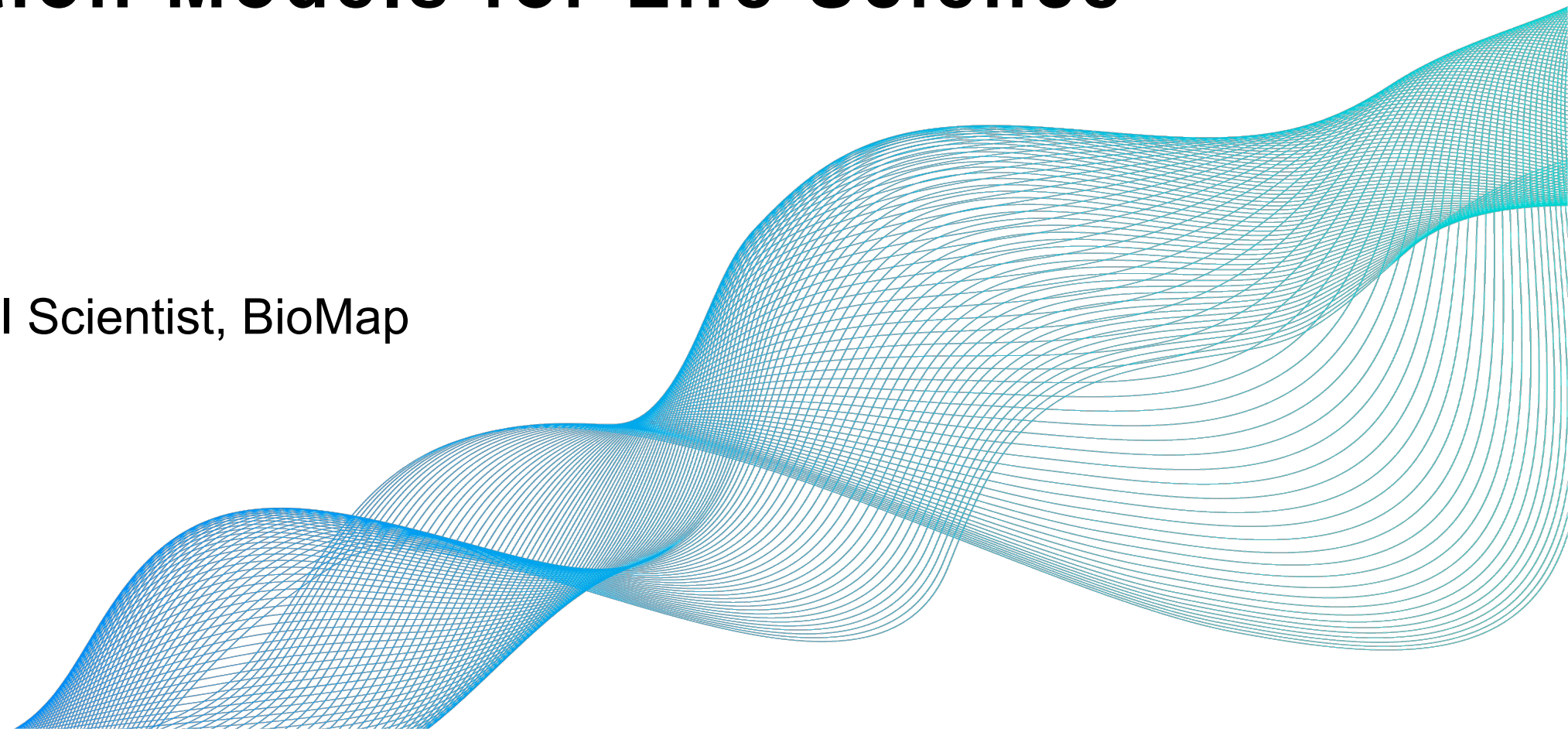
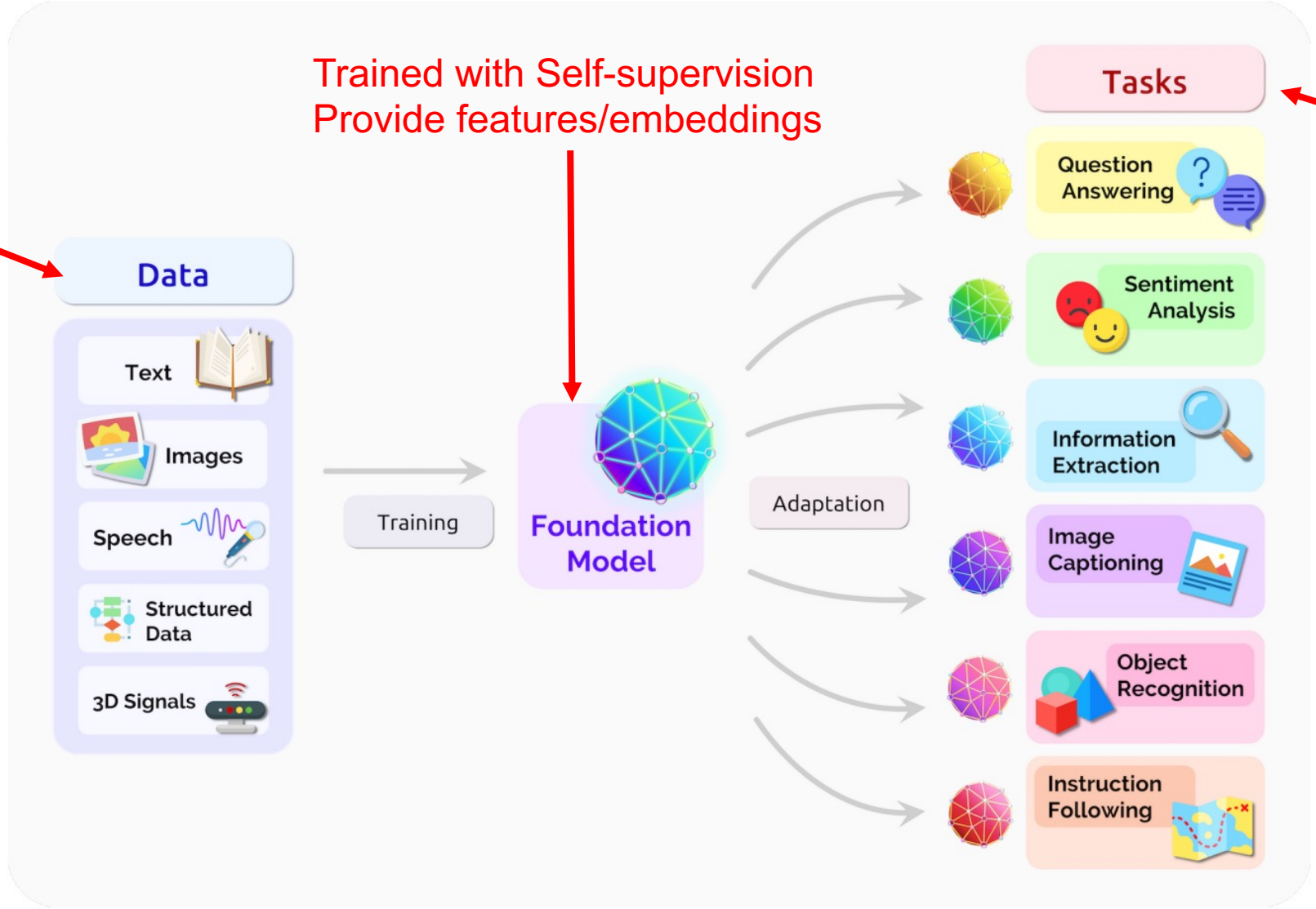
Foundation Models for Life Science

