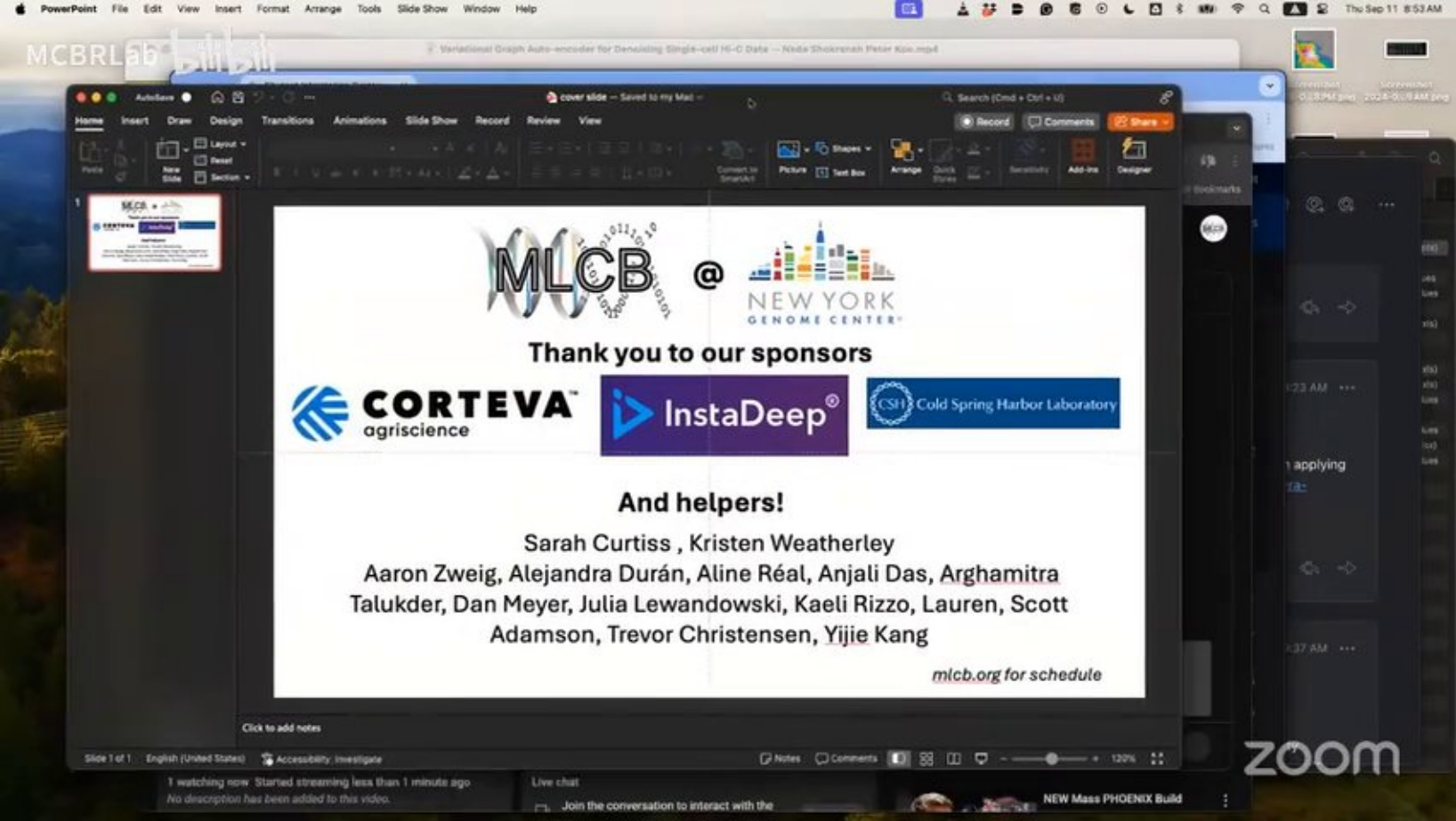
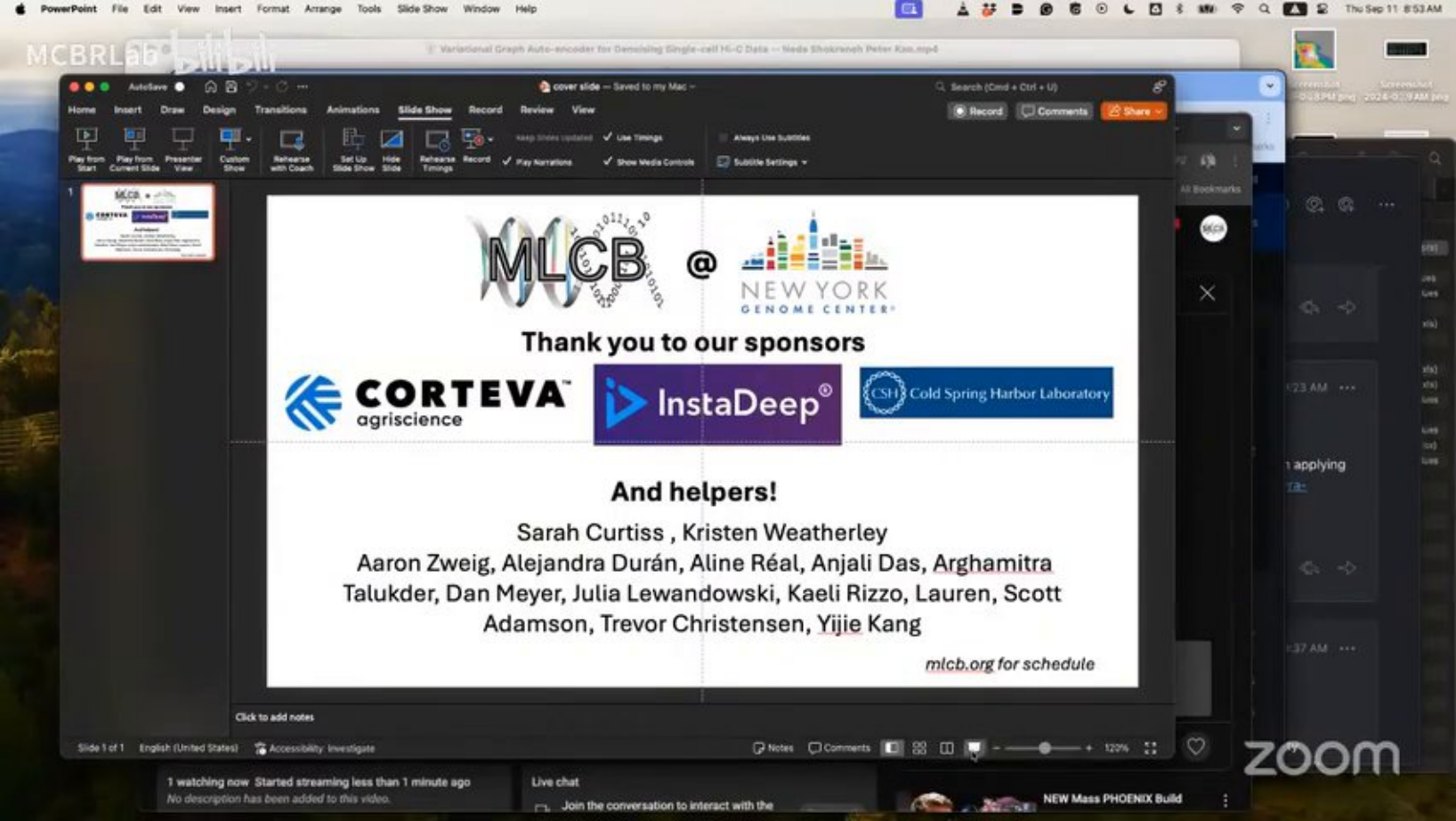
MCBRLab bilibili

NYGC Events

zoom


MCBRLab
MLCB @ NEW YORK GENOME CENTER
Thank you to our sponsors
CORTEVA agriscience
InstaDeep
Cold Spring Harbor Laboratory
And helpers!
Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang
mlcb.org for schedule
zoom


MLCB @ NEW YORK GENOME CENTER
Thank you to our sponsors
CORTEVA agriscience
InstaDeep
Cold Spring Harbor Laboratory
And helpers!
Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang
mlcb.org for schedule
Click to add notes
1 watching now Started streaming less than 1 minute ago
No description has been added to this video.
Live chat
Join the conversation to interact with the
NEW Mass PHOENIX Build
zoom

MLCB @ NEW YORK GENOME CENTER

Thank you to our sponsors

CORTEVA agriscience

InstaDeep®

CSH Cold Spring Harbor Laboratory

And helpers!

Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang

mlcb.org for schedule

MCBRLab bilibili

NYGC Events

zoom

MLCB @ NEW YORK GENOME CENTER

Thank you to our sponsors

CORTEVA™ agriscience

InstaDeep®

CSH Cold Spring Harbor Laboratory

And helpers!

Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang

mlcb.org for schedule

MCBRLab bilibili

zoom

MLCB @ NEW YORK GENOME CENTER

Thank you to our sponsors

CORTEVA agriscience

InstaDeep®

CSH Cold Spring Harbor Laboratory

And helpers!

Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang

mlcb.org for schedule

MCBRLab bilibili

NYGC Events

zoom

Understanding and designing regulatory DNA using machine learning

Jacob Schreiber
Genomics and Computational Biology
UMass Chan Medical School

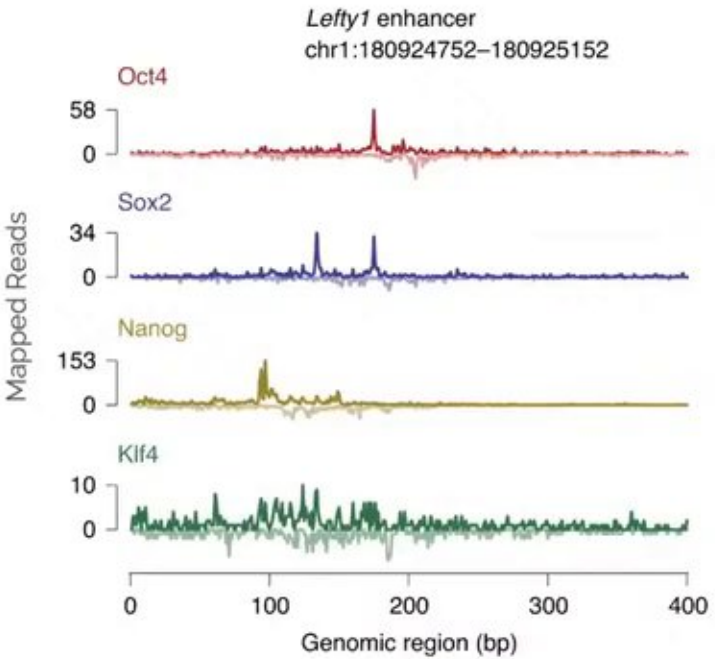

jmschreiber91


@jmschrei


@jmschreiber91

The first assembly of a human reference genome left us with ...

These assemblies were augmented with high-throughput assays measuring cell type-specific signals

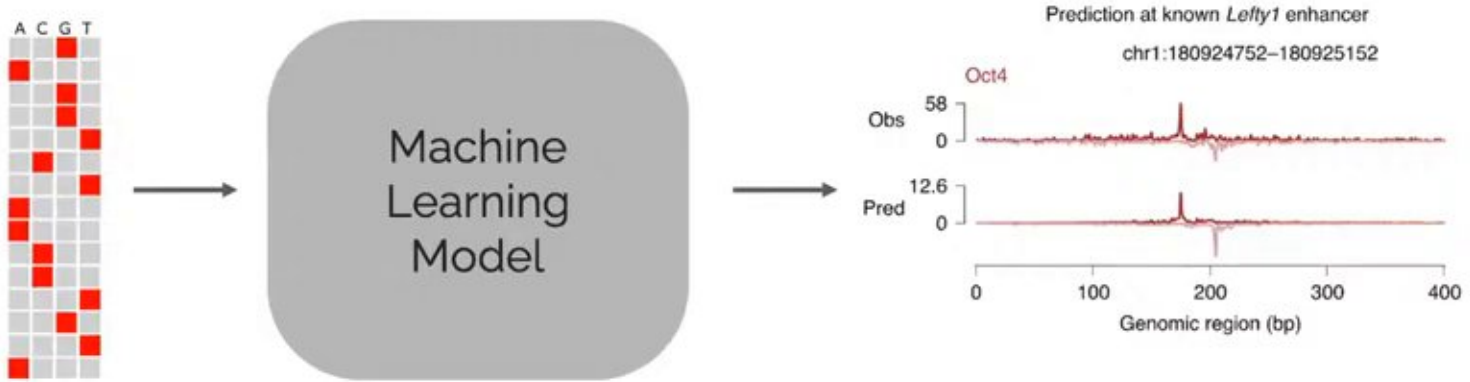


Avsec *et al.* (2021)

Sequence-based machine learning models can be trained to make **predictions**...

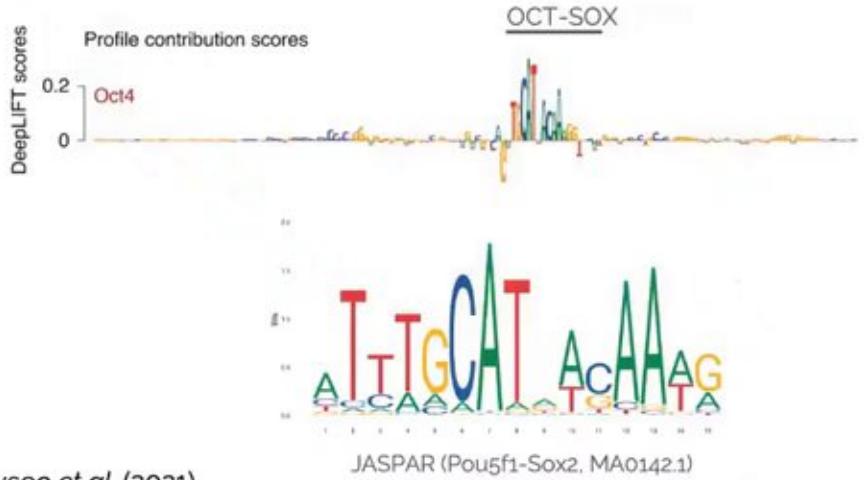
Avsec *et al.* (2021)

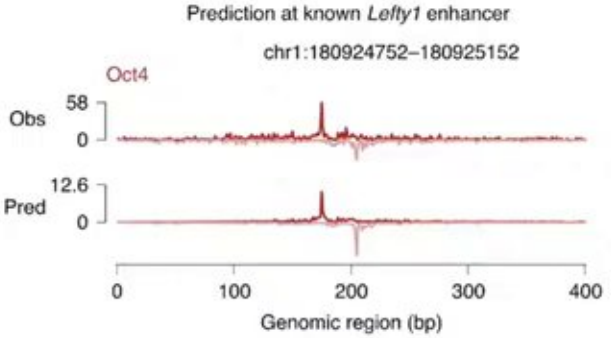
Sequence-based machine learning models can be trained to make **predictions**...



Avsec *et al.* (2021)

Sequence-based machine learning models can be trained to make **predictions**... and highlight nucleotides driving activity with **attributions**





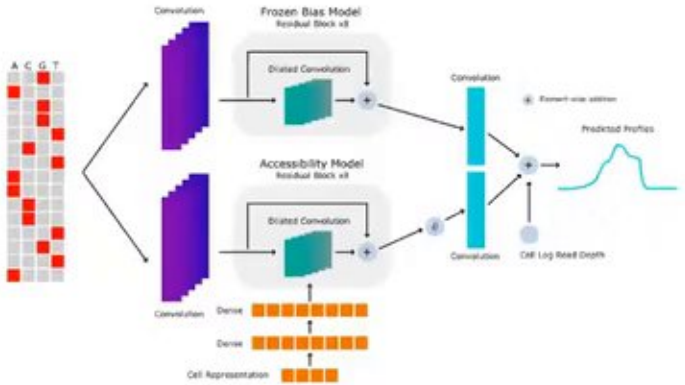
Avsec *et al.* (2021)

The same idea can then be used to **design** sequences with desired properties

Machine Learning Model

Desired Output

Understanding *cis*-regulatory elements at single-cell resolution with DragoNNFruit



Improvements in single-cell tech allow us to investigate continuous biological processes such as cellular reprogramming

Donor-derived dermal fibroblasts

Log Expression (tp10k)

Fibroblasts

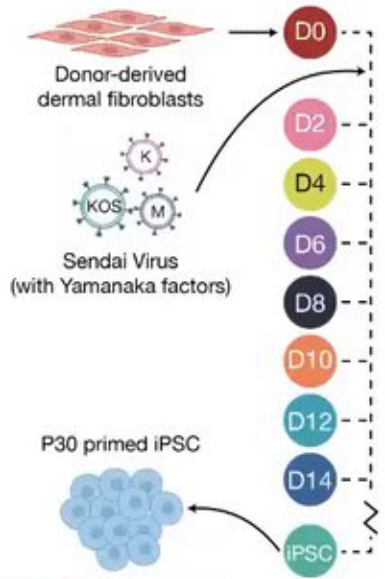
Fibroblasts

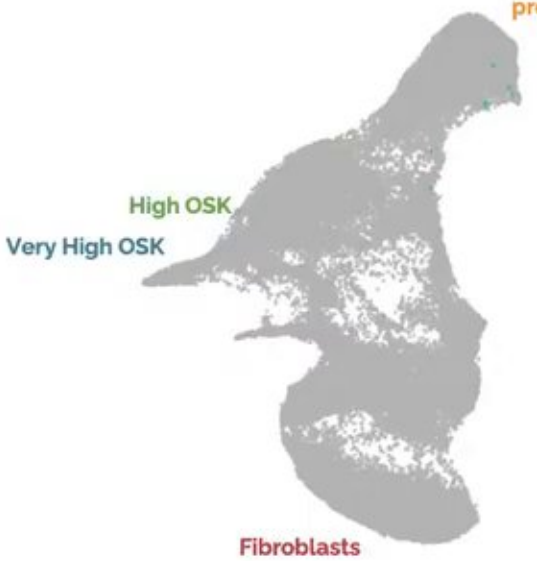
Ȯ Ṡ K̇ Ṁ

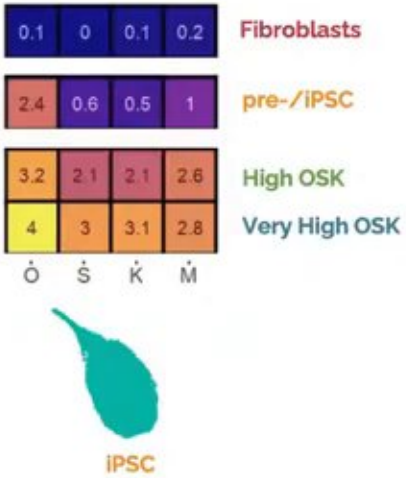
Surag Nair, Mo Ameen

Gene expression from harmonized parallel scRNA-seq, not an exact mapping

Improvements in single-cell tech allow us to investigate continuous biological processes such as cellular reprogramming

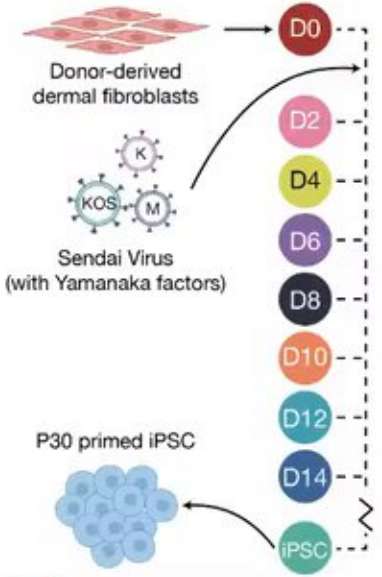




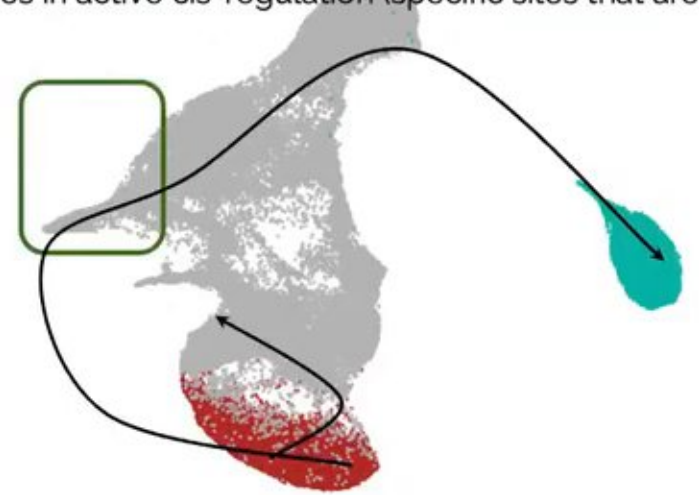


Gene expression from harmonized parallel scRNA-seq, not an exact mapping

Most processes are driven by a complex interplay between cis- (sequence-based) and trans- (protein- and other molecule-based) regulation

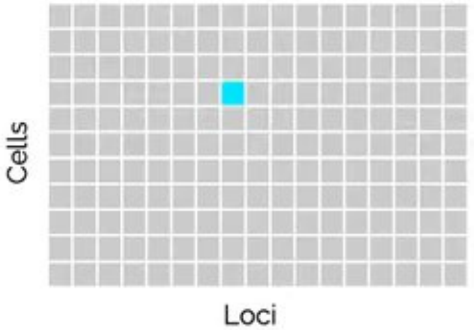


High ectopic expression of OSKM trans-factors drive rapid changes in active cis-regulation (specific sites that are bound)

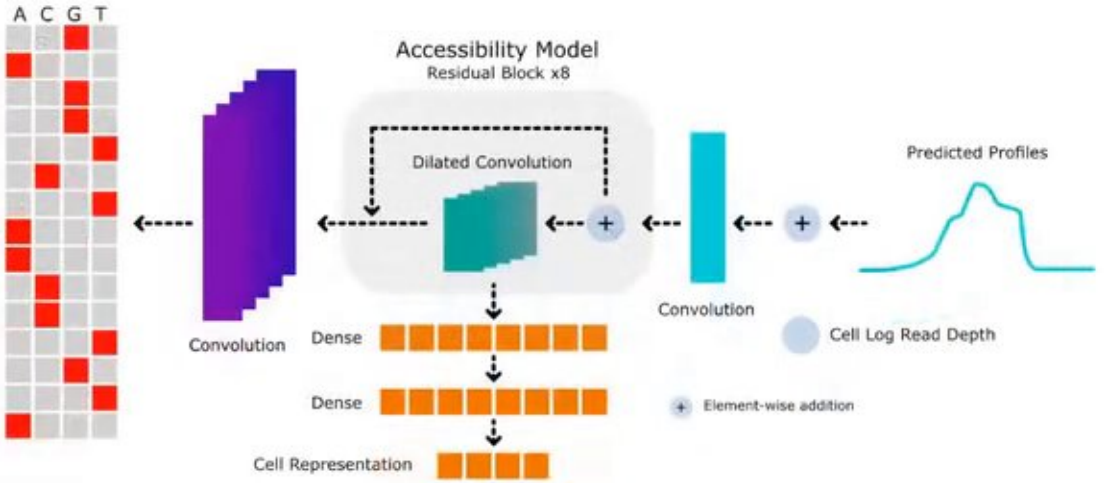


However, most analysis pipelines involve clustering+pseudobulking the data and treating it as a set of discrete states

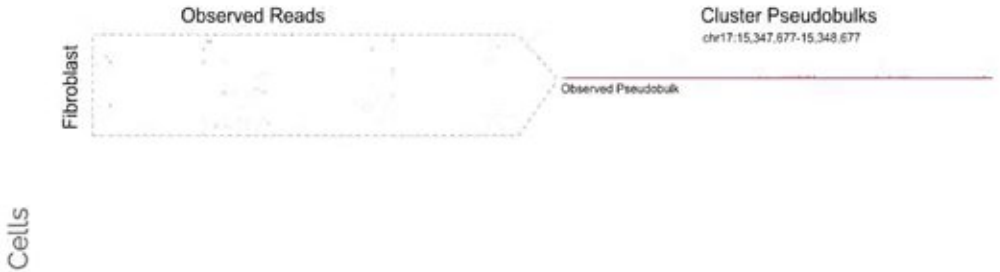
DragoNNFruit extends sequence-based modeling frameworks to single-cell data



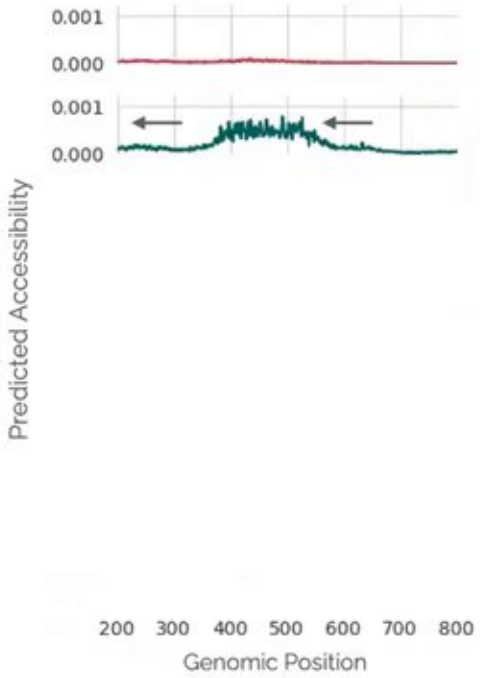
DragoNNFruit extends sequence-based modeling frameworks to single-cell data



DragoNNFruit accurately predicts accessibility as it changes across reprogramming



Some peaks appear to shift coordinates as proteins alter their binding



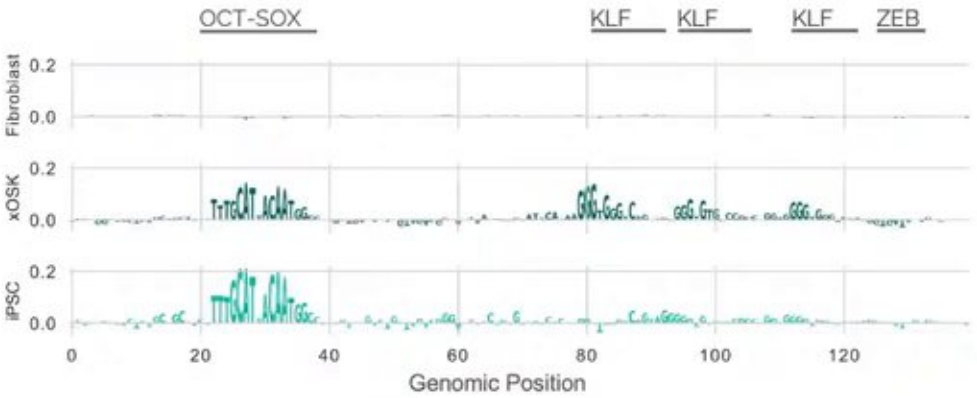
Some sites exhibit persistent accessibility but change their reasons why

Observed Reads

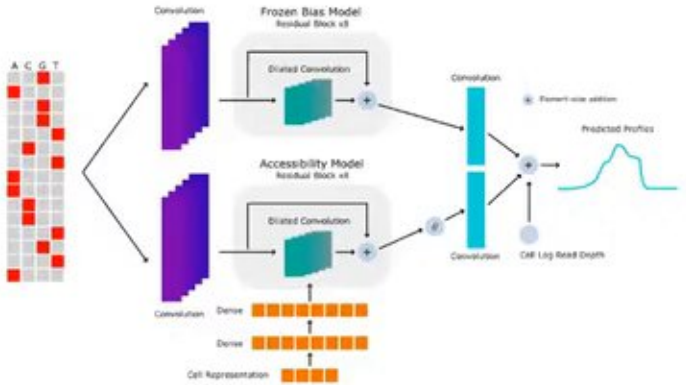
Predicted Accessibility

Accessibility at the NANOG promoter is driven by the usual suspects…

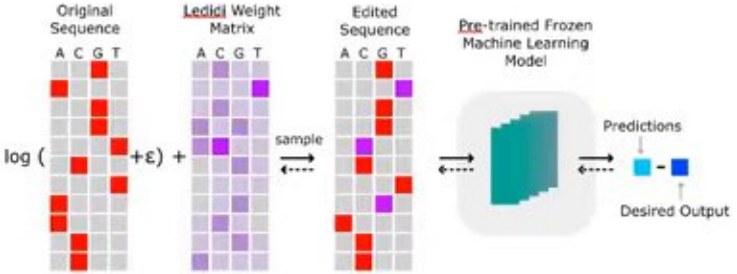
chr12:7,789,855-7,790,035 Nanog Promoter



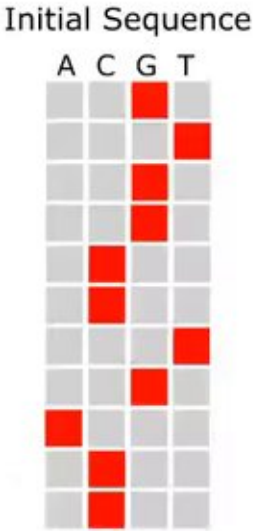
Understanding *cis*-regulatory elements at single-cell resolution with DragoNNFruit



Designing *cis*-regulatory elements using Ledidi



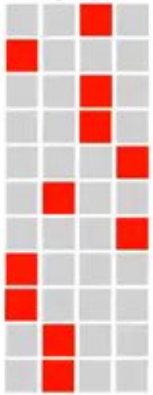
Inverting the normal machine learning paradigm to **design** sequences with desired properties



Ledidi turns the edit design task into a simple optimization problem

Initial Sequence

A C G T



Ledidi Weight Matrix

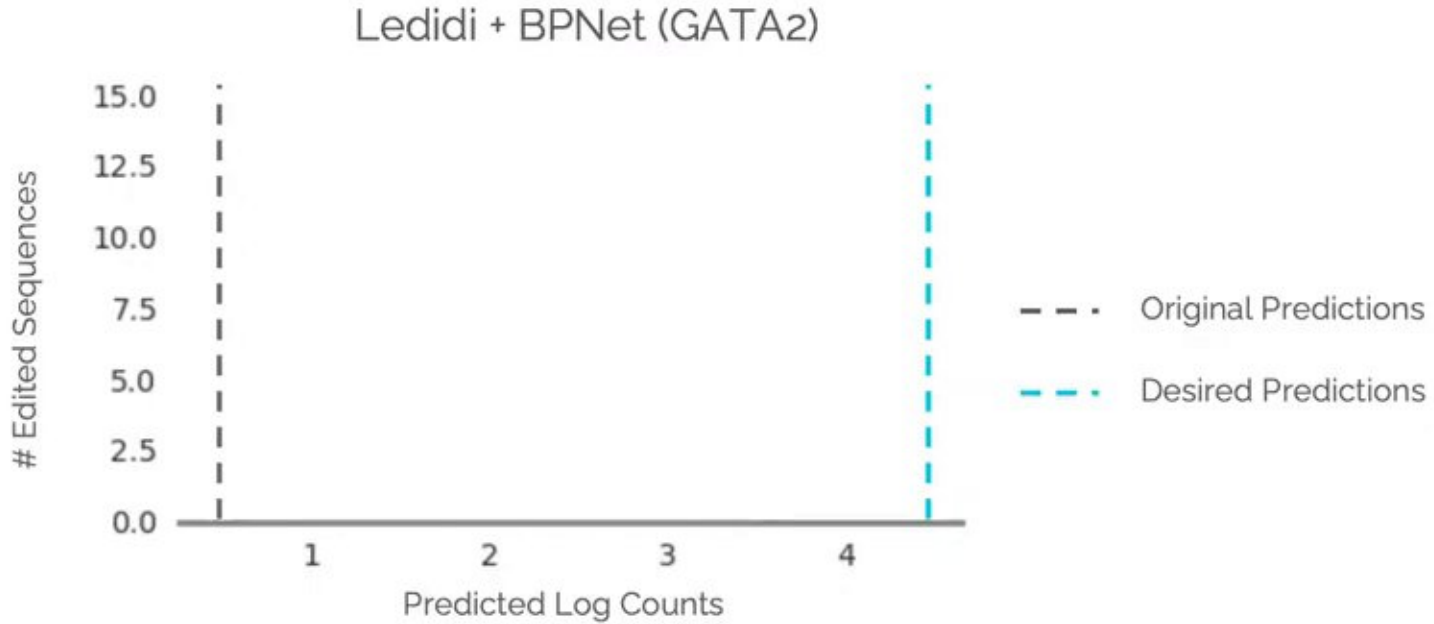
A C G T

Edited Sequence

A C G T

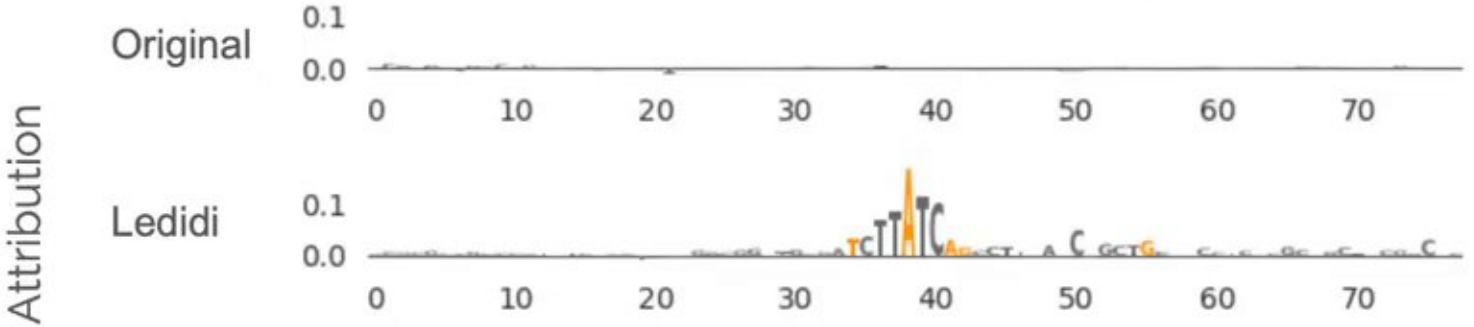
Pre-trained Frozen Machine Learning Model

An initial test shows Ledidi can design edits predicted to greatly increase GATA binding

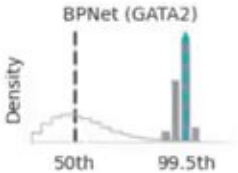


The number of edits is significantly lower than expected by precisely targeting regions that are "poised" to become binding sites

(colored characters show edits)



Ledidi can design edits yielding desired characteristics in a variety of settings



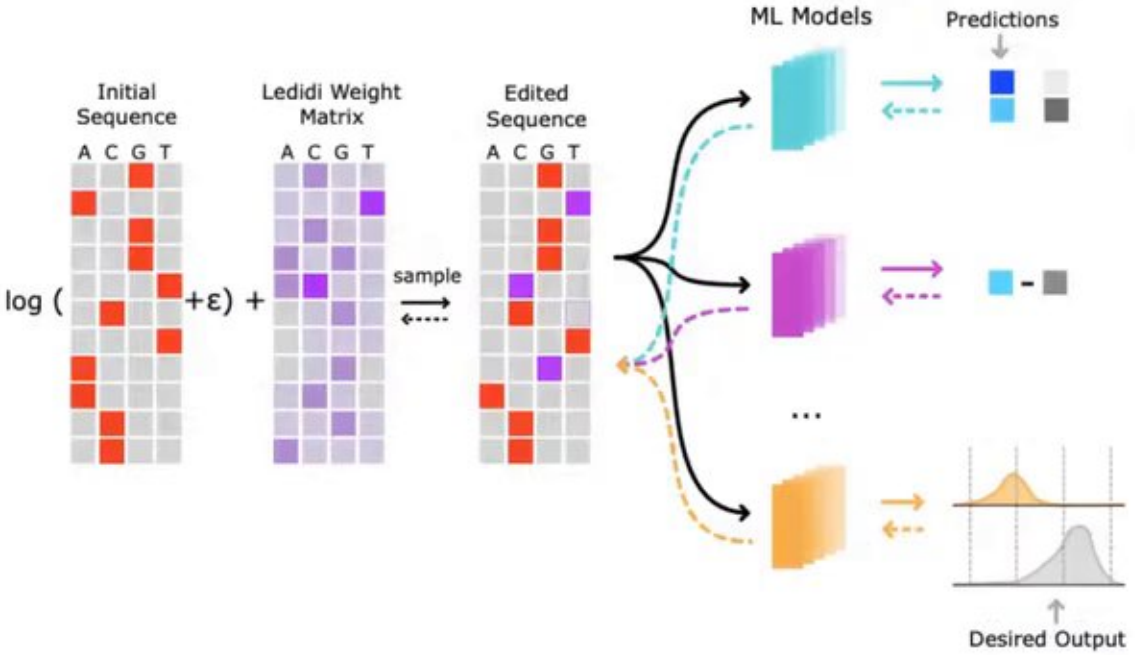
Original Predictions

Desired Predictions

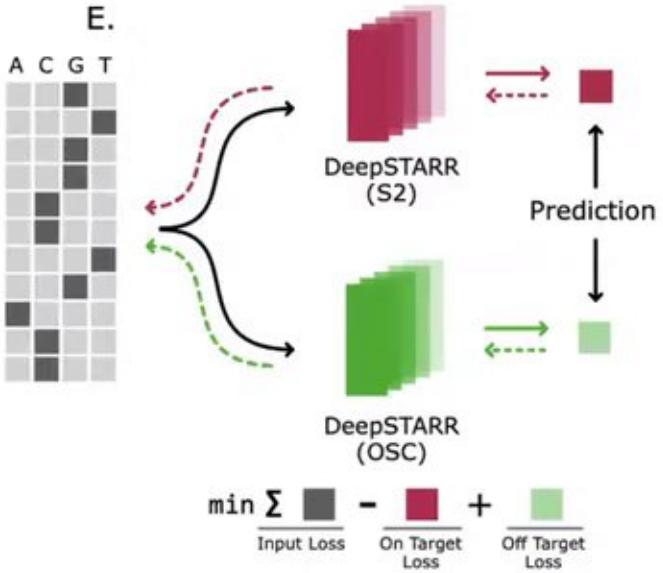
Predicted Genomic Distribution

Predictions on Ledidi-edited Sequences

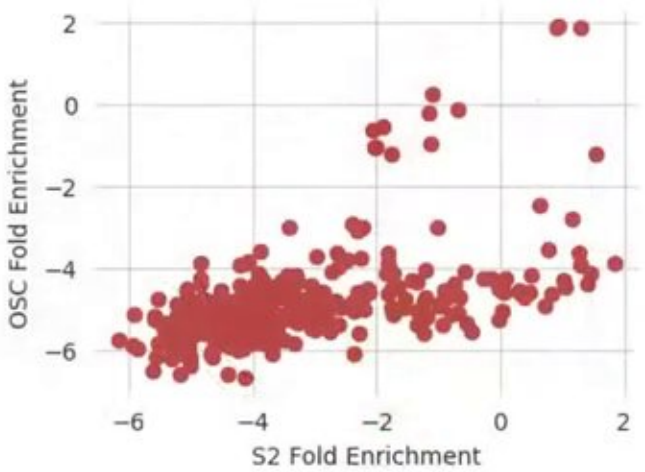
Ledidi allows multiple models to be used simultaneously for design

