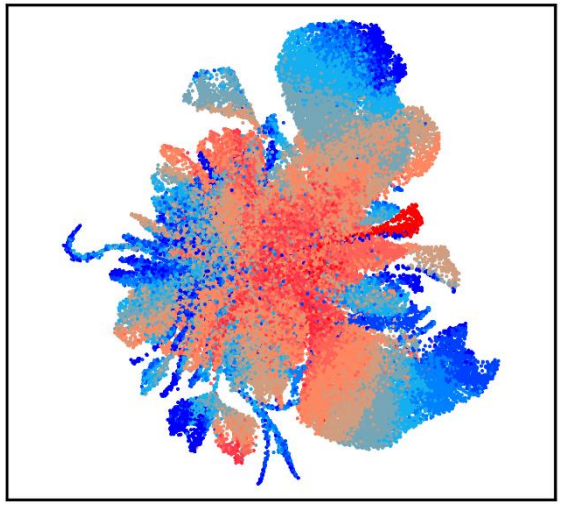




Embryo Time







- < 100
- 100-130
- 130-170
- 170-210
- 210-270
- 270-330
- 330-390
- 390-450
- 450-510
- 510-580
- 580-650
- > 650

DMT (Hyperbolic)

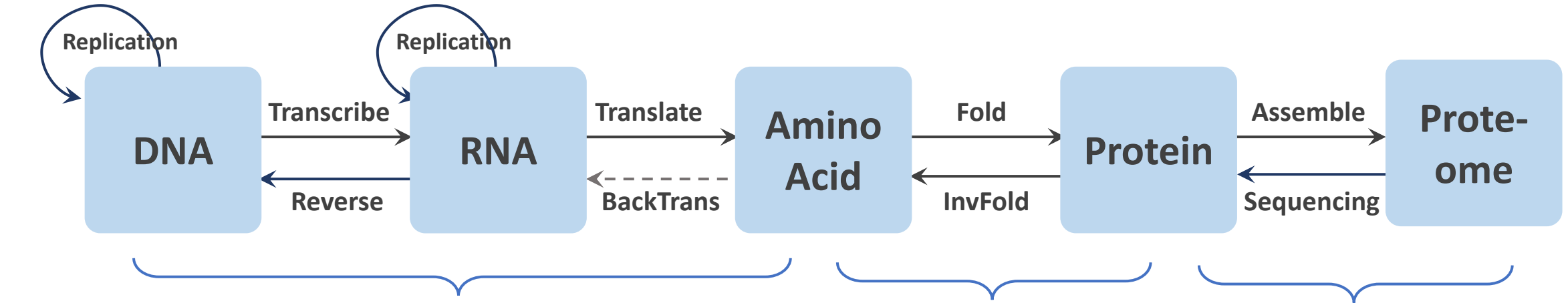
t-SNE (Euclidean)

UMAP (Euclidean)

Outline

1. Introduction
2. Proteomics Models
   - Proteomics Empowered by Protein Language Models
   - De Novo Protein Sequencing
3. Single-Cell Data Analysis
   - Dimension Reduction
   - Visualization
4. Some Future Directions

To Innovate Life Science Research

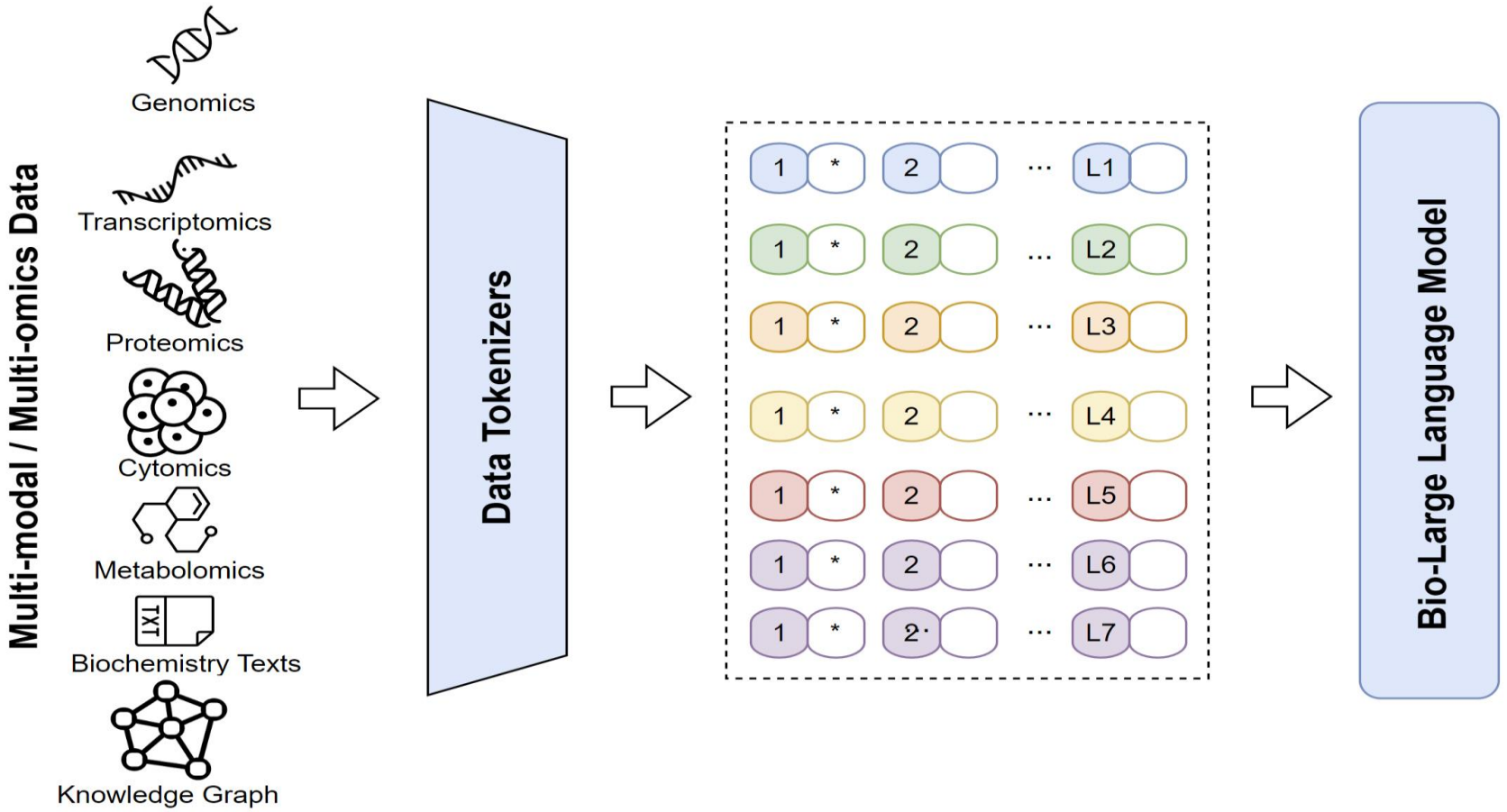


Principles:

- Central Dogma
- First Principles
- Bio-Chemistry

→ **Functions / Phenotypes**

Large Language Models of Biology



Bio-Data (Multi-Modal)

Bio-Tokenizer

Bio-Tokens

Bio-LLM

Thank You