Le Song

CTO & Chief AI Scientist, BioMap
MBZUAI

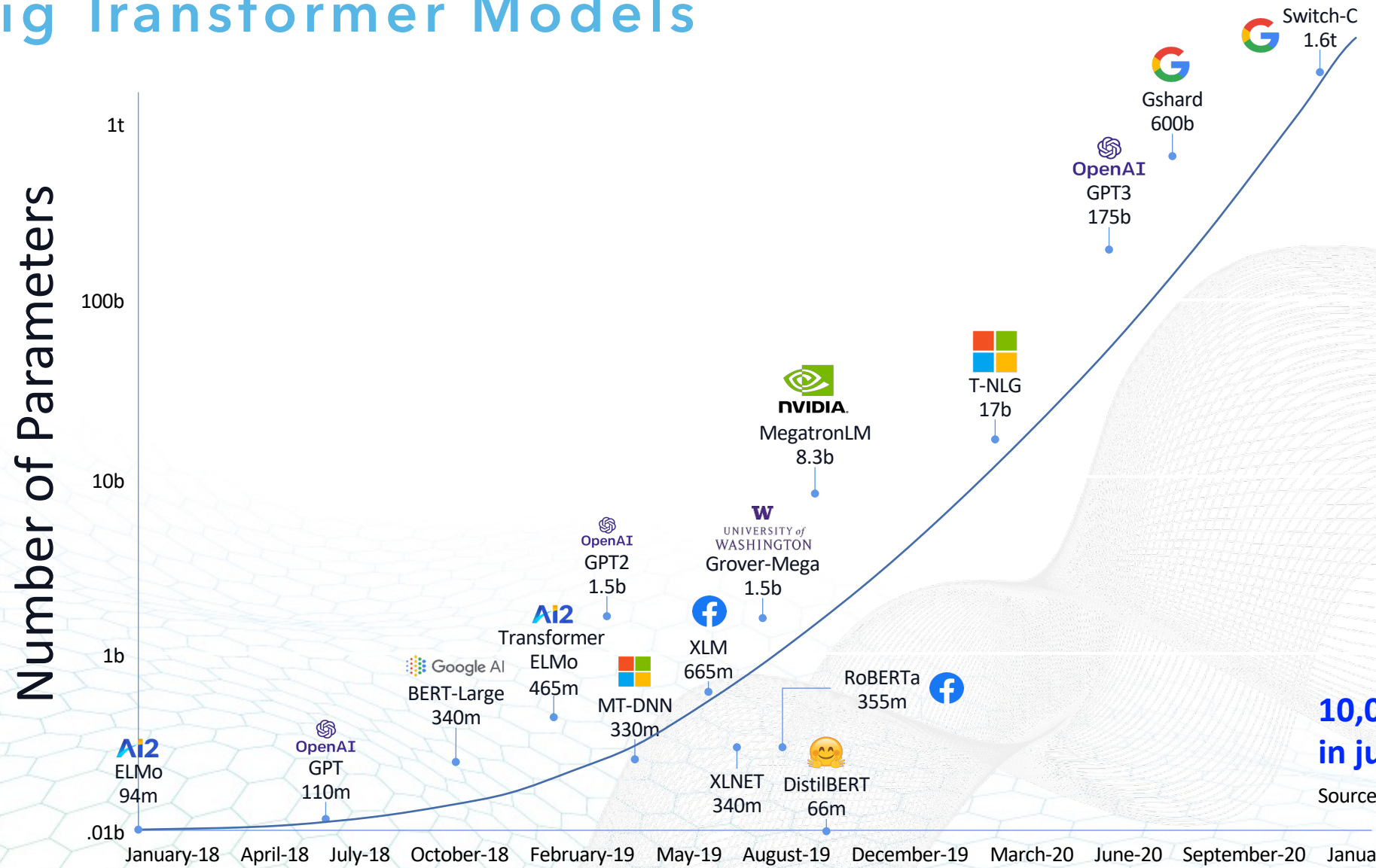


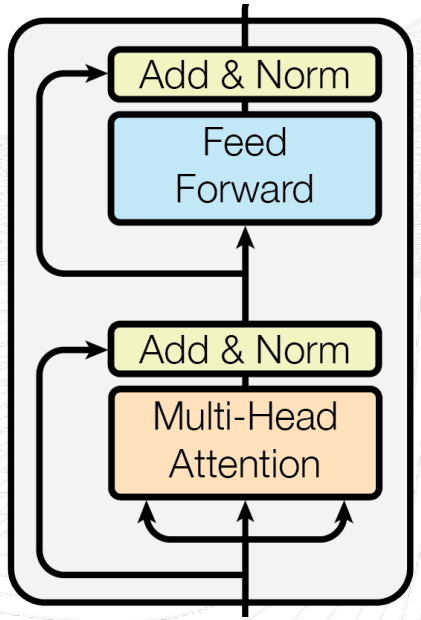
The Paradigm of Foundation Models in AI



Bommasani et al. On the Opportunities and Risks of Foundation Models. Arxiv

Foundation Models are Increasingly Big Transformer Models

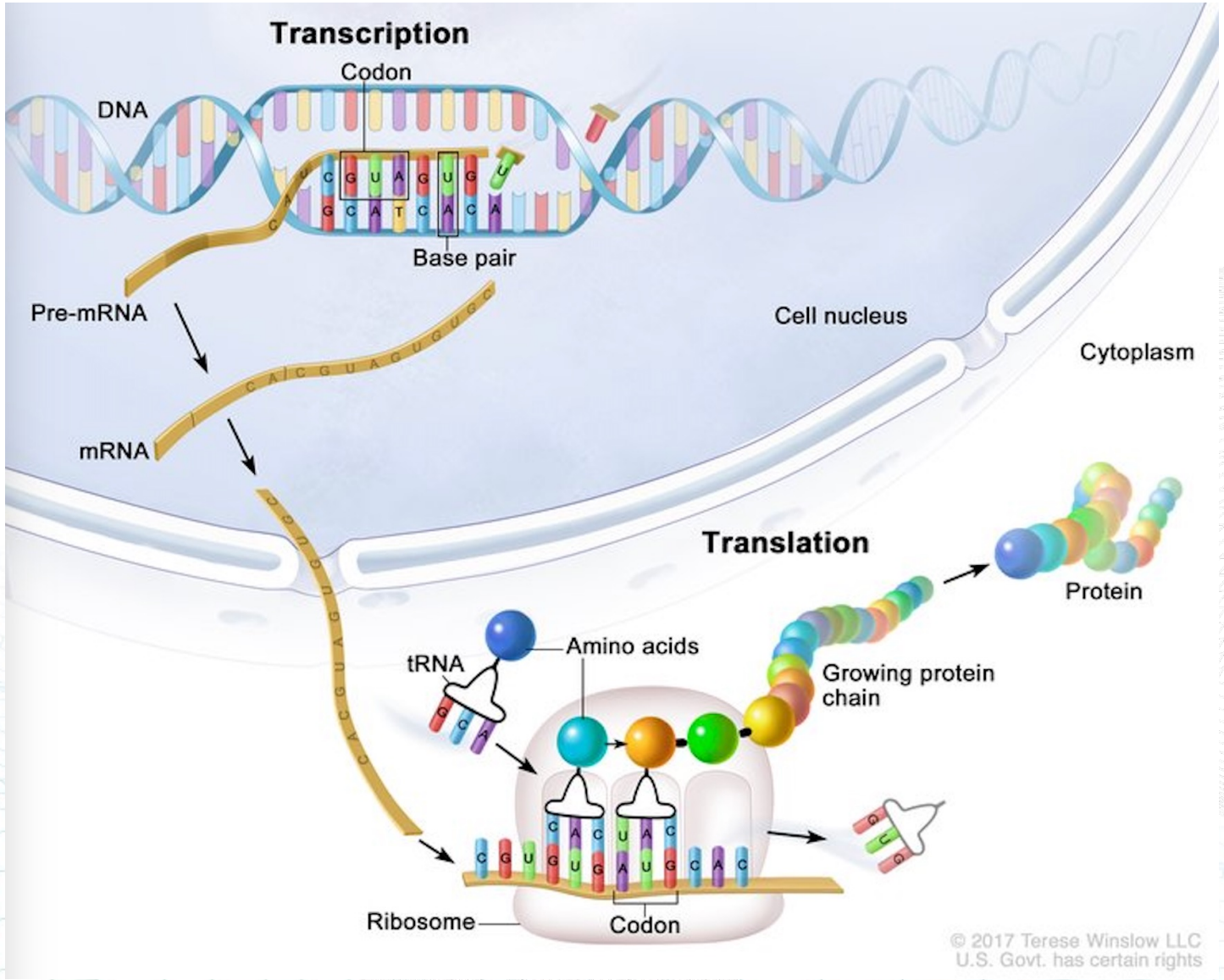




10,000x model size increase in just 3 years

Source: Gadi Singer, VP at Intel Labs

The Language of Life is Unique

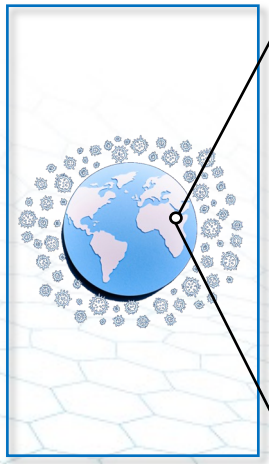


https://visualsonline.cancer.gov/details.cfm?imageid=11683

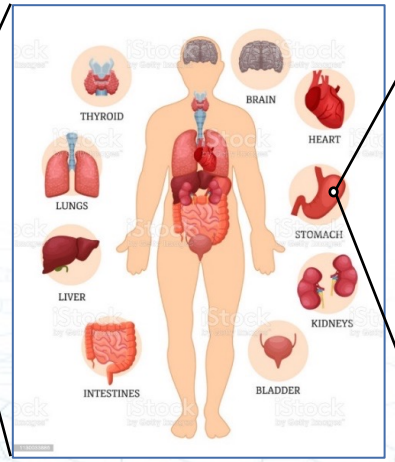
Multiscale Heterogeneous Networks

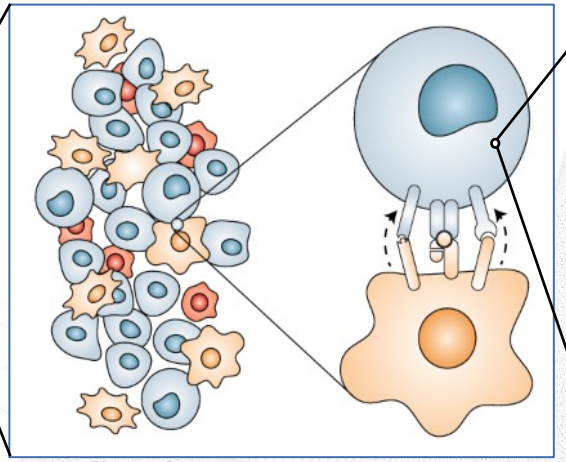
Public Safety, Vaccine, etc. | **Treatment Personalization and Repurposing** | **Clinical Diagnosis** | **Target Discovery** | **Drug Design**

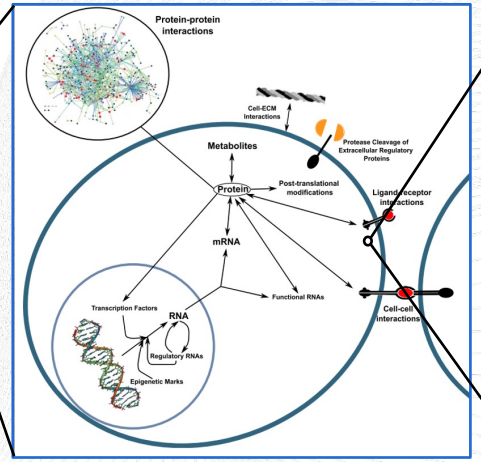
Social Networks with pathogen networks | Patient Networks | Tissue/Organ Networks | Intercellular Interaction Networks | Intracellular Interaction Networks | Molecular Structure/Graph

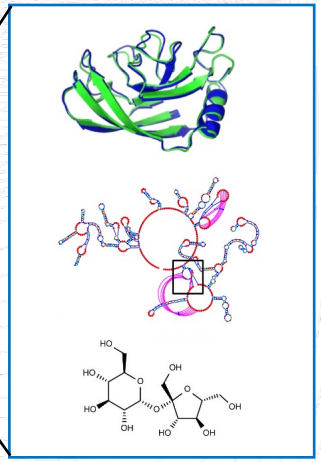






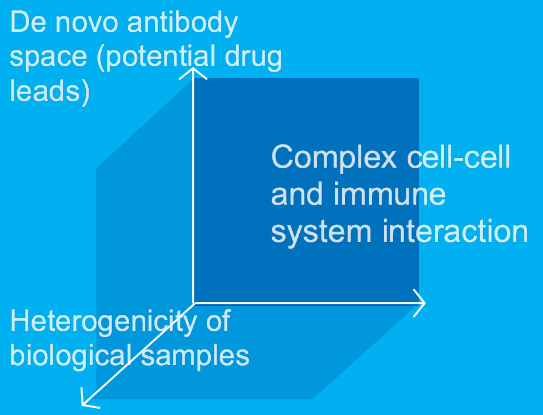






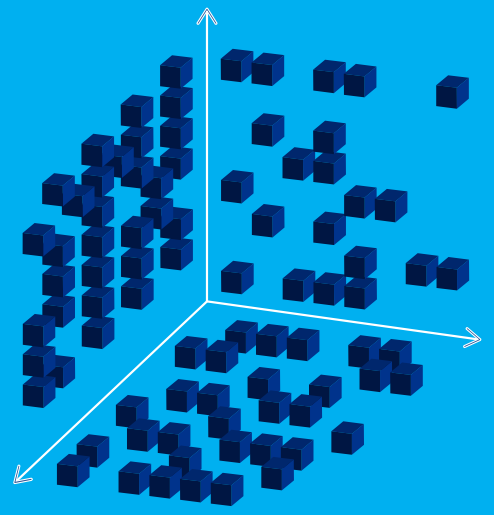
Pandemic | Public health | Medicine & Clinical | Cell Biology | Molecular Biology | Biochemistry

High Dimensional Space to Explore but Small Data



Immune modulation has many dimensions of complexity

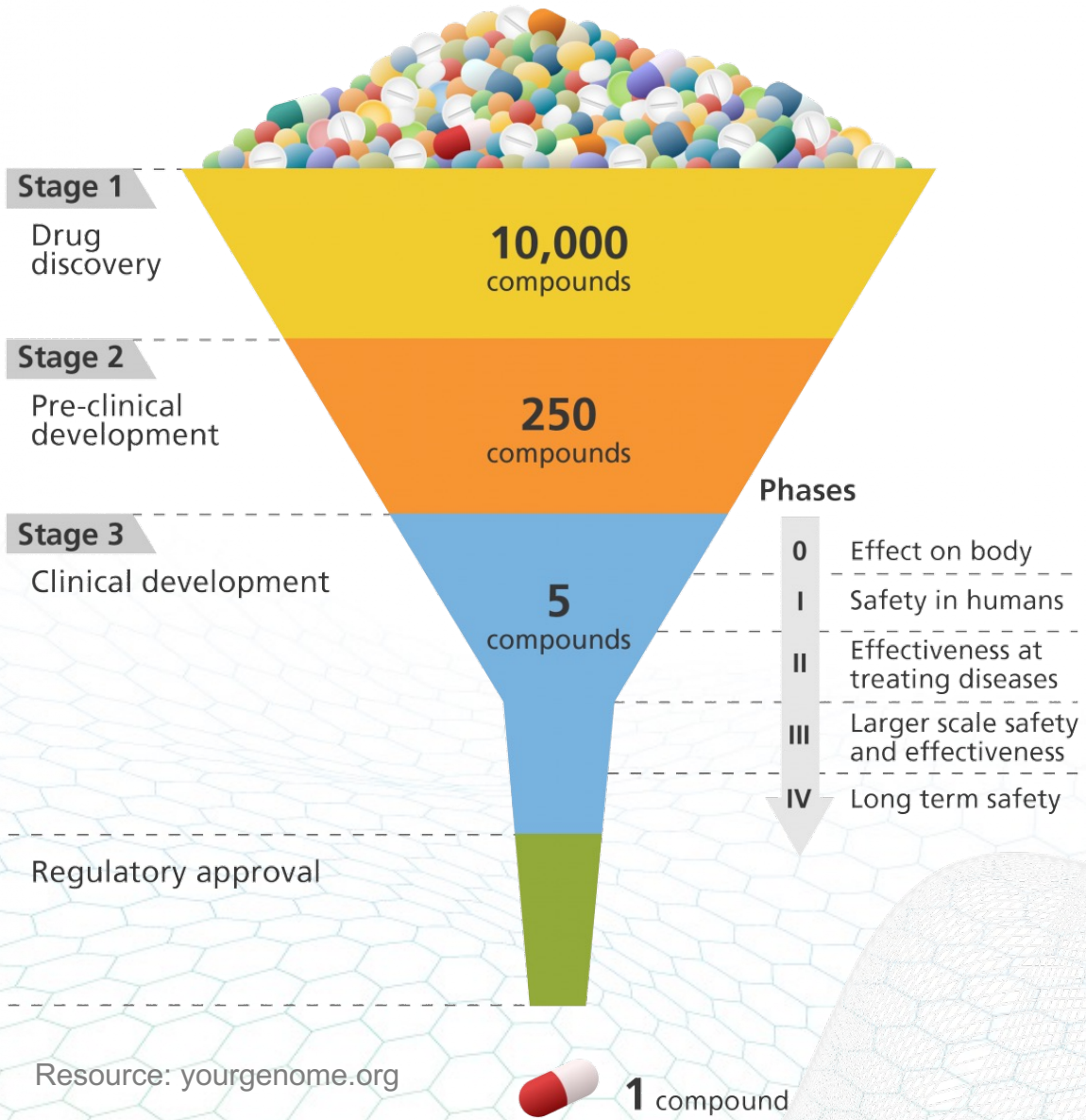
Developing immune mediated therapies requires an understanding of this complexity for both target identification and antibody development.