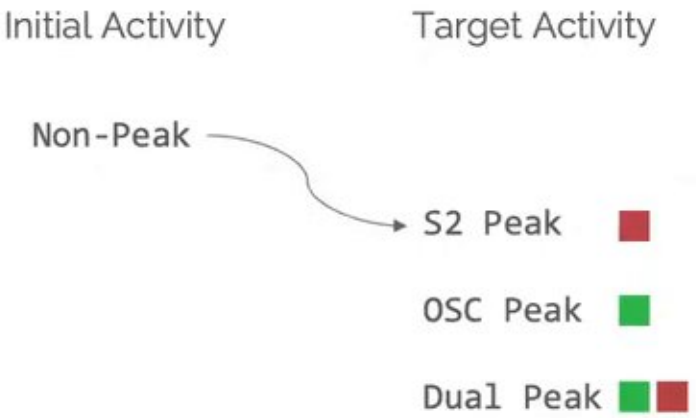


Ledidi can use multiple DeepSTARR models to design cell type-specific enhancers



Ledidi can use multiple DeepSTARR models to design cell type-specific enhancers



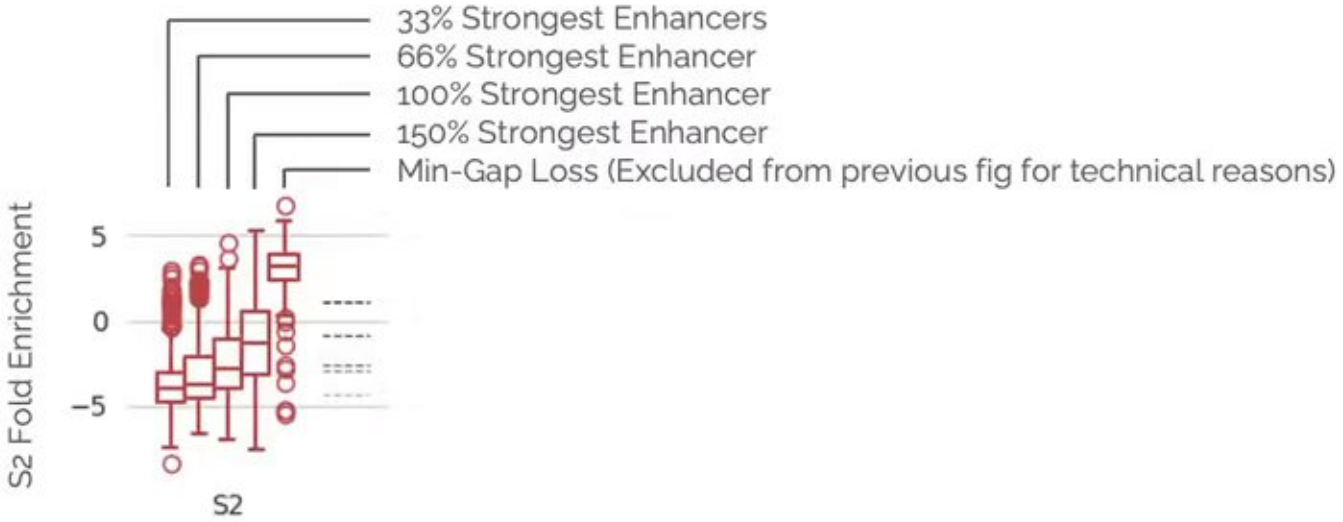


Franziska Lorbeer

Real STARR-seq data!

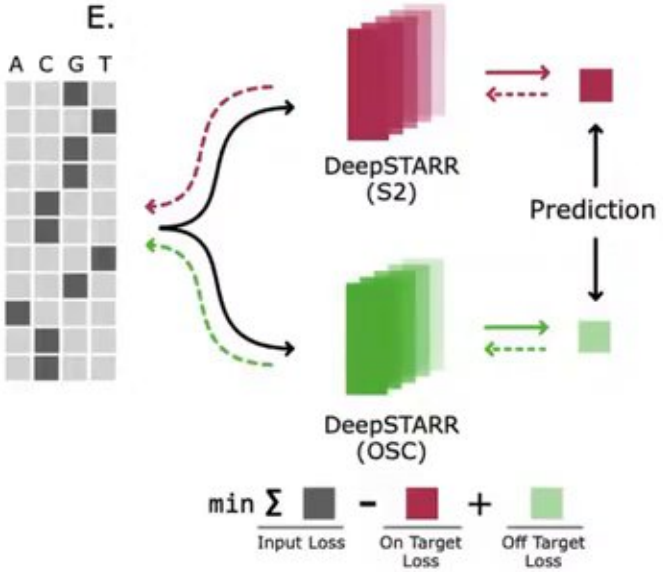
Each dot is a designed regulatory element

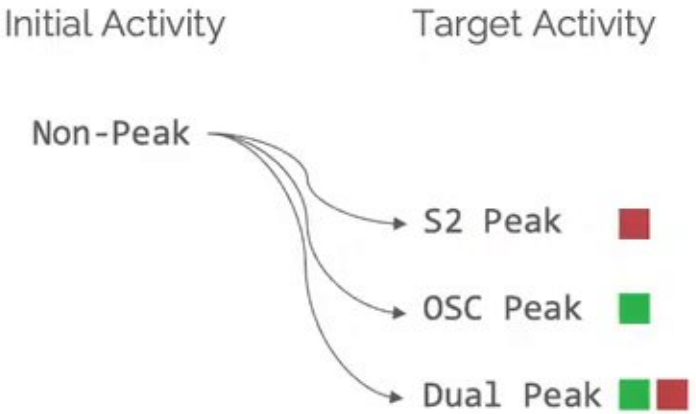
These designed enhancers intentionally exhibit a range of activities



Franziska Lorbeer

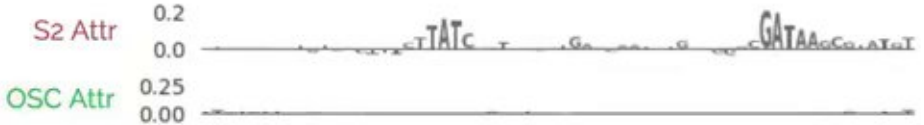
Ledidi can use multiple DeepSTARR models to design cell type-specific enhancers



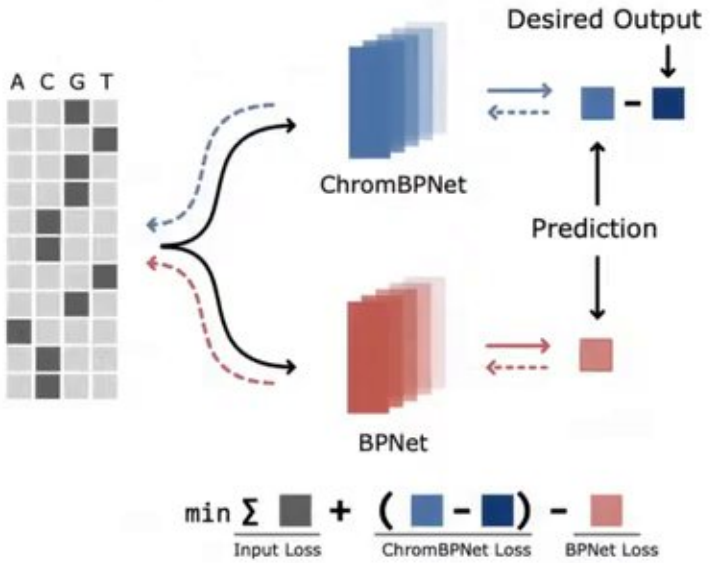


Ledidi can turn a weak S2-specific enhancer into an strong OSC-enhancer

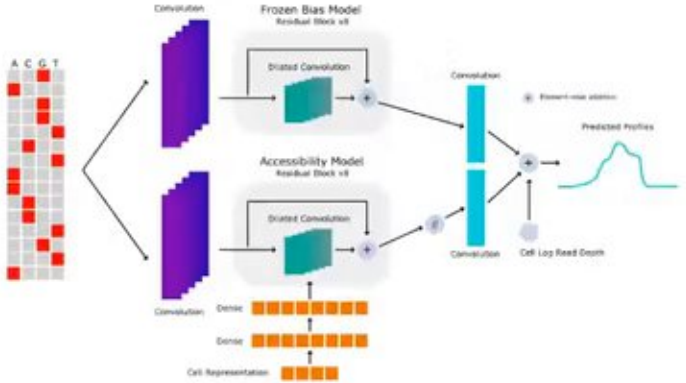
Original S2 Enhancer



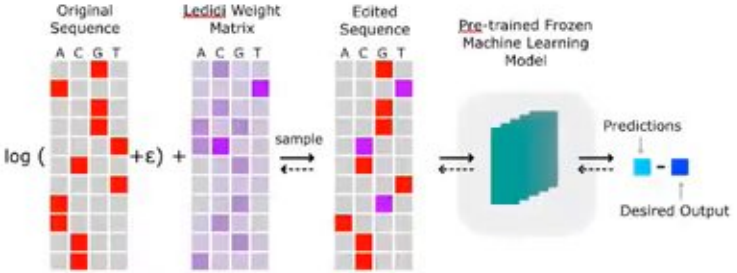
We can use multiple models to programmatically control which TFs drive accessibility



Understanding *cis*-regulatory elements at single-cell resolution with DragoNNFruit



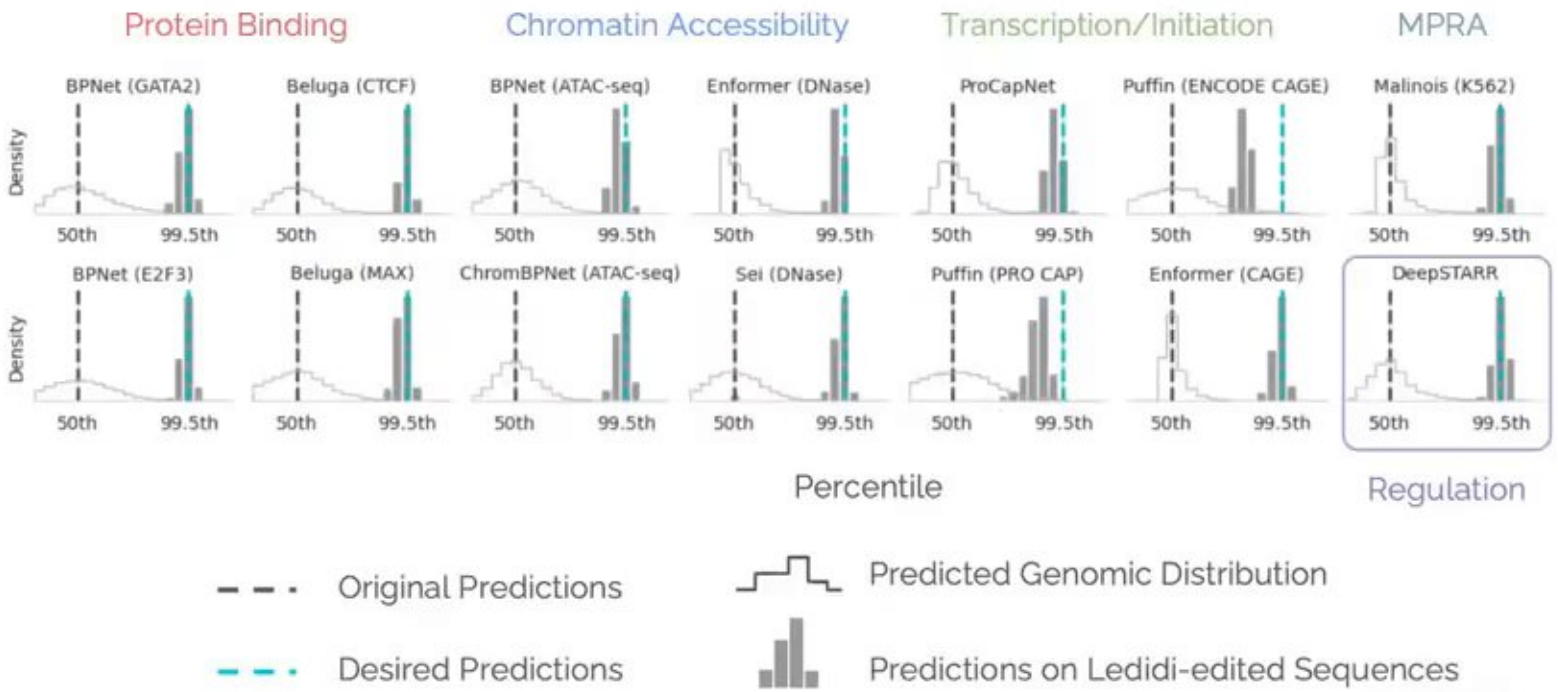
Designing *cis*-regulatory elements using Ledidi



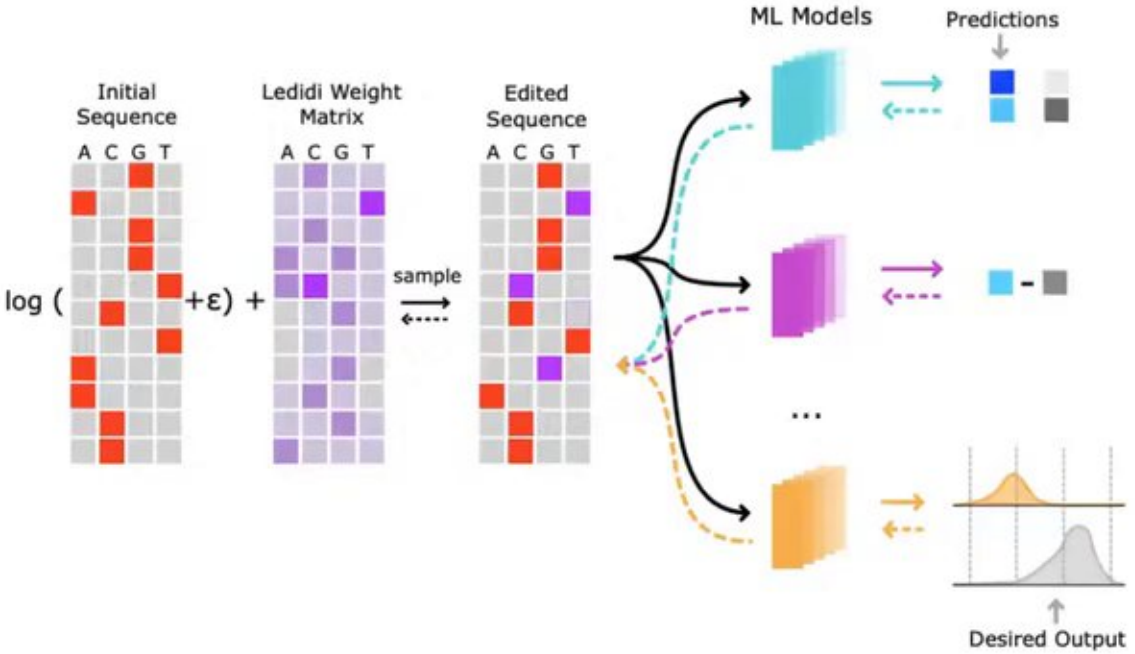

MCBRLab
Now recruiting post-docs and software engineers to join me at UMass Chan Medical School!!
If interested, please reach out with your CV and reference contact info!
UMASS CHAN MEDICAL SCHOOL
zoom


MCBRLab
bilibili
zoom

Ledidi can design edits yielding desired characteristics in a variety of settings

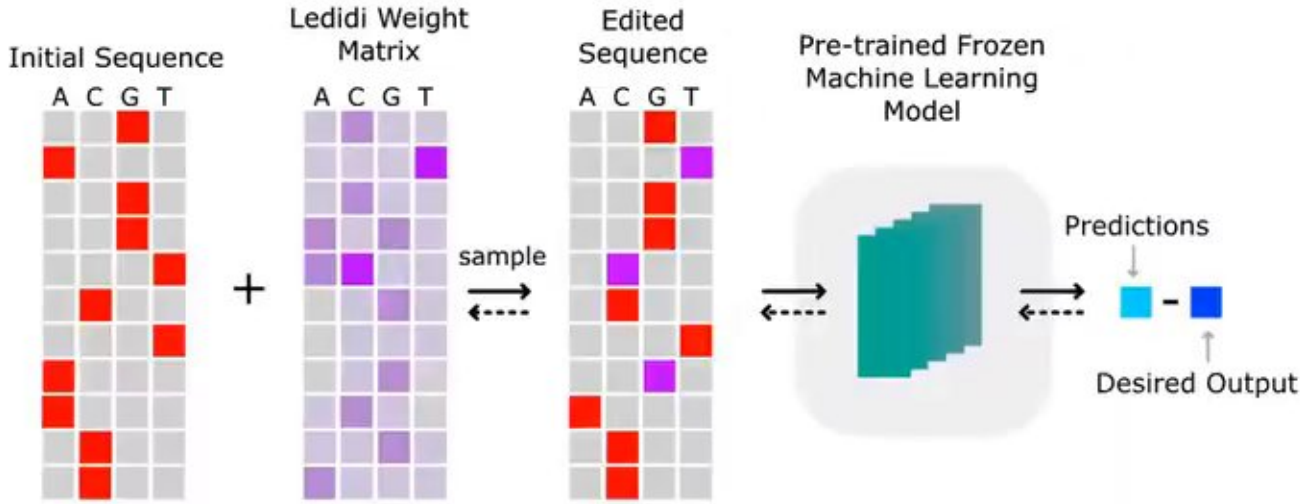


Ledidi allows multiple models to be used simultaneously for design

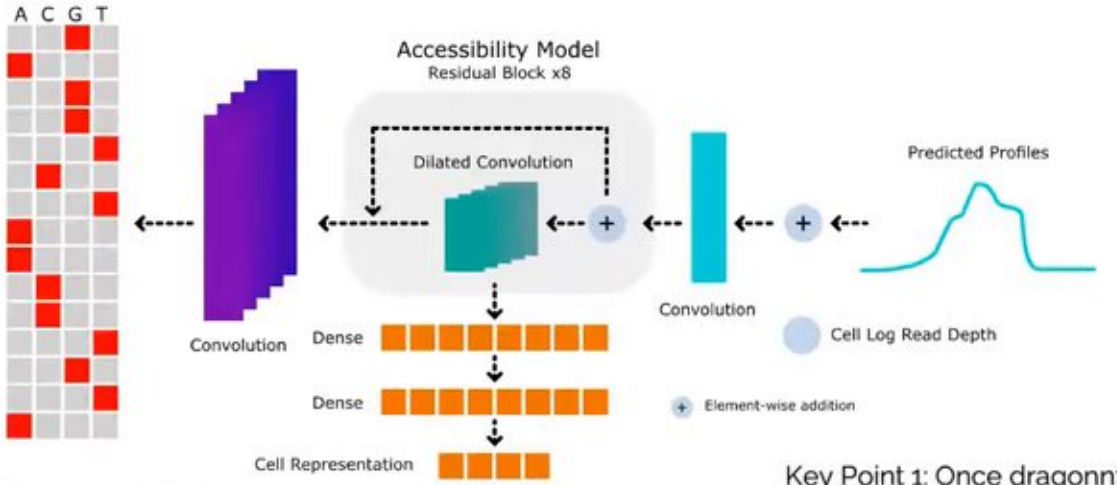


Ledidi turns the edit design task into a simple optimization problem

$$\text{ledidi loss} = \textbf{input\_loss} * \lambda + \textbf{output\_loss}$$



DragoNNFruit extends sequence-based modeling frameworks to single-cell data



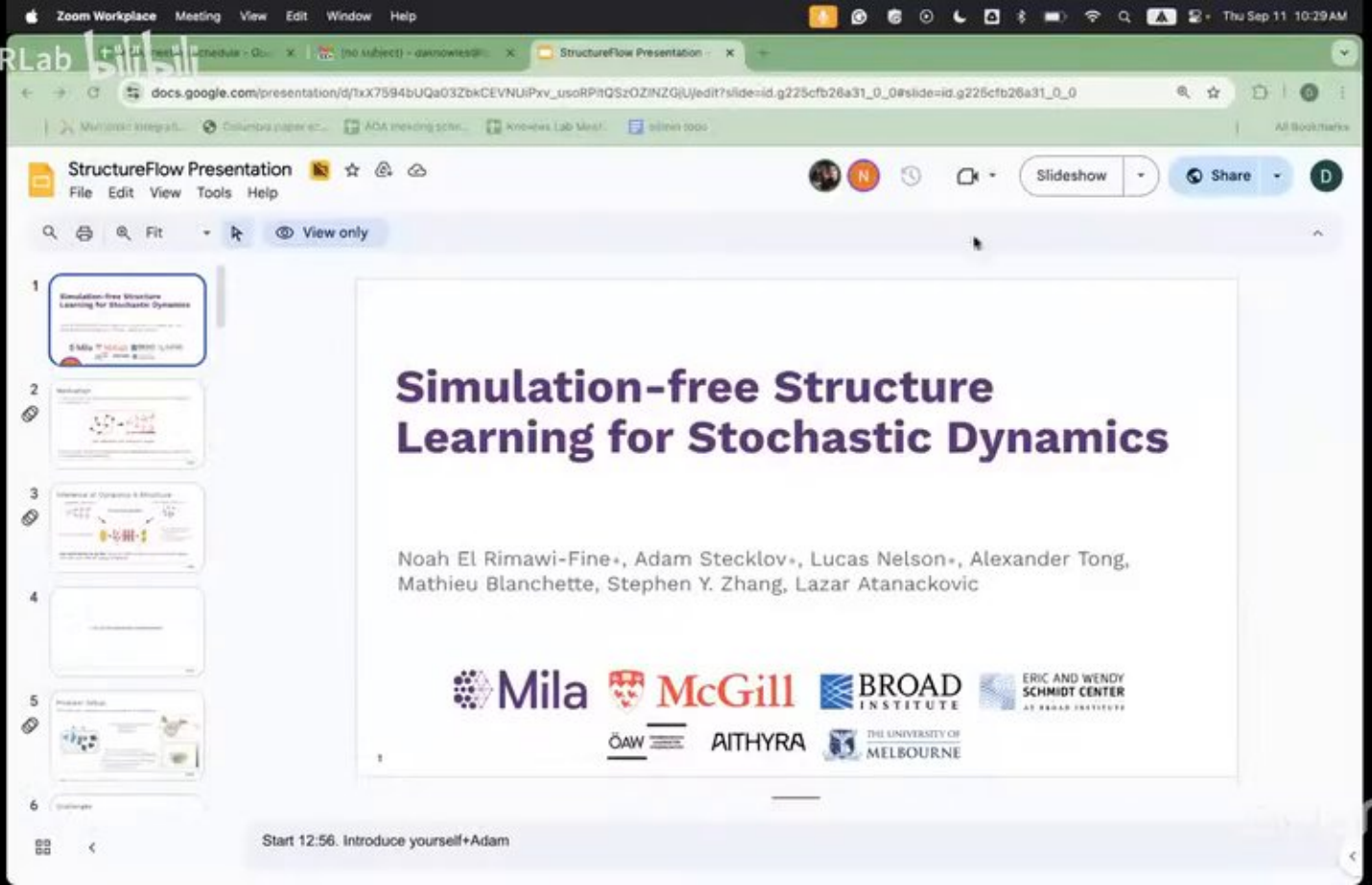
Key Point 2: The cell representation can be anything, e.g. scATAC LSI, gene expression, spatial coordinates, protein abundances, batch-effect corrected embeddings..

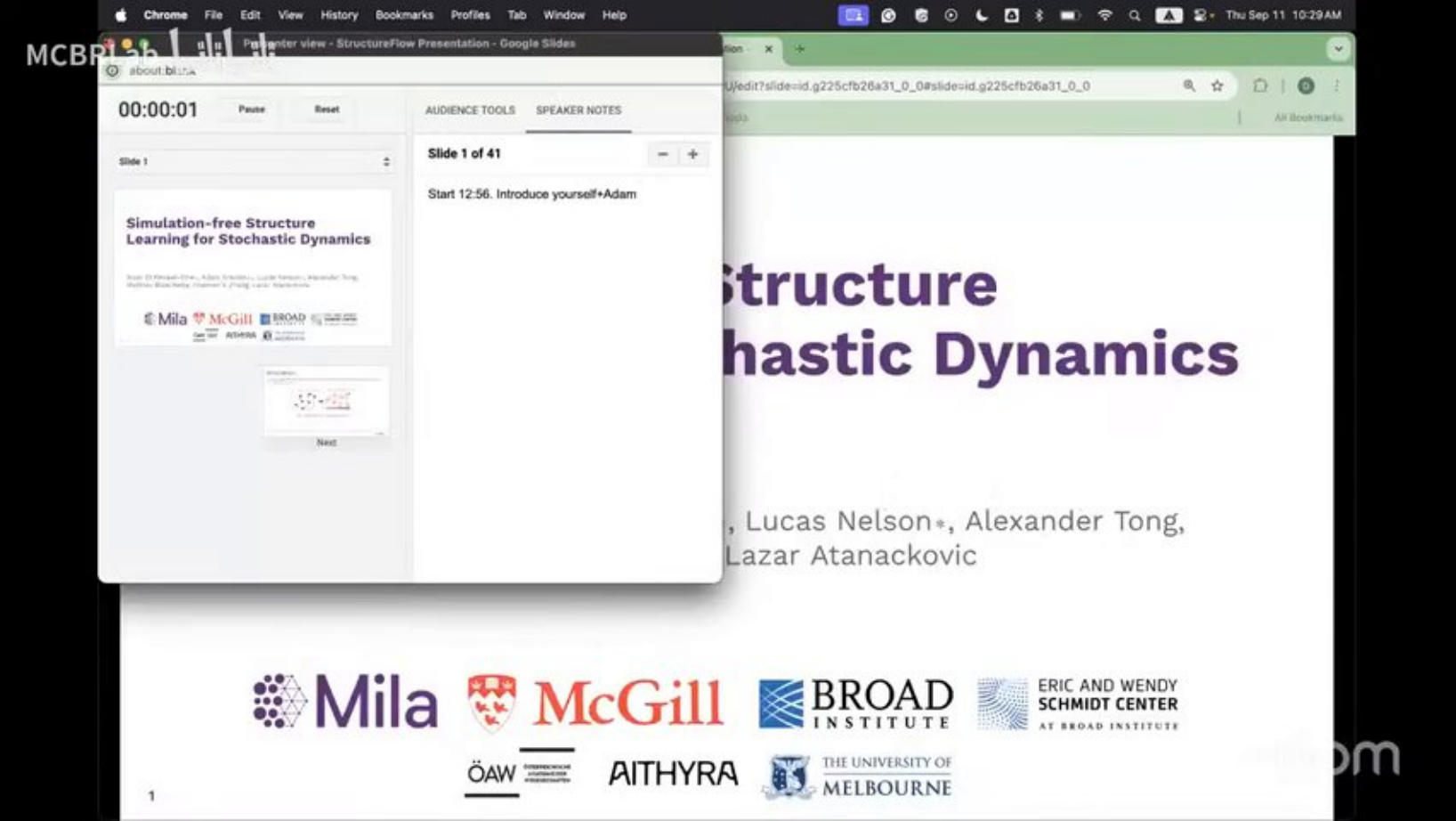
Key Point 1: Once dragonnfruit is trained it literally produces a model for each cell and do anything a (pseudo-)bulk model can do.

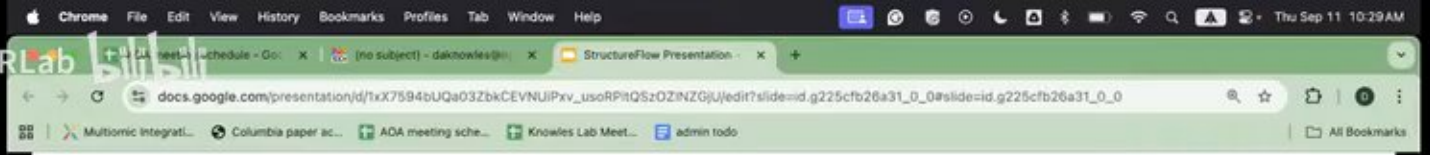
MCBRLab bilibili

NYGC Events

zoom


Zoom Workplace Meeting View Edit Window Help
Thu Sep 11 10:29 AM
StructureFlow Presentation
docs.google.com/presentation/d/1xX7594bUQa03ZbkCEVNUiPxv_usoRPltQSzOZlNZGjU/edit?slide=id.g225cfb26a31_0_0#slide=id.g225cfb26a31_0_0
StructureFlow Presentation
File Edit View Tools Help
Slideshow
Share
Fit
View only
Simulation-free Structure Learning for Stochastic Dynamics
Noah El Rimawi-Fine*, Adam Stecklov*, Lucas Nelson*, Alexander Tong, Mathieu Blanchette, Stephen Y. Zhang, Lazar Atanackovic
Mila
McGill
BROAD INSTITUTE
ERIC AND WENDY SCHMIDT CENTER
ÖAW
AITHYRA
THE UNIVERSITY OF MELBOURNE
Start 12:56. Introduce yourself+Adam


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Thu Sep 11 10:29 AM
Presenter view - StructureFlow Presentation - Google Slides
about:blank
00:00:01
Pause
Reset
AUDIENCE TOOLS
SPEAKER NOTES
Slide 1
Slide 1 of 41
Start 12:56. Introduce yourself+Adam
Simulation-free Structure Learning for Stochastic Dynamics
Next
Structure
hastic Dynamics
, Lucas Nelson*, Alexander Tong,
Lazar Atanackovic
Mila
McGill
BROAD INSTITUTE
ERIC AND WENDY SCHMIDT CENTER AT BROAD INSTITUTE
ÖAW
AITHYRA
THE UNIVERSITY OF MELBOURNE
1



Simulation-free Structure Learning for Stochastic Dynamics

Noah El Rimawi-Fine*, Adam Stecklov*, Lucas Nelson*, Alexander Tong, Mathieu Blanchette, Stephen Y. Zhang, Lazar Atanackovic

Mila McGill BROAD INSTITUTE ERIC AND WENDY SCHMIDT CENTER AT BROAD INSTITUTE

ÖAW AITHYRA THE UNIVERSITY OF MELBOURNE

Chrome File Edit View History Bookmarks Profiles Tab Window Help

2025 MLCB Keynote Jacob Schreiber — Saved to my Mac

Draw Design Transitions Animations Slide Show Record Review View

AOA meeting schedule - Go | (no subject) - daknowles@m | StructureFlow Presentation

docs.google.com/presentation/d/1xX7594bUQa03ZbkCEVNUlPxv_usoRPitQSzOZINZGjU/edit?slide=id.g225cfb26a31_0_0#slide=id.g225cfb26a31_0_0

Multiomic Integrati... Columbia paper ac... AOA meeting sche... Knowles Lab Meet... admin todo

Dragon
data

Key Point 2:
can be anyth
gene expres
protein abur
corrected er

Simulation-free Structure Learning for Stochastic Dynami

Noah El Rimawi-Fine*, Adam Stecklov*, Lucas Nelson*, Alexander Ton
Mathieu Blanchette, Stephen Y. Zhang, Lazar Atanackovic

Mila McGill BROAD INSTITUTE ERIC AND WEND SCHMIDT CENTE AT BROAD INSTITU

ÖAW AITHYRA THE UNIVERSITY OF

Chrome File Edit View History Bookmarks Profiles Tab Window Help

2025 MLCB Keynote Jacob Schreiber — Saved to my Mac

docs.google.com – To exit full screen, press esc

Simulation-free Structure Learning for Stochastic Dyn

Noah El Rimawi-Fine*, Adam Stecklov*, Lucas Nelson*, Alexar
Mathieu Blanchette, Stephen Y. Zhang, Lazar Atanackovic

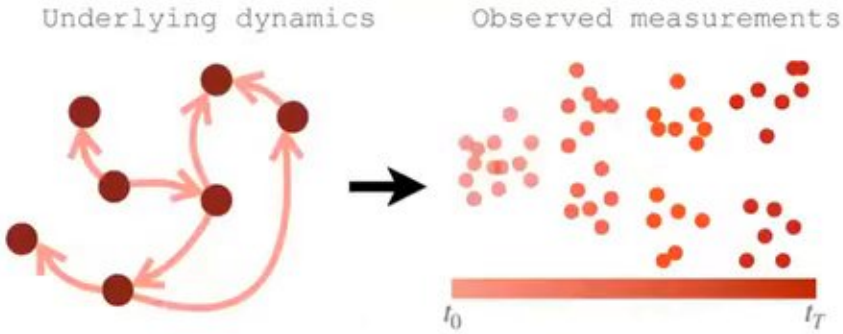
protein abur
corrected er

Simulation-free Structure Learning for Stochastic Dyn

Noah El Rimawi-Fine*, Adam Stecklov*, Lucas Nelson*, Alexar
Mathieu Blanchette, Stephen Y. Zhang, Lazar Atanackovic

Motivation

In many scientific problems we want to model the dynamic evolutic
(e.g. cells) over time



Data generated via dyna-ical system

MCBF

Inference of Dynamics & Structur

How can we address both problems jointly?

with the ones that do having limitations

Challenges

There are still significant challenges for addressing the problems of mod
dynamics and inferring the network structure of these systems

Simulation Free Training (flow + s

SF2M (Tong et al, 2024) infers dynamics from noisy snapshots by le
population flow

StructureFlow (our method)

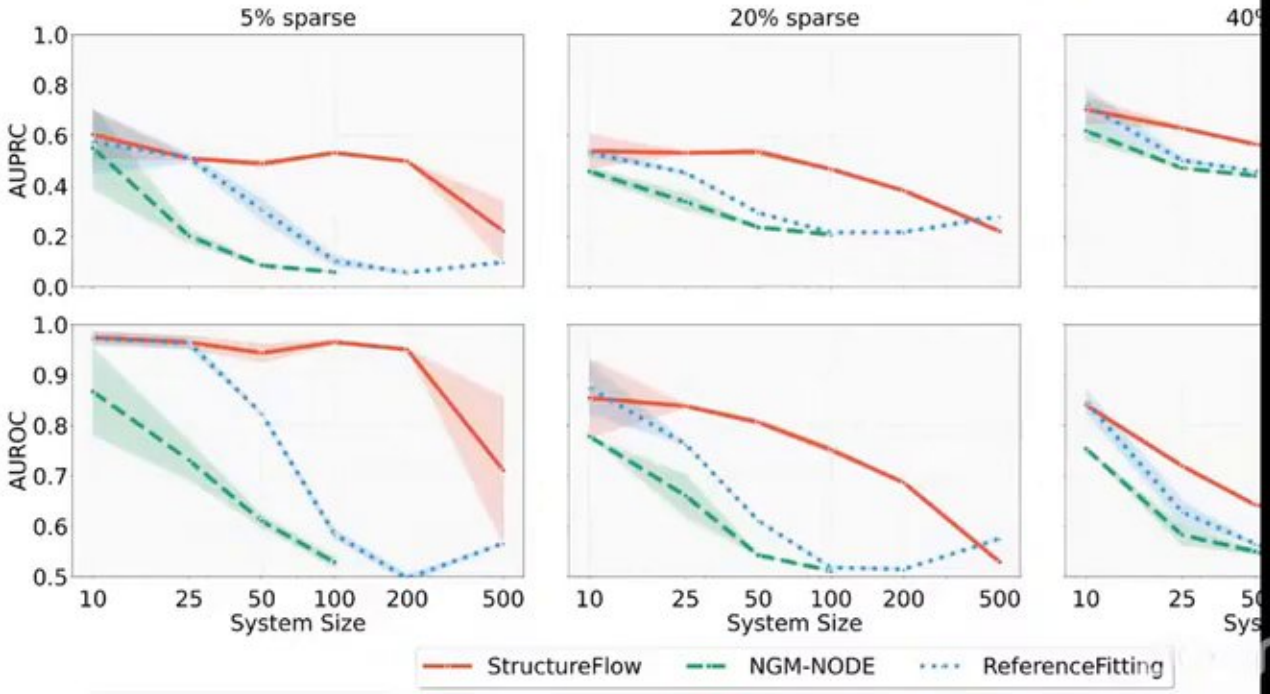
However, Tong et al. doesn't exactly address this joint learning task, or mode

$$\mathcal{L}_{[\mathrm{SF}]^2\mathrm{M}}(\theta) = \mathbb{E}_{t,\boldsymbol{z},\boldsymbol{x}}\left[\|\boldsymbol{v}_t^\theta(\boldsymbol{x}) - \boldsymbol{v}_t^\circ(\boldsymbol{x}|\boldsymbol{z})\|^2 + \lambda(t)\|\boldsymbol{s}_t^\theta(\boldsymbol{x}) - \nabla\right.$$

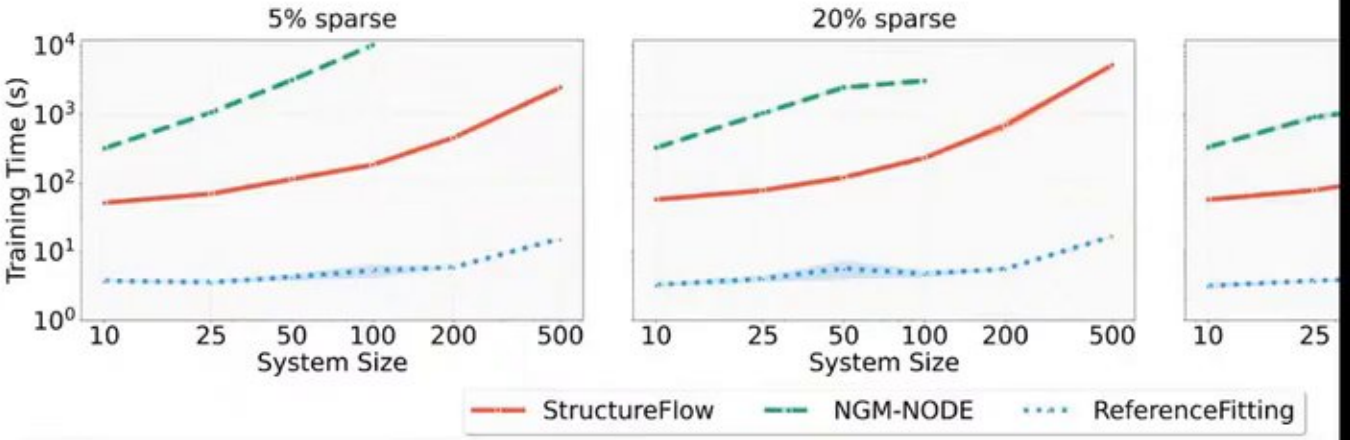
MCBR

How well does this work?