Limited exploration of space by experimental and public data

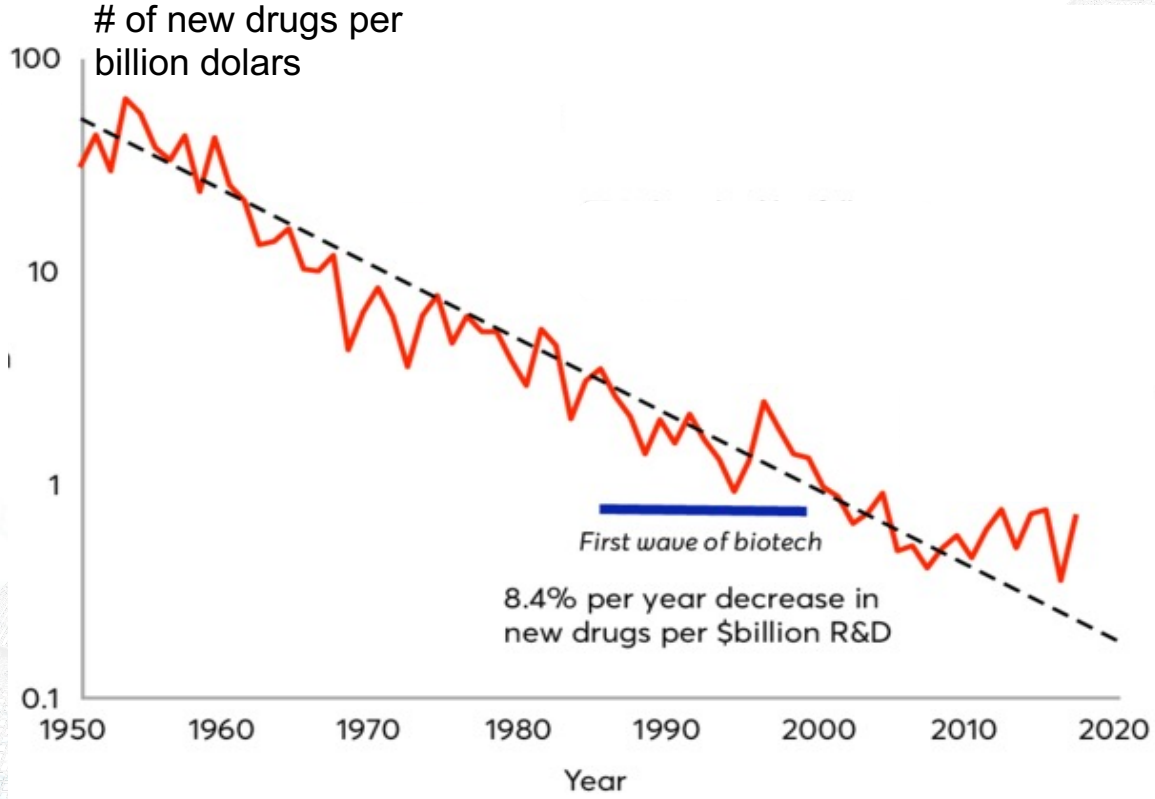
Currently available data is disorganized, heterogeneous, noisy and gives only a partial observations of the parameters involved.

Drug Discovery is a Costly Process



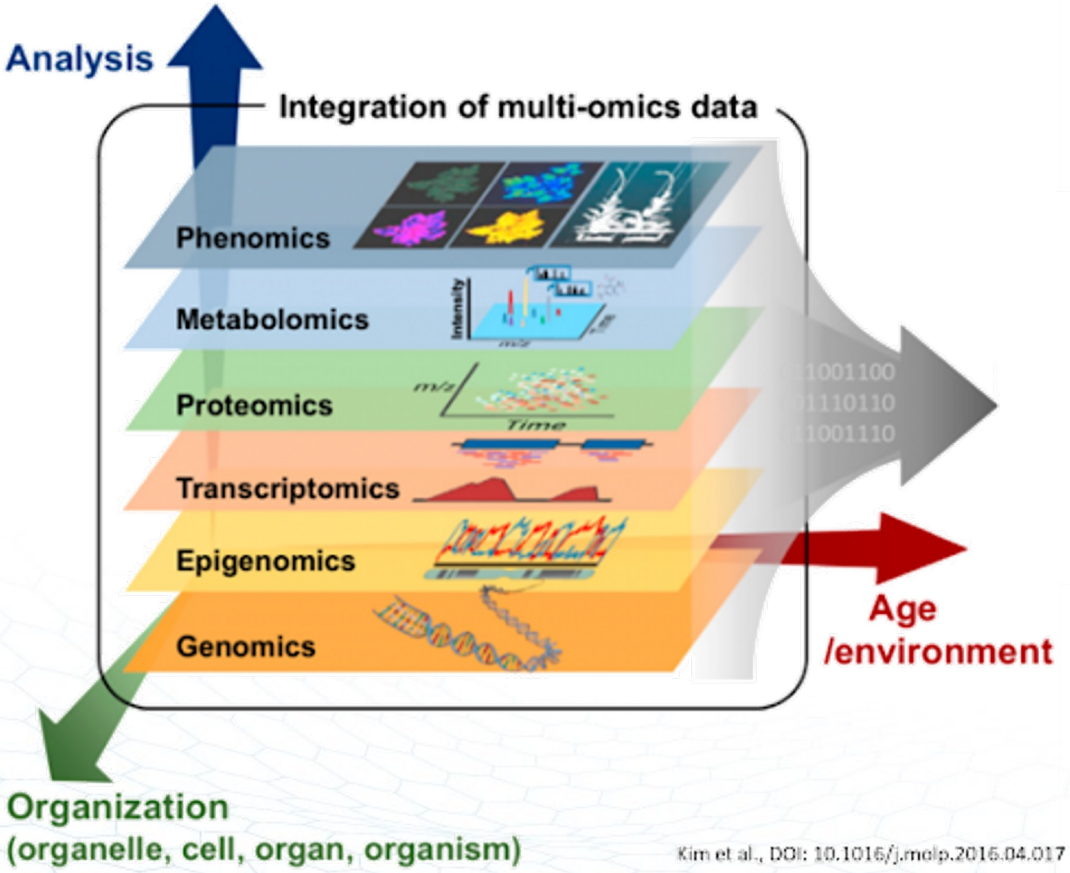
Resource: yourgenome.org

~10 Years, 1 Billion Dolloars for 1 Drug

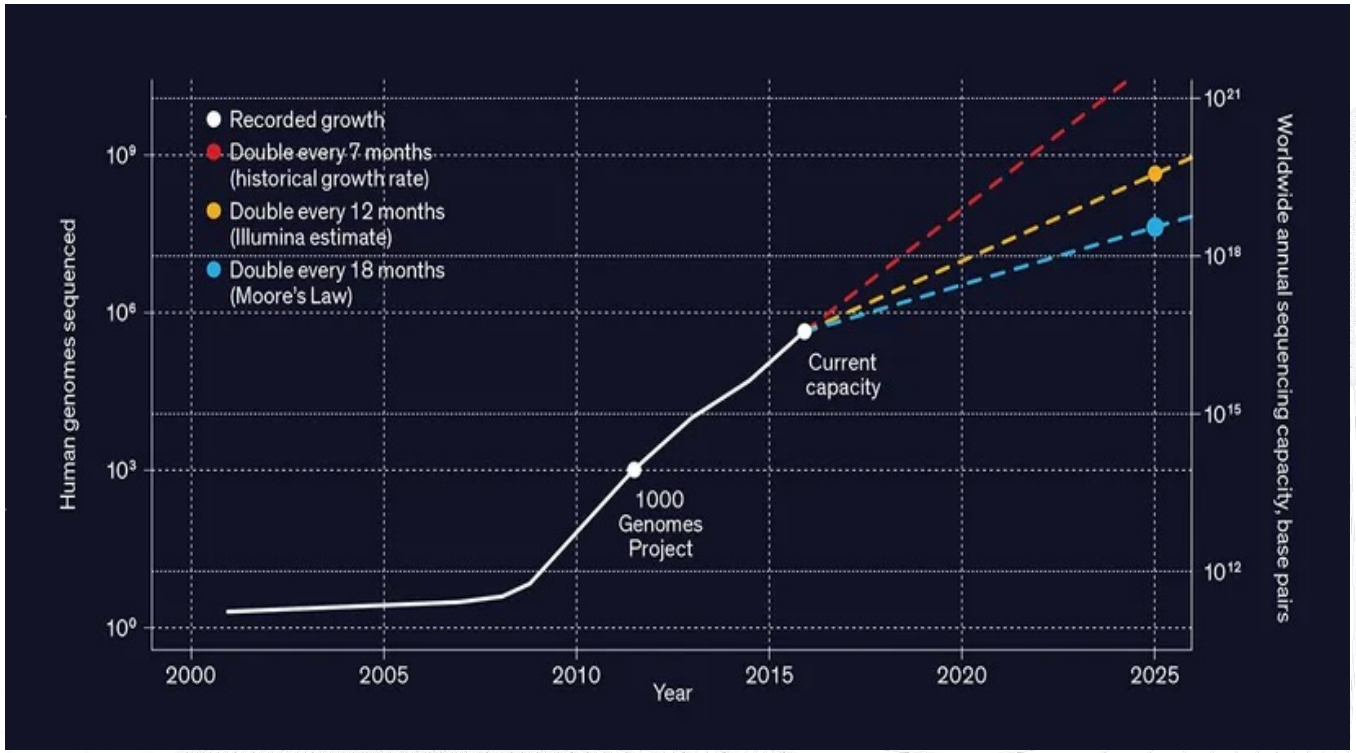


Resource: researchgate

Exponential Data Volume, Diversity & Complexity



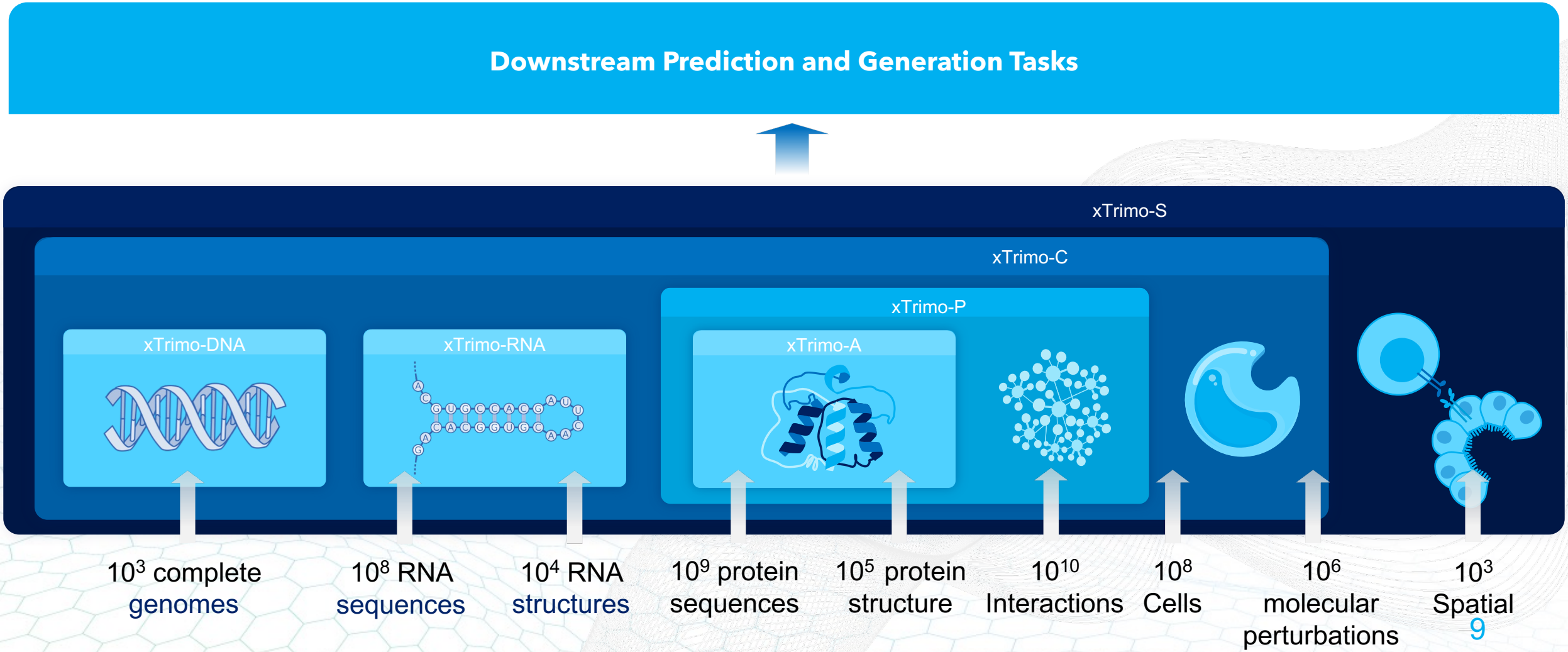
Kim et al., DOI: 10.1016/j.molp.2016.04.017



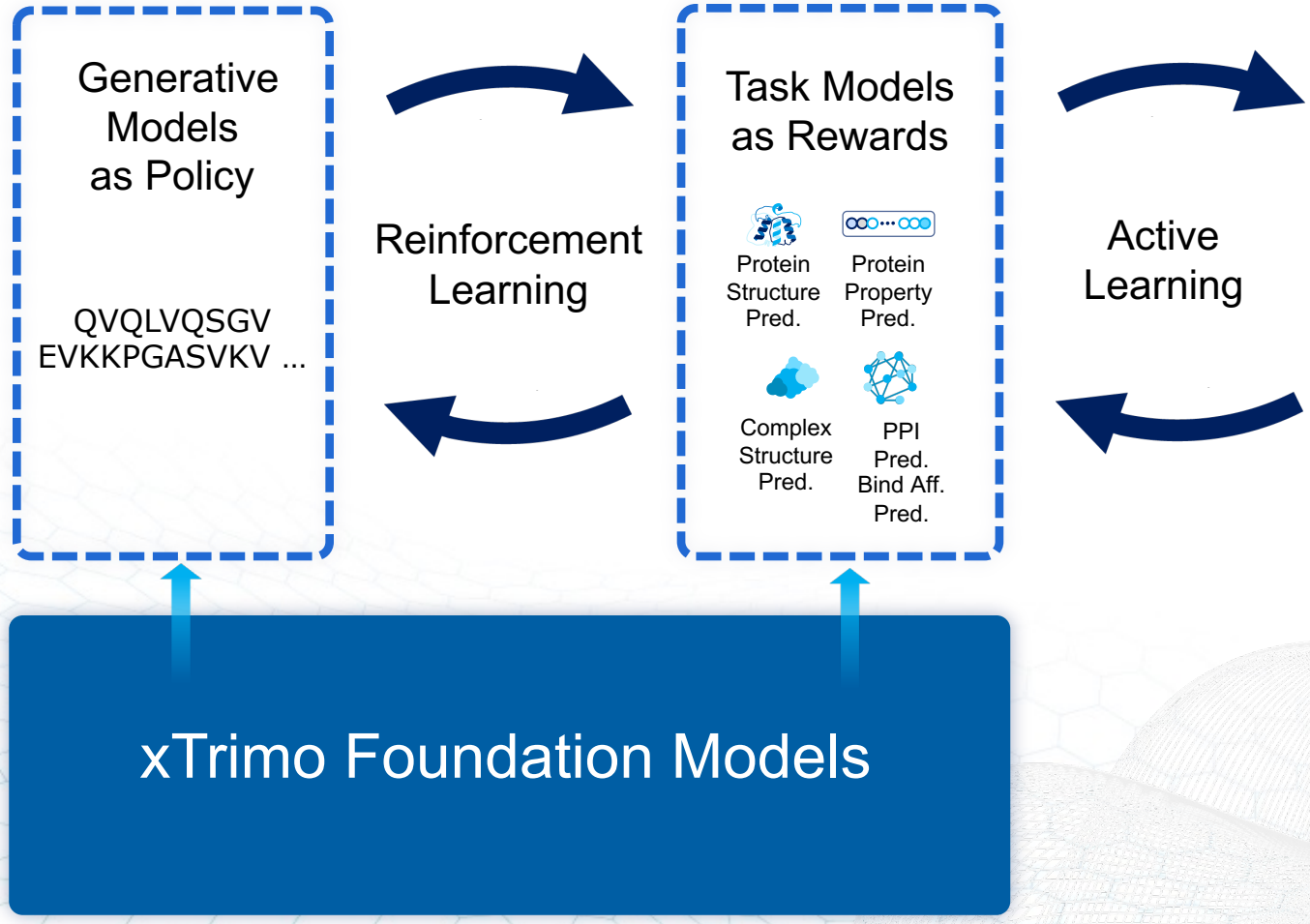
10^8 published paper texts | 10^3 complete genomes | 10^8 RNA sequences | 10^4 RNA structures | 10^9 protein sequences | 10^5 protein structure | 10^{10} Interactions | 10^8 Cells | 10^6 molecular perturbations | 10^3 Spatial

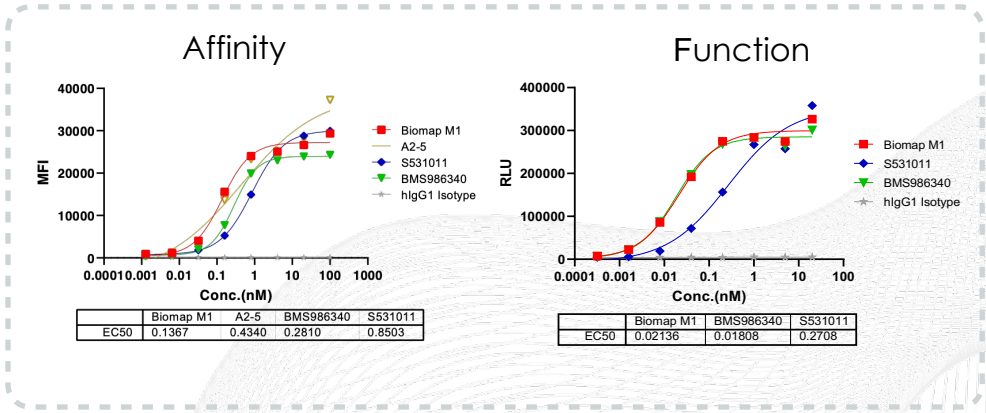
xTrimo

The Cross-modal Transformer Representation of Interactome and Multi-Omics



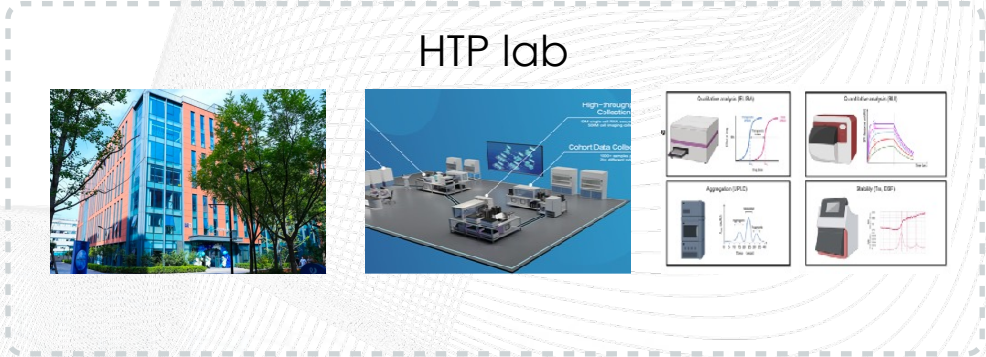
Iterative Computational Design & Wet-lab Validation



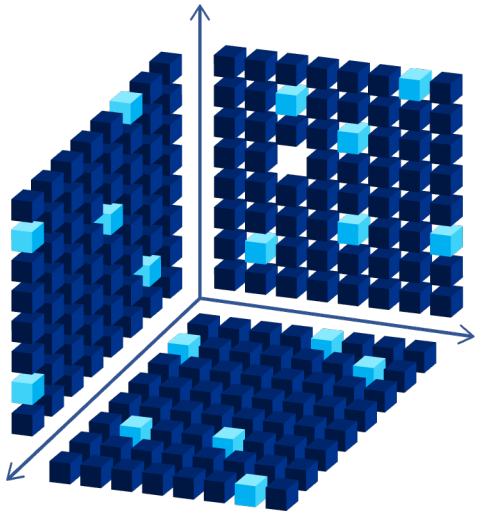


	Biomap M1	A2-5	BMS986340	S531011
EC50	0.1367	0.4340	0.2810	0.8503

	Biomap M1	BMS986340	S531011
EC50	0.02136	0.01808	0.2708



Efficient Exploration of Biological Complexity and Design Space



Full space exploration in silico inference and HT in-vitro assays

BioMap systematically curated all available data in single cell gene expression as well as protein structure and interaction. These are used to train AI models and inform high throughput in vitro experiments to explore the full space for target ID and antibody design



AI based decoding of systems, cells and proteins to rank targets and antibodies

Immune targets and antibody designs can be ranked and top hits selected for validation with a higher rate of success

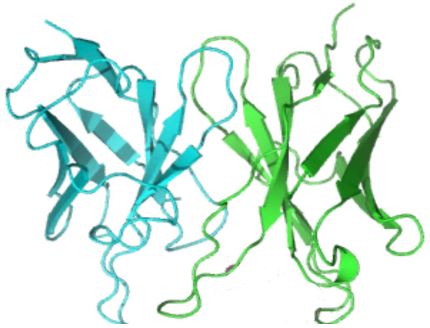
Protein Language is Different from Natural Language

1D sequence

QVQLVQSGVEVKKPGASVKV ...

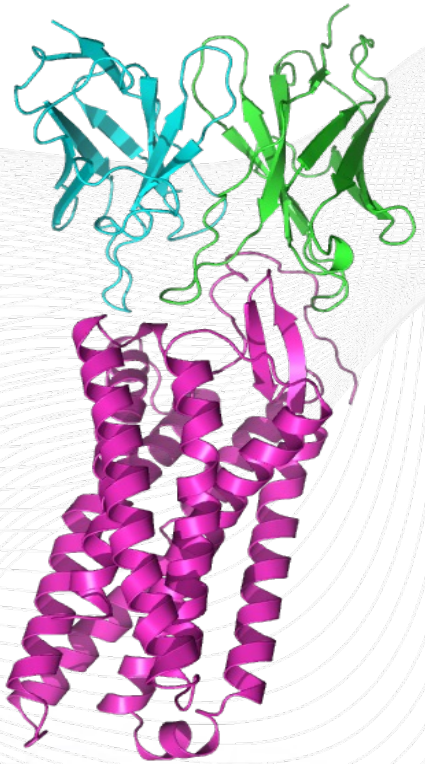
determine

3D Structure



enable

Function

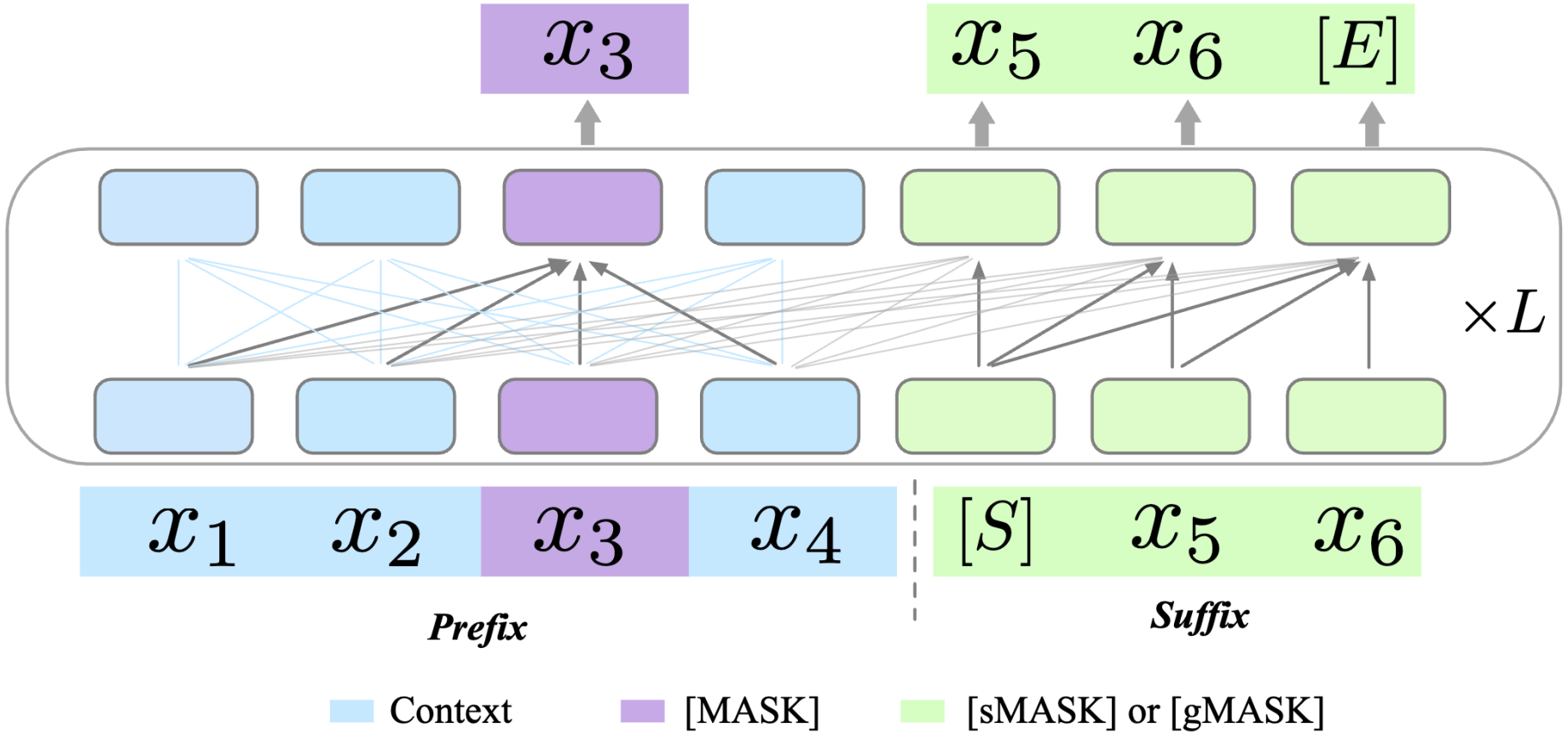


Analogy to natural language

amino acid	⇔	"letter"
protein		"sentence"
ppi		"paragraph"
...		...

xTrimoPGLM: 100B Parameter Unified Model

Unified Mask Recovery + Generative Pretraining



[MASK]: xTrimoPGLM acts like BERT.
[sMASK] OR [gMASK]: xTrimoPGLM acts like GPT.