StructureFlow scales better to larger syst

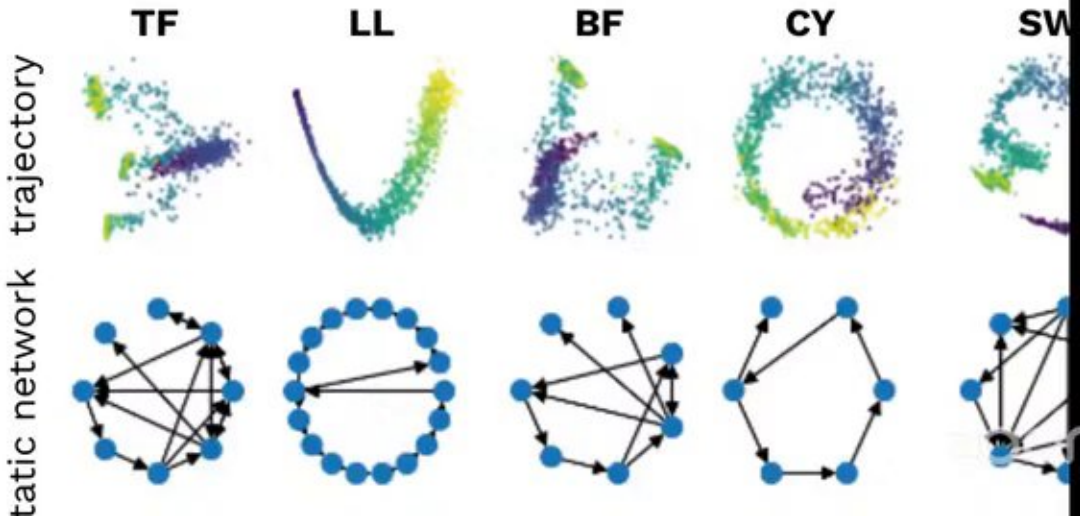


StructureFlow can be trained efficiently



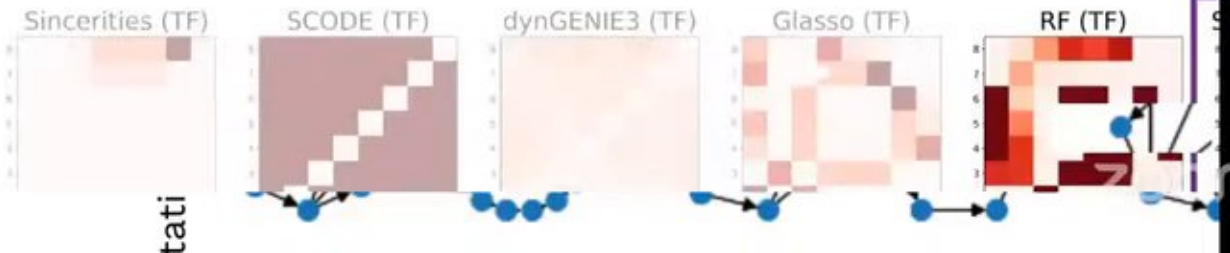
II - Simulated Biological Systems

Joint Trajectory Inference and GRN inference on small, curated, s
systems, each representing common gene regulatory mechanics
Data generation with BoolODE (Pratapa et al., 2020)

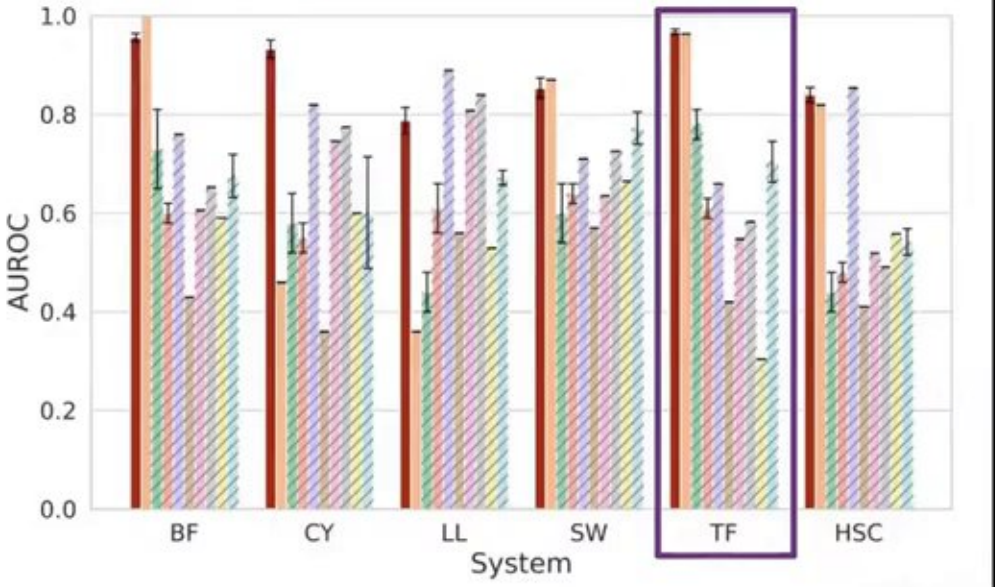


II - Simulated Biological Systems (*Structure Learning*)

Joint Trajectory Inference and GRN inference on small syntheti
→ GRN inference measures AP and AUROC of the recovered gra
ground truth



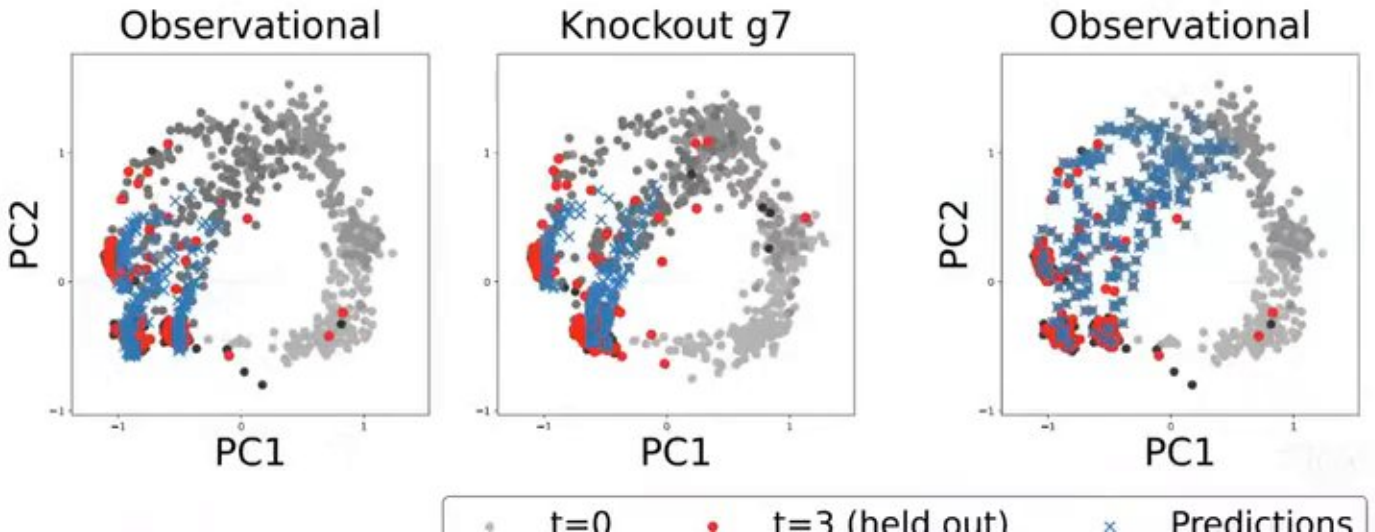
II - Simulated Biological Systems (*Structure Learning*)



II - Simulated Biological Systems (*Trajectory inference*)

StructureFlow (ours)

Referenc



II - Simulated Biological Systems (*Trajectory inference*)

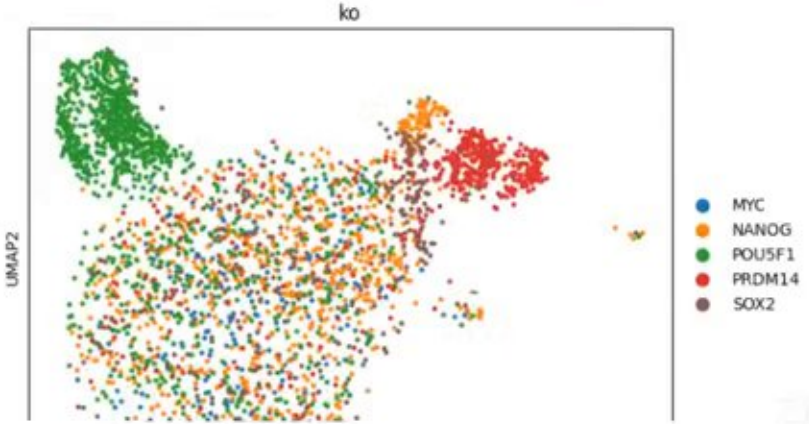
→ Trajectory inference measures W2 and MMD, where one time
hidden from the model during training

	TF		CY		LL		HSC (Curated)		
	$W_2\downarrow$	MMD↓	$W_2\downarrow$	MMD↓	$W_2\downarrow$	MMD↓	$W_2\downarrow$	MMD↓	$W_2\downarrow$
$[SF]^2M$ (ODE)	**0.76 ± 0.02**	0.13 ± 0.01	**0.52 ± 0.01**	0.15 ± 0.00	0.85 ± 0.04	0.19 ± 0.01	0.67 ± 0.01	0.08 ± 0.00	0.38 ± 0.0
$[SF]^2M$ (SDE)	0.79 ± 0.02	**0.07 ± 0.01**	0.53 ± 0.01	0.06 ± 0.00	1.11 ± 0.05	0.13 ± 0.01	**0.64 ± 0.01**	**0.02 ± 0.00**	0.45 ± 0.0
RF	1.03 ± 0.00	**0.07 ± 0.00**	0.76 ± 0.00	**0.05 ± 0.00**	1.19 ± 0.00	**0.10 ± 0.00**	0.72 ± 0.00	**0.02 ± 0.00**	0.81 ± 0.0
StructureFlow (ODE)	0.79 ± 0.01	0.13 ± 0.00	0.57 ± 0.01	0.14 ± 0.00	**0.83 ± 0.02**	0.19 ± 0.01	0.68 ± 0.01	0.07 ± 0.00	**0.37 ± 0.0**
StructureFlow (SDE)	0.82 ± 0.02	0.08 ± 0.01	0.59 ± 0.01	0.06 ± 0.00	1.05 ± 0.03	0.12 ± 0.01	0.66 ± 0.01	**0.02 ± 0.00**	0.42 ± 0.0

Top in **Bold**
Second in Blue

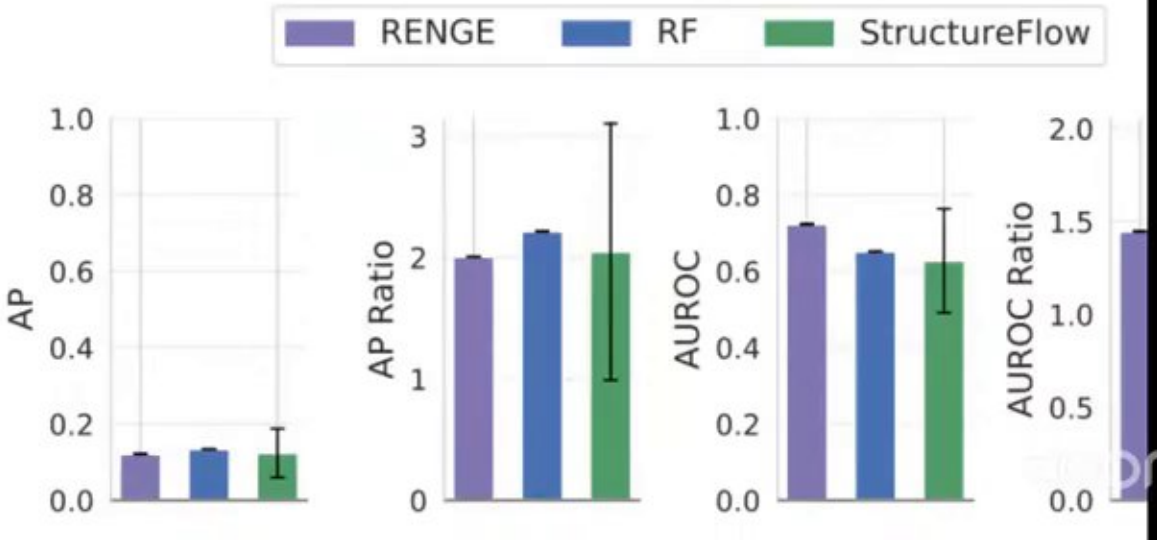
II - Real Data (Renge, CRISPR KO)

Multi-marginal dataset (4 time-points) of iPSCs with CRISPR in
(knockouts) for 23 transcription factors (Tokumasu et al., 2023)
→ Higher dimensionality (103 genes)
→ Downside: ground truth network is a subset, experimental es



II - Real Data (Renge, CRISPR KO)

→ Gene Regulatory Network Inference Results



II - Real Data (Renge, CRISPR KO)

→ Trajectory Inference performance (prediction of unseen time

	Average	
	$W_2\downarrow$	MMD$\downarrow$
$[SF]^2M$ (ODE)	5.75 ± 0.09	0.019 ± 0.004
$[SF]^2M$ (SDE)	6.13 ± 0.11	**0.018 ± 0.004**
RF	6.53 ± 0.06	**0.018 ± 0.004**
StructureFlow (ODE)	**5.63 ± 0.04**	0.020 ± 0.004
StructureFlow (SDE)	5.94 ± 0.04	0.019 ± 0.004

Top in **Bold**
Second in Blue

Conclusion & Future Work

This is still a hard (and unsolved) problem!

We introduce an (efficient) method for jointly inferring network
and dynamics *(with a single model and end-to-end optimizatio*

Still some limitations and things to improve ... future work?
→ We don't think the model is learning *true* causal relationship
→ Still challenging to *systematically* select good hyperparamete

Second in Blue

Acknowledgements

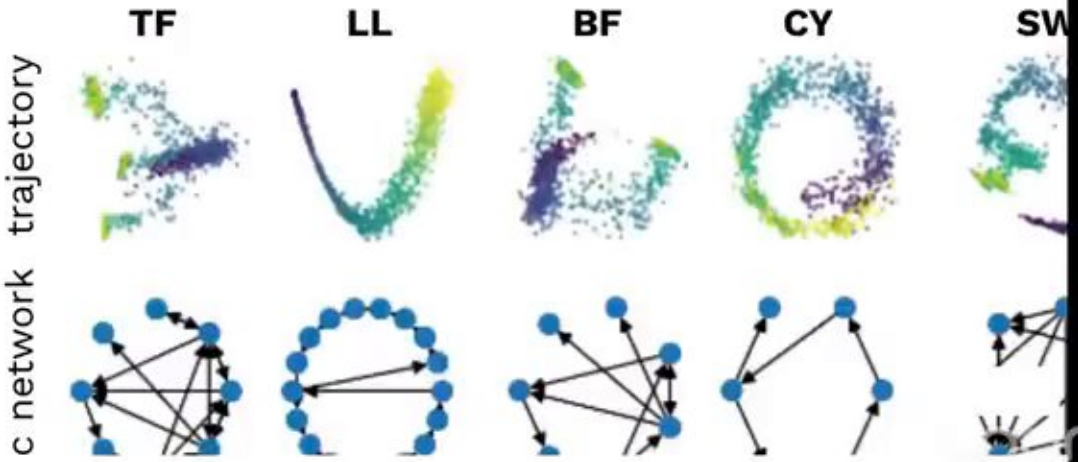
McGill

AITHYRA

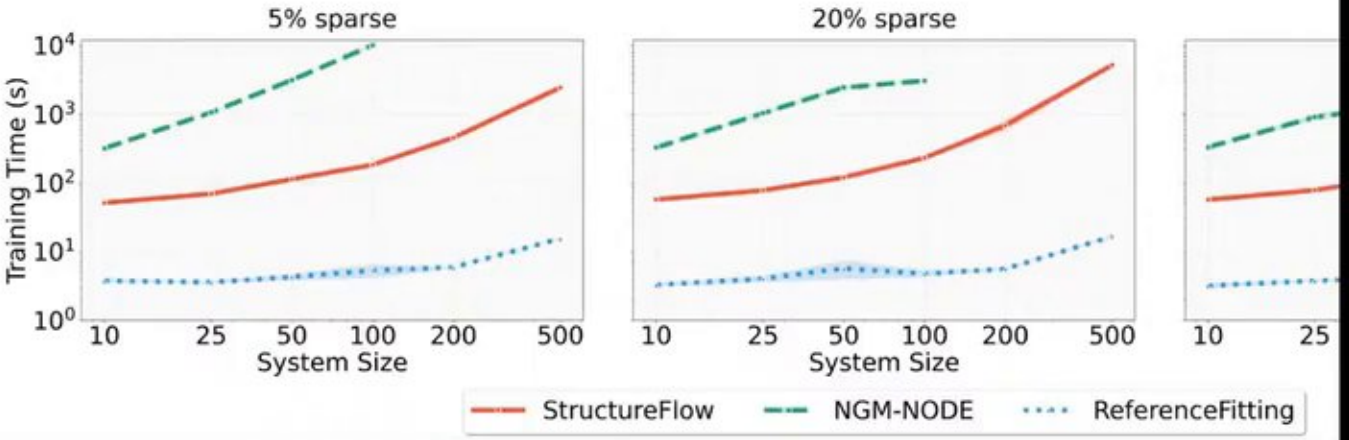
Questions?

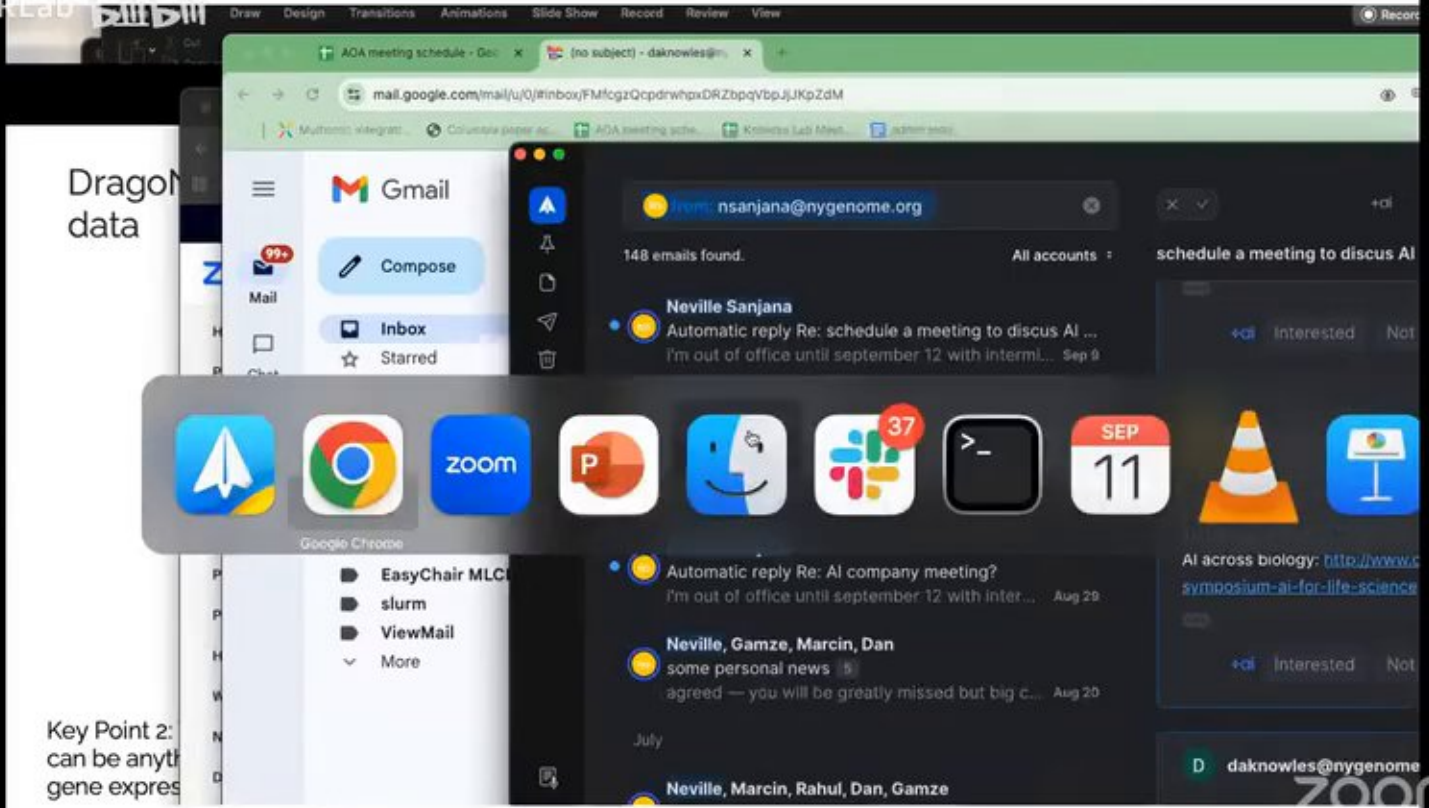
II - Simulated Biological Systems

Joint Trajectory Inference and GRN inference on small, curated, s
systems, each representing common gene regulatory mechanics
Data generation with BoolODE (Pratapa et al., 2020)



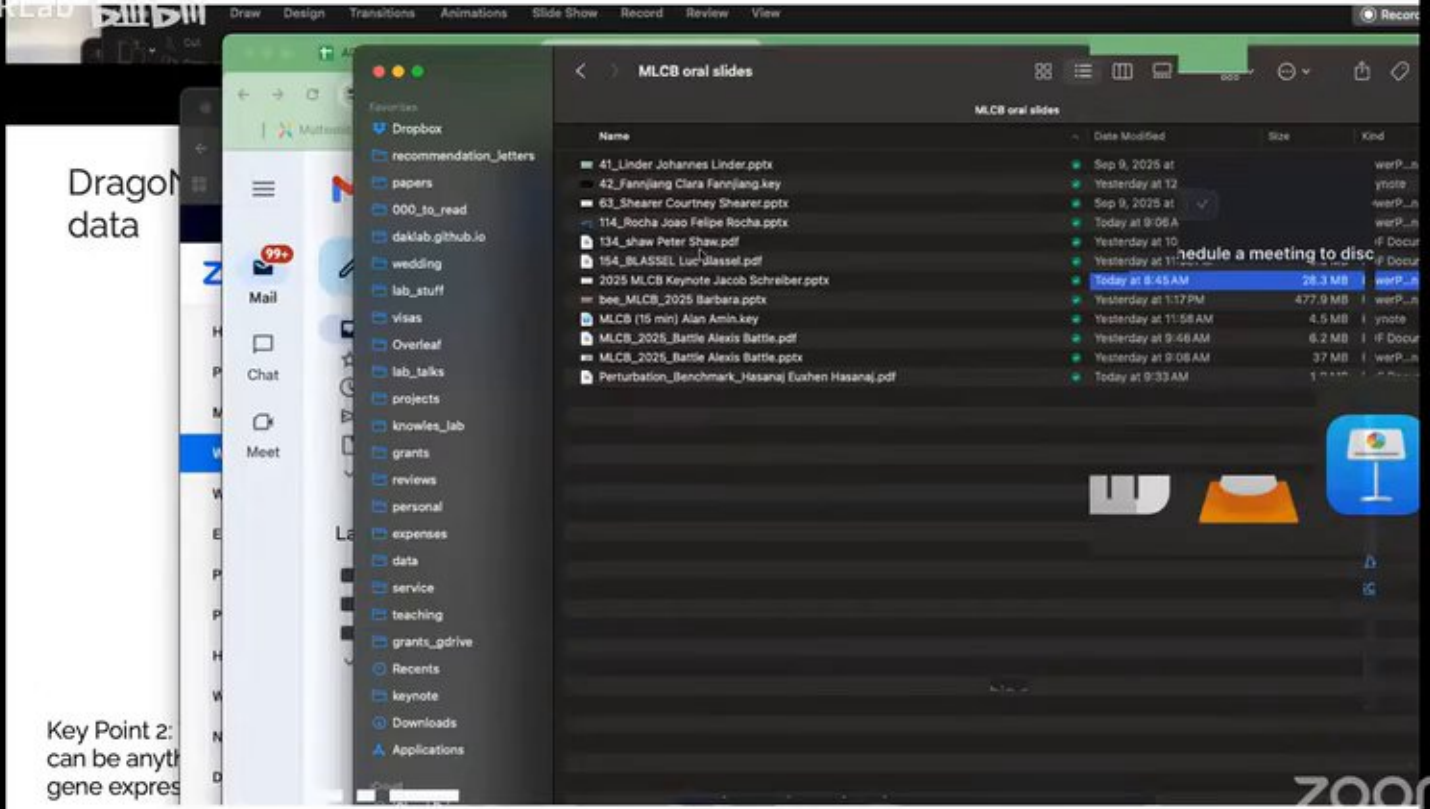
StructureFlow can be trained efficiently

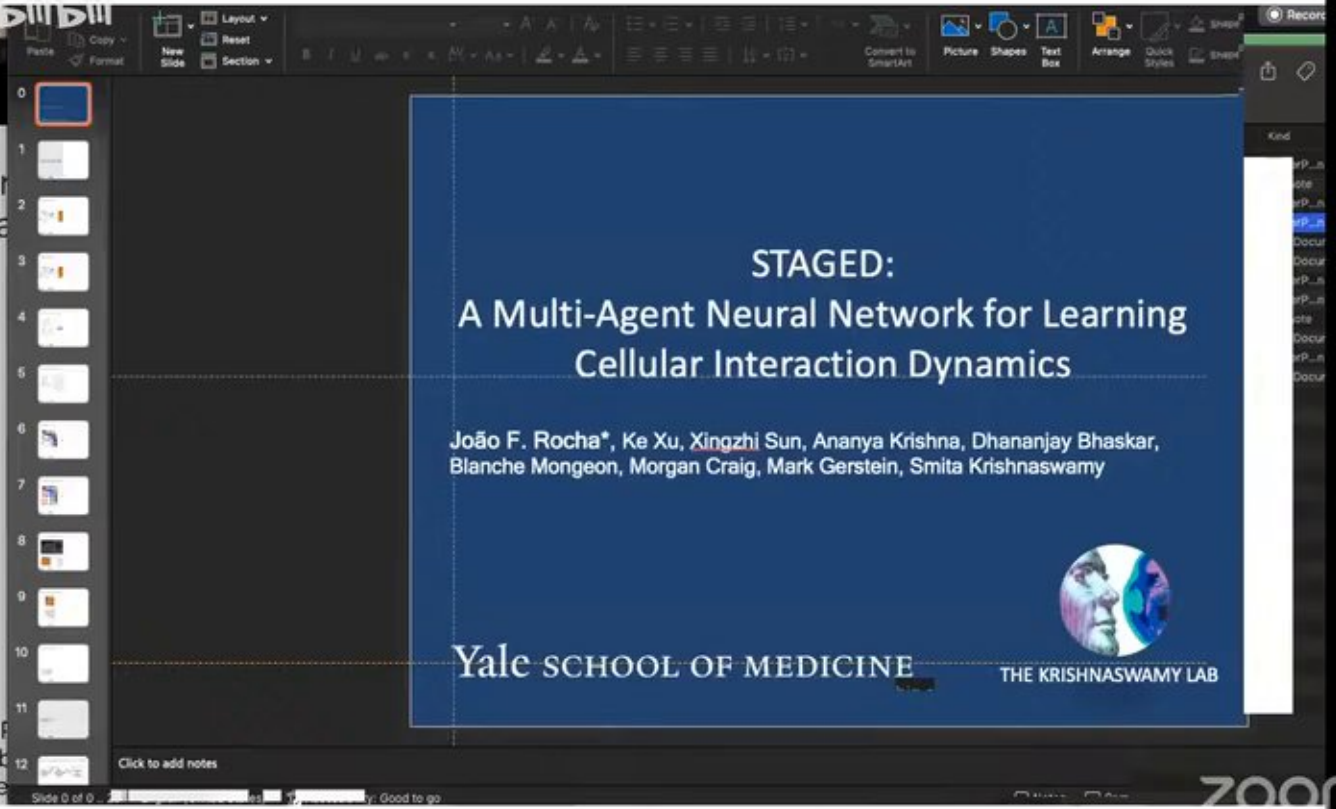




Dragon
data

Key Point 2:
can be anyth
gene expres


Draw
Design
Transitions
Animations
Slide Show
Record
Review
View
MLCB oral slides
Dropbox
recommendation_letters
papers
000_to_read
daklab.github.io
wedding
lab_stuff
visas
Overleaf
lab_talks
projects
knowles_lab
grants
reviews
personal
expenses
data
service
teaching
grants_gdrive
Recents
keynote
Downloads
Applications
Name
Date Modified
Size
Kind
41_Linder Johannes Linder.pptx
42_Fannjiang Clara Fannjiang.key
63_Shearer Courtney Shearer.pptx
114_Rocha Joao Felipe Rocha.pptx
134_shaw Peter Shaw.pdf
154_BLASSEL Lucie Blassel.pdf
2025 MLCB Keynote Jacob Schreiber.pptx
bee_MLCB_2025 Barbera.pptx
MLCB (15 min) Alan Amin.key
MLCB_2025_Battle Alexis Battle.pdf
MLCB_2025_Battle Alexis Battle.pptx
Perturbation_Benchmark_Hasanaj Evahen Hasanaj.pdf
Mail
Chat
Meet
Key Point 2:
gene expres
zoom



MCBRLab bilibili

zoom

STAGED:
A Multi-Agent Neural Network for Learning Cellular Interaction Dynamics

João F. Rocha*, Ke Xu, Xingzhi Sun, Ananya Krishna, Dhananjay Bhaskar, Blanche Mongeon, Morgan Craig, Mark Gerstein, Smita Krishnaswamy

Yale SCHOOL OF MEDICINE

THE KRISHNASWAMY LAB

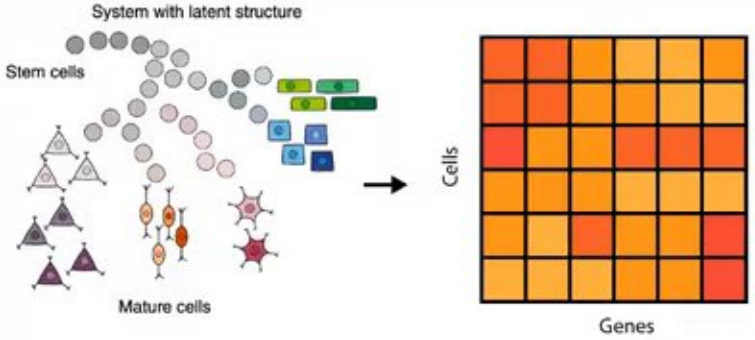
MCBRLab

Transcriptomics Data

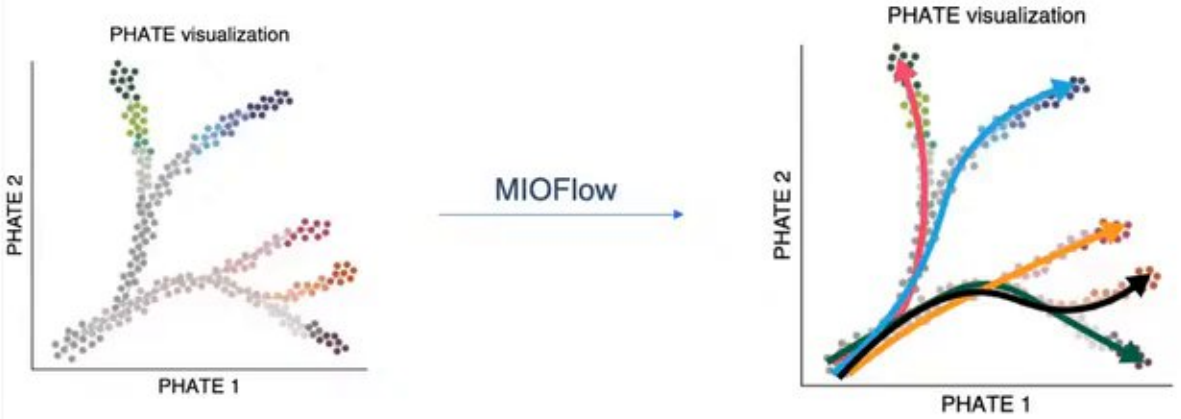
Yale SCHOOL OF MEDICINE

zoom

Single Cell Data

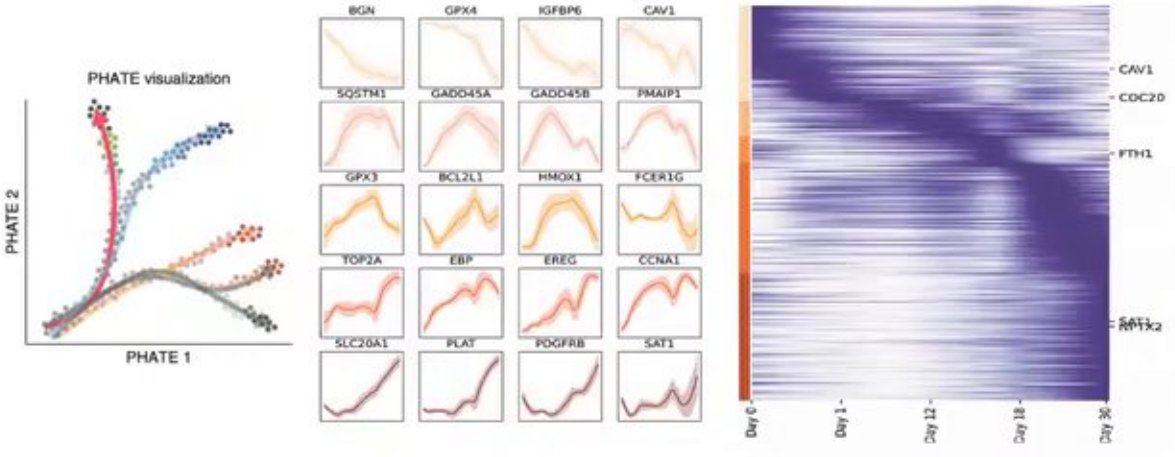


Trajectories in the data

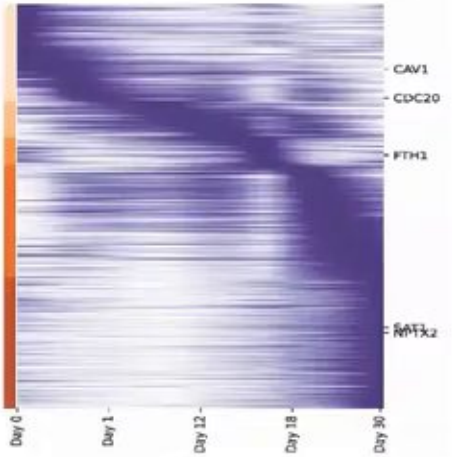


Huguet, G., et al. Manifold Interpolating Optimal-Transport Flows for Trajectory Inference. NEURIPS (2022).

Trajectories in the data

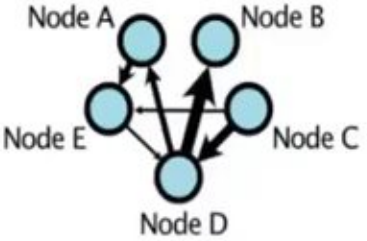


Gene Dynamics

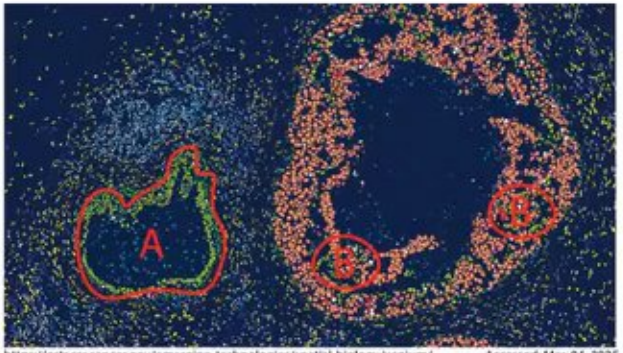


Granger Causality

Gene Regulatory Networks

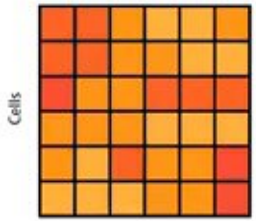


Spatial Transcriptomics Data



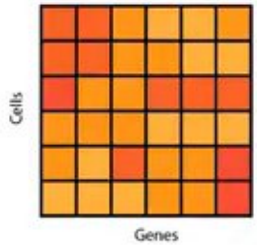
https://ostr.ccr.cancer.gov/emerging-technologies/spatial-biology/xenium/ Accessed: Mar. 04, 2025

- **Spatial Transcriptomics**: Gives information on both the space and gene-level.

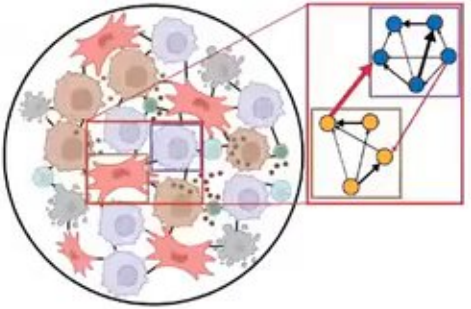


Cells	Position
0.1	1.1
0.2	4.1
2.1	2.9
3.5	7.6
5.6	0.7
0.6	1.2

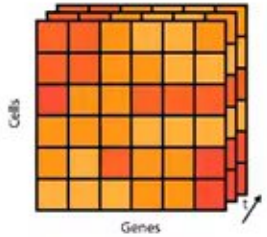
Spatial Transcriptomics Data: Hierarchical Graphs

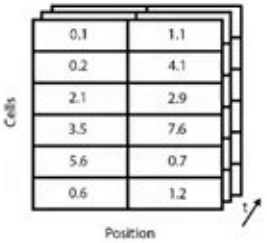


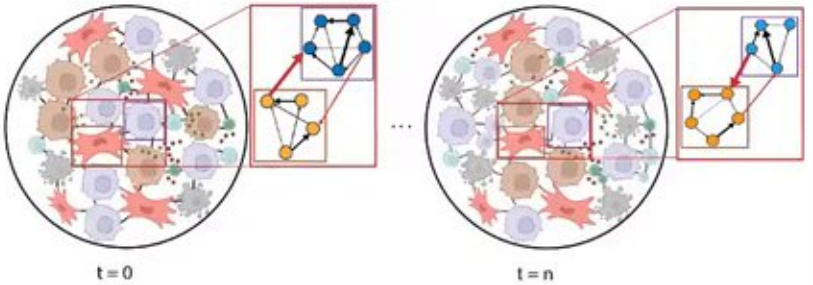
Cells	Position
0.1	1.1
0.2	4.1
2.1	2.9
3.5	7.6
5.6	0.7
0.6	1.2



Time evolving Spatial Transcriptomics Data

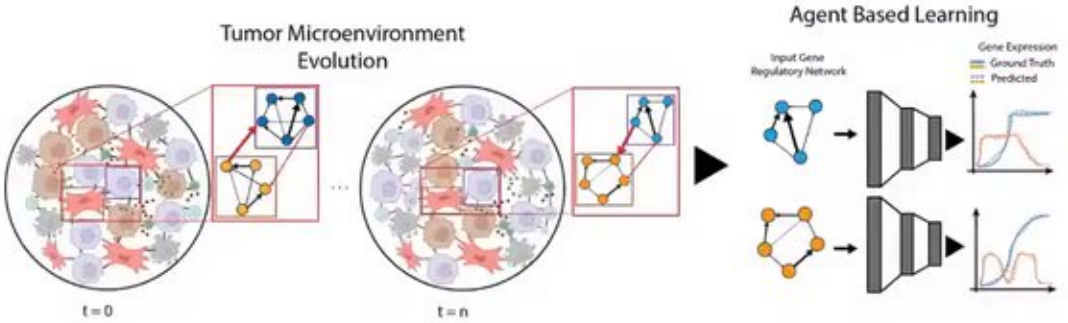






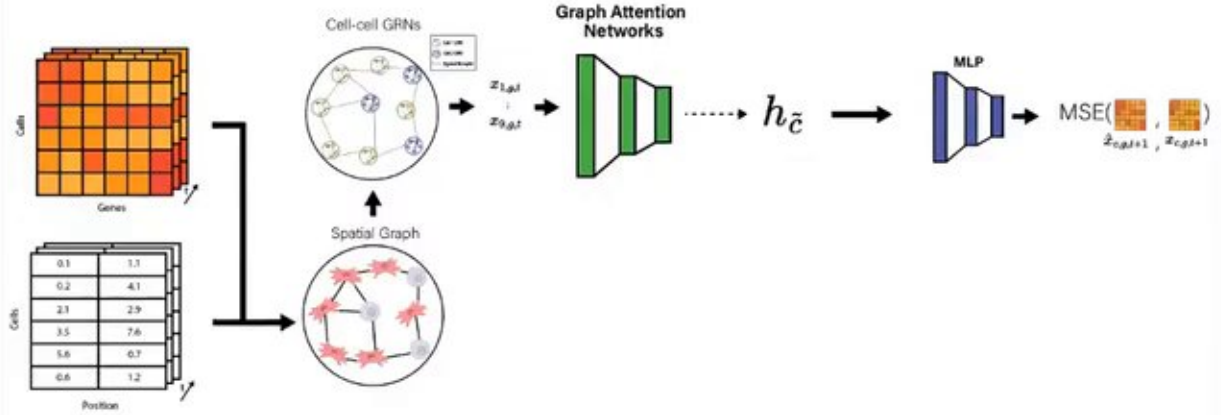
STAGED

Spatio-Temporal Agent-Based Graph Evolution Dynamics (STAGED)

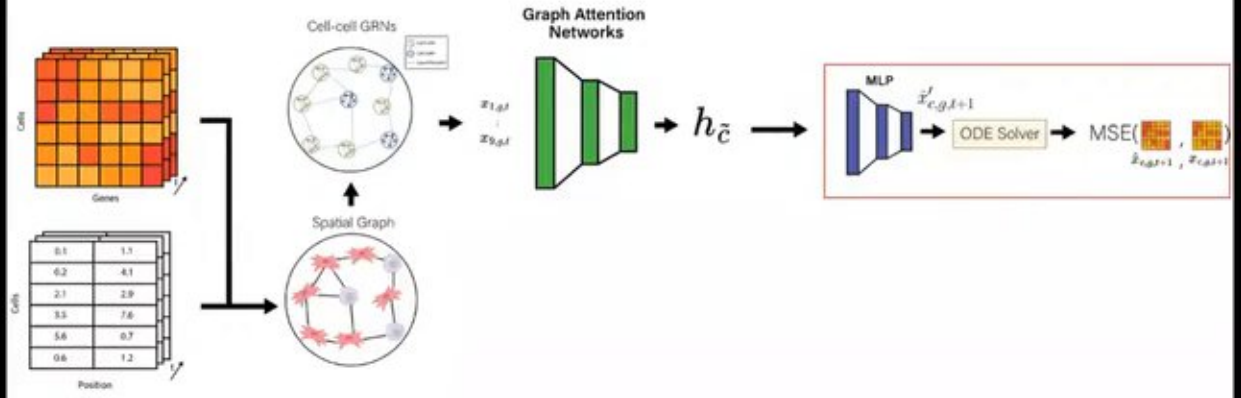


- **Level 1**: Cell-cell communication (agents talking)
- **Level 2**: Gene regulatory networks (within each cell)

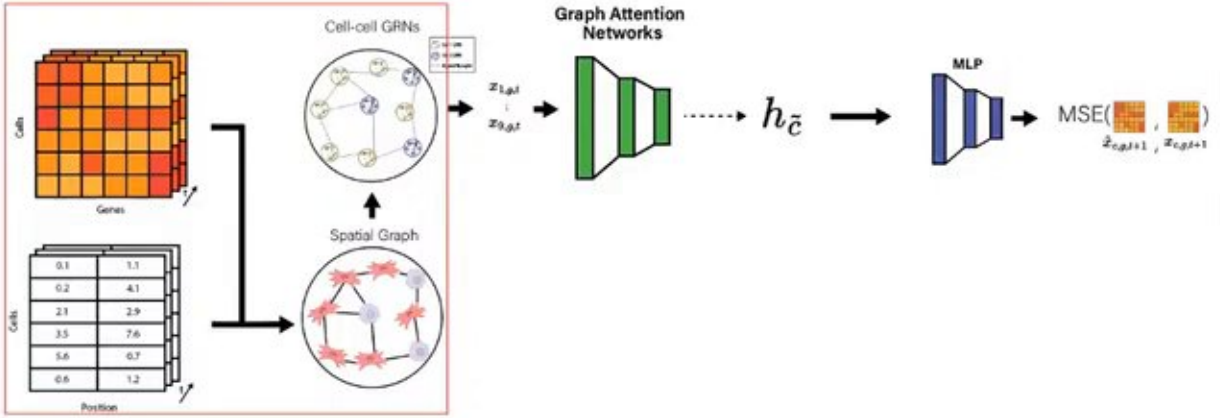
High-level Architecture



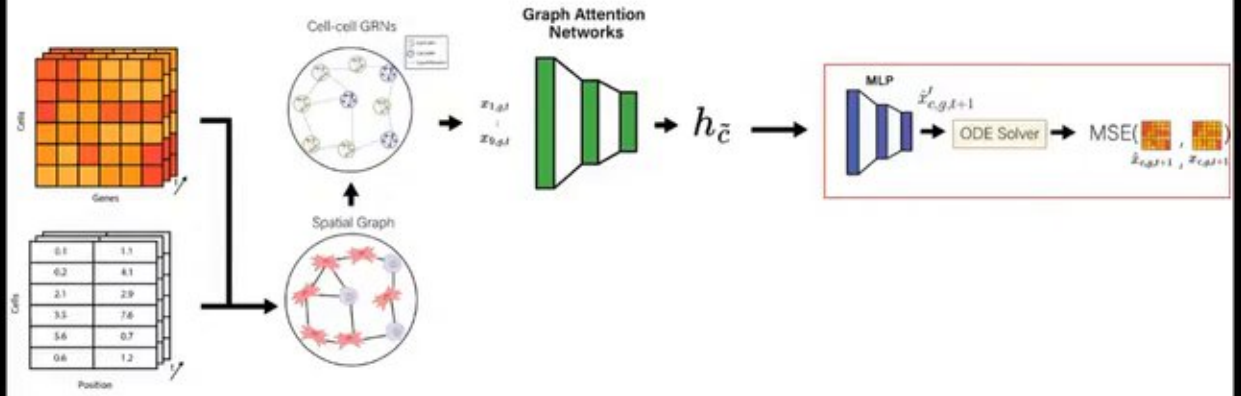
Better dynamics inference



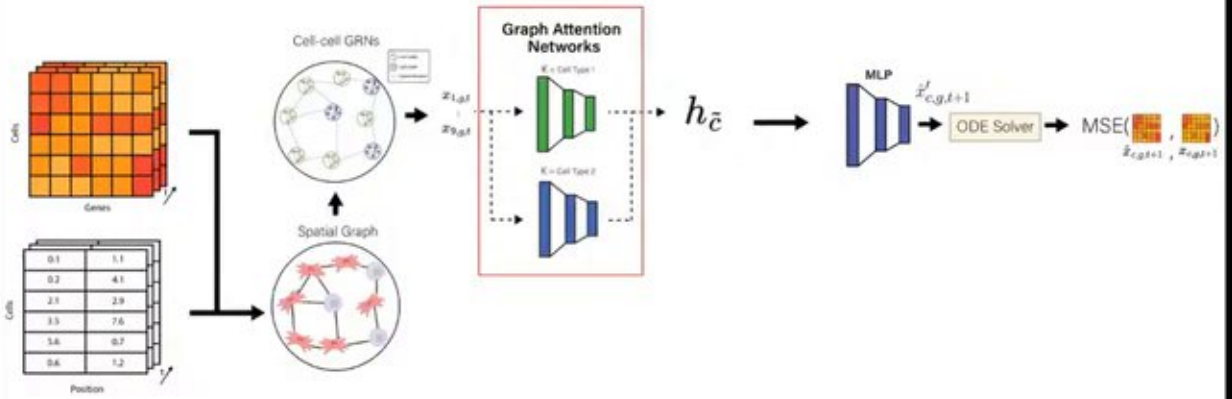
Biological Prior I: Ligand Receptor Pairs



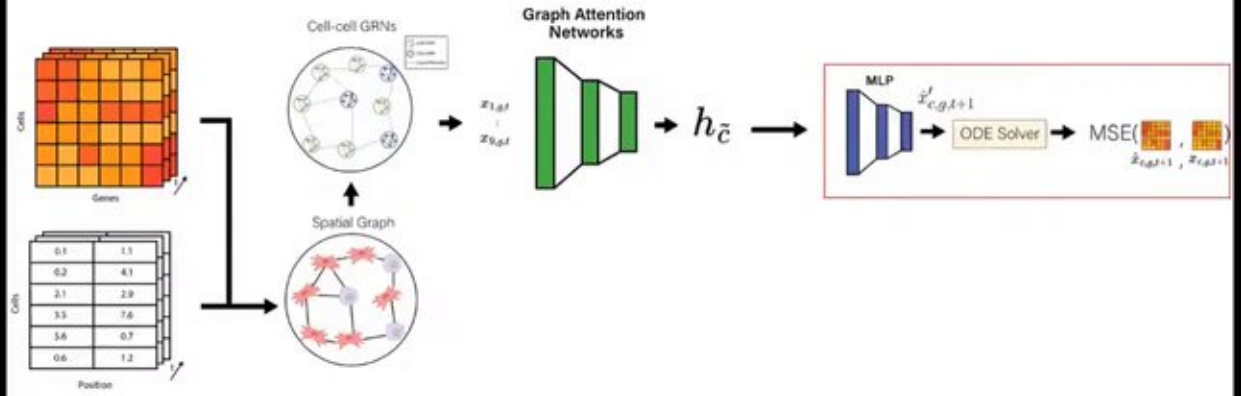
Better dynamics inference



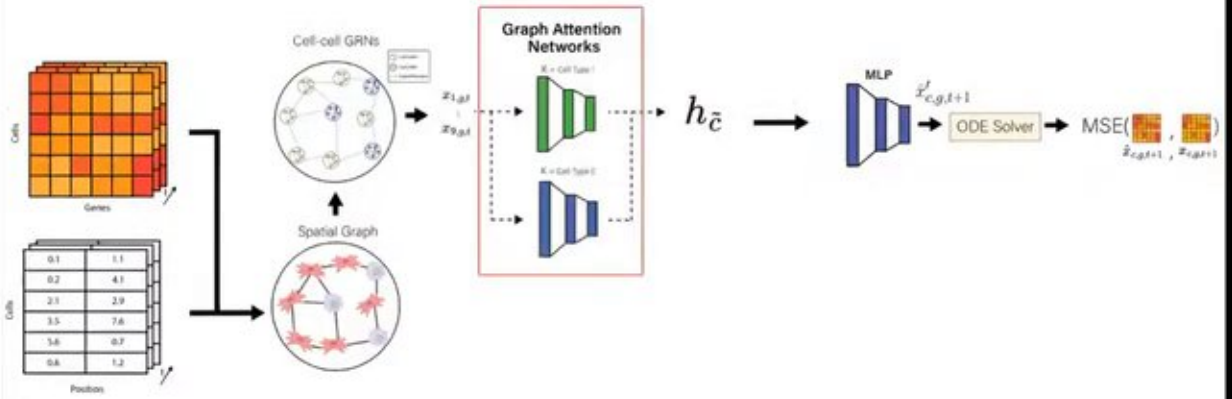
Biological Prior II: Agent-based learning



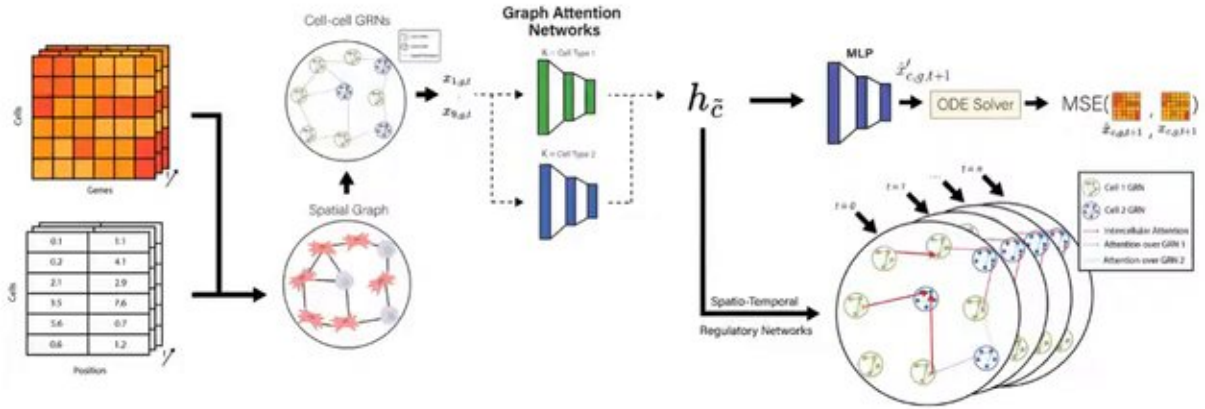
Better dynamics inference



Biological Prior II: Agent-based learning



Obtaining the cell-cell interactions from attention



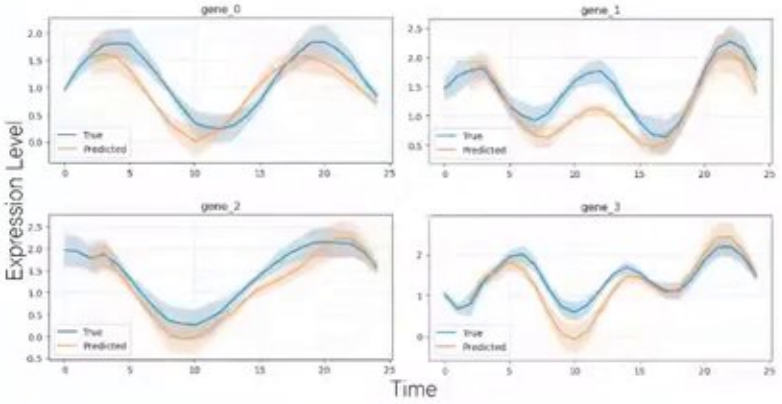
Spatial neighborhoods → Attention mechanism → Gene dynamics

Each agent in **STAGED** is a cell-type–specific model composed of a graph attention network (GAT) and a neural ODE.

Preliminary Results

Simulating Oscillatory Dynamics

Mimicking oscillatory dynamics

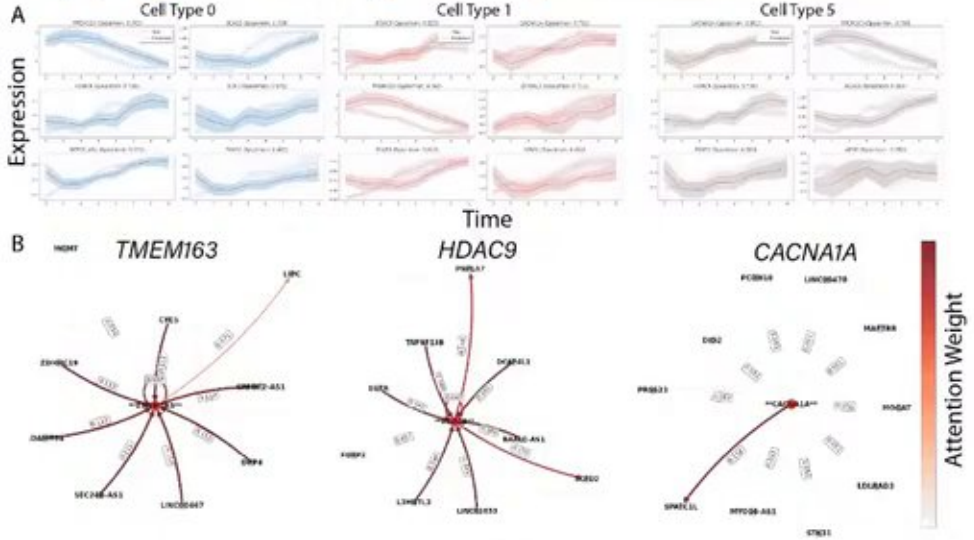


STAGED can accurately capture these dynamics, shown here one cell type for a subset of genes.

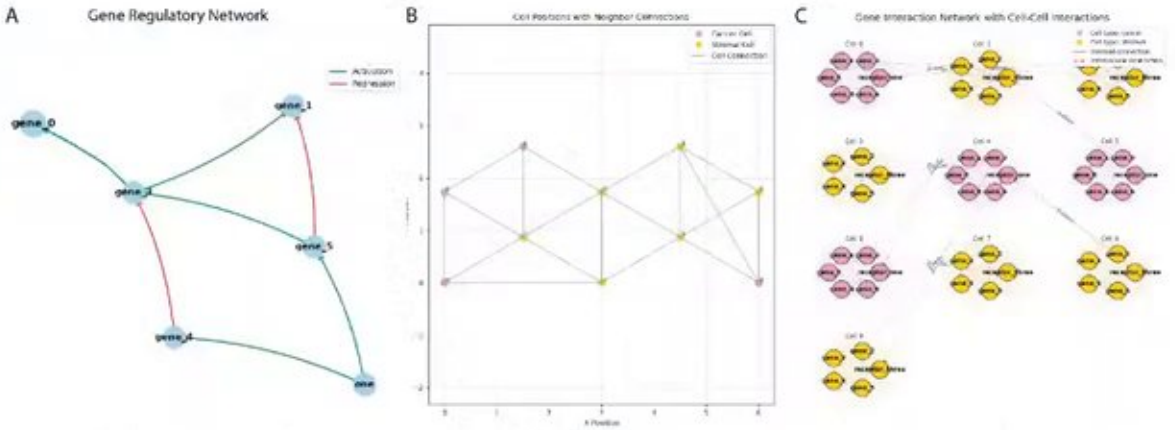
Results – Real Data

Microglia during Alzheimer's disease

Regulatory programs in microglia during Alzheimer's disease



Simulations using simulated GRNs

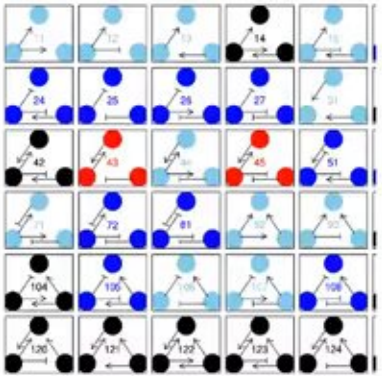


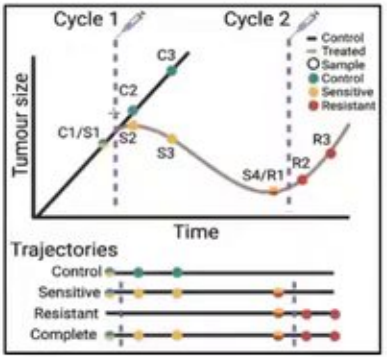
Spatially resolved stochastic simulation of gene expression dynamics in a randomly constructed gene regulatory network (GRN) using GillesPy2.

MCBRLab

Next Steps

- Use small simulation of GRNs
- Improve benchmarking and simulations
- Apply Multi-Agent RL techniques
- Application to new cancer data





Acknowledgements

Ke Xu

Xingzhi Sun

STAGED Authors

Ananya Krishna
Dhananjay Bhaskar
Blanche Mongeon
Morgan Craig
Mark Gerstein
Smita Krishnaswamy

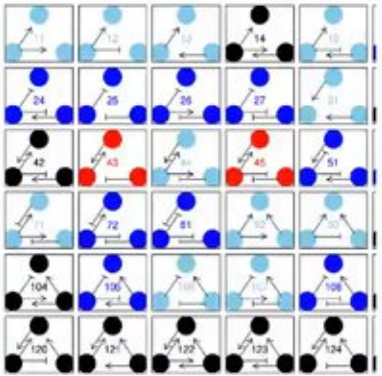
THE KRISHNASWAMY LAB

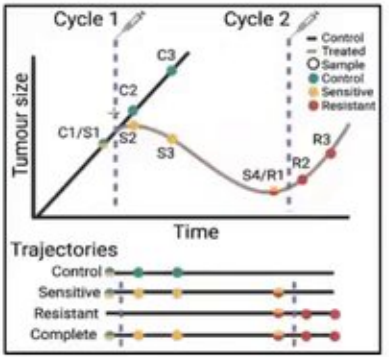
Yale SCHOOL OF MEDICINE

MCBRLab

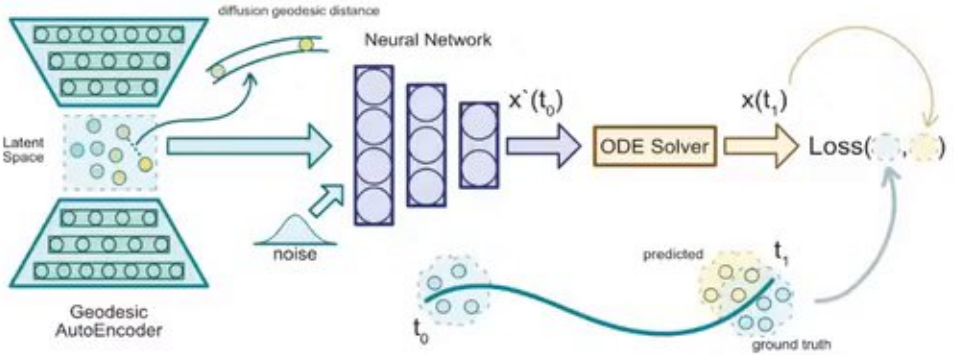
Next Steps

- Use small simulation of GRNs
- Improve benchmarking and simulations
- Apply Multi-Agent RL techniques
- Application to new cancer data





MIOFlow



MIOFlow models continuous population dynamics by learning optimal transport-driven flows in a manifold-aware latent space to interpolate between static data snapshots.

MCBRLab bilibili

NYGC Events

zoom

Motivation

- Why perturbation response matters?
- **Current challenge**: FMs often don't outperform simple baselines.
- **Key gap**: we don't understand which part (embeddings, downstream models, datasets) drive performance.

Brief Communication | Open access | Published: 04 August 2025

Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines

Constantin Ahlmann-Eltze, Wolfgang Huber & Simon Anders

Nature Methods 22, 1657–1661 (2025) | Cite this article

49k Accesses | 8 Citations | 181 Altmetric | Metrics

A systematic comparison of computational methods for expression forecasting

choice of metric, and especially for simple metrics like mean squared error, it is uncommon for expression forecasting methods to out-perform simple baselines. Our platform will serve as a


GenBio
Multimodal Benchmarking of Foundation Model Representations for Cellular Perturbation Response Prediction
Euxhen Hasanaj[*1], Elijah Cole[*1], Shahin Mohammadi[1], Sohan Addagudi[1,2], Xingyi Zhang[1,3], Le Song[1,3], Eric Xing[1,2,3]
[1] GenBio AI
[2] Carnegie Mellon University
[3] Mohamed bin Zayed University of Artificial Intelligence
Presenter: Euxhen Hasanaj
zoom

Motivation

- Why perturbation response matters?
- **Current challenge**: FMs often don't outperform simple baselines.
- **Key gap**: we don't understand which part (embeddings, downstream models, datasets) drive performance.

Brief Communication | Open access | Published: 04 August 2025

Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines

Constantin Ahlmann-Eltze, Wolfgang Huber & Simon Anders

Nature Methods 22, 1657–1661 (2025) | Cite this article

49k Accesses | 8 Citations | 181 Altmetric | Metrics

A systematic comparison of computational methods for expression forecasting

[1]Eric Kernfeld, ekernfe1@jh.edu, https://orcid.org/0000-0002-2310-6191
[2]Yunxiao Yang, yyang117@jh.edu, https://orcid.org/0000-0003-3429-3174
[3]Joshua S. Weinstock, jweins17@jhu.edu, https://orcid.org/0000-0001-7013-1899
[1,2,3,4,5]Alexis Battle, ajbattle@jhu.edu (corresponding author), https://orcid.org/0000-0002-5287-627X
[1,4,5]Patrick Cahan, patrick.cahan@jhmi.edu (corresponding author), https://orcid.org/0000-0003-3652-2540

choice of metric, and especially for simple metrics like mean squared error, it is uncommon for expression forecasting methods to out-perform simple baselines. Our platform will serve as a

Goal

- What makes an embedding useful for perturbation response?
    1. We **fix the perturbation response model** to decouple embeddings from downstream predictors.
    2. We study a **diverse collection of modalities** and representation learning strategies.
    3. We utilize **simple metrics** (e.g., L2), but also more **biologically interpretable scoring** methods.

Methods: Problem Formulation

Methods: Embeddings Benchmarked

- Categories
  - Baselines: Random, No Change, PCA, Train Mean, Idealized
    - No Change: Predict the expression of control cells
    - Idealized: Embeddings obtained by fitting PCA on *all* pseudobulk data (train + test)
    - Random: Embeddings drawn at random from a Gaussian distribution
  - **Expression FMs**: AIDO.Cell (3M/10M/100M), scGPT, Geneformer, scPRINT, Transcriptformer
  - **Protein FMs**: AIDO.ProteinIF, ESM2, AIDO.StructureTokenizer, STRING-SPACE
  - **DNA FMs** AIDO.DNA
  - **Network embeddings**: STRING (network, NBFNet, WaveGC)
  - **Prior Knowledge**: GenePT (GO BP/CC/MF/NCBI+UniProt), GenotypeVAE

- Stress Diversity
  - Expression, sequence, structure, prior knowledge

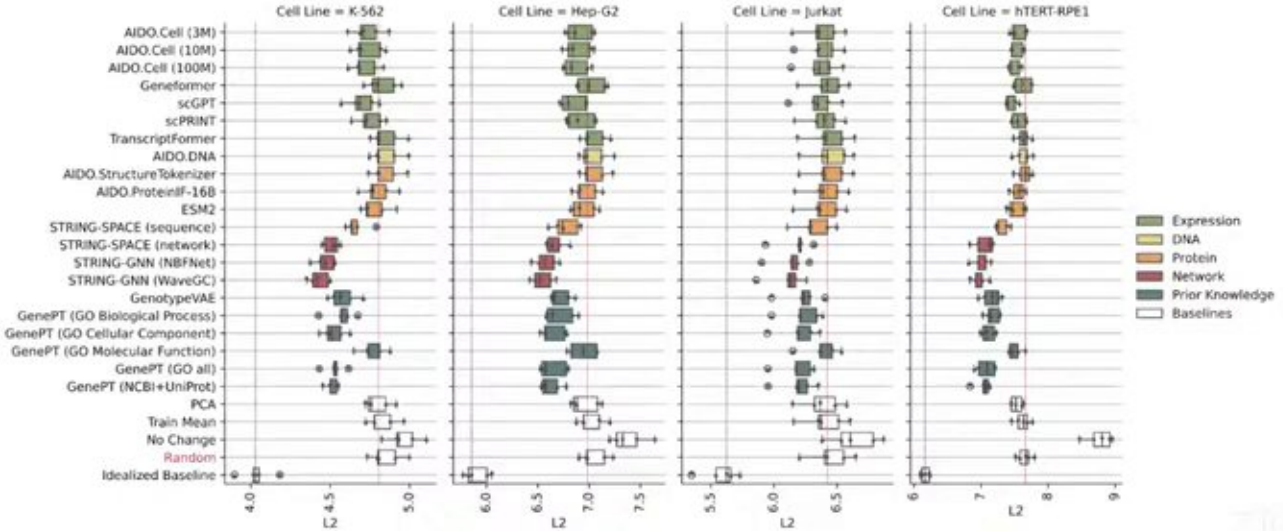
Datasets

- **K-562** [1]
  - lymphoblast from individual with chronic myelogenous leukemia
  - 1534 perturbations
- **RPE1** [1]
  - retinal pigment epithelial cell from a healthy individual
  - 1752 perturbations
- **Jurkat** [2]
  - T cell from individual with leukemia
  - 1752 perturbations
- **Hep-G2** [2]
  - epithelial cell from an individual with hepatocellular carcinoma
  - 1752 perturbations

[1] Replogle, J. M. *et al.* Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559-2575.e28 (2022).
[2] Nadig, A. *et al.* Transcriptome-wide analysis of differential expression in perturbation atlases. *Nat. Genet.* **57**, 1228–1237 (2025).

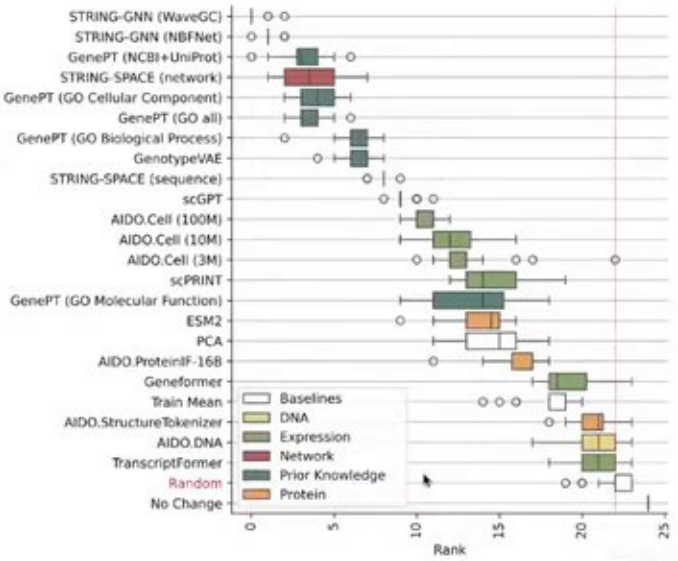
Results: Per-Cell Line Performance

Split-to-split variability: robust cross-validation is important!