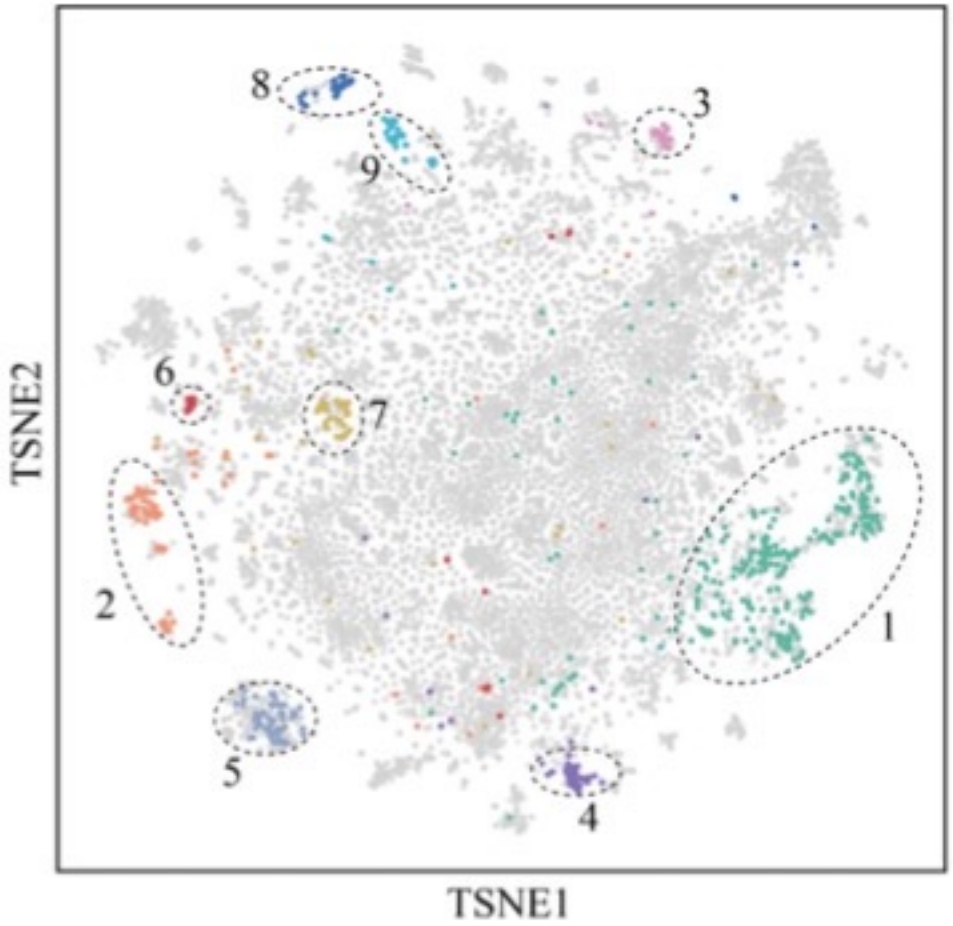
xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. biorxiv

Comparison with Other Large-scale Language Models

Model	Pre-trained Models				Pretraining Techniques
	Parameters	Layers	Datasets	Embedding Dim	
xTrimoPGLM-100B	100B	72	Uniref+ ColabfoldDB	10240	Mixture Obj Masked Language Modeling + Causal Language Modeling
ESM-15B	15B	48	UR50/S 2018_03	5120	Masked Language Modeling
ESM2(3B)	3B	36	UR50/D 2021_04	2560	Masked Language Modeling
ESM2(650M)	650M	33	UR50/D 2021_04	1280	Masked Language Modeling
ESM2(150M)	150M	30	UR50/D 2021_04	640	Masked Language Modeling
OmegaPLM	670M	66(head=1)	Uniref50 2021_04	1280	Masked Language Modeling
ProtBERT	420M	/	/	/	Masked Language Modeling
Prot-T5-XL-Ur50 (3B)	3B	/	/	/	Masked Language Modeling
ProGen	1.2B	36	280M	1028	Casual Language Modeling

xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. biorxiv

Learned Embedding is Informative



1: DNA-binding transcription repressor activity, RNA polymerase II-specific (GO:0001227, n=324)

2: transmembrane transporter activity (GO:0022857, n=107)

3: neuropeptide hormone activity (GO:0005184, n=30)

4: cytoplasmic translation (GO:0002181, n=87)

5: odorant binding (GO:0005549, n=118)

6: delayed rectifier potassium channel activity (GO:0005251, n=32)

7: protein glycosylation (GO:0006486, n=89)

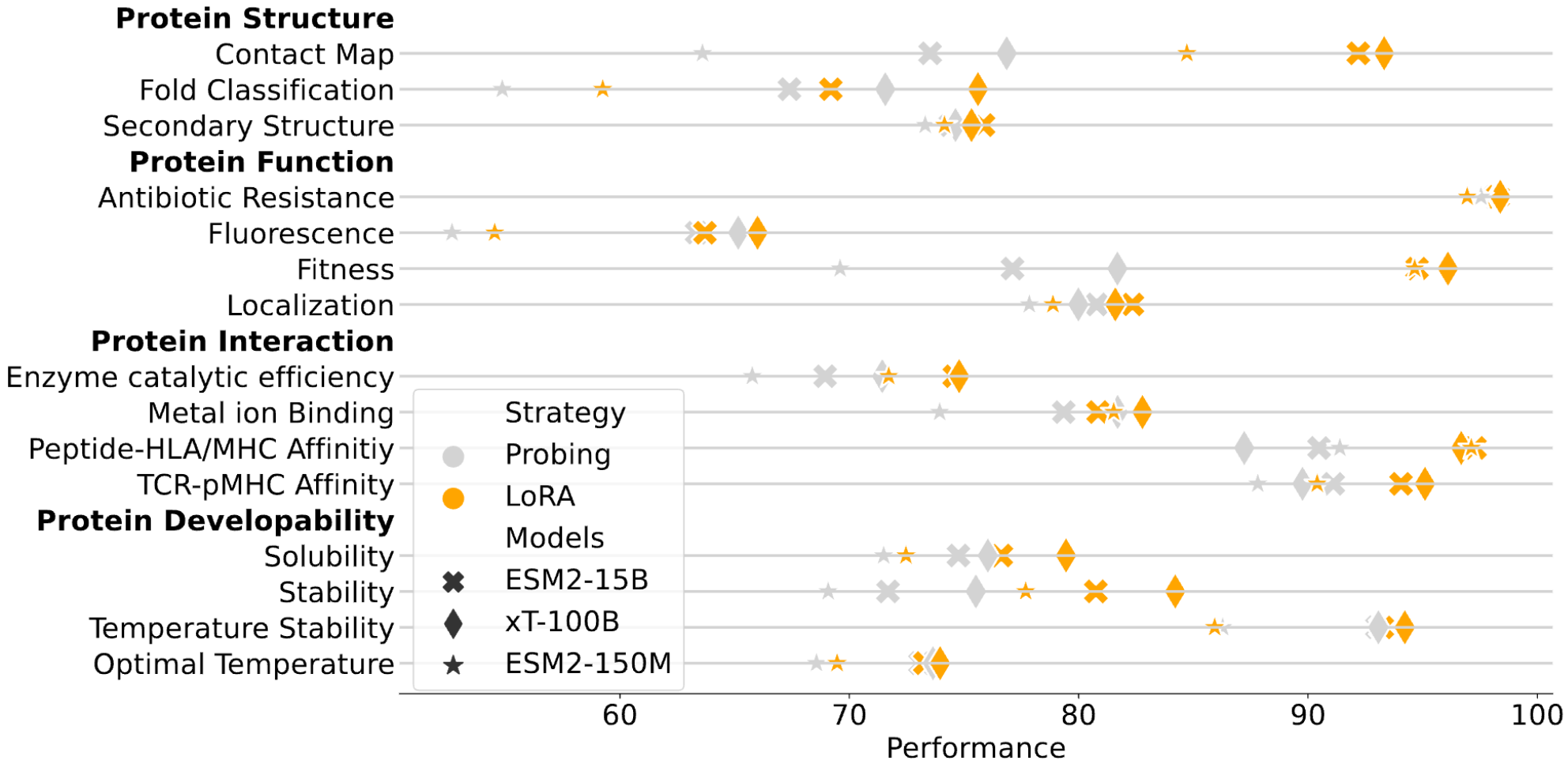
8: immunoglobulin complex (GO:0019814, n=90)

9: collagen trimer (GO:0005581, n=60)

xTrimoPGLM-100B embeddings can discriminative different proteins

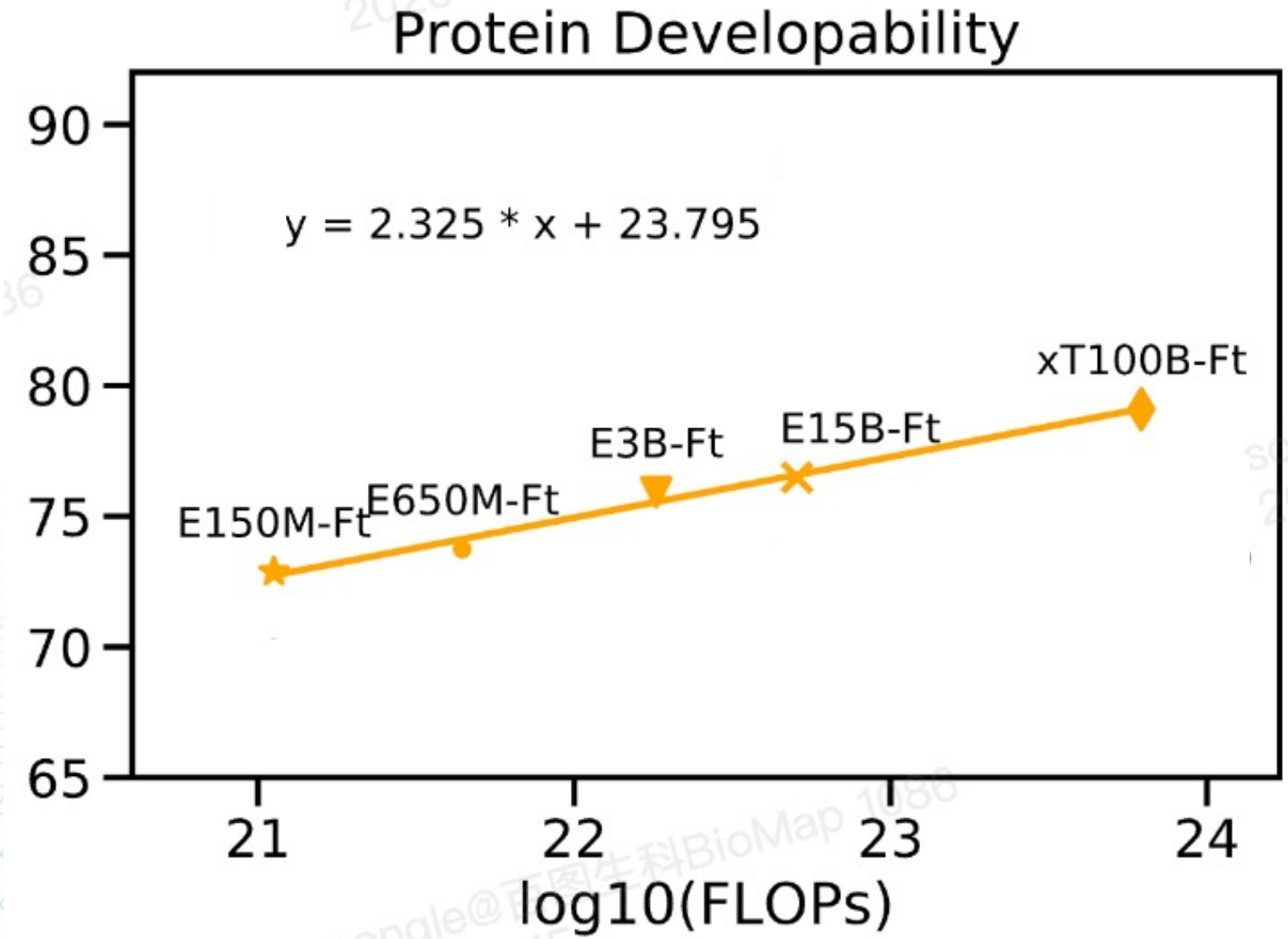
xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. biorxiv

SOTA across Diverse Downstream Tasks



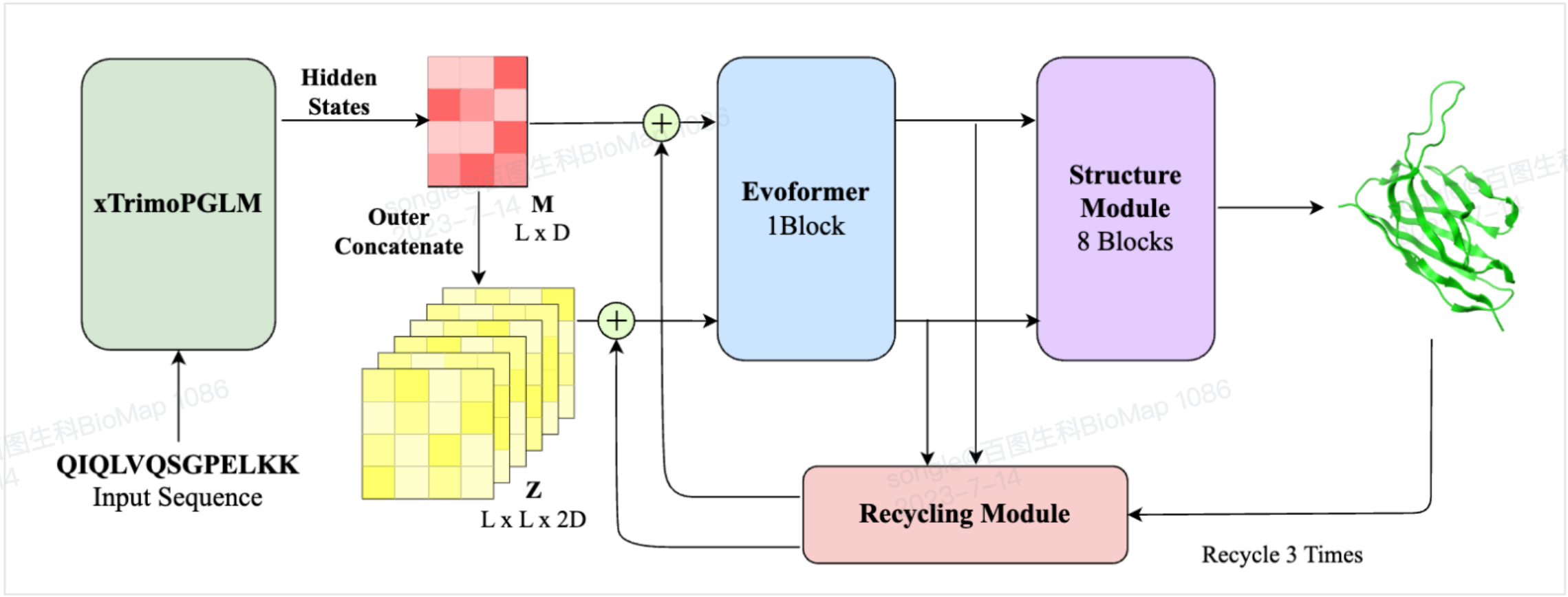
xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. biorxiv

Scaling Law for Protein Language Model



xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. biorxiv

xTrimoPGLM for Antibody Structure Prediction



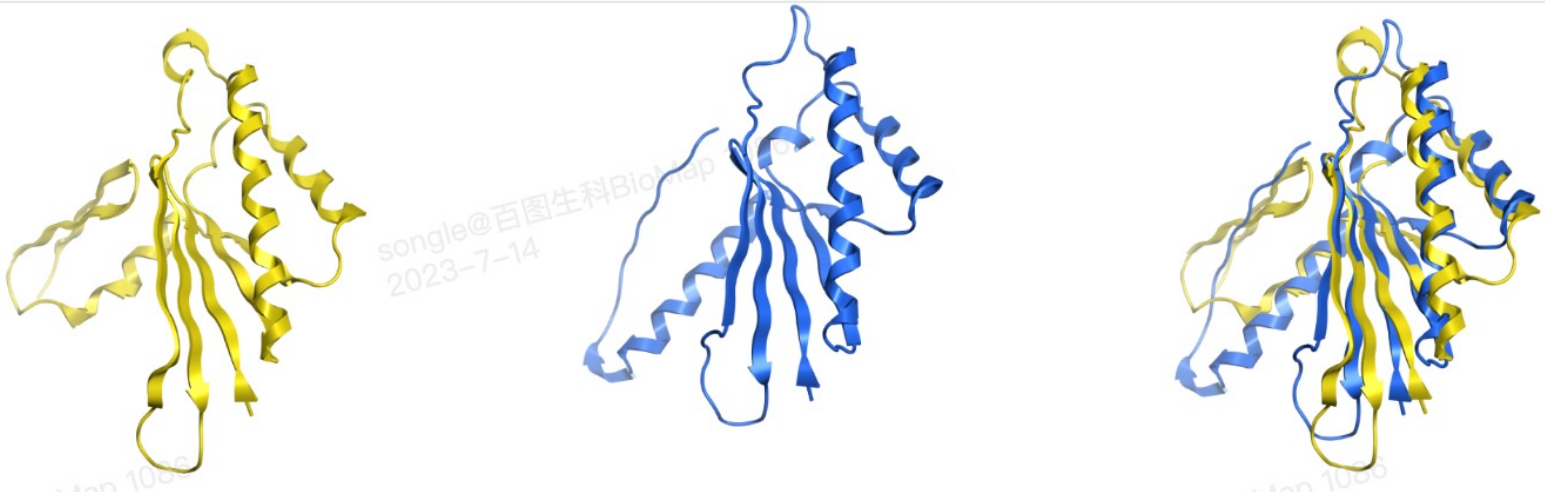
xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. biorxiv

Accurate and Fast Antibody Structure Prediction

	RMSD↓	TM-SCORE↑	DOCKQ ↑	INFERENCE TIME ↓
ZDock	10.982	0.596	0.108	28.8h
ClusPro	5.899	0.792	0.404	34.8h
EquiDock	18.293	0.559	0.032	2m
HDOCK	2.032	0.926	0.705	1.7h
AlphaFold-Multimer	1.325	0.962	0.765	41h (original) 40m (faster MSA)
xTrimoPGLM-AbFold (evo 1)	1.304	0.962	0.762	**32s**
xTrimoPGLM-AbFold (evo 16)	**1.234**	**0.966**	**0.770**	82s

xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. biorxiv

Protein Generation



Case-1: TM-score=0.735, sequence identity=25.1%



Case-2: TM-score=0.594, sequence identity=13.4%

xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. biorxiv

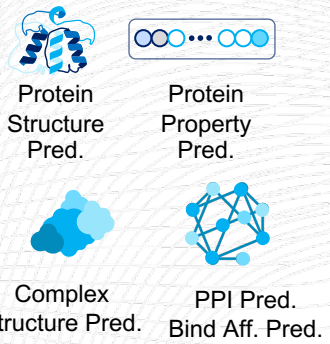
Protein Design with Foundation Models

Conditional Generative Models as Policy

QVQLVQSGV
EVKKPGASVKV ...