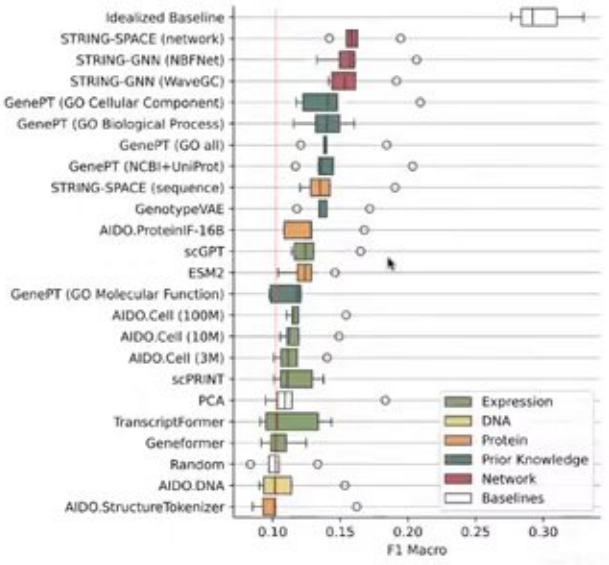
Results: Overall Ranking

1. **Knowledge-based and network embeddings** perform best.
2. Some cell **FMs outperform PCA**.
3. **Scaling helps** (AIDO.Cell 100M > 10M > 3M).
4. Protein FMs — **STRING-sequence best**.
5. GenePT using **"GO Cellular Component"** *outperforms* "GO Biological Process".



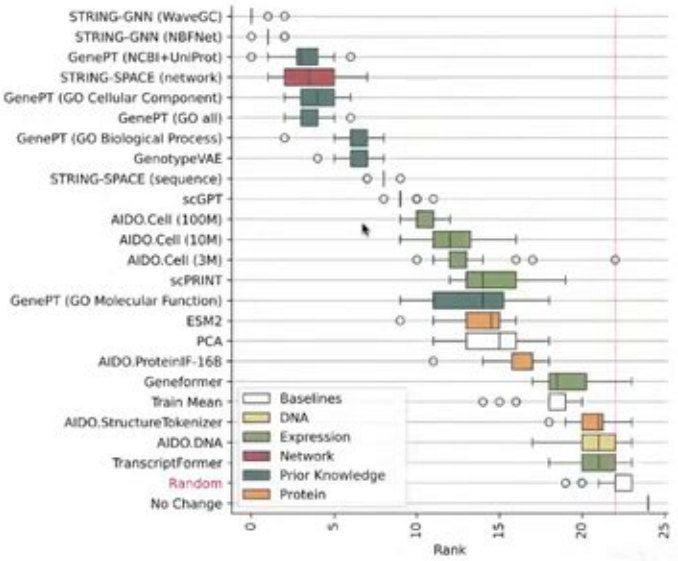
Results: Function Prediction

1. Models **ranked similarly** as for expression prediction.
2. Knowledge-driven embeddings again strongest.
3. **Protein embeddings slightly better** here vs expression.
4. There is still room for improvement (idealized baseline).



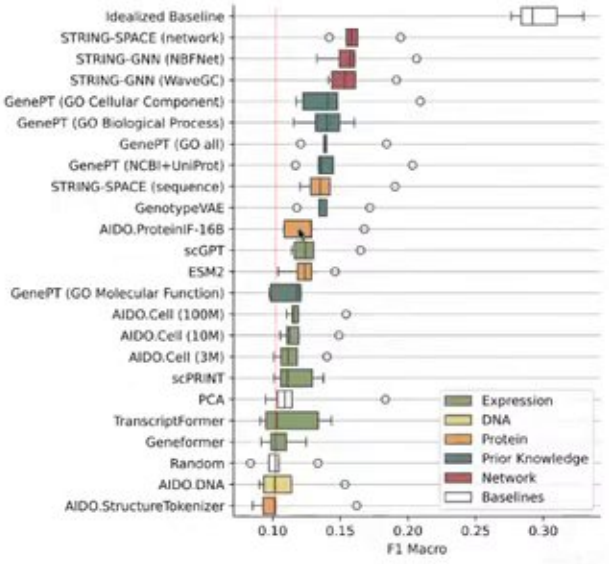
Results: Overall Ranking

1. **Knowledge-based and network embeddings** perform best.
2. Some cell **FMs outperform PCA**.
3. **Scaling helps** (AIDO.Cell 100M > 10M > 3M).
4. Protein FMs — **STRING-sequence best**.
5. GenePT using **"GO Cellular Component"** *outperforms* "GO Biological Process".



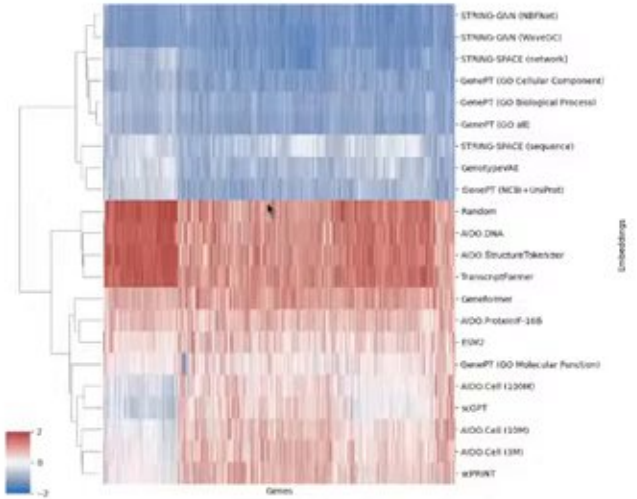
Results: Function Prediction

1. Models **ranked similarly** as for expression prediction.
2. Knowledge-driven embeddings again strongest.
3. **Protein embeddings slightly better** here vs expression.
4. There is still room for improvement (idealized baseline).



Limitations & Future Directions

- **Limitations**
  - Only one model type (kNN/logistic)
  - Pseudobulk, not single-cell
  - Fixed embeddings (no fine-tuning)
- **Next steps**
  - Fusion across scales / modalities
    - Extend beyond single-modality cell FMs by integrating diverse biological data types
  - Multi cell-line generalization



Contributions

- Conducted a **systematic pan-modal study of perturbation embeddings**, and evaluated their predictive utility separately from downstream predictors.
- Evaluated embeddings from **20+ biological foundation models** and rigorously benchmarked them against simpler baselines.
- Introduced a **novel, biologically interpretable** formulation of perturbation response modeling based on predicting functional GO terms.
- Knowledge-based and network embeddings perform best.

Multimodal Benchmarking of Foundation Model Representations for Cellular Perturbation Response Prediction

Euxhen Hasanaj
Research Scientist
GenBio AI
Contact: euxhen.hasanaj@genbio.ai

Elijah Cole
Shahin Mohammadi
Sohan Addagudi
Xingyi Zhang
Le Song
Eric Xing

GenBio

Carnegie Mellon University

Mohamed bin Zayed University of Artificial Intelligence

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Disclaimer: All authors are affiliated with GenBio AI.

MCBRLab bilibili

NYGC Events

zoom

@

Thank you to our sponsors

And helpers!

Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang

Restarting at 11.55am

mlcb.org for schedule

MLCB @ NEW YORK GENOME CENTER

Thank you to our sponsors

CORTEVA™ agriscience

InstaDeep®

CSH Cold Spring Harbor Laboratory

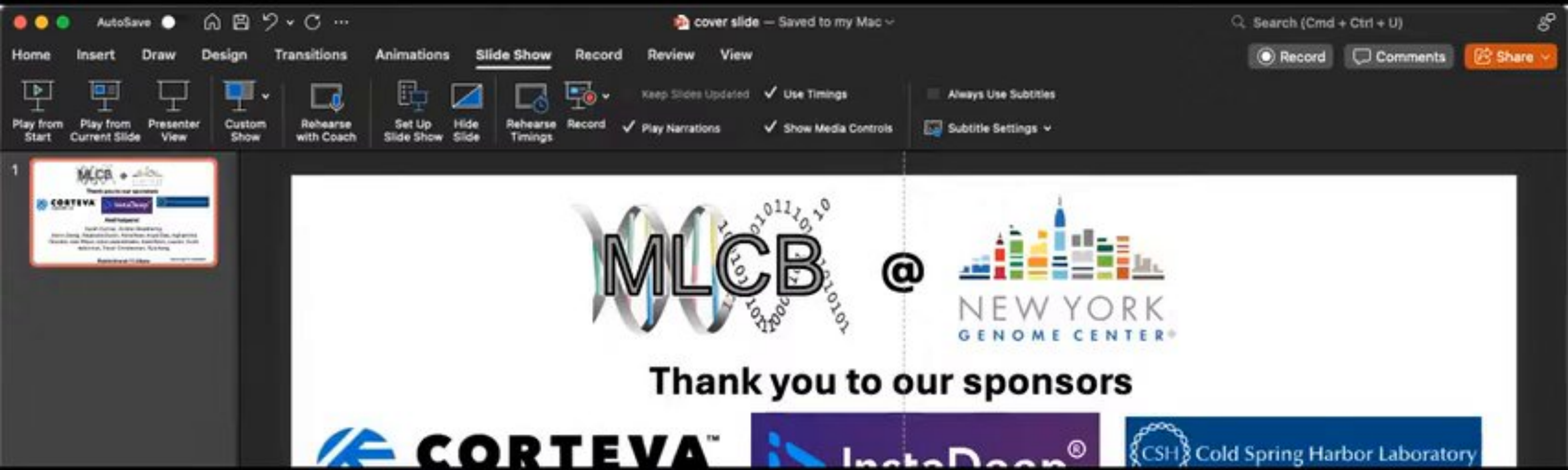
And helpers!

Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang

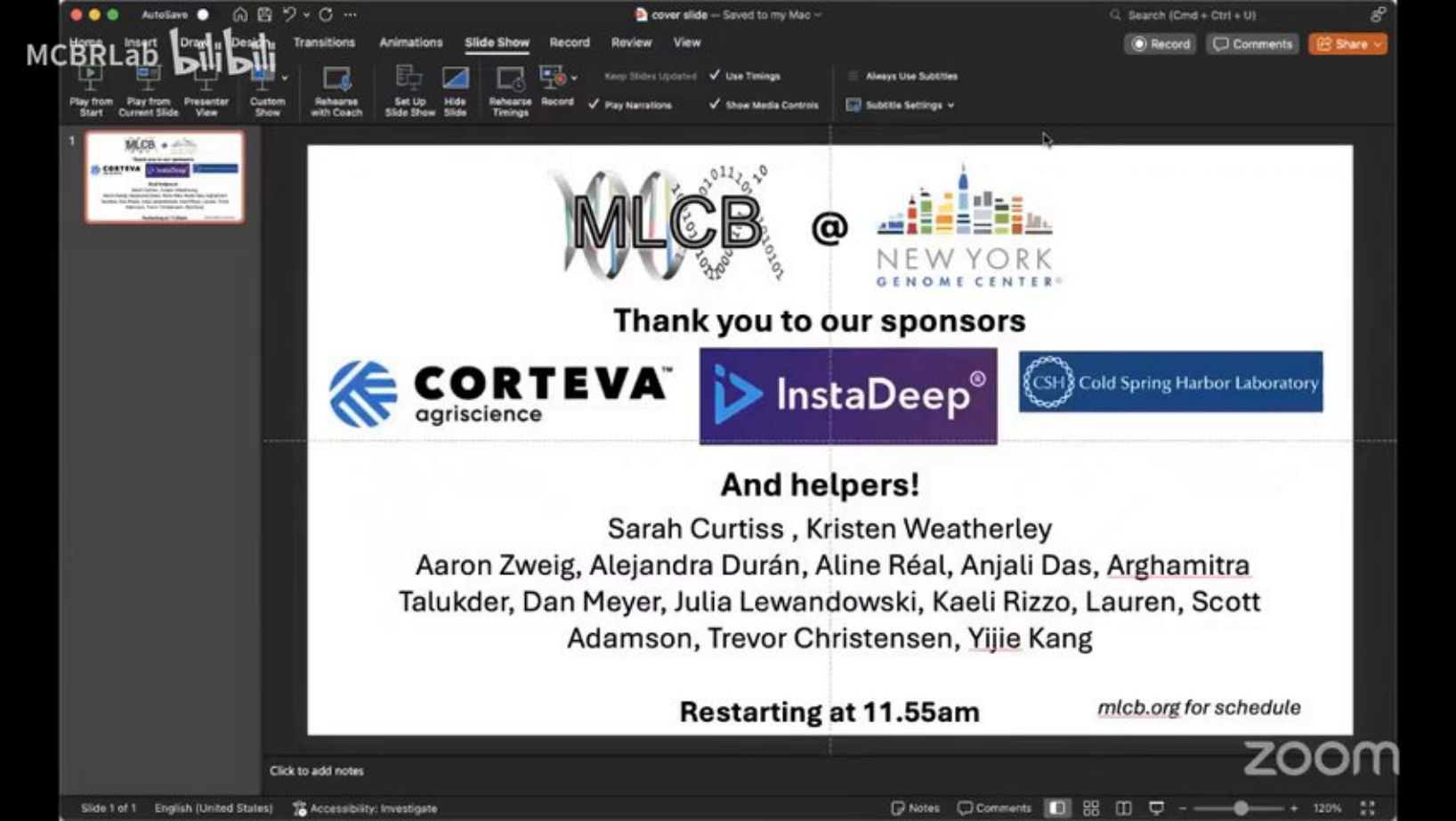
Restarting at 11.55am

mlcb.org for schedule

MCBRLab bilibili



zoom



MCBF

MLCB @ NEW YORK GENOME CENTER

Thank you to our sponsors

CORTEVA™ agriscience

InstaDeep®

CSH Col

And helpers!

Sarah Curtiss , Kristen Weatherley

Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das,

Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, La

Adamson, Trevor Christensen, Yijie Kang

MCBRLab bilibili

NYGC Events

zoom

MLCB @ NEW YORK GENOME CENTER

Thank you to our sponsors

CORTEVA™ agriscience

InstaDeep®

CSH Cold Spring Harbor Laboratory

And helpers!

Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang

Restarting at 11.55am

mlcb.org for schedule

Click to add notes

@

Thank you to our sponsors

And helpers!

Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang

Restarting at 11.55am

mlcb.org for schedule

MCBRLab bilibili

NYGC Events

zoom

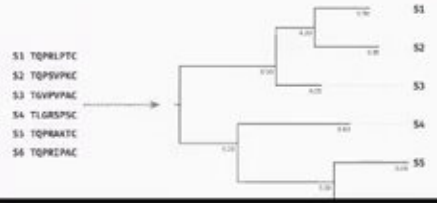
MCBRLab

Deep end-to-end likelihood-free inference of phylogenetic trees

Luc Blassel, Nicolas Lartillot, Bastien Boussau, Laurent Jacob

MLCB - September 11th, 2025

Context - Phylogenetic inference



28	Day 1	27	
29	Day 1	28	Enhancing Bulk RNA-Seq Deconvolution Using Atlas-Level Deep Learning Embeddings
30	Day 1	29	Decoding translation regulation at the 5' untranslated region of mRNA with augmented multi-species massively parallel reporter assay and active learning

Schedule · Spotlights & Posters ·

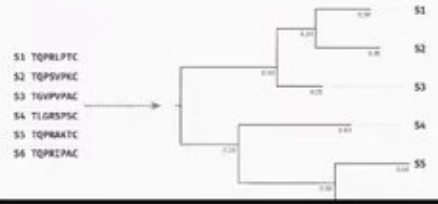
Convert to table

Deep end-to-end likelihood-free inference of phylogenetic trees

Luc Blassel, Nicolas Lartillot, Bastien Boussau, Laurent Jacob

MLCB - September 11th, 2025

Context - Phylogenetic inference



Day 1 27 Enhancing Bulk RNA-Seq Deconvolution Using Atlas-Level Deep Learning Embeddings Ali Oku (New York Genome Cente

Day 1 28 Decoding translation regulation at the 5' untranslated region of mRNA with augmented multi-species massively parallel reporter assay and active learning

Day 1 29

Convert to table

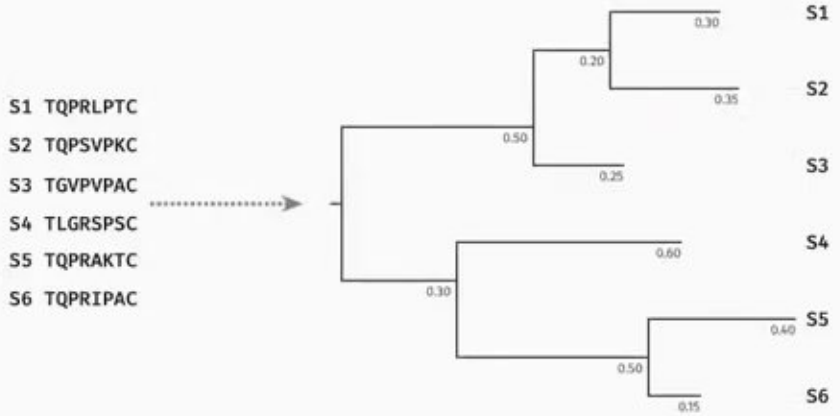
Schedule Spotlights & Posters

Deep end-to-end likelihood-free inference of phylogenetic trees

Luc Blassel, Nicolas Lartillot, Bastien Boussau, Laurent Jacob

MLCB - September 11th, 2025

Context - Phylogenetic inference



Goal: describe **evolutionary-history** of MSA

MCBRLab

Deep end-to-end likelihood-free inference of phylogenetic trees

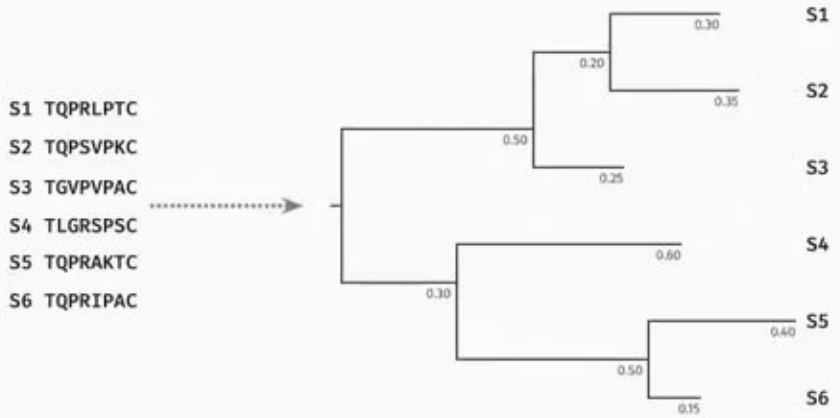
Luc Blassel, Nicolas Lartillot, Bastien Boussau, Laurent Jacob

MLCB - September 11th, 2025

LCQB SORBONNE UNIVERSITÉ LBBE cnrs

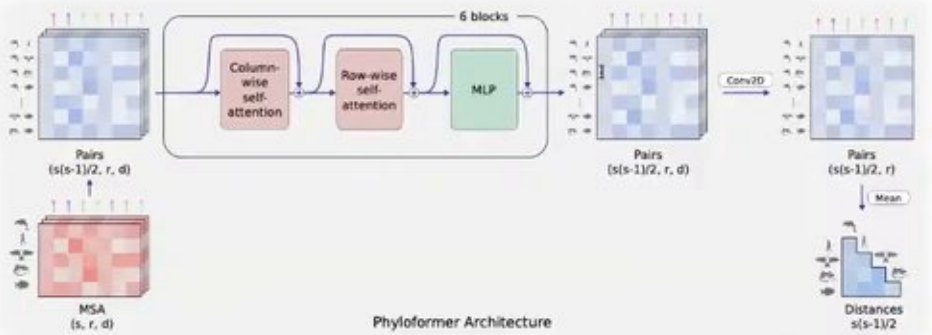
zoom

Context - Phylogenetic inference



Goal: describe **evolutionary-history** of MSA

Related Work - Phyloformer, our first approach

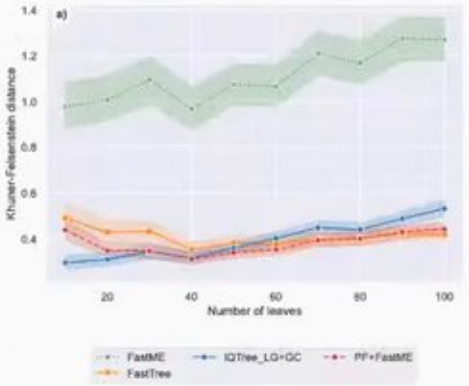


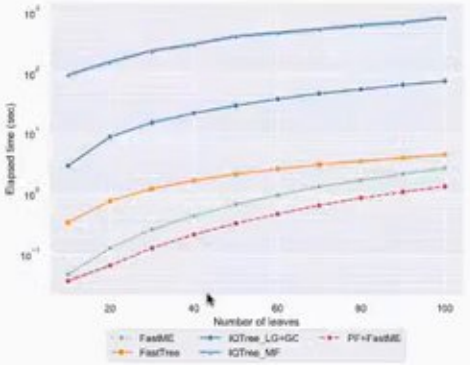
Phyloformer Architecture

- Input an **MSA**, get a **Distance matrix**
- Feed Distance matrix to **FastME** to get **tree**

Nesterenko et al. 2025; Lefort et al. 2015

Related Work - Phyloformer is good!



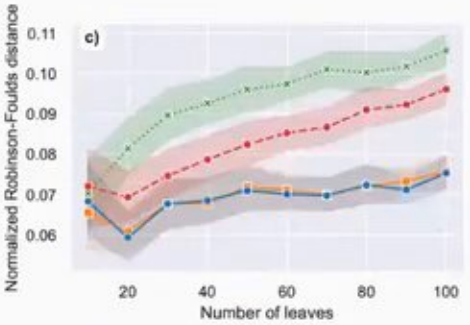


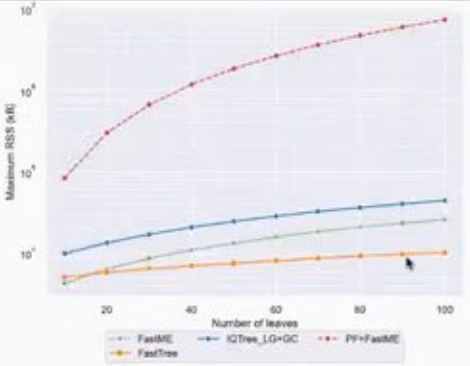
Tree inference accuracy (KF) *Runtime*

- Fairly **competitive** even on simple LG+GC model
- **Fast** because we use GPUs [1]

Nesterenko et al. 2025, [1] Jean-Zay

Related Work - But also sometimes less good...



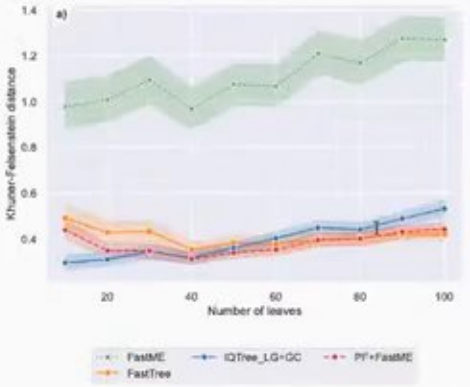


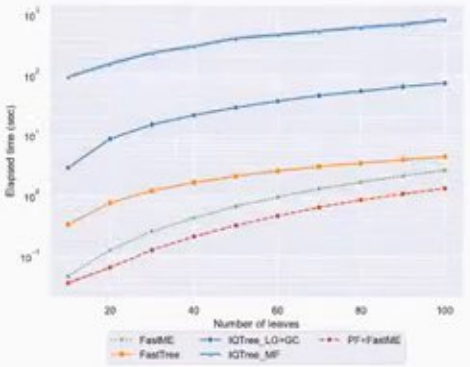
Topological accuracy (RF)

Memory usage

- **Gap** between PF and **ML methods**
- PF is **by far** the most **memory intensive**

Related Work - Phyloformer is good!



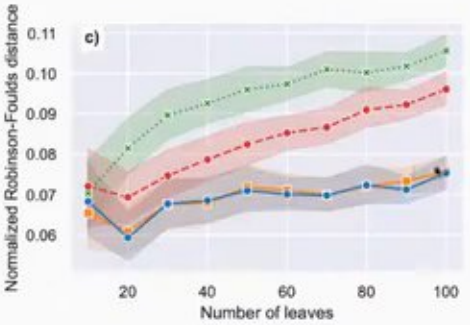


Tree inference accuracy (KF) *Runtime*

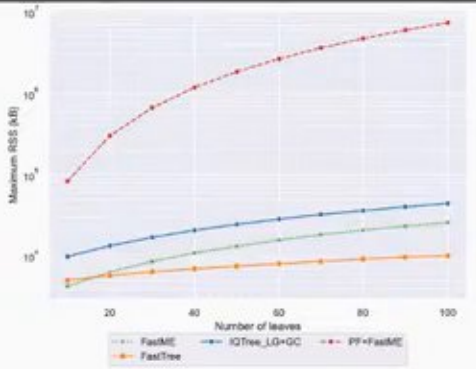
- Fairly **competitive** even on simple LG+GC model
- **Fast** because we use GPUs [1]

Nesterenko et al. 2025, [1] Jean-Zay

Related Work - But also sometimes less good...



Topological accuracy (RF)



Memory usage

- **Gap** between PF and **ML methods**
- PF is **by far** the most **memory intensive**

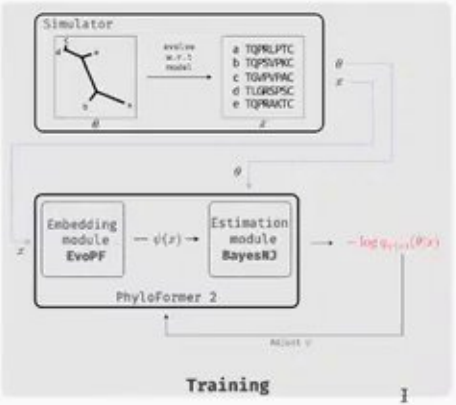
How to do phylogenetic inference end-to-end ?

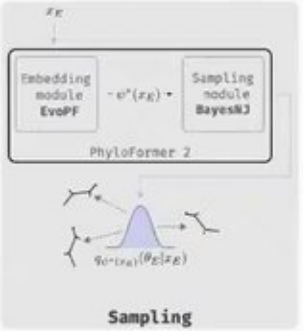
Methods - Neural Posterior Estimation (NPE)

- Given a **probabilistic model** $p(x|\theta)$ with some prior $p(\theta)$
- We want to **estimate the posterior**: $p(\theta|x)$
- We build $q_\psi(\theta|x)$ a **family** of distributions **parametrized** by ψ (our NN)
- We find $q_{\psi^*} = \underset{\psi}{\text{argmin}}\ \mathbb{E}_{p(x)}[KL(q_\psi(\theta|x)||p(\theta|x)]$
- In practice we **maximize** $\mathbb{E}_{p(x,\theta)}[\log q_{\psi(x)}(\theta|x)]$ by **sampling** from $p(x, \theta)$

x : MSA, $\theta = (\tau, \ell)$: Phylogenetic tree, $\psi(x)$: NN applied to x

Methods - How do we do NPE?





- During **training** find $\psi^* = \underset{\psi}{\text{argmin}} \; -\sum_i \log q_{\psi(x_i)}(\theta_i|x_i)$
- At **inference** time **sample** from: $q_{\psi^*(x_E)}(\theta_E|x_E)$ [1]

[1] WIP: so for now only point-estimation

Methods - The EvoPF module, intro

the EvoPF module is an **adaptation** of the **EvoFormer** module from **AlphaFold2**. The tasks are **transpositions** of each other:

given input MSA ($n \times r$)

EvoFormer represent $r \times r$ relationships between sites

EvoPF represent $n \times n$ relationships between sequences

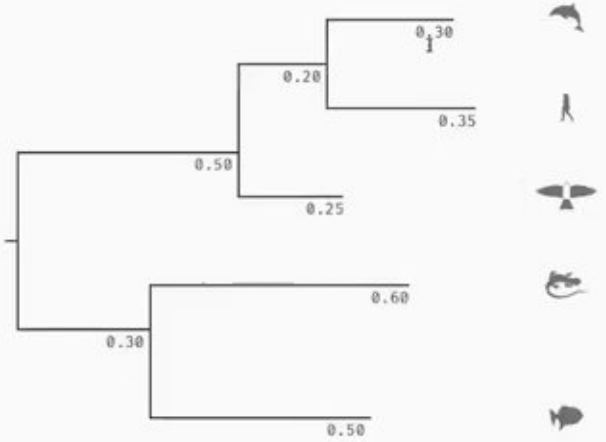
More expressive than MSA transformer
More lightweight than PF

Jumper et al. 2021; Rao et al. 2021

MCBRLab bilibili

Methods - A tree is a series of merges

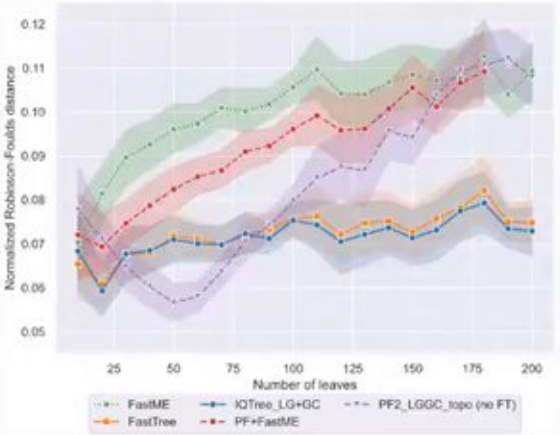
We want to describe the following tree:



zoom

Does it work ?

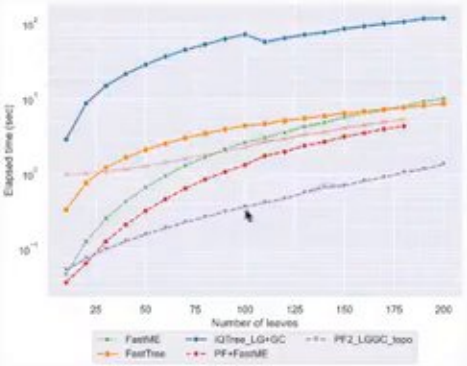
Results - Training topology only



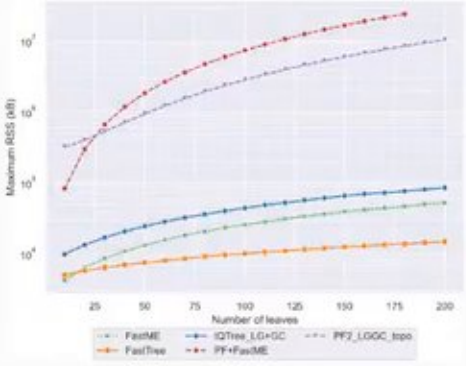
- **overfitting** on tree-size is an **issue**

Same train set as PF1 paper: ≈ 170*k* 50 seq LG+GC MSAs on rescaled BD trees

Results - Scalability



Execution time



Memory usage[1]

[1] With 2× bigger sequence, and 4× bigger pair embeddings...

Perspectives - Short term challenges

This is very much still a **work in progress...**

- Training on **more complex** data *(e.g. indels)* **increases** length-**overfitting**
- Learning **topology** and **branch-lengths** is also **challenging**
- How can we move **away** from **point-estimation** ?
- We might need to **adjust** our **priors** to compare with MCMC tools

Conclusion

Takeaways

- **Topologically** we manage to **beat** ML-methods[1] on LG
- While being **more scalable** than PF1
- Still needs some **work** for a fully **end-to-end** phylogenetic **inference** tool

What next ?

- Can we do **better** where computing $p(\theta|x)$ is **difficult** or **intractable**? *(e.g. Potts, epistatis, Selection, ...)*
- **Confident** this can **work** given our experience with **PF**

Prillo et al. 2023; Duchemin et al. 2023; Latrille et al. 2021

[1] Yay!

MCBRLab bilibili

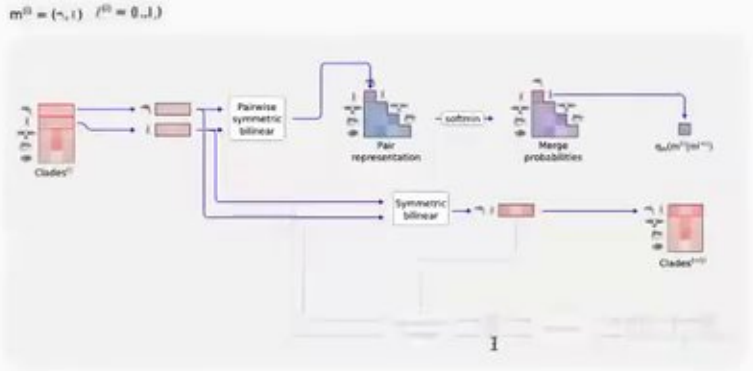
Thanks to:

- **Luca Nesterenko**
- **Laurent Jacob**
- **Bastien Boussau**
- **Nicolas Lartillot**
- **Philippe Veber**
- **Vincent Garot**
- **Amélie Leroy**
- **Anybody that listened to me!**

Special thanks to Jean-Zay for all the GPUs!

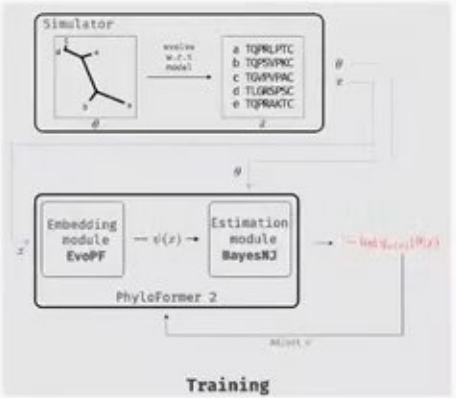
zoom

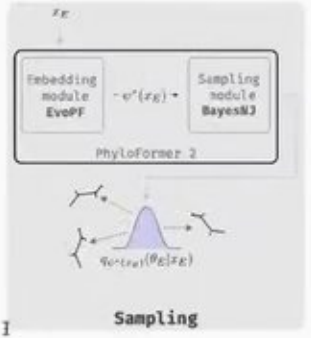
Methods - BayesNJ, evaluating topological probabilities



Update clade **representation** for next merge

Methods - How do we do NPE?





- During **training** find $\psi^* = \underset{\psi}{\text{argmin}} \; -\sum_i \log q_{\psi(x_i)}(\theta_i|x_i)$
- At **inference** time **sample** from: $q_{\psi^*(x_E)}(\theta_E|x_E)$ [1]

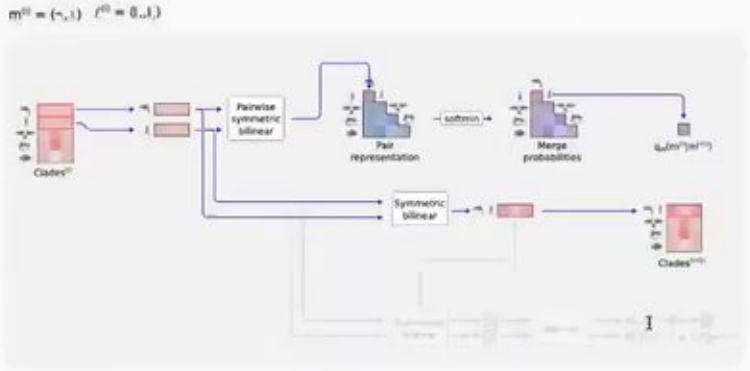
[1] WIP: so for now only point-estimation

Methods - Neural Posterior Estimation (NPE)

- Given a **probabilistic model** $p(x|\theta)$ with some prior $p(\theta)$
- We want to **estimate the posterior**: $p(\theta|x)$
- We build $q_\psi(\theta|x)$ a **family** of distributions **parametrized** by ψ (our NN)
- We find $q_{\psi^*} = \underset{\psi}{\text{argmin}}\ \mathbb{E}_{p(x)}[KL(q_\psi(\theta|x)||p(\theta|x)]$
- In practice we **maximize** $\mathbb{E}_{p(x,\theta)}[\log q_{\psi(x)}(\theta|x)]$ by **sampling** from $p(x, \theta)$

x : MSA, $\theta = (\tau, \ell)$: Phylogenetic tree, $\psi(x)$: NN applied to x

Methods - BayesNJ, evaluating topological probabilities

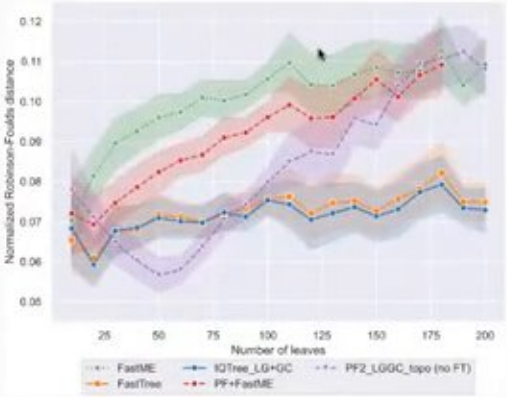


Update clade **representation** for next merge

MCBRLab

154_BLASSEL Luc Blassel.pdf

Results - Training topology only

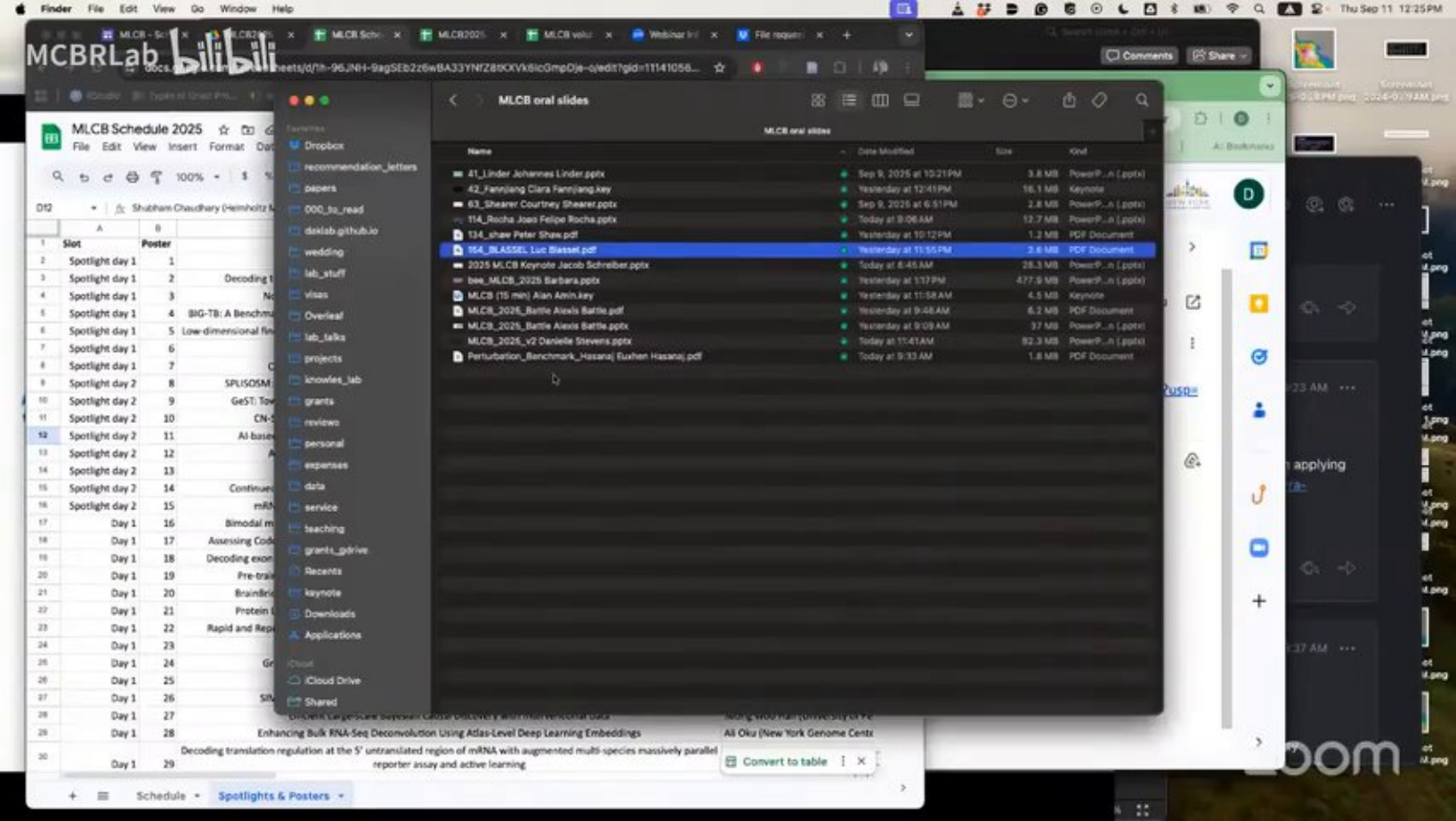


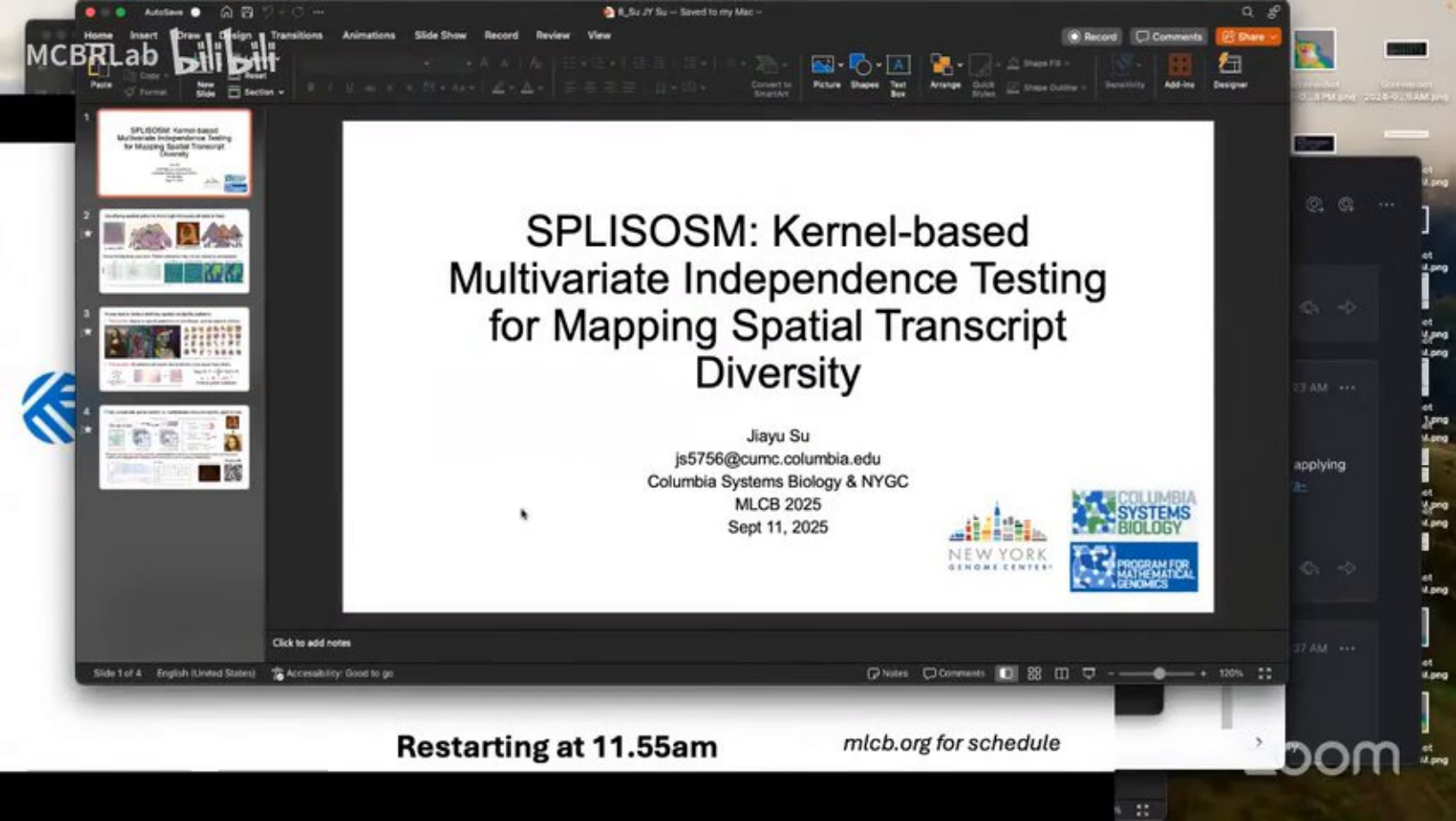
- **overfitting** on tree-size is an **issue**

28	Day 1	27	
29	Day 1	28	Enhancing Bulk RNA-Seq Deconvolution Using Atlas-Level Deep Learning Embeddings
30	Day 1	29	Decoding translation regulation at the 5' untranslated region of mRNA with augmented multi-species massively parallel reporter assay and active learning

Convert to table

Schedule · Spotlights & Posters


MCBRLab bilibili
MLCB Schedule 2025
MLCB oral slides
Schedule
Spotlights & Posters


MCBRLab
SPLISOSM: Kernel-based Multivariate Independence Testing for Mapping Spatial Transcript Diversity
Jiayu Su
js5756@cumc.columbia.edu
Columbia Systems Biology & NYGC
MLCB 2025
Sept 11, 2025
NEW YORK GENOME CENTER
COLUMBIA SYSTEMS BIOLOGY
PROGRAM FOR MATHEMATICAL GENOMICS
Click to add notes
Restarting at 11.55am
mlcb.org for schedule

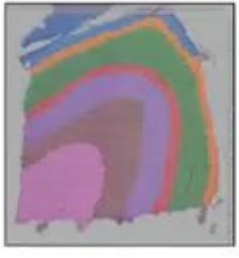
SPLISOSM: Kernel-based Multivariate Independence Testing for Mapping Spatial Transcript Diversity

Jiayu Su
js5756@cumc.columbia.edu
Columbia Systems Biology & NYGC
MLCB 2025
Sept 11, 2025

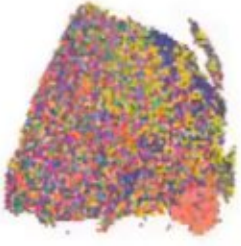
Identifying spatial patterns from high-dimensional data is hard

Human DLPFC

Identifying spatial patterns from high-dimensional data is hard

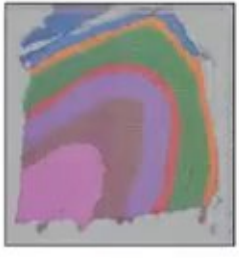


Human DLPFC

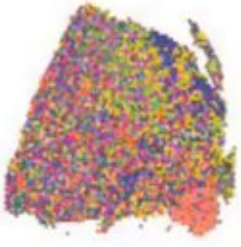


Human glioblastoma

Identifying spatial patterns from high-dimensional data is hard

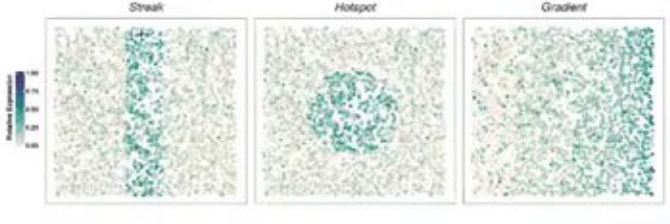


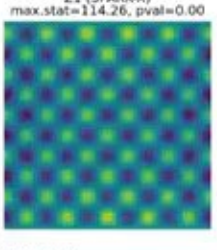
Human DLPFC

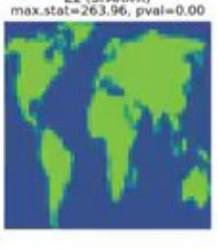


Human glioblastoma

Never blindly trust your lens: Pattern detection may not be robust to perturbation

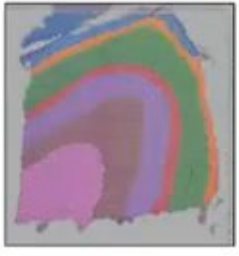




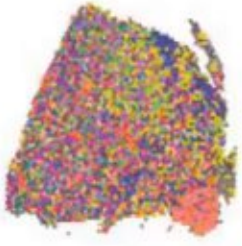


Zhu, Jiaqiang, Shiquan Sun, and Xiang Zhou. *Genome biology* 22.1 (2021): 184.

Identifying spatial patterns from high-dimensional data is hard

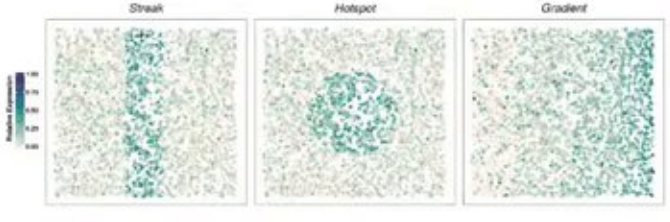


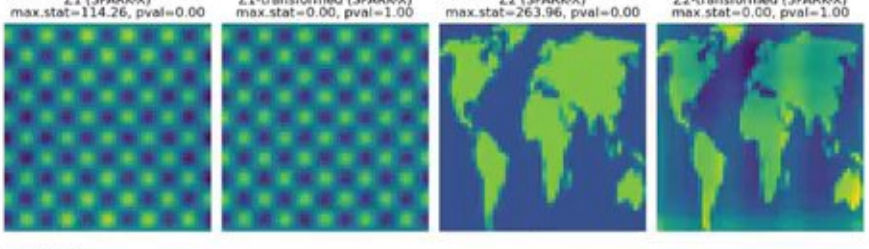
Human DLPFC



Human glioblastoma

Never blindly trust your lens: Pattern detection may not be robust to perturbation



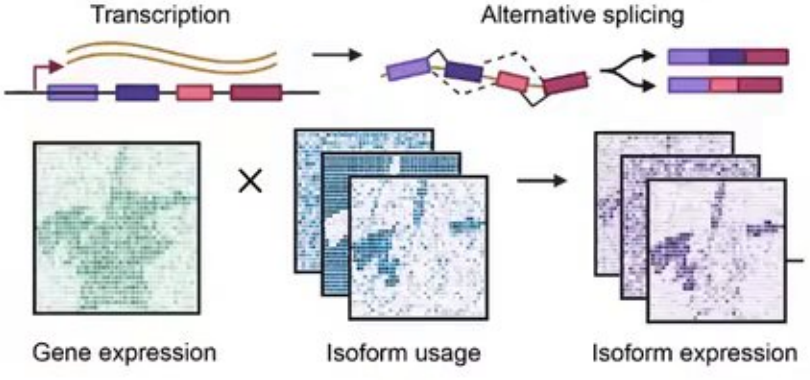


Zhu, Jiaqiang, Shiquan Sun, and Xiang Zhou. *Genome biology* 22.1 (2021): 184.

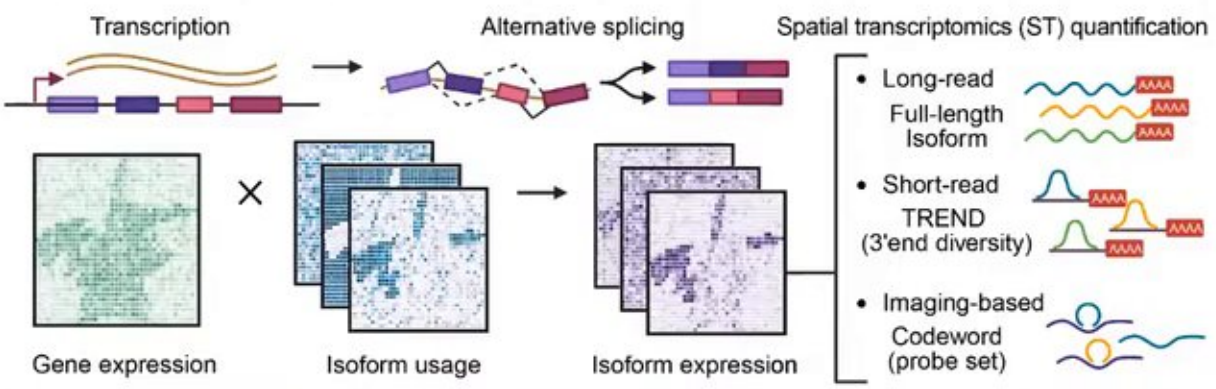
A new test to detect arbitrary spatial variability patterns

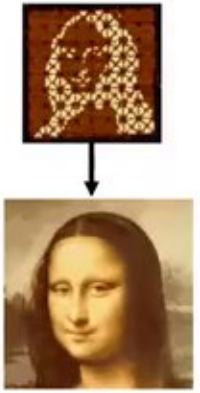
- The conflict: Space of spatial patterns is of dim #spots, and we want all of them

From univariate gene-centric to multivariate isoform-centric gold mines

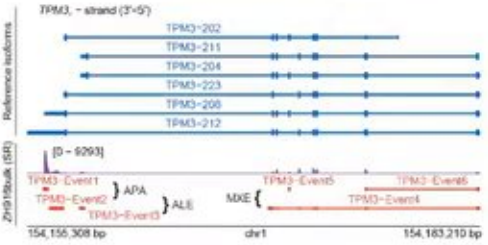


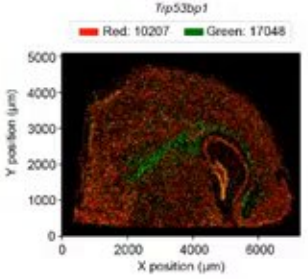
From univariate gene-centric to multivariate isoform-centric gold mines

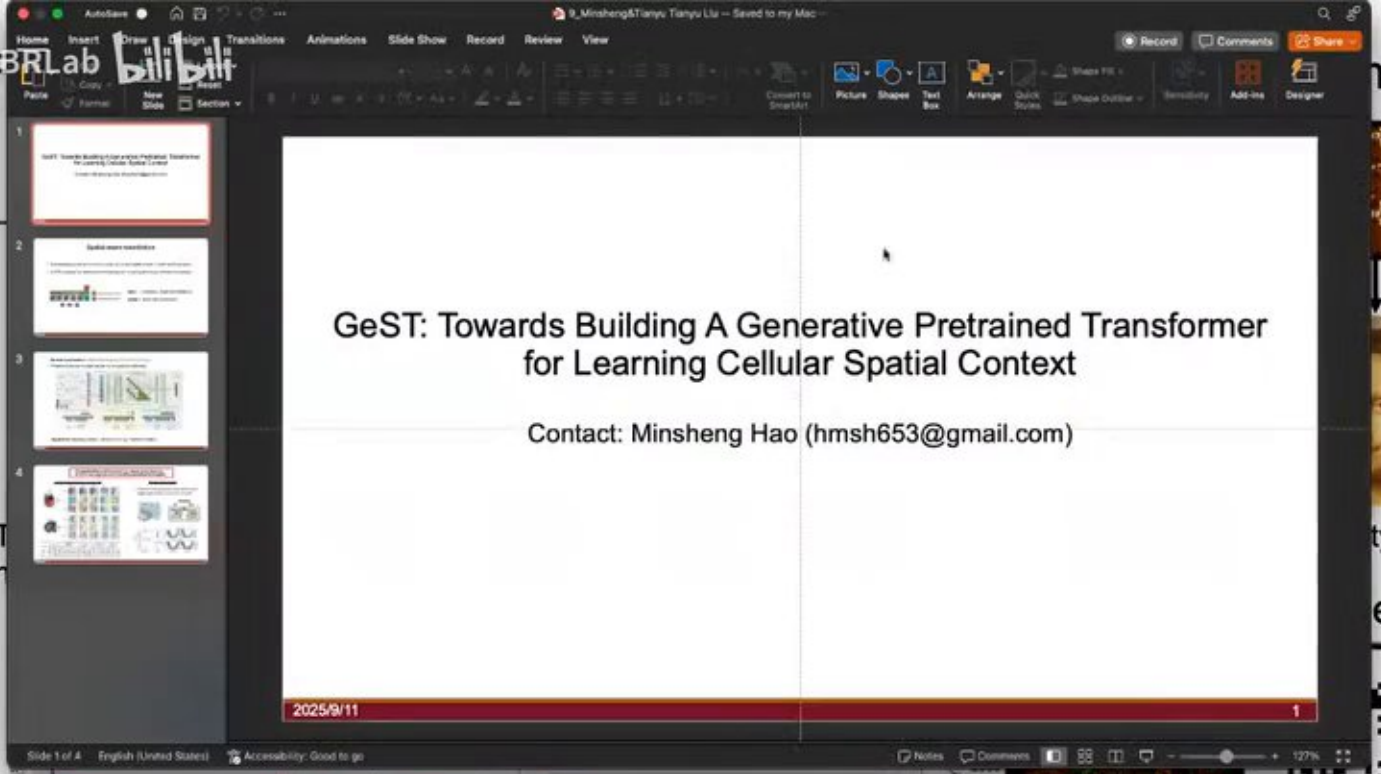




Through reanalysis of existing datasets, we revealed thousands of conserved spatial transcript diversity in healthy and malignant brain tissues and uncovered complex regulatory relationships





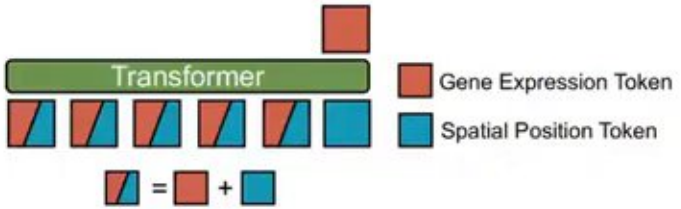


GeST: Towards Building A Generative Pretrained Transformer for Learning Cellular Spatial Context

Contact: Minsheng Hao (hmsh653@gmail.com)

Spatial-aware tokenization

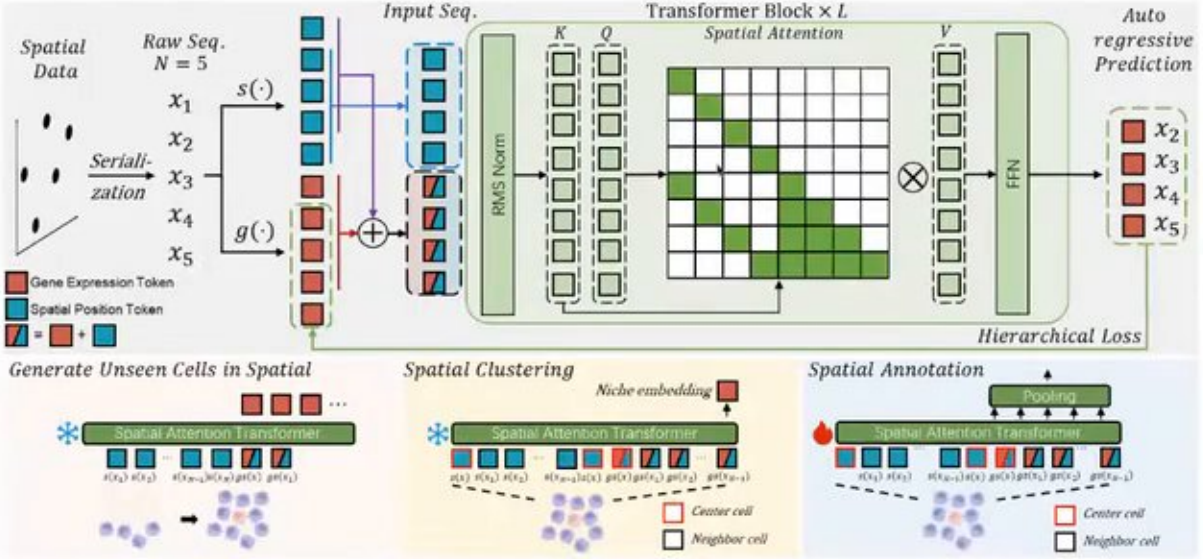
- Understanding spatial cell context is crucial, but current models is hard to model spatial dynamics
- GeST is a generative transformer that learns spatial context by predicting a cell from its neighbors.



Input = neighbors, target cell's spatial loc.

output = target cell's gene profile

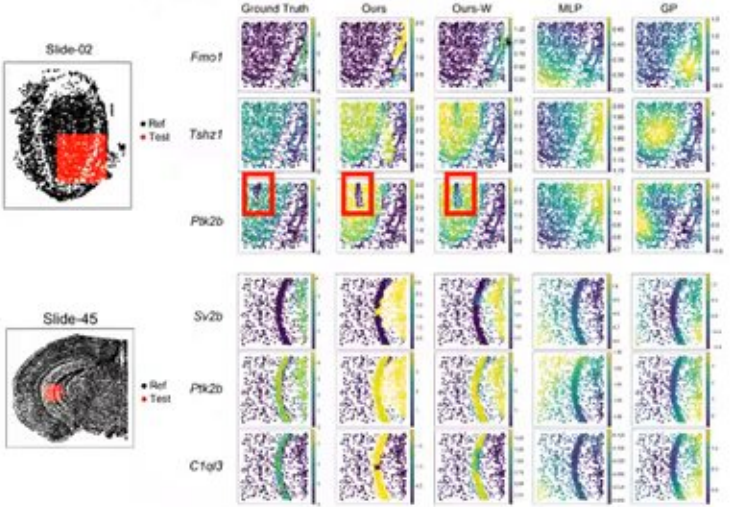
- **Spatial serialization + Cell tokenization + Spatial attention**
- Pretraining across multiple spatial transcriptomics datasets



- **Application**: Spatial generation + Spatial clustering + Spatial Annotation

- First **generative** pretrained model for spatial transcriptomics.
- Enables new applications like **in-silico perturbation analysis**

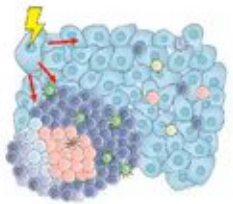
Generating unseen region (in red)

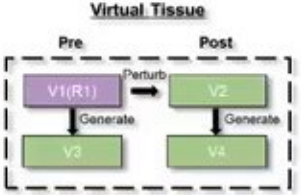


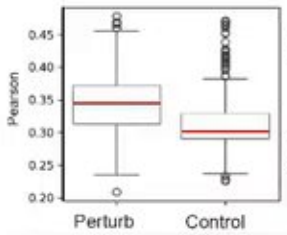
Method	MERFISH Anterior Brain R.all↓	R.200↓	ρ↑	MERFISH Mid Brain R.all↓	R.200↓	ρ↑	MERFISH Posterior Brain R.all↓	R.200↓	ρ↑	Visium Primary Liver Cancer R.all↓	R.200↓	ρ↑	Stereo-seq Sagittal Brain R.all↓	R.200↓	ρ↑
Ours	1.374	1.296	0.301	1.379	1.319	0.265	1.386	1.340	0.242	**1.303**	**1.126**	**0.540**	1.405	1.357	**0.326**
Ours-W	**1.352**	**1.244**	**0.340**	**1.367**	**1.288**	**0.302**	**1.376**	**1.322**	**0.275**	1.320	1.150	0.499	**1.399**	**1.340**	0.323
GP	1.379	1.289	0.241	1.389	1.337	0.241	1.393	1.354	0.214	1.357	1.264	0.272	1.413	1.405	0.073
MLP	1.399	1.369	0.319	1.404	1.393	0.283	1.406	1.397	0.267	1.347	1.174	0.491	1.403	1.357	0.314

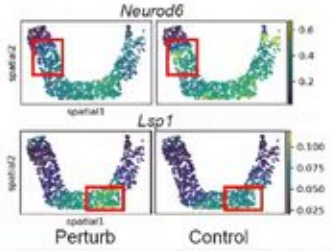
Spatial Perturbation

- ➢ Reference: A real experiment to study ischemic brain
- ✓ **Regenerate** the ischemic perturbation via GeST

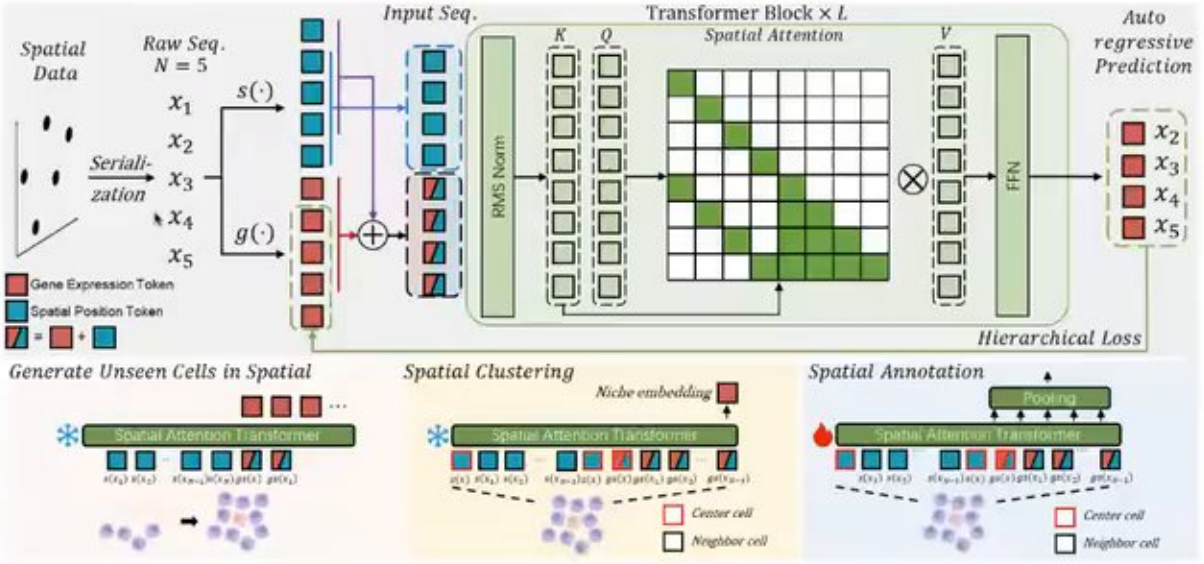








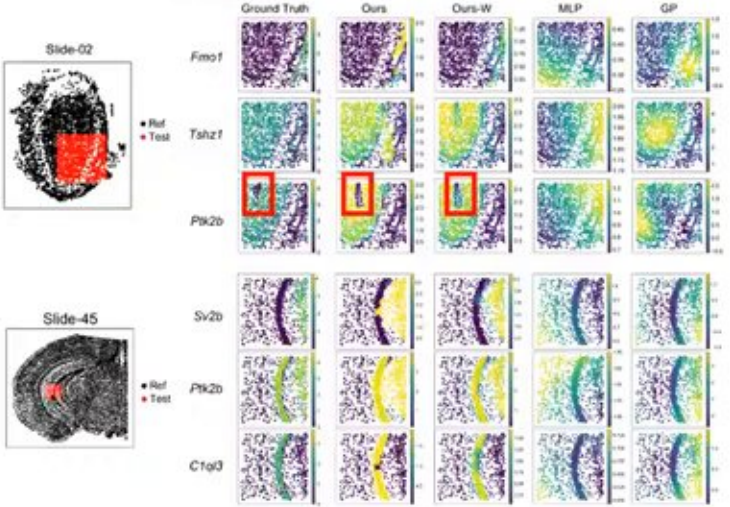
- **Spatial serialization** + **Cell tokenization** + **Spatial attention**
- Pretraining across multiple spatial transcriptomics datasets



- **Application**: Spatial generation + Spatial clustering + Spatial Annotation

- First **generative** pretrained model for spatial transcriptomics.
- Enables new applications like **in-silico perturbation analysis**

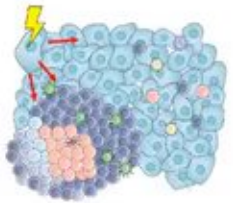
Generating unseen region (in red)

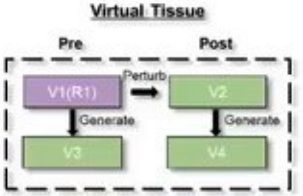


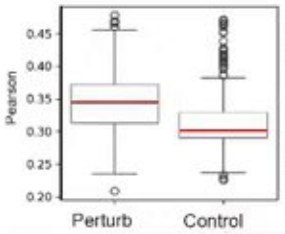
Method	MERFISH Anterior Brain			MERFISH Mid Brain			MERFISH Posterior Brain			Visium Primary Liver Cancer			Stereo-seq Sagittal Brain		
	R.all↓	R.200↓	ρ↑	R.all↓	R.200↓	ρ↑	R.all↓	R.200↓	ρ↑	R.all↓	R.200↓	ρ↑	R.all↓	R.200↓	ρ↑
Ours	1.374	1.296	0.301	1.379	1.319	0.265	1.386	1.340	0.242	**1.303**	**1.126**	**0.540**	1.405	1.357	**0.326**
Ours-W	**1.352**	**1.244**	**0.340**	**1.367**	**1.288**	**0.302**	**1.376**	**1.322**	**0.275**	1.320	1.150	0.499	**1.399**	**1.340**	0.323
GP	1.379	1.289	0.241	1.389	1.337	0.241	1.393	1.354	0.214	1.357	1.264	0.272	1.413	1.405	0.073
MLP	1.399	1.369	0.319	1.404	1.393	0.283	1.406	1.397	0.267	1.347	1.174	0.491	1.403	1.357	0.314

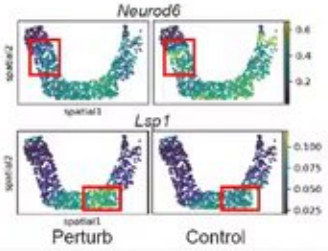
Spatial Perturbation

- ➢ Reference: A real experiment to study ischemic brain
- ✓ **Regenerate** the ischemic perturbation via GeST









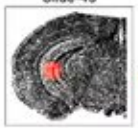
- First **generative** pretrained model for spatial transcriptomics.
- Enables new applications like **in-silico perturbation analysis**

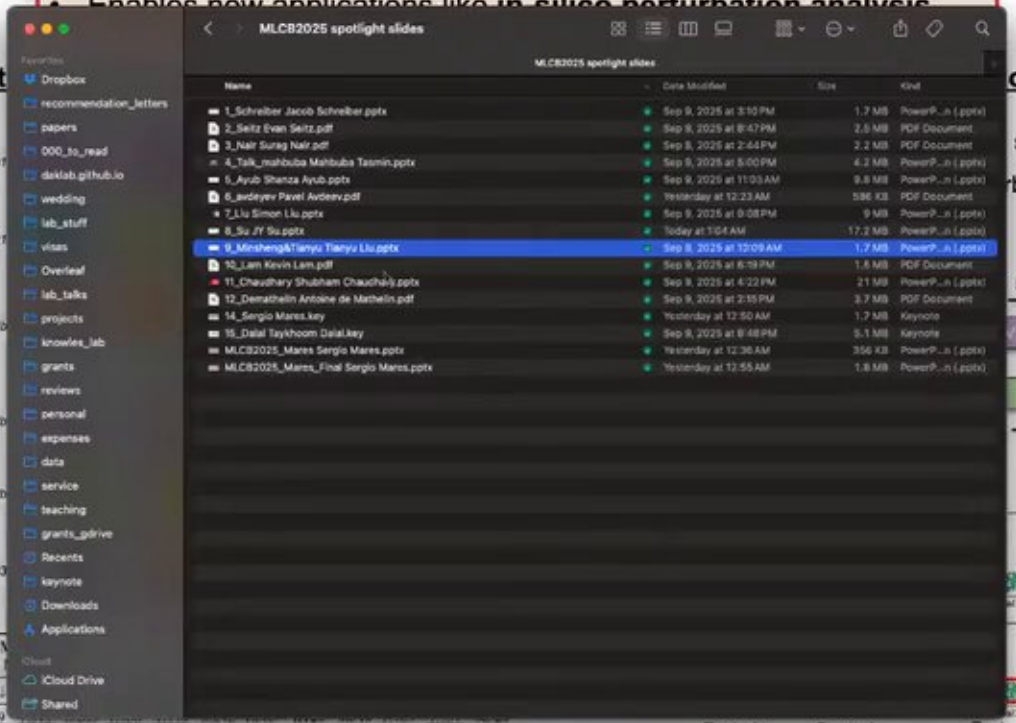
Generat

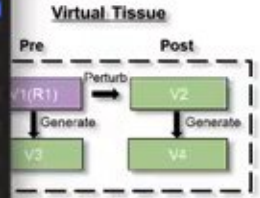
on

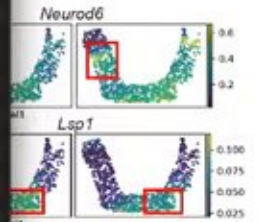
study ischemic brain

rbation via GeST







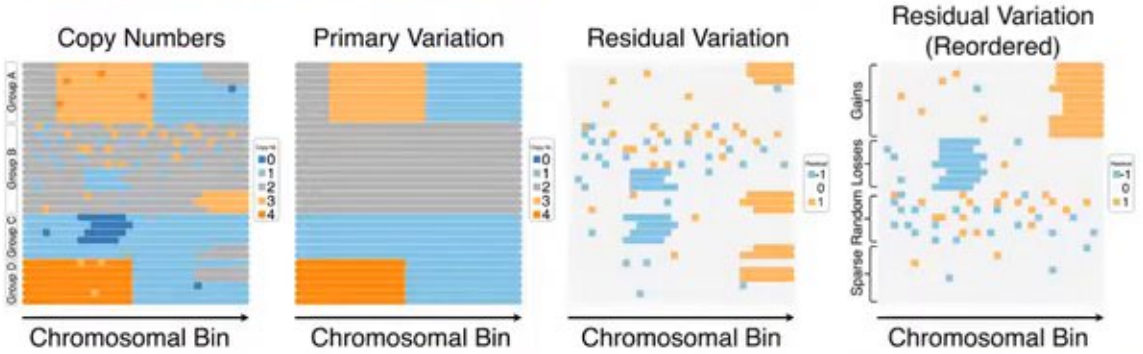


Method	Anterior Brain				Posterior Brain				Whole Brain				Perturb	Control	Perturb	Control
	R.all↓	R.200↓	ρ ↑	R.all↓	R.all↓	R.200↓	ρ ↑	R.all↓	R.all↓	R.200↓	ρ ↑	R.all↓				
Ours	1.374	1.296	0.301	1.379												
Ours-W	**1.352**	**1.244**	**0.340**	**1.367**	**1.288**	**0.302**	**1.376**	**1.322**	**0.275**	1.320	1.150	0.499	1.399	1.340	0.321	
GP	1.379	1.289	0.241	1.389	1.337	0.241	1.393	1.354	0.214	1.357	1.264	0.272	1.413	1.405	0.073	
MLP	1.399	1.369	0.319	1.404	1.393	0.283	1.406	1.397	0.267	1.347	1.174	0.491	1.403	1.357	0.314	

CN-SBM: Categorical Block Modelling For Primary and Residual Copy Number Variation

Kevin Lam[1,2], William Daniels[2], J Maxwell Douglas[2], Daniel Lai[2], Samuel Aparicio[2], Benjamin Bloem-Reddy[1], Yongjin Park[1,2]

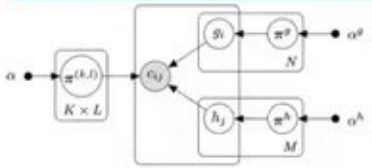
[1] Department of Statistics, University of British Columbia; [2] Department of Molecular Oncology, BC Cancer Research Centre



Copy Number Variation (CNV): alterations in the number of copies of DNA segments (amplifications and deletions)

- **Primary (main) variation**: large-scale, recurrent alterations
- **Residual variation**: finer sample-specific deviations

CN-SBM Model



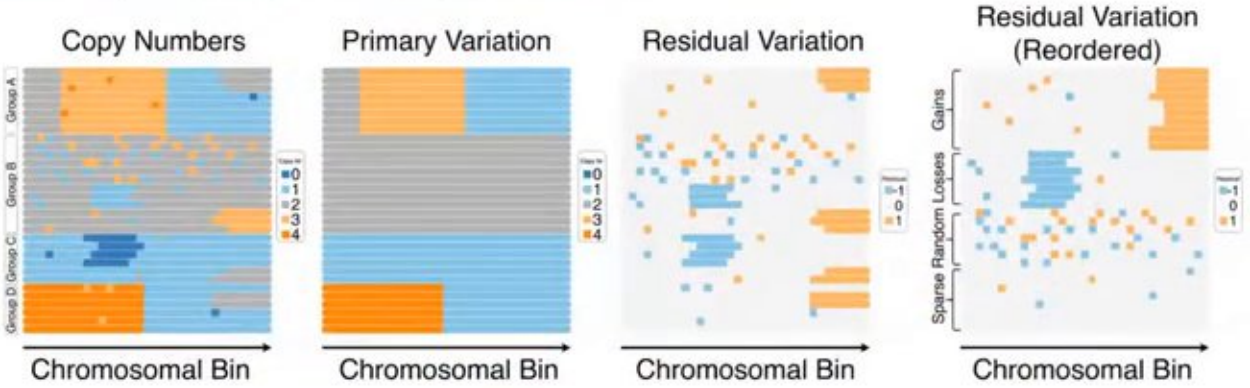
- Explicitly models categorical CN states ($c_{ij} \in \{0, 1, \ldots, 10, \geq 11\}$)
- Jointly clusters samples and genomic bins

THE UNIVERSITY OF BRITISH COLUMBIA BC CANCER

CN-SBM: Categorical Block Modelling For Primary and Residual Copy Number Variation

Kevin Lam[1,2], William Daniels[2], J Maxwell Douglas[2], Daniel Lai[2], Samuel Aparicio[2], Benjamin Bloem-Reddy[1], Yongjin Park[1,2]

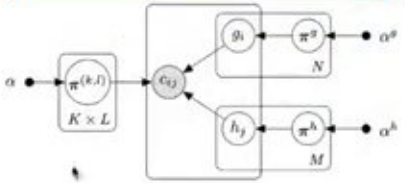
[1] Department of Statistics, University of British Columbia; [2] Department of Molecular Oncology, BC Cancer Research Centre



Copy Number Variation (CNV): alterations in the number of copies of DNA segments (amplifications and deletions)

- **Primary (main) variation**: large-scale, recurrent alterations
- **Residual variation:** finer sample-specific deviations

CN-SBM Model



- Explicitly models categorical CN states ($c_{ij} \in \{0, 1, \ldots, 10, \geq 11\}$)
- Jointly clusters samples and genomic bins

CN-SBM Detects Clinically Relevant Clusters in Low-Grade Glioma

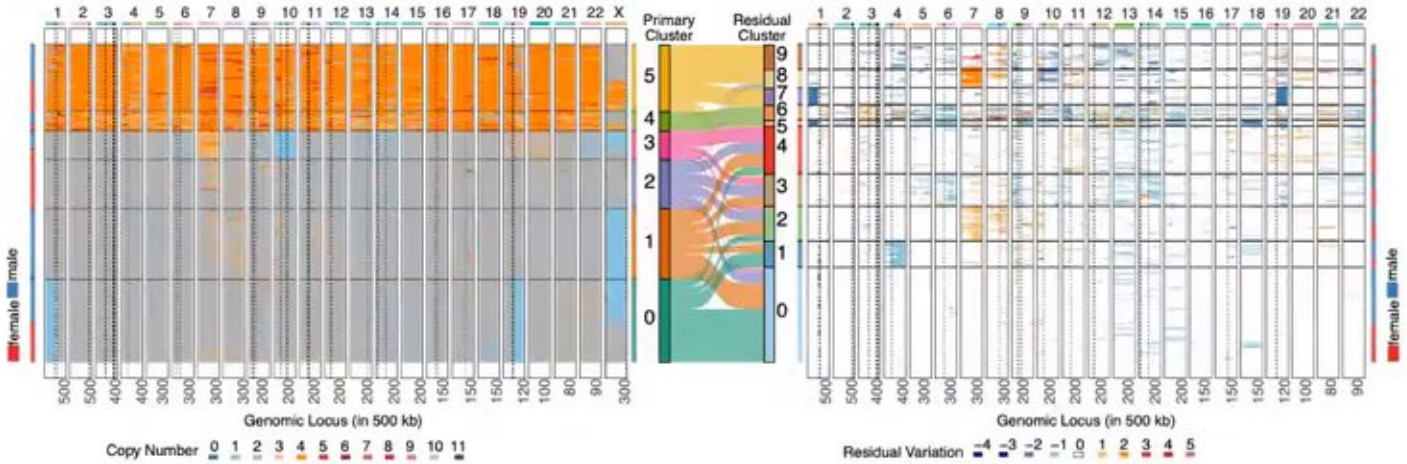
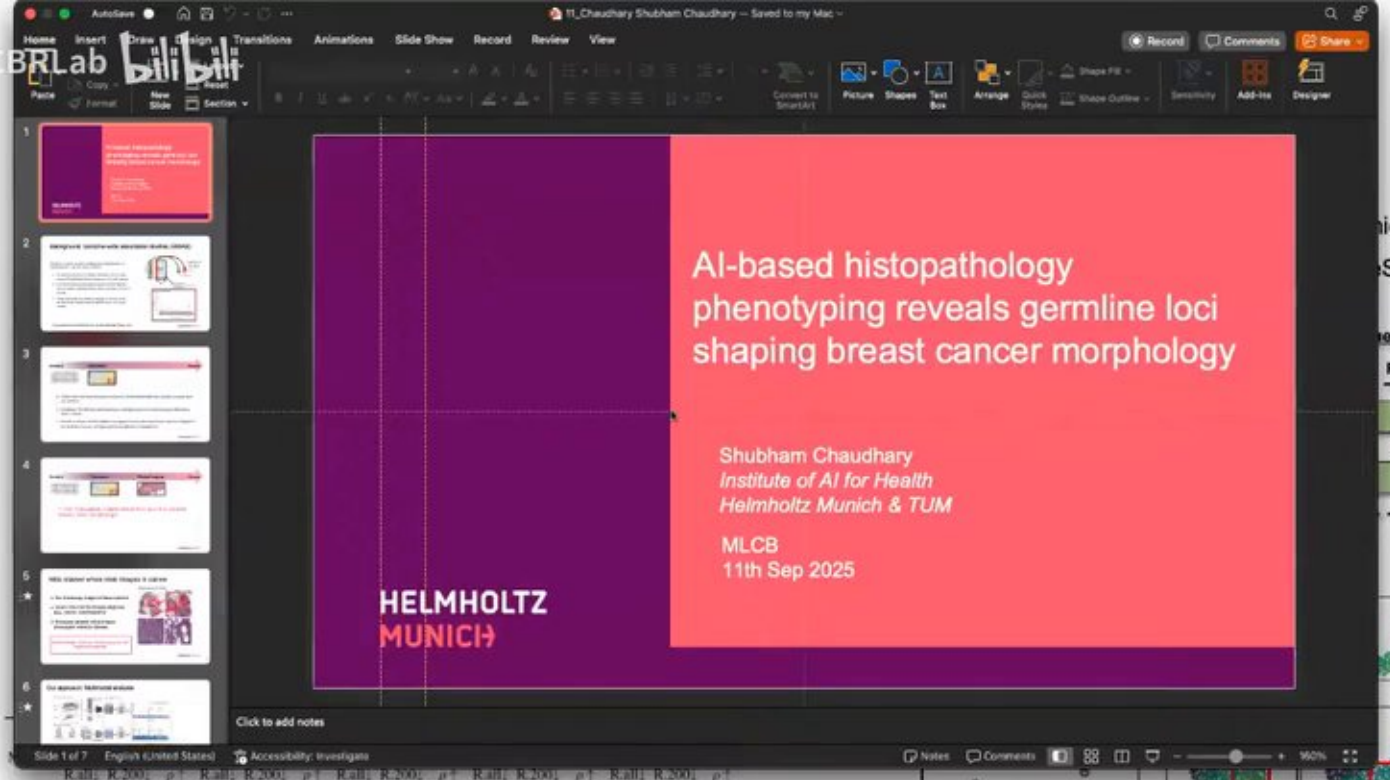


Figure: Primary and residual CNVs in low-grade glioma TCGA-cohort ($N = 490$)

- **Primary clusters**: arm- or whole-chromosome alterations (chr. 1, 7, 10, 19, X)
- **Residual clusters:** sparse, fine alterations



	R.all↓	R.200↓	ρ↑	R.all↓	R.200↓	ρ↑	R.all↓	R.200↓	ρ↑	R.all↓	R.200↓	ρ↑	R.all↓	R.200↓	ρ↑
Ours	1.374	1.296	0.301	1.379	1.319	0.265	1.386	1.340	0.242	**1.303**	**1.126**	**0.540**	1.405	1.357	**0.326**
Ours-W	**1.352**	**1.244**	**0.340**	**1.367**	**1.288**	**0.302**	**1.376**	**1.322**	**0.275**	1.320	1.150	0.499	**1.399**	**1.340**	0.323
GP	1.379	1.289	0.241	1.389	1.337	0.241	1.393	1.354	0.214	1.357	1.264	0.272	1.413	1.405	0.073
MLP	1.399	1.369	0.319	1.404	1.393	0.283	1.406	1.397	0.267	1.347	1.174	0.491	1.403	1.357	0.314

MCBRLab bilibili

AI-based histopathology phenotyping reveals germline loci shaping breast cancer morphology

Shubham Chaudhary
Institute of AI for Health
Helmholtz Munich & TUM

MLCB
11th Sep 2025

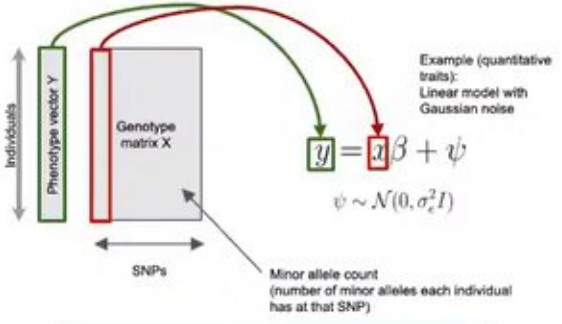
HELMHOLTZ MUNICH

zoom

Background: Genome-wide association studies (GWAS)

GWAS is a method used to identify genetic variants that are associated with a specific trait or disease.

- The study compares the genetic variations across large groups of individuals with the presence of a trait / disease
- A statistical testing procedure is used to identify specific genetic variants (typically SNPs) associated with the trait or disease
- These SNPs are then used as markers to identify other genetic variations that may be contributing to the trait or disease



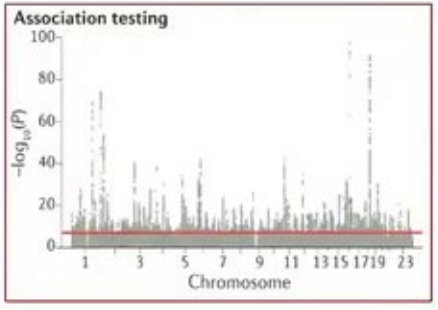
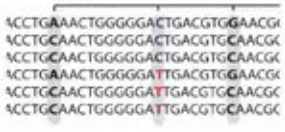


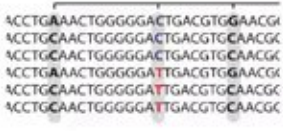
Figure adapted from *Uffelmann et al, Nat Rev Methods Primers, 2021*

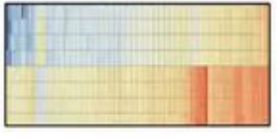
Genetics **Expression** **Disease**

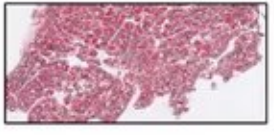


- GWAS have identified thousands of genetic variants associated with human complex traits and disease
- Challenge: We still lack understanding of biological processes and pathways affected by these variants
- Genetic analyses of intermediate phenotypes such as gene expression can help mitigate this by identifying tissues, cell types and genes affected by disease loci

Genetics → Expression → Medical imaging → Disease



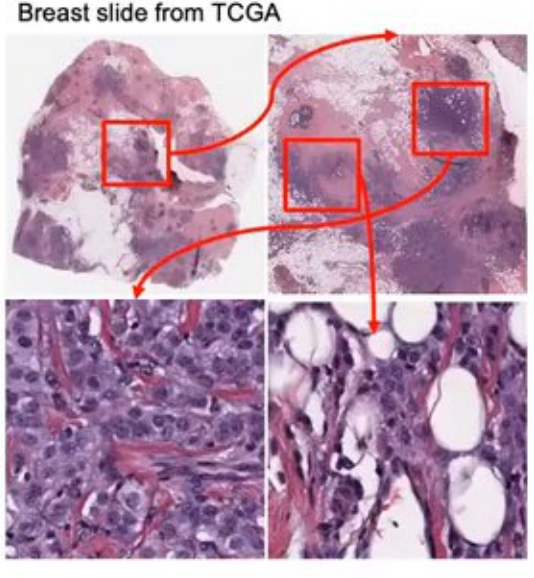




↦ Can histological images reveal how germline variants shape tumor morphology?

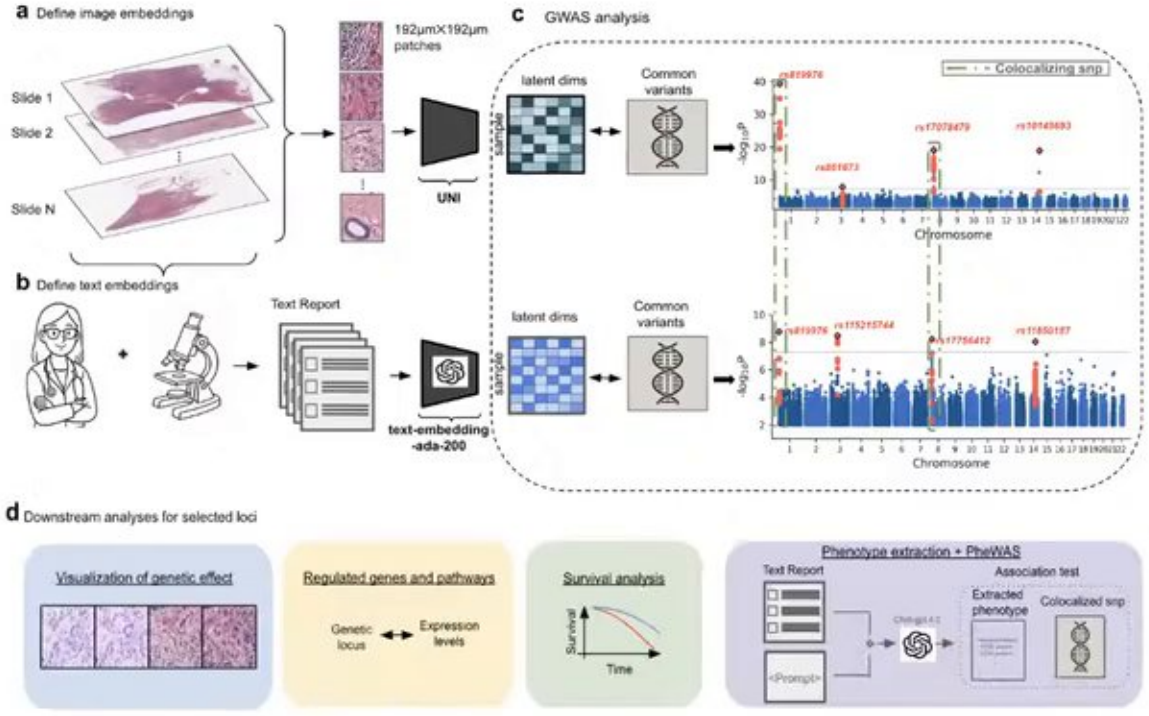
H&E-stained whole slide images in cancer

- Are microscopy images of tissue sections
- Used in the clinic for disease diagnosis (e.g., cancer, steatohepatitis)
- Showcase detailed cell and tissue phenotypes related to disease

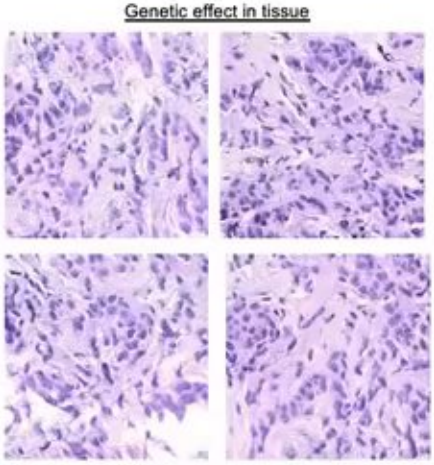


Our approach: Multimodal analysis

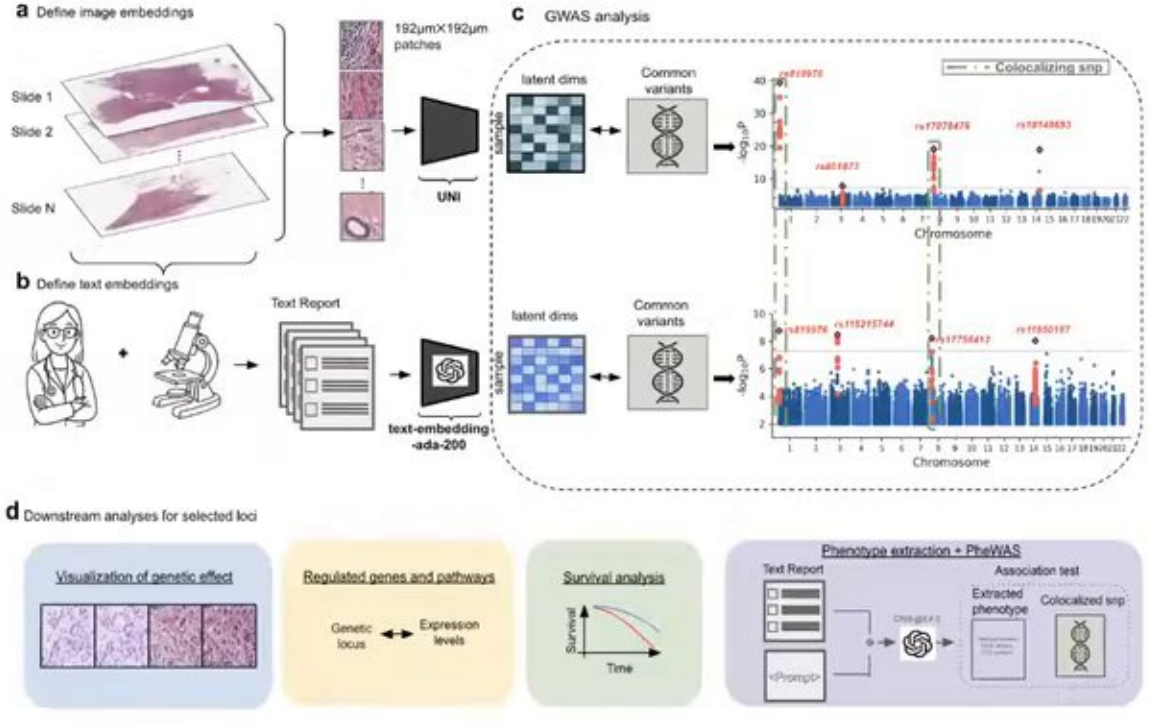
Our approach: Multimodal analysis



Characterization of genetic effect on tissue Histology

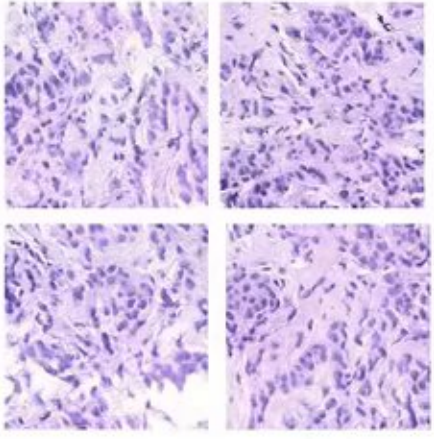


Our approach: Multimodal analysis



Characterization of genetic effect on tissue Histology

Genetic effect in tissue



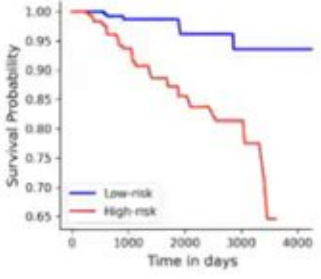
Regulated genes and pathways

MSigDB Term	P value	Adjusted P-value	genes
G2-M Checkpoint	1.1e−4	0.013	PLK1;E2F1;CDC6
Mitotic Spindle	9.4e−4	0.05	*PLK1;SAC3D1*
E2F Targets	1.5e−3	0.06	*PLK1;TK1*
Myc Targets V2	3.0e−2	0.102	*PLK1*
DNA Repair	4.5e−2	0.108	*SAC3D1*

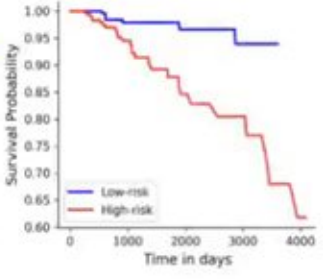
Genetic locus ⟷ Expression levels

Survival analysis

(i) KM curve (Gene:MIR205HG)



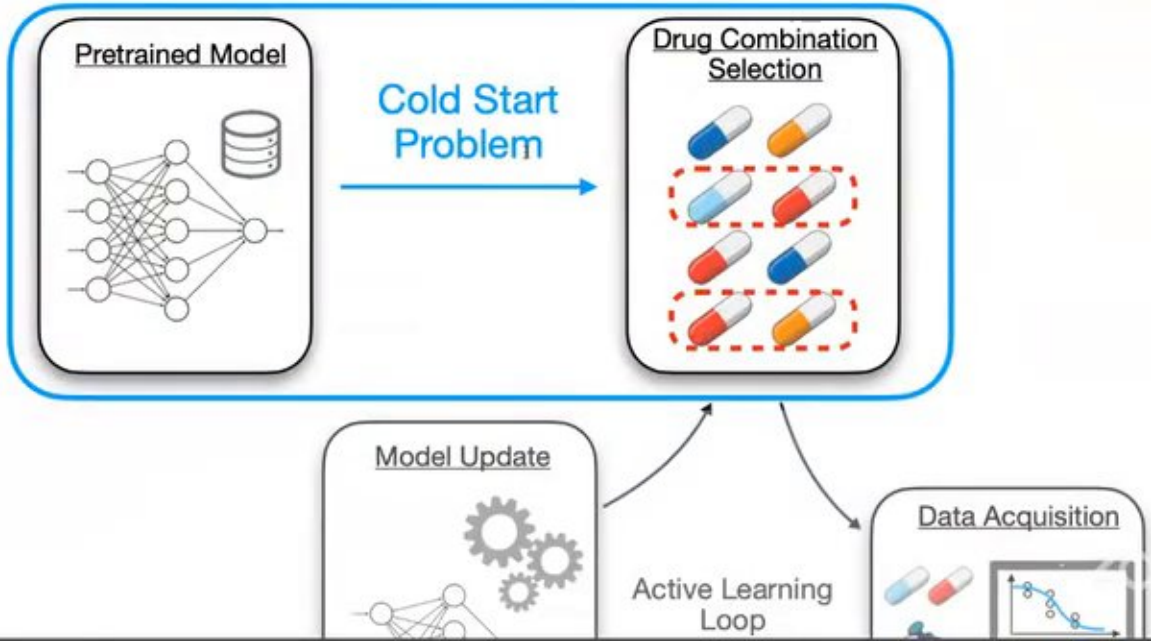
(ii) KM curve (Gene:C4BPA)



Addressing the Cold-Start Problem for Personalized Combination Drug Screening

Antoine de Mathelin, Christopher Tosh and Wesley Tansey

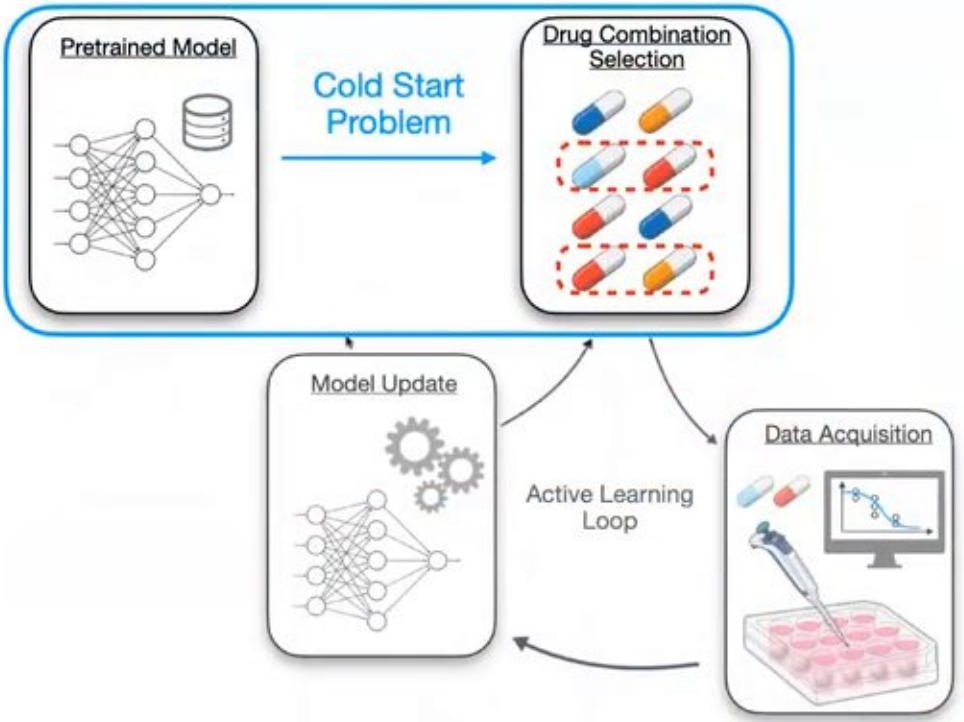
Memorial Sloan Kettering Cancer Center



Addressing the Cold-Start Problem for Personalized Combination Drug Screening

Antoine de Mathelin, Christopher Tosh and Wesley Tansey

Memorial Sloan Kettering Cancer Center

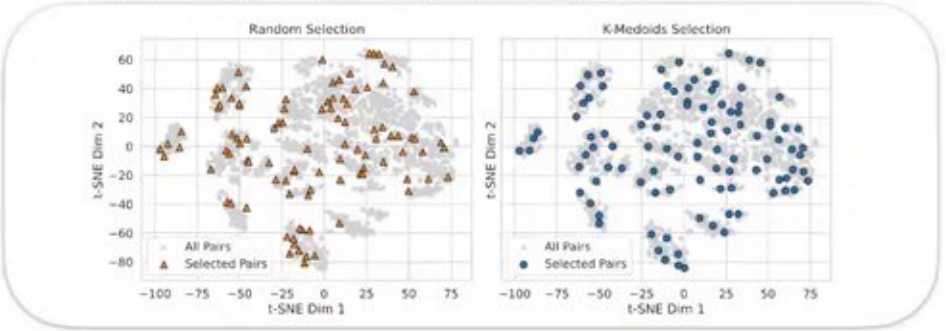


Addressing the Cold-Start Problem for Personalized Combination Drug Screening

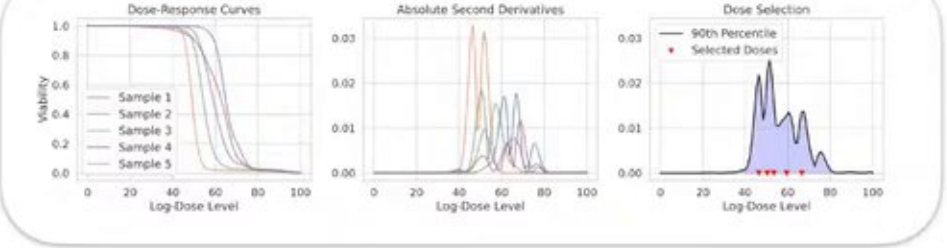
Antoine de Mathelin, Christopher Tosh and Wesley Tansey

Memorial Sloan Kettering Cancer Center

Drug Pair Selection: *K-medoids in AUC Space*



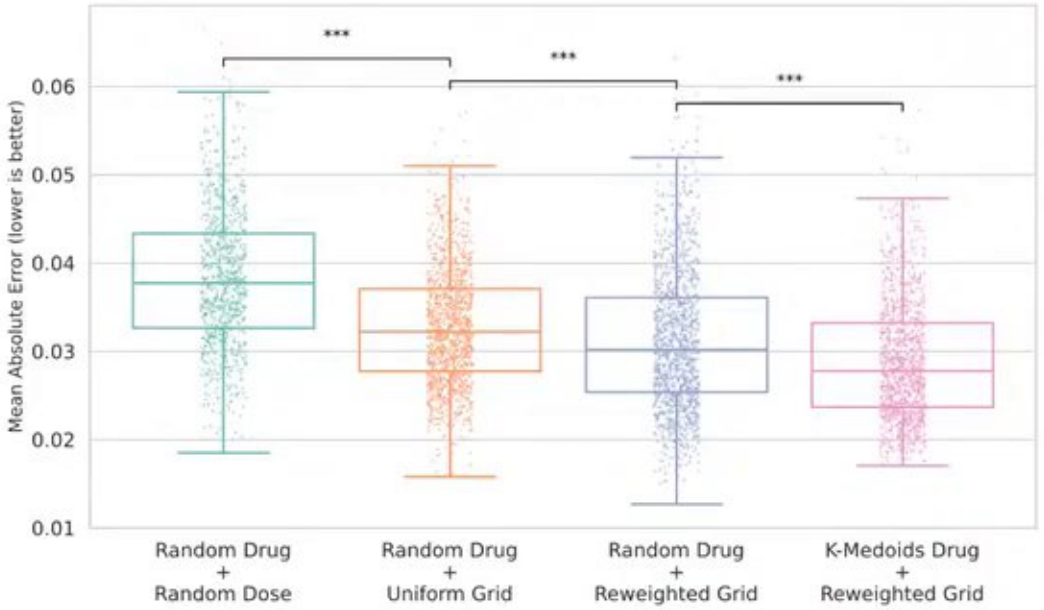
Dose Selection: *Curvature-based Importance Sampling*



Addressing the Cold-Start Problem for Personalized Combination Drug Screening

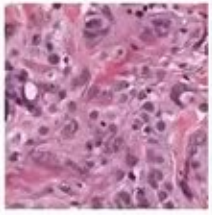
Antoine de Mathelin, Christopher Tosh and Wesley Tansey

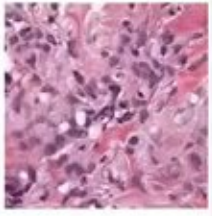
Memorial Sloan Kettering Cancer Center



Characterization of genetic effect on tissue Histology

Genetic effect



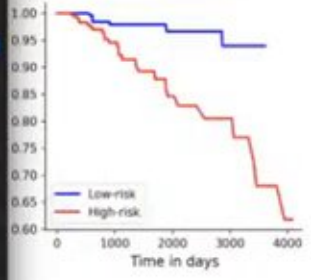


"If you're i

MLCB2025 spotlight slides

Name	Date Modified	Size	Kind
1_Schreiber Jacob Schreiber.pptx	Sep 9, 2025 at 3:10 PM	1.7 MB	PowerP...n (.pptx)
2_Seitz Evan Seitz.pdf	Sep 9, 2025 at 8:47 PM	2.6 MB	PDF Document
3_Nair Suraj Nair.pdf	Sep 8, 2025 at 2:44 PM	2.2 MB	PDF Document
4_Talk_mahbuba Mahbuba Tasmin.pptx	Sep 9, 2025 at 5:00 PM	4.2 MB	PowerP...n (.pptx)
5_Ayub Shanza Ayub.pptx	Sep 9, 2025 at 11:03 AM	9.8 MB	PowerP...n (.pptx)
6_avdeyev Pavel Avdeev.pdf	Yesterday at 12:23 AM	586 KB	PDF Document
7_Liu Simon Liu.pptx	Sep 9, 2025 at 9:08 PM	9 MB	PowerP...n (.pptx)
8_Su JY Su.pptx	Today at 1:04 AM	17.2 MB	PowerP...n (.pptx)
9_Minsheng&Tianyu Tianyu Liu.pptx	Sep 8, 2025 at 10:09 AM	1.7 MB	PowerP...n (.pptx)
10_Lam Kevin Lam.pdf	Sep 9, 2025 at 6:19 PM	1.5 MB	PDF Document
11_Chaudhary Shubham Chaudhary.pptx	Sep 9, 2025 at 4:22 PM	21 MB	PowerP...n (.pptx)
12_Dematheln Antoine de Mathelin.pdf	Sep 9, 2025 at 2:15 PM	3.7 MB	PDF Document
14_Sergio Mares.key	Yesterday at 12:50 AM	1.7 MB	Keynote
15_Dalal Taykhoom Dalal.key	Sep 9, 2025 at 8:48 PM	5.1 MB	Keynote
MLCB2025_Mares Sergio Mares.pptx	Yesterday at 12:36 AM	356 KB	PowerP...n (.pptx)
MLCB2025_Mares_Final Sergio Mares.pptx	Yesterday at 12:55 AM	1.8 MB	PowerP...n (.pptx)

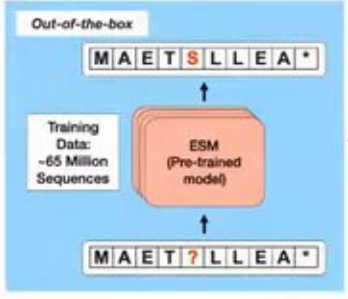
(ii) KM curve (Gene:C4BPA)



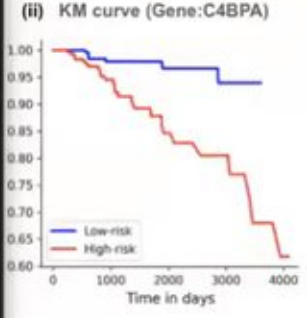
HELMHOLTZ MUNICH

Characterization of genetic effect on tissue Histology

We apply additional domain pre-training to test whether this improves predictions



Training Data for additional pre-training: curated database of MHC I binder sequences



"If you're i

HELMHOLTZ MUNICH

It's infeasible to test all neoantigens experimentally—we need accurate predictions for prioritization