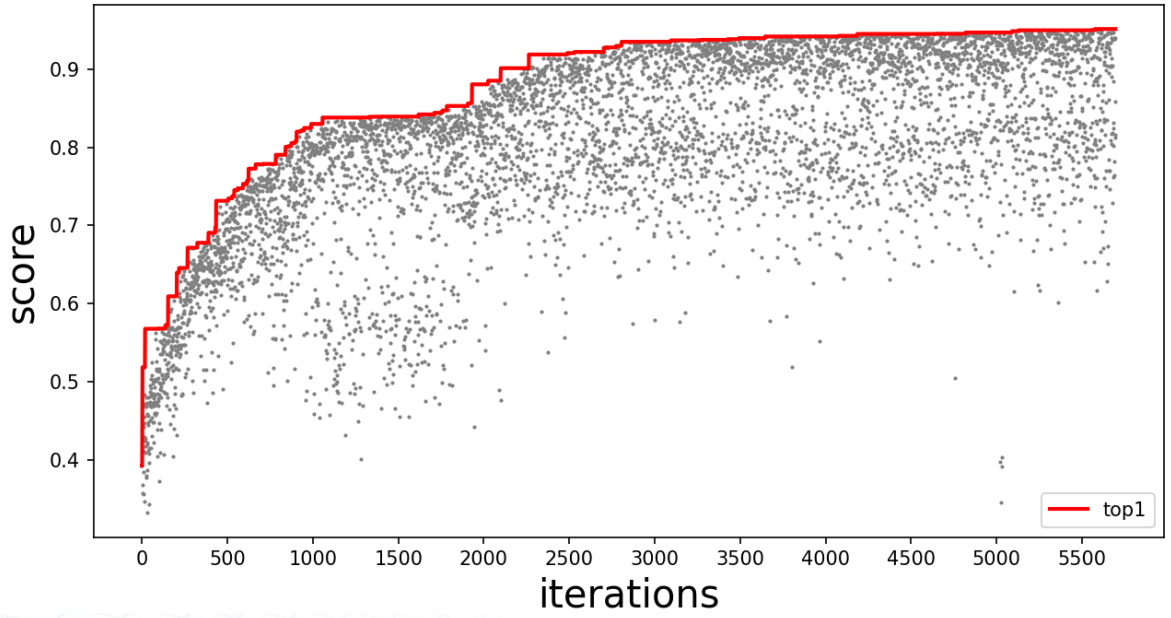
Reinforcement Learning

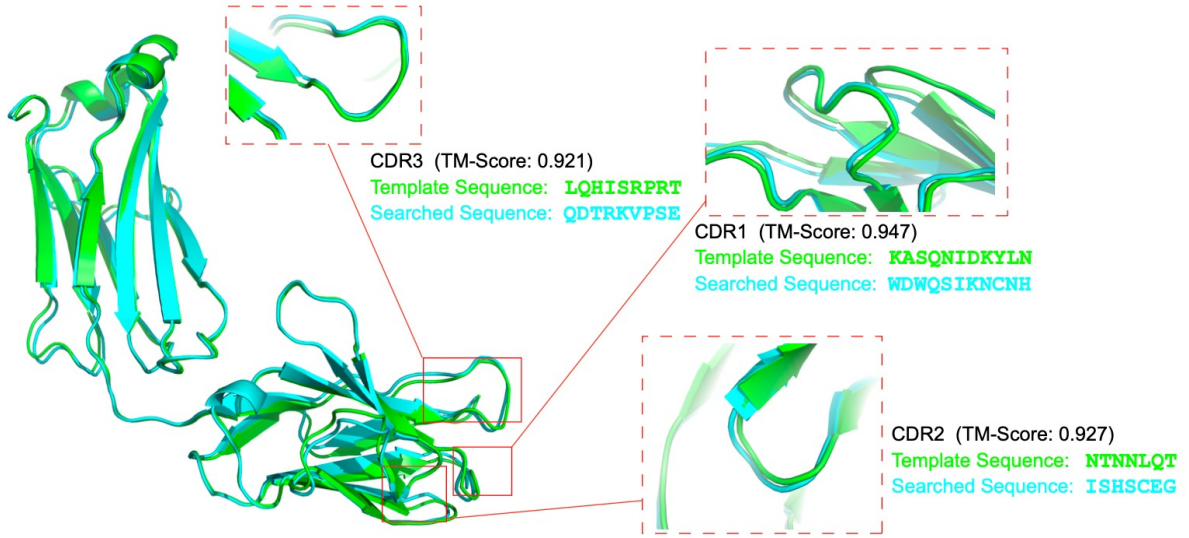
Downstream Task Models as Rewards



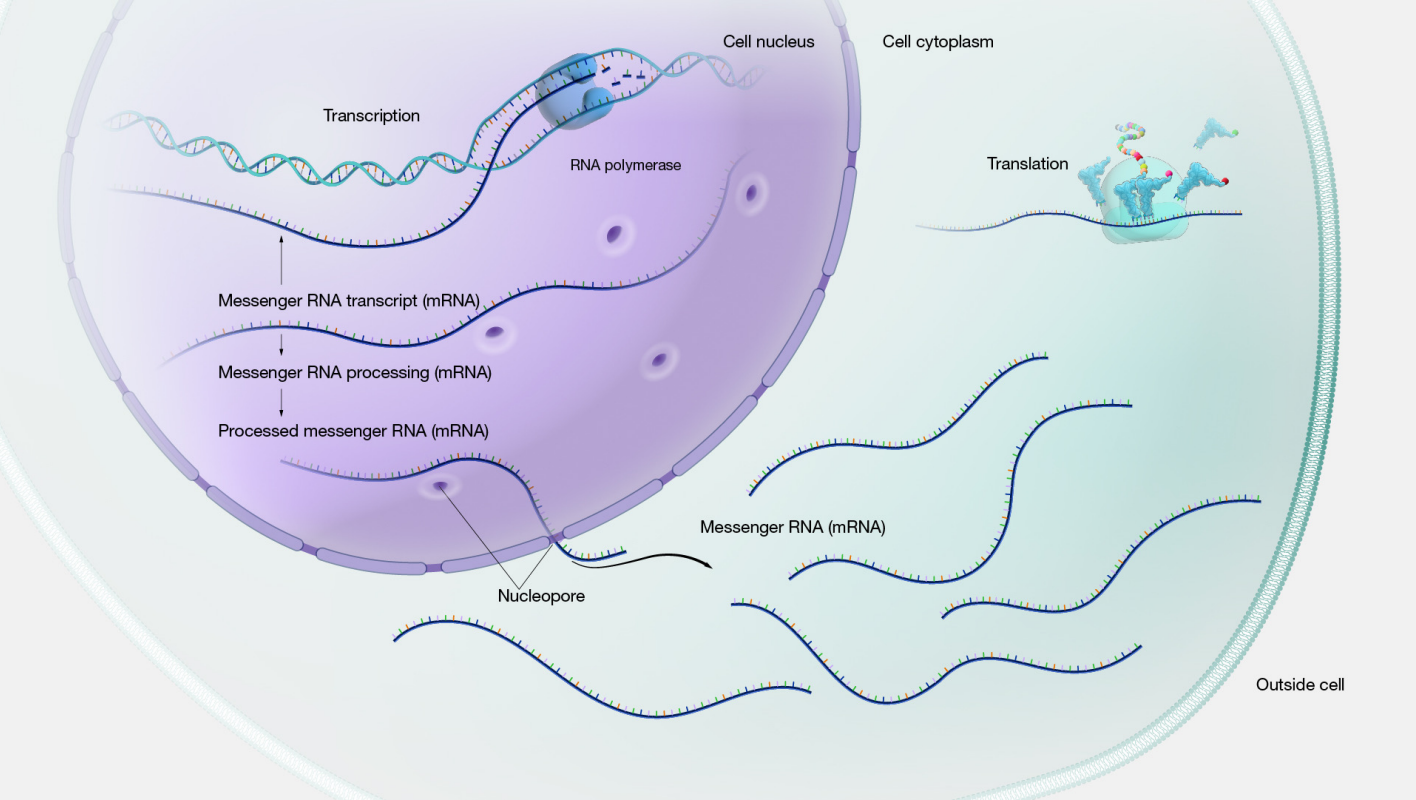
xTrimoPGLM

Antibody Binder Design





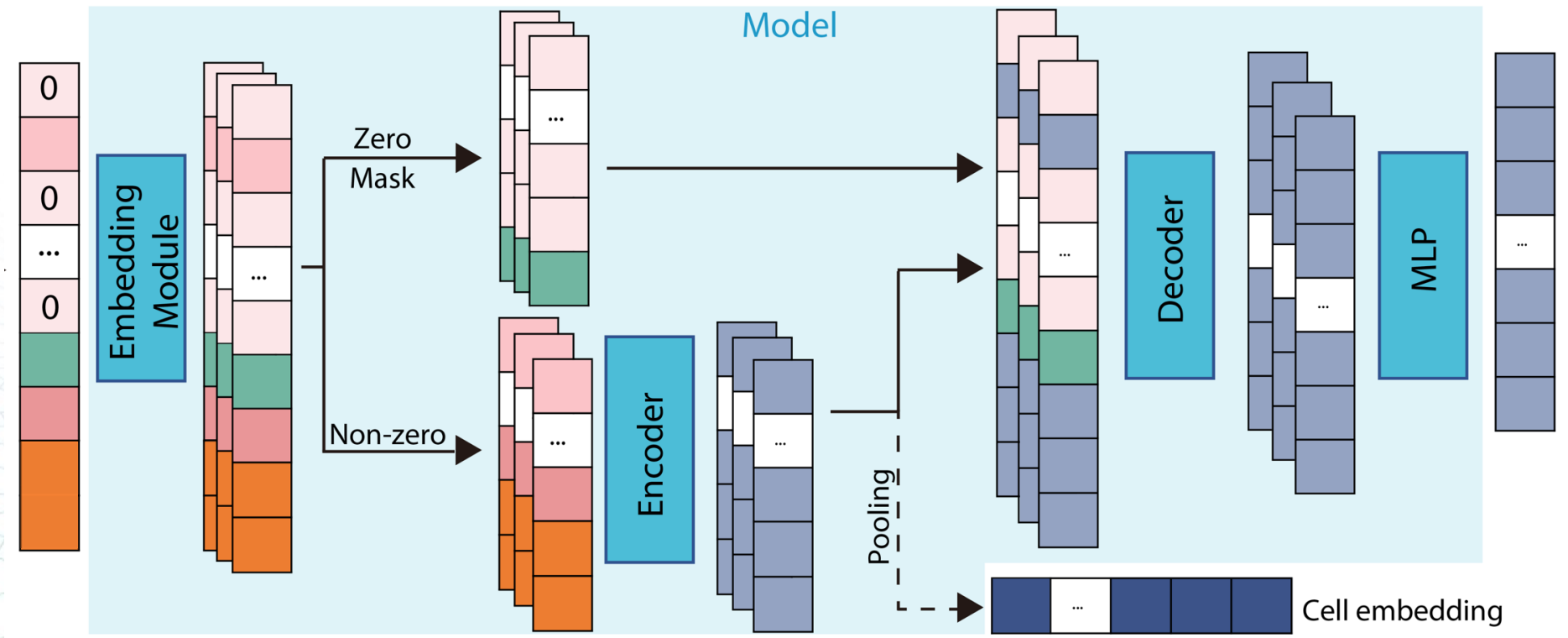
Pretrained Single-Cell Models



Cell \ Gene	1	2	3	...	19262	19263	19264
1	0	1.5	2.5		3.5	4.5	6.5
2	1.4	3.4	0		0	8.4	5.4
3	2.6	0	0		6.6	3.6	1.6
...							
C-2	5.7	2.7	1.7		0	3.7	8.7
C-1	6.8	7.8	8.8		0	1.8	0
C	2.9	0	3.9		7.9	0	1.9

xTrimoGene Transformer for ~20K Genes

- Asymmetric encoder-decoder transformer leveraging sparsity
  - Encoder processes non-zero and non-mask values (non-zero values contain more information)
  - Decoder processes all values

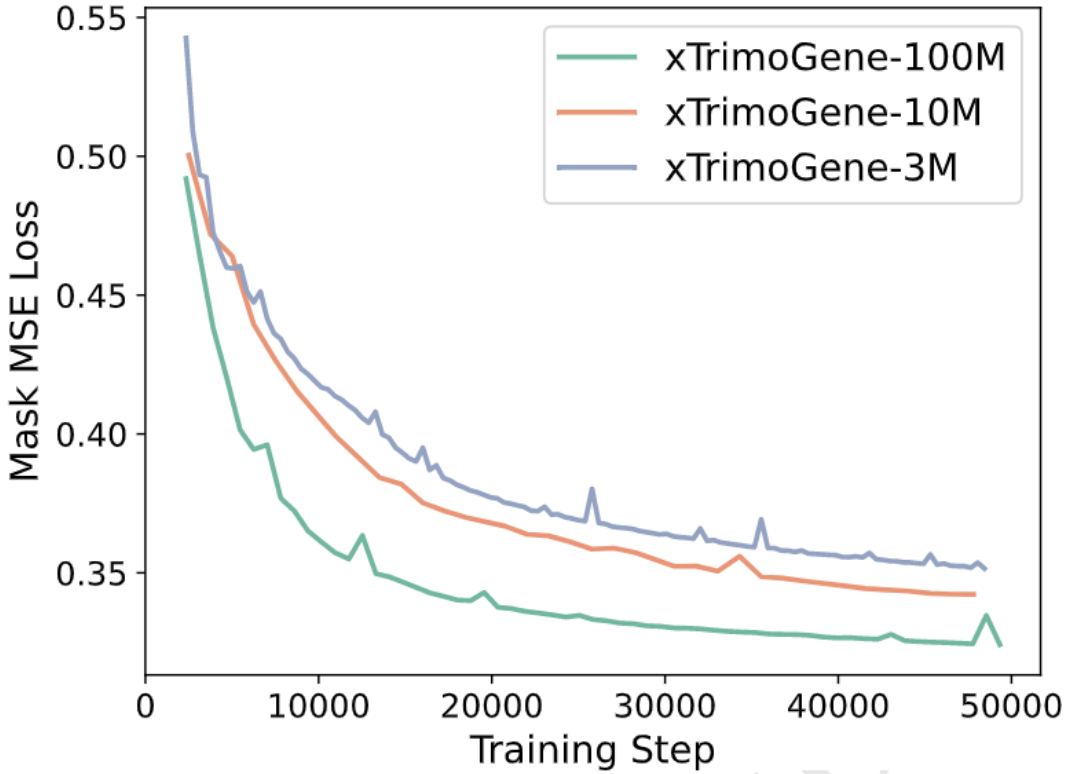


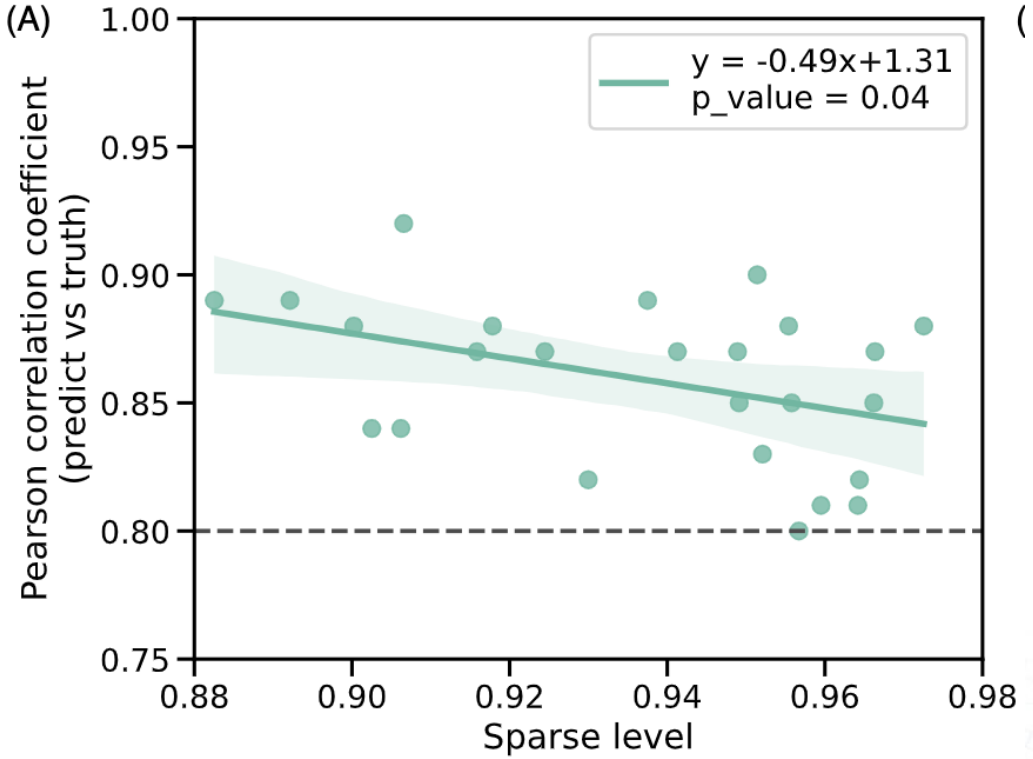
Large Scale Foundation Model on Single-cell Transcriptomics. biorxiv

Model		xTrimoGene	scBERT	scGPT	Geneformer
Data	Size	**50M**	1M	10M	30M
	Resources	Public data: HCA, GEO, hECA…	PanglaoDB	CELLxGENE portal	Public data
	Variability	**264 tissues**	74 tissues	6 tissues	39 organs (tissue coverage not reported)
Pre-training	Model size	**~100M**	~10M	~50M	~30M
	Model architecture	**Asymmetric Encoder-Decoder**	Encoder Only	Encoder Only	Encoder Only
	Design (Layer-Head-Dim)	Encoder Transformer: 12-12-768 Decoder Performer: 6-8-512	Performer: 6-10-200	Transformer: 12-8-512	Transformer: 6-4-512
	Input value	**continuous** normalized expression values	binned normalized expression values	binned normalized expression values	ranked normalized expression values
	Number of input genes (L)	**19,264** protein-coding or mt genes*	16,906 genes	~2,000 non-zero genes	2,048 genes with different ranks
	token number (in 1 epoch training)	50,000,000×19,264 = 963,200,000,000 **(~963B)**	1,000,000×16,906 =16,906,000,000 (~16B)	10,000,000×2,000 =20,000,000,000 (~20B)	30,000,000×2,048 =61,440,000,000 (~61B)
	Model scalability	**Yes**	No (architecture limited)	No (length limited)	No (length limited)
	Training acceleration	MMF	vanilla Pytorch		Huggingface (with Deepspeed)
	Pretraining Task	Read-depth aware masked value prediction**	Masked value prediction	Masked value prediction	Masked value prediction

Properties of the Model

Model name	Parameter (M)	Forward + backward (FLOPs/sample)	Total train (FLOPs)	Resource
Transformer	11.3	9.86E+12	2.46E+20	100%
Performer	8.9	1.06E+12	2.65E+19	10.8%
xTrimoGene	9.8	3.35E+11	8.38E+18	3.4%

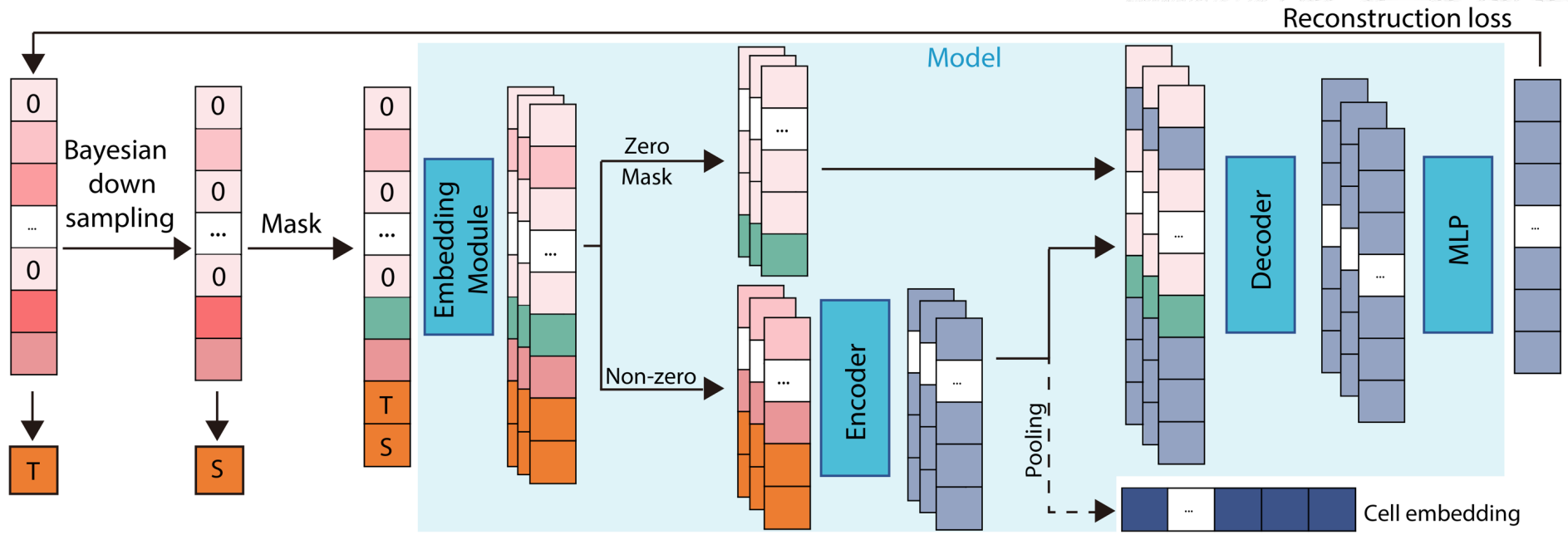




xTrimoGene: An Efficient and Scalable Representation Learner for Single-Cell RNA-Seq Data. Biorxiv