MCBRLab bilibili

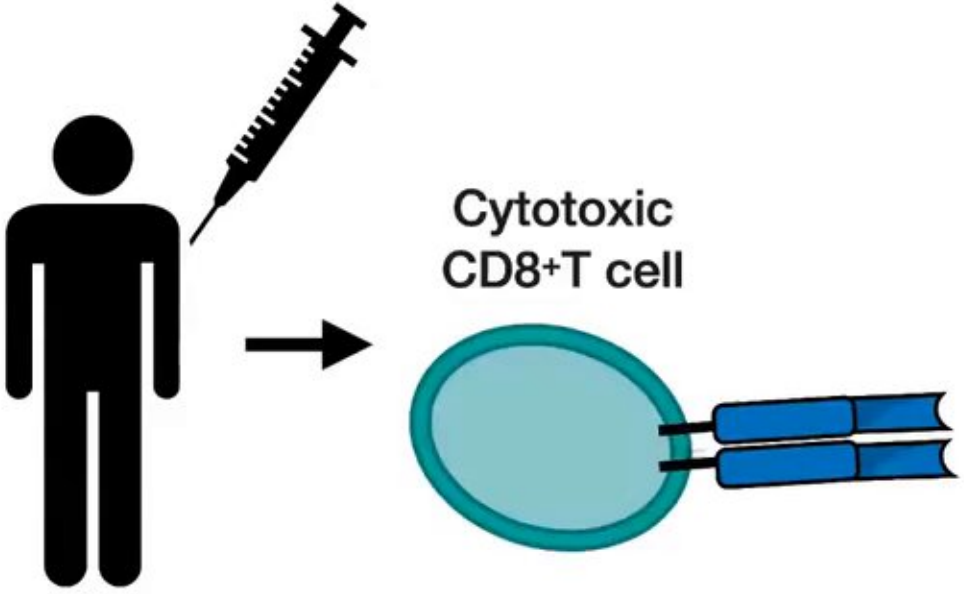
Continued domain-specific pre-training of protein language models for pMHC-I binding prediction

Sergio E. Mares
PhD Student
Center for Computational Biology

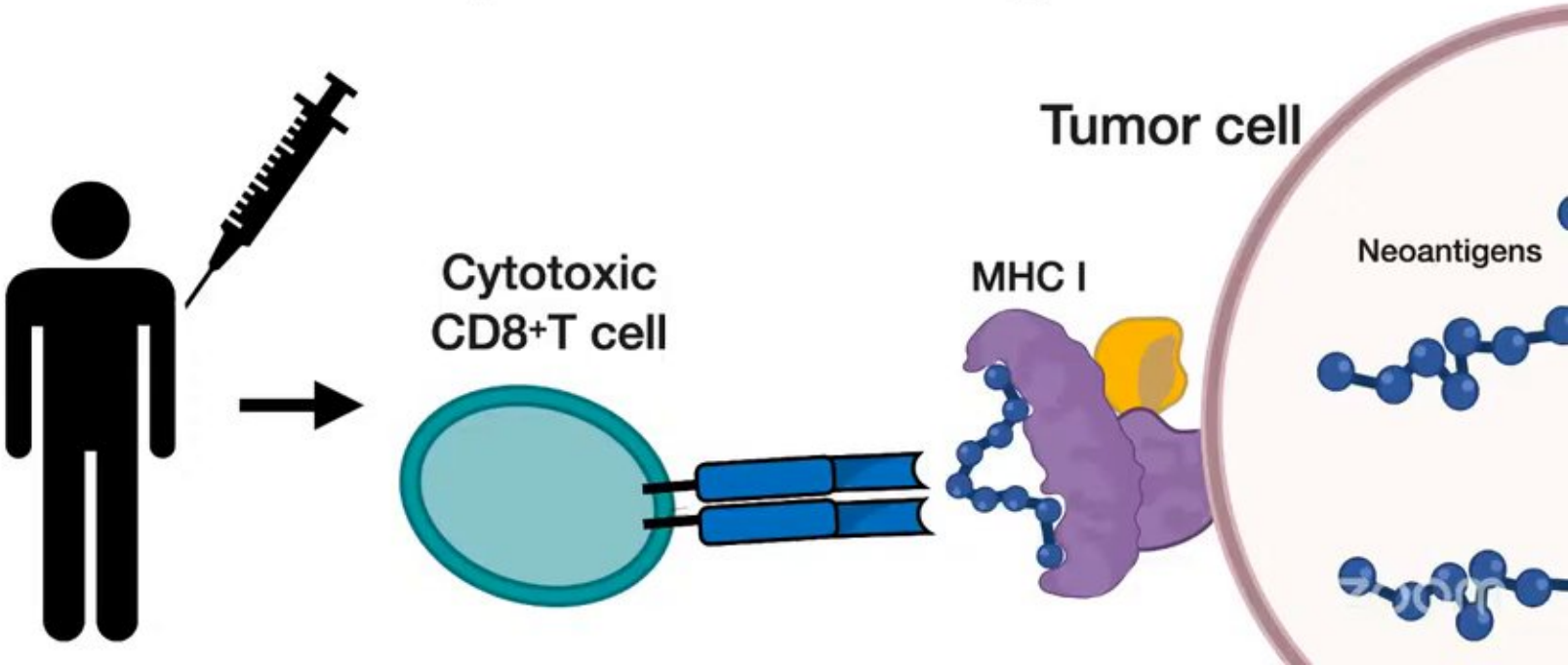
Berkeley
UNIVERSITY OF CALIFORNIA

Vaccines stimulate the immune system to respond to neoantigens

Vaccines stimulate the immune system to respond to neoantigens

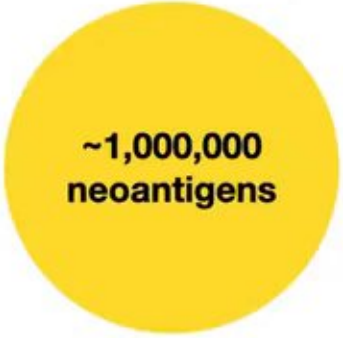


Vaccines stimulate the immune system to respond to neoantigens



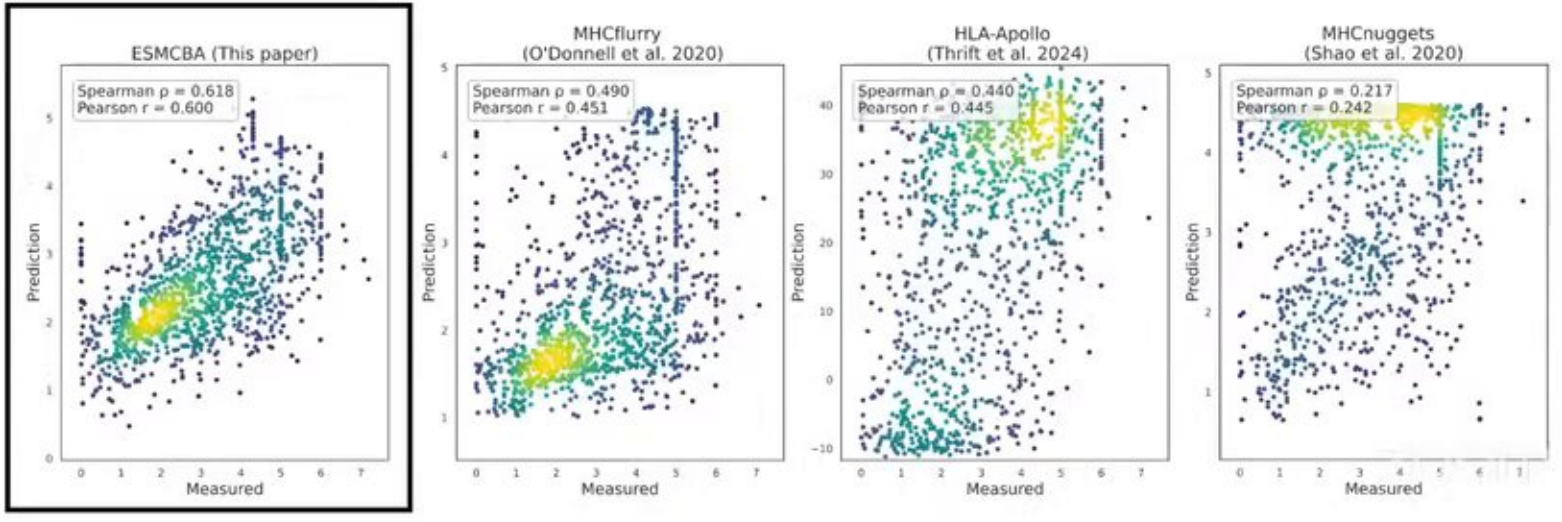
It's infeasible to test all neoantigens experimentally—we need accurate predictions for prioritization

The Cancer Genome Atlas



We outperform SOTA models in binding affinity prediction leveraging pre-trained language models

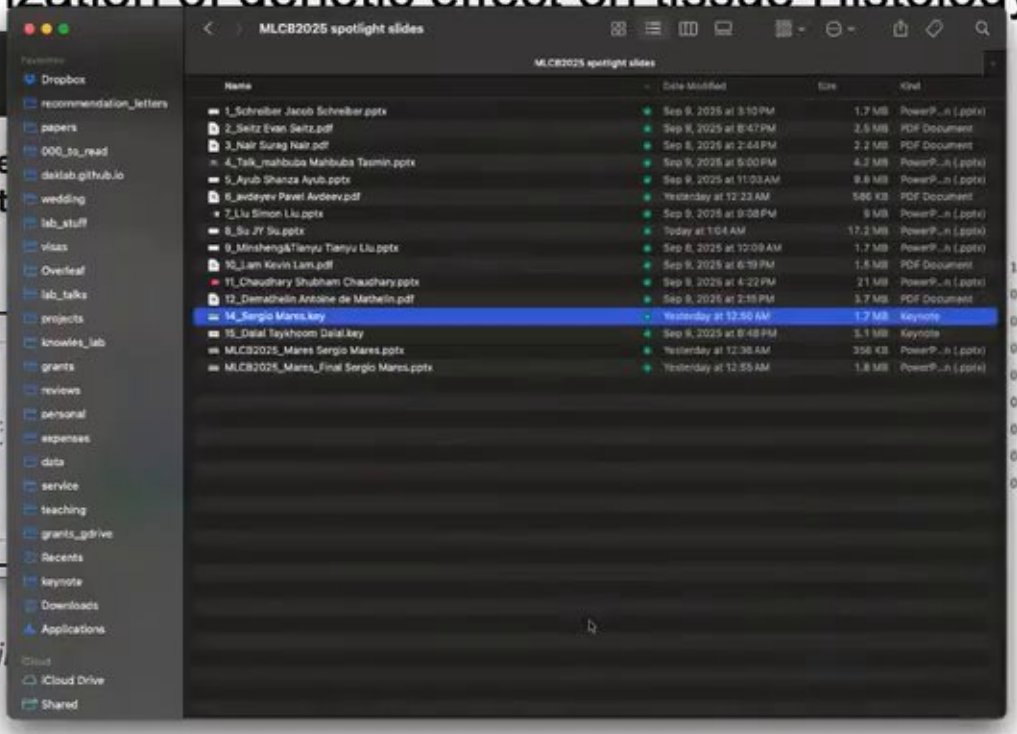
Poster #14

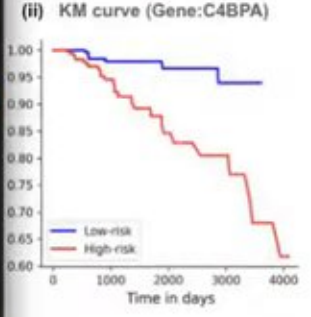


Finder File Edit View Go Window Help

Thu Sep 11 12:42 PM

Characterization of genetic effect on tissue Histology



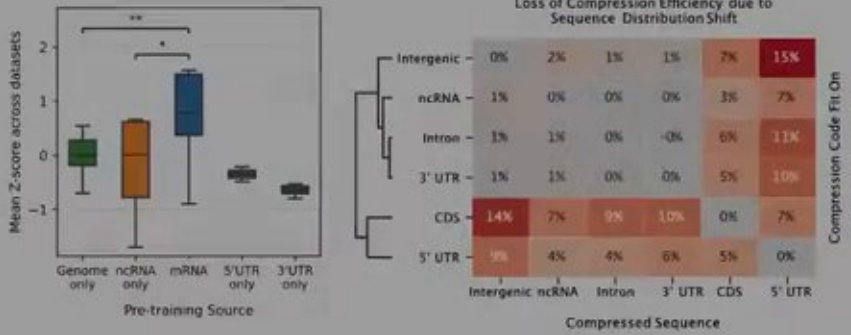


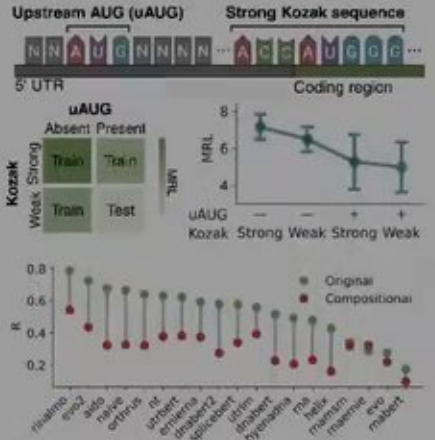
"If you're i

HELMHOLTZ MUNICH

MCBRLab bilibili

What else did we do?



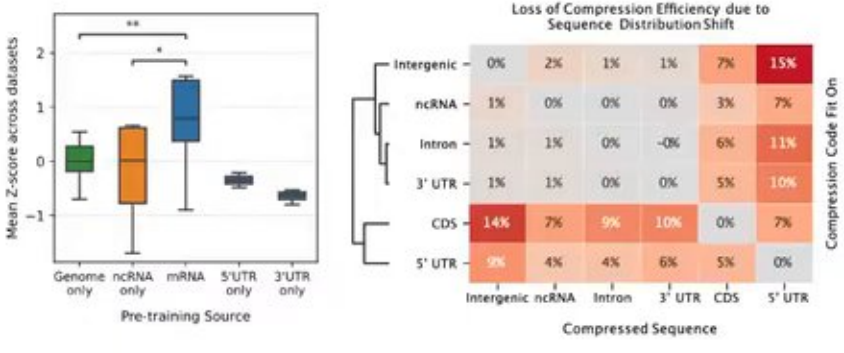


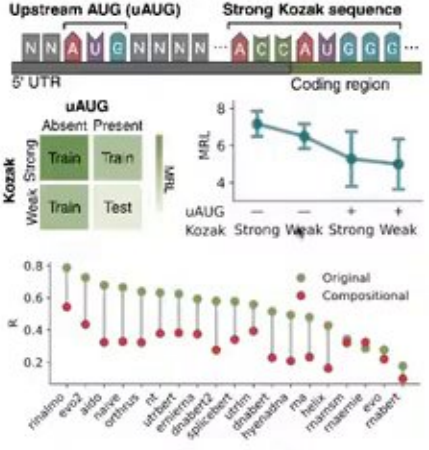
→ Quantified why DNA / ncRNA models perform poorly at mRNA tasks

→ Assessed if nucleotide FMs can generalize to novel combinations of familiar sequence features

zoom

What else did we do?





→ **Quantified why DNA / ncRNA models perform poorly at mRNA tasks**

→ **Assessed if nucleotide FMs can generalize to novel combinations of familiar sequence features**

mRNABench: A curated benchmark for mature mRNA property and function prediction

Ruian Shi*, **Taykhoom Dalal***, **Philip Fradkin***, Divya Koyyalagunta, Simran Chhabria, Andrew Jung, Cyrus Tam, Defne Ceyhan, Jessica Lin, Kaitlin U. Laverty, Ilyes Baali, Bo Wang, Quaid Morris

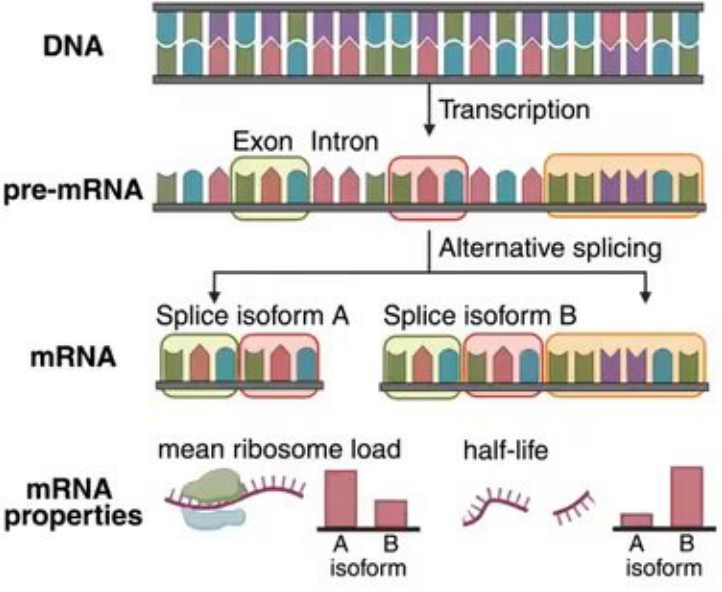
Machine Learning in Computational Biology 2025

Memorial Sloan Kettering Cancer Center

UNIVERSITY OF TORONTO

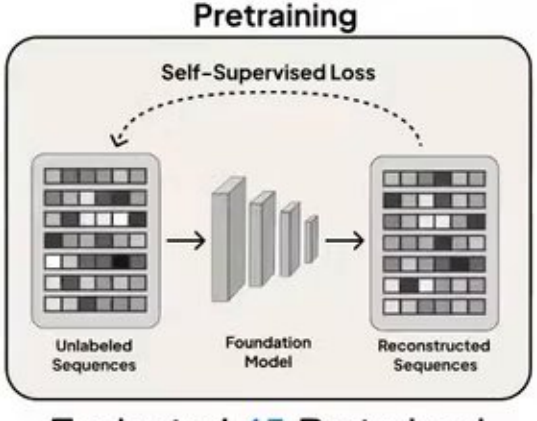
VECTOR INSTITUTE

Motivation

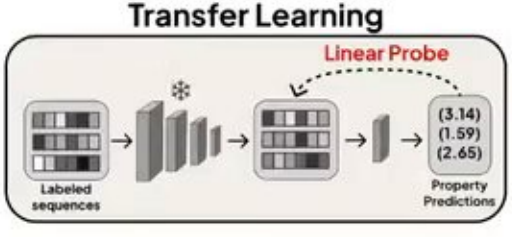


- Alternative splicing can generate multiple splice isoforms with distinct properties and functions
- Self-supervised nucleotide FMs can be used to predict these properties but there are no existing mRNA-focused benchmarks

mRNABench is the first mRNA-specific nucleotide FM benchmarking suite



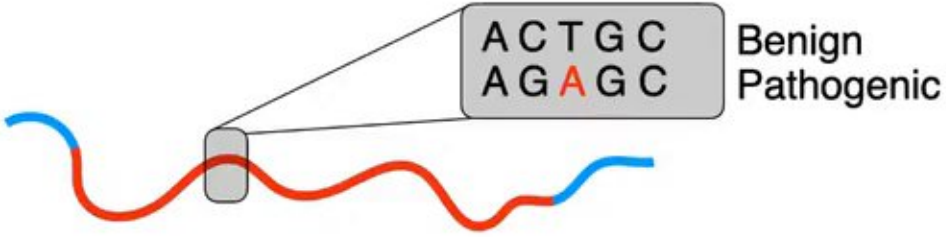
Evaluated 45 Pretrained Foundation Models



Variant effect prediction
RBP-mRNA binding
mRNA localization
Half-life
Mean ribosome load
Protein localization
Gene-ontology classification

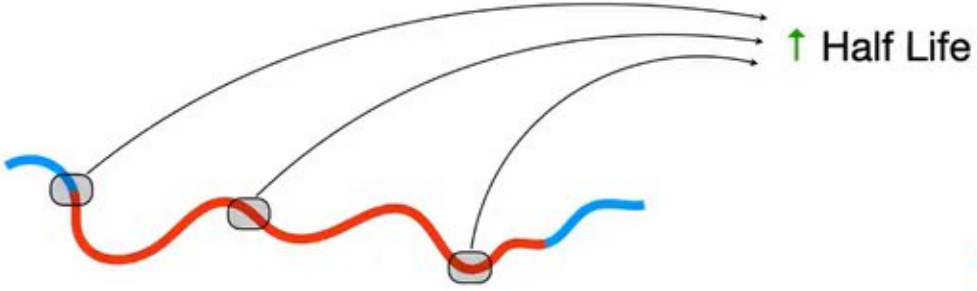
59 total subtasks

We categorized tasks in our benchmark as either being "local" or "global"



Local

Variant effect prediction
RBP-mRNA binding
Mean Ribosome Load (MPRA)



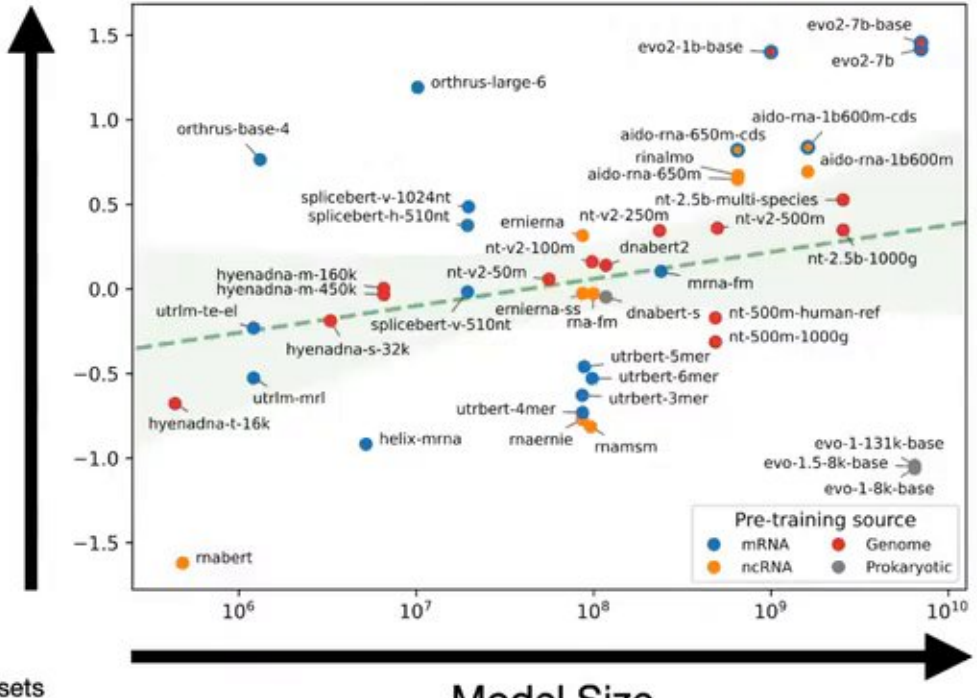
Global

mRNA localization
Half-life
Mean ribosome load
Protein localization
Gene-ontology classification

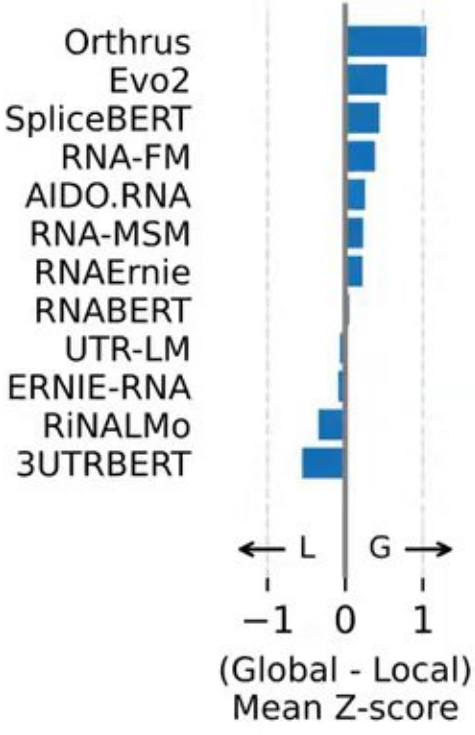
* UTR CDS

The top performing models are Evo2 and Orthrus

*Average Performance Across Tasks

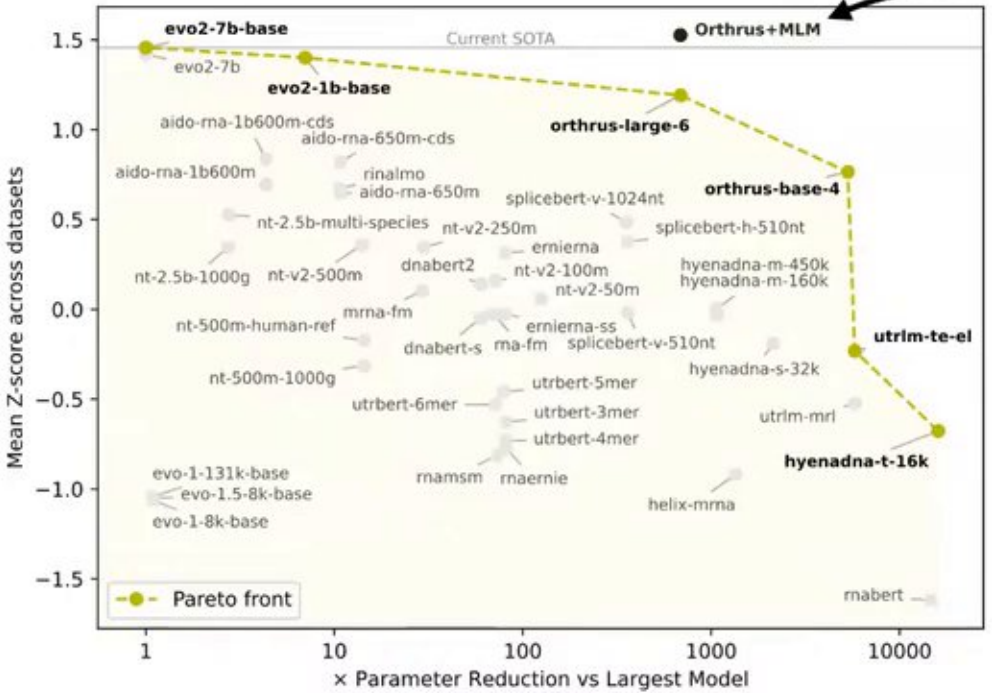


*Mean Z-score across datasets



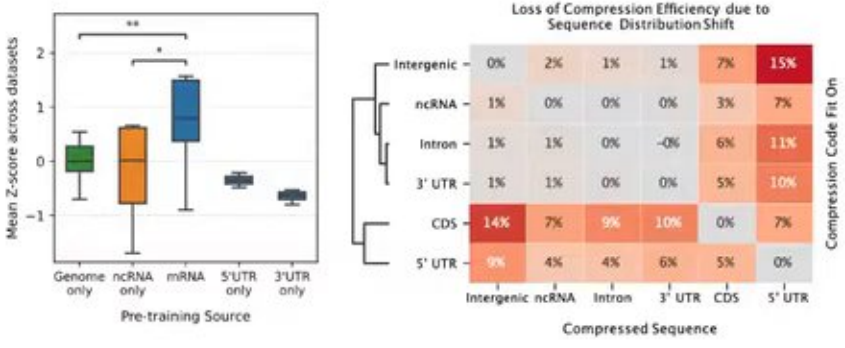
Orthrus does much better on "global" tasks than "local" ones due to its contrastive learning objective

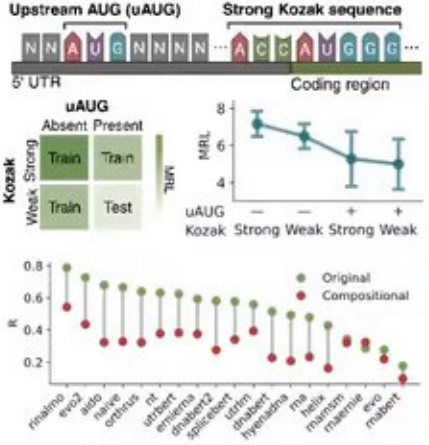
What did we do with this information?



Trained a SOTA Orthrus variant using a joint contrastive learning and masked language modeling objective

What else did we do?





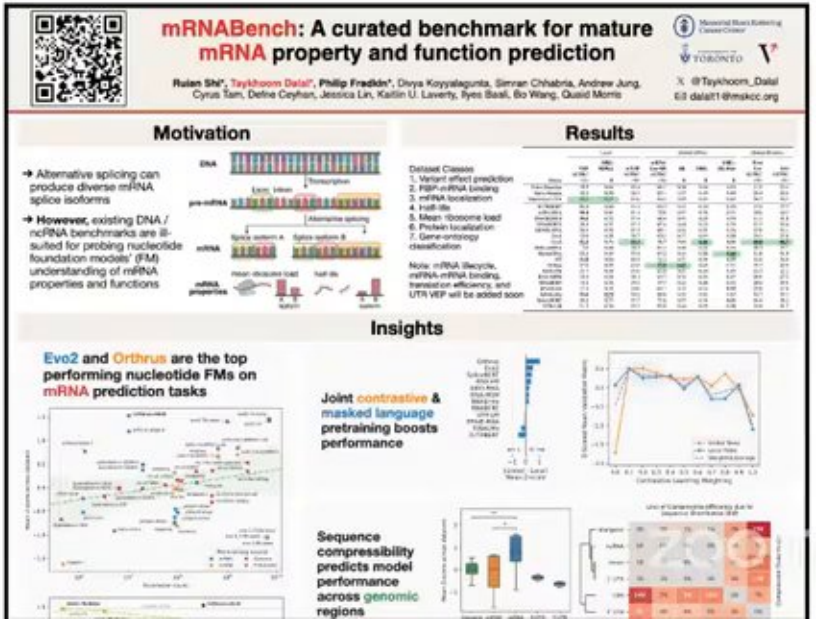
→ **Quantified why DNA / ncRNA models perform poorly at mRNA tasks**

→ **Assessed if nucleotide FMs can generalize to novel combinations of familiar sequence features**

If any of this sounds interesting…

Preprint

Poster 15



MCBRLab bilibili

NYGC Events

zoom

@

Thank you to our sponsors

And helpers!

Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang

Restarting at 1.30pm ET

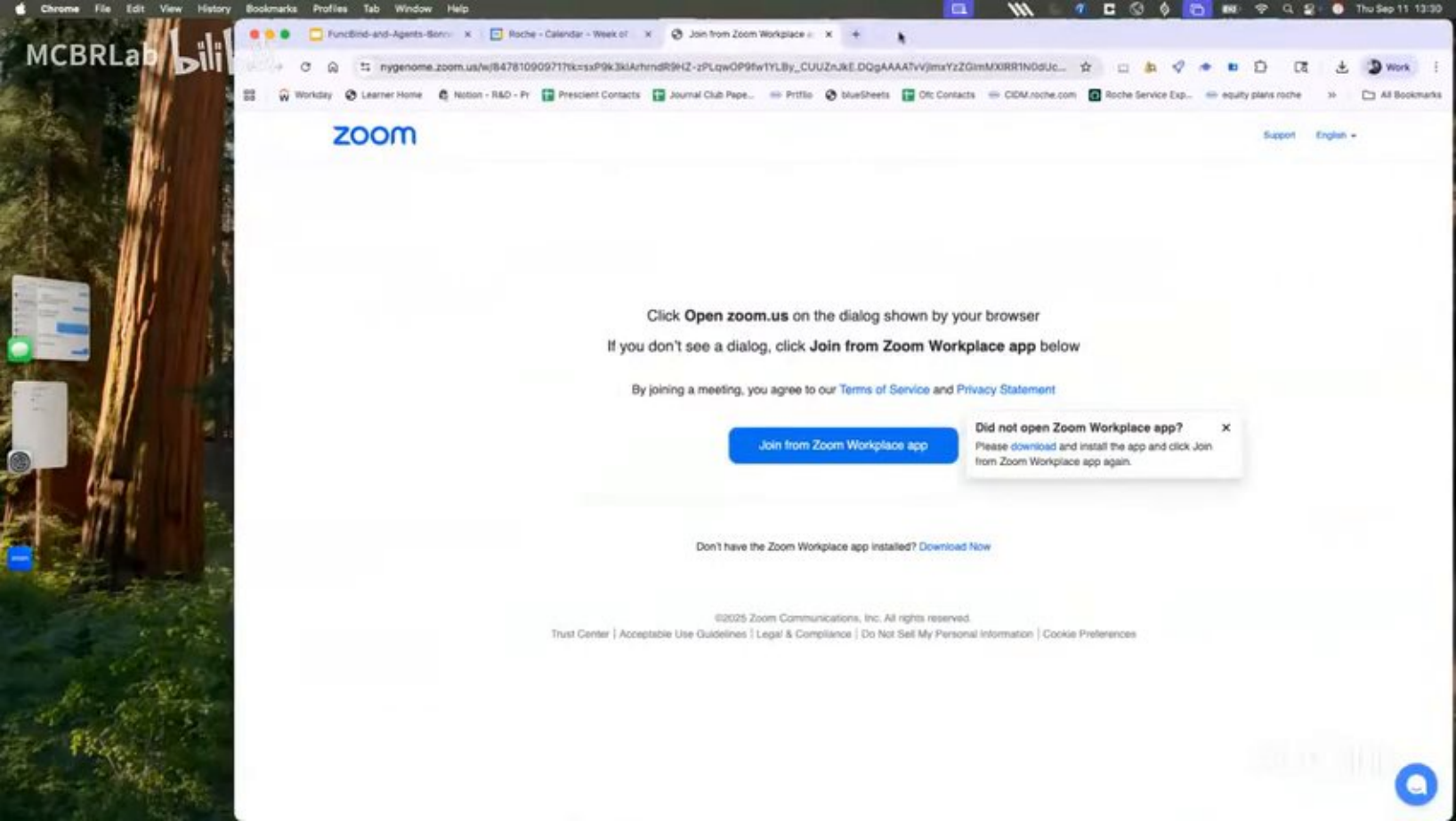
mlcb.org for schedule

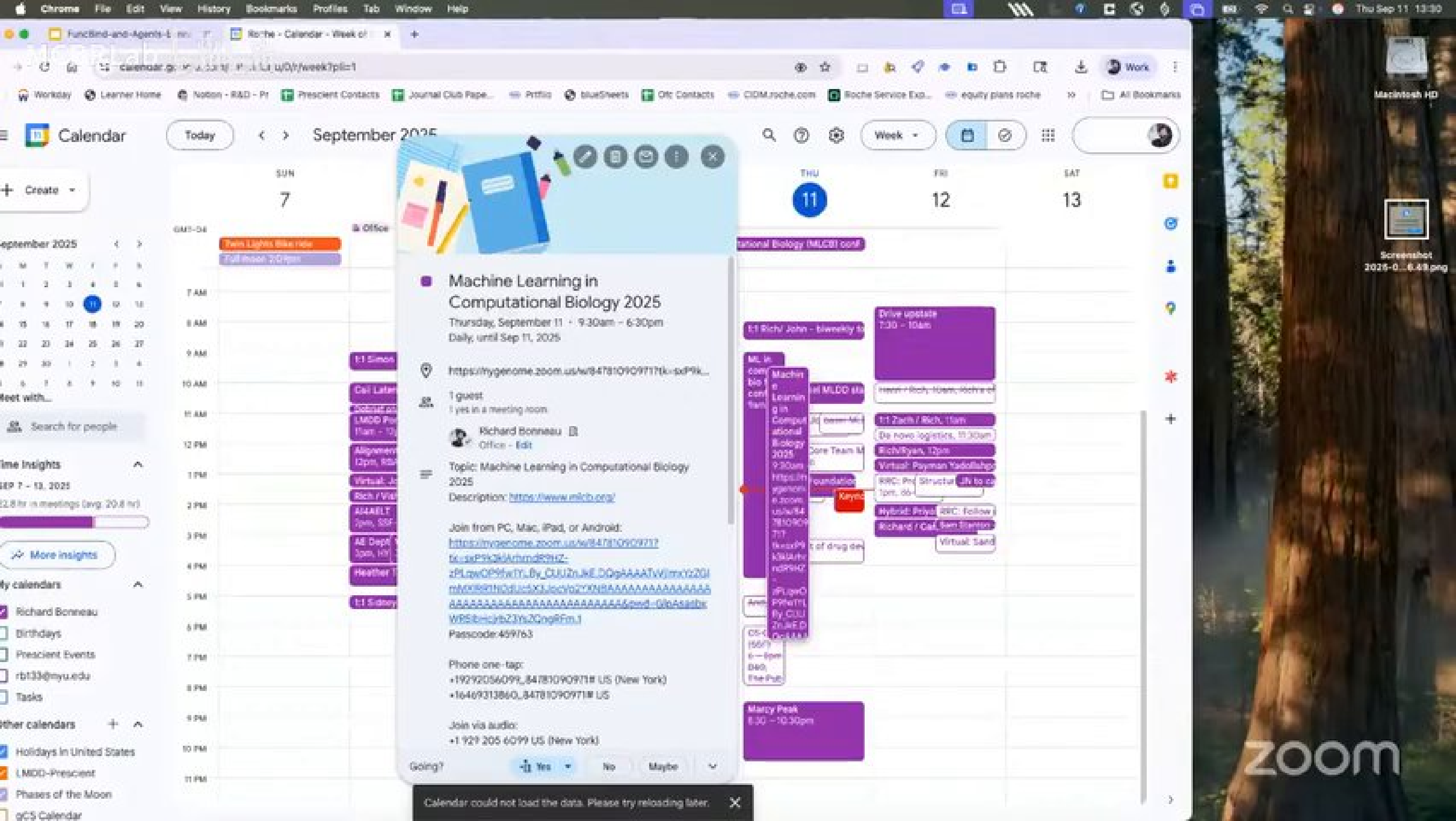
MCBRLab bilibili

NYGC Events