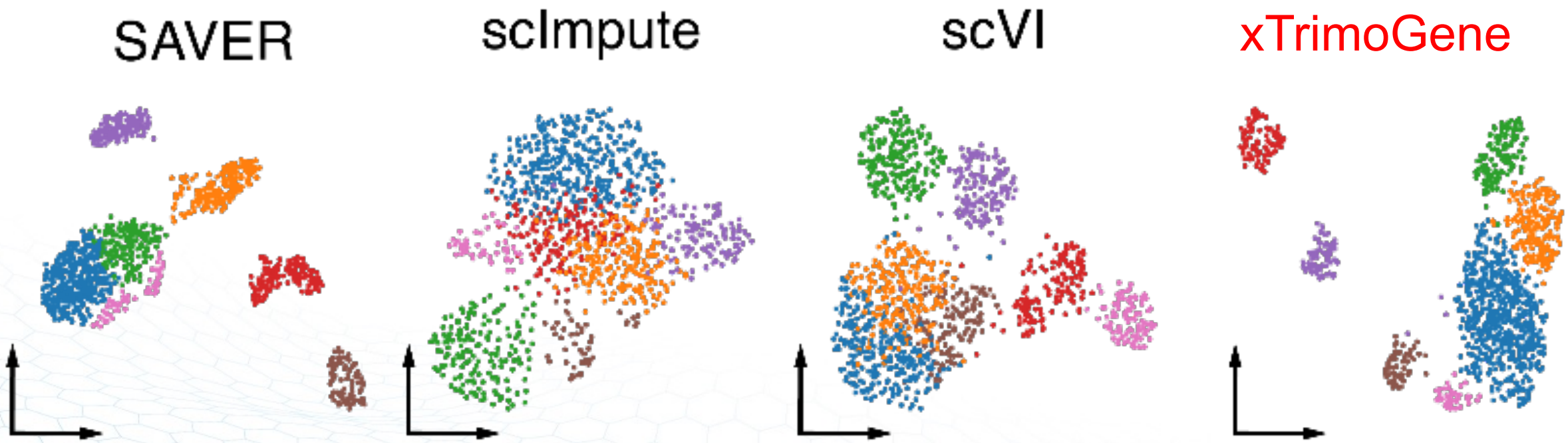
Read Depth Aware Pretraining for Reducing Batch Effects

- Predicting a cell's masked gene expression based on other genes' context
- The context is from a **low read-depth variant** of that cell's gene expression profile
  - Target: original gene expression
  - Input: generated by hierarchical Bayesian downsampling



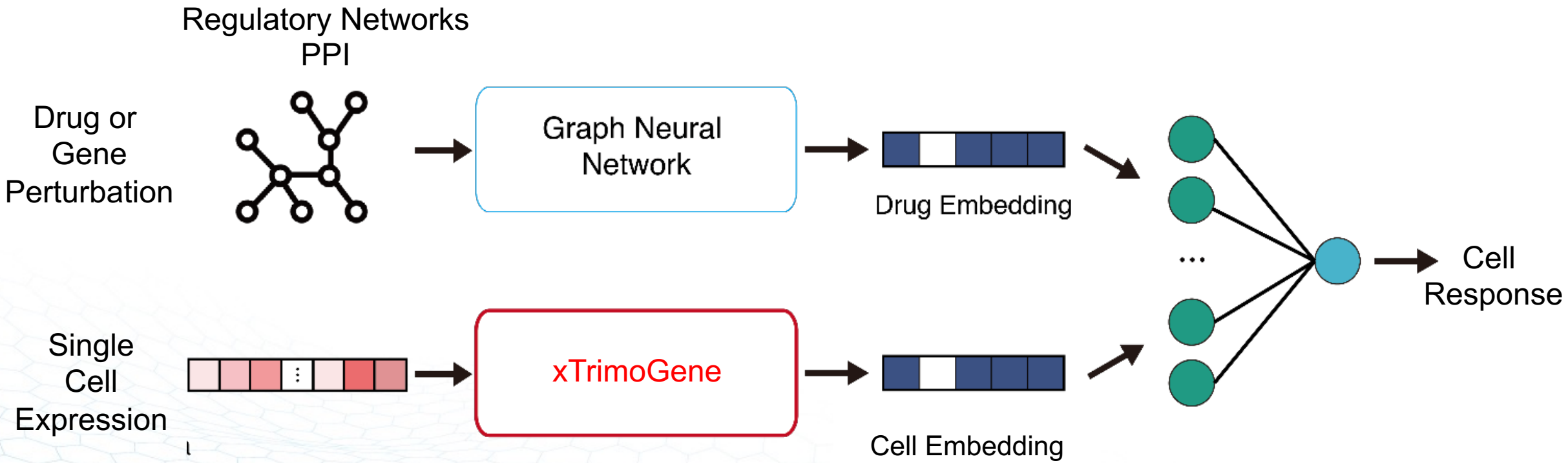
Large Scale Foundation Model on Single-cell Transcriptomics. biorxiv

Transformer Embedding Better Separate Different Cell Types



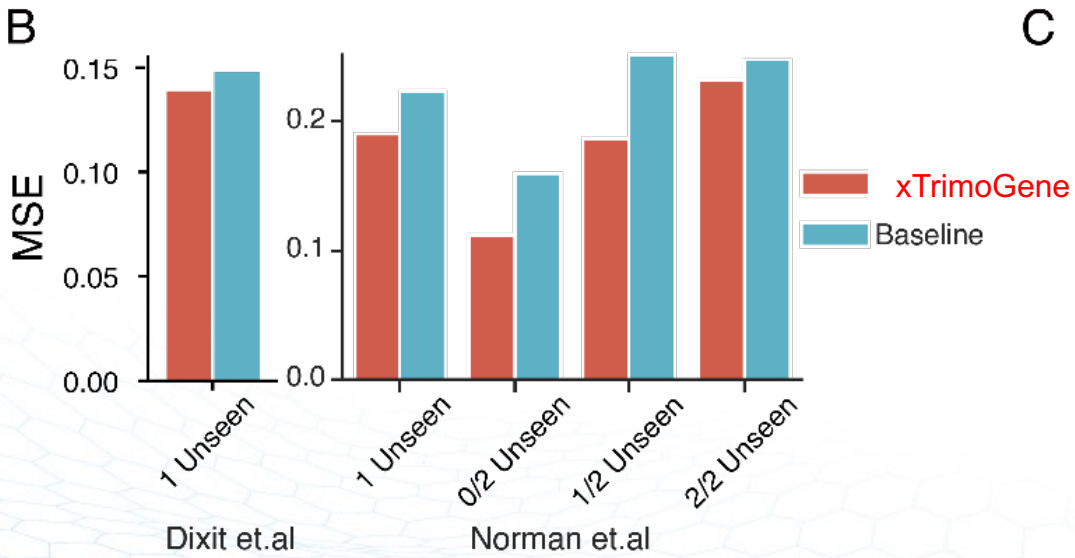
Large Scale Foundation Model on Single-cell Transcriptomics. biorxiv

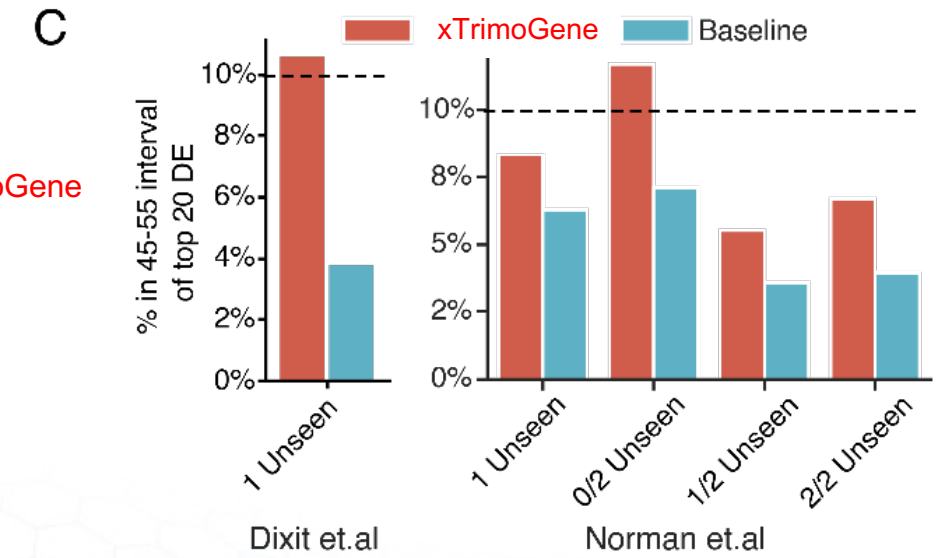
Perturbation Response Prediction Model



Large Scale Foundation Model on Single-cell Transcriptomics. biorxiv

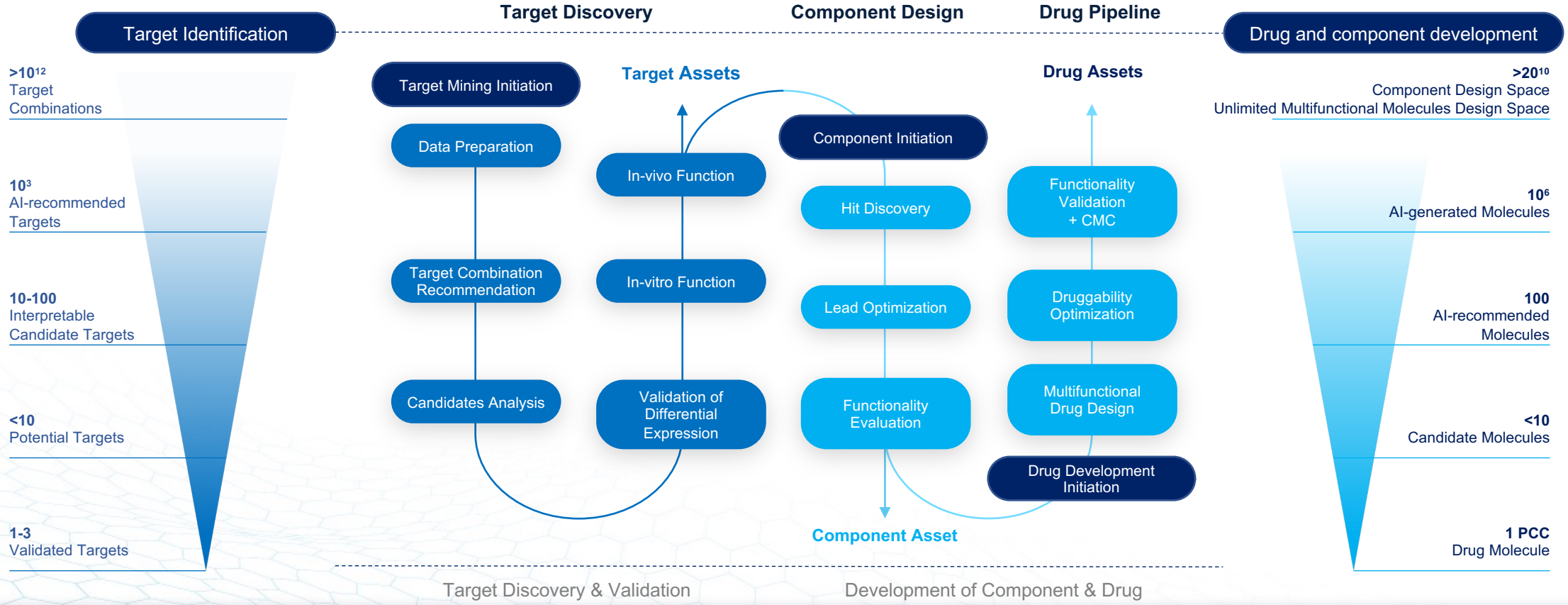
Prediction for CRISPR Perturbation Data





Large Scale Foundation Model on Single-cell Transcriptomics. biorxiv

End-to-End Efficiency Improvement Powered by xTrimo Foundation Models



xTrimo Foundation Models

BioMap 百图生科

xTrimo Life Science Leaderboard

xTrimo, short for Cross-modal Transformer Representation of Interactome and Multi-Omics, is a large-scale multimodal AI system. It consists of a series of pre-trained large language models for life science and the corresponding generative models with hundreds of billions of parameters, and multiple families of growing number of downstream tasks.

CHALLENGE

28 Tasks
https://www.biomap.com/sota/

State-of-the-Art Overview

All | General Protein | Antibody Related | New Drug Modality Related | Enzyme Related | Novel Protein Design | Cell Related

Task	Leading Model	Metric	BioMap Score	Prev.Sota	Prev.Sota Method
Amino Acid Contact Prediction	xTrimo-Biomap	Accuracy	93.32	92.19	ESM-2 (15B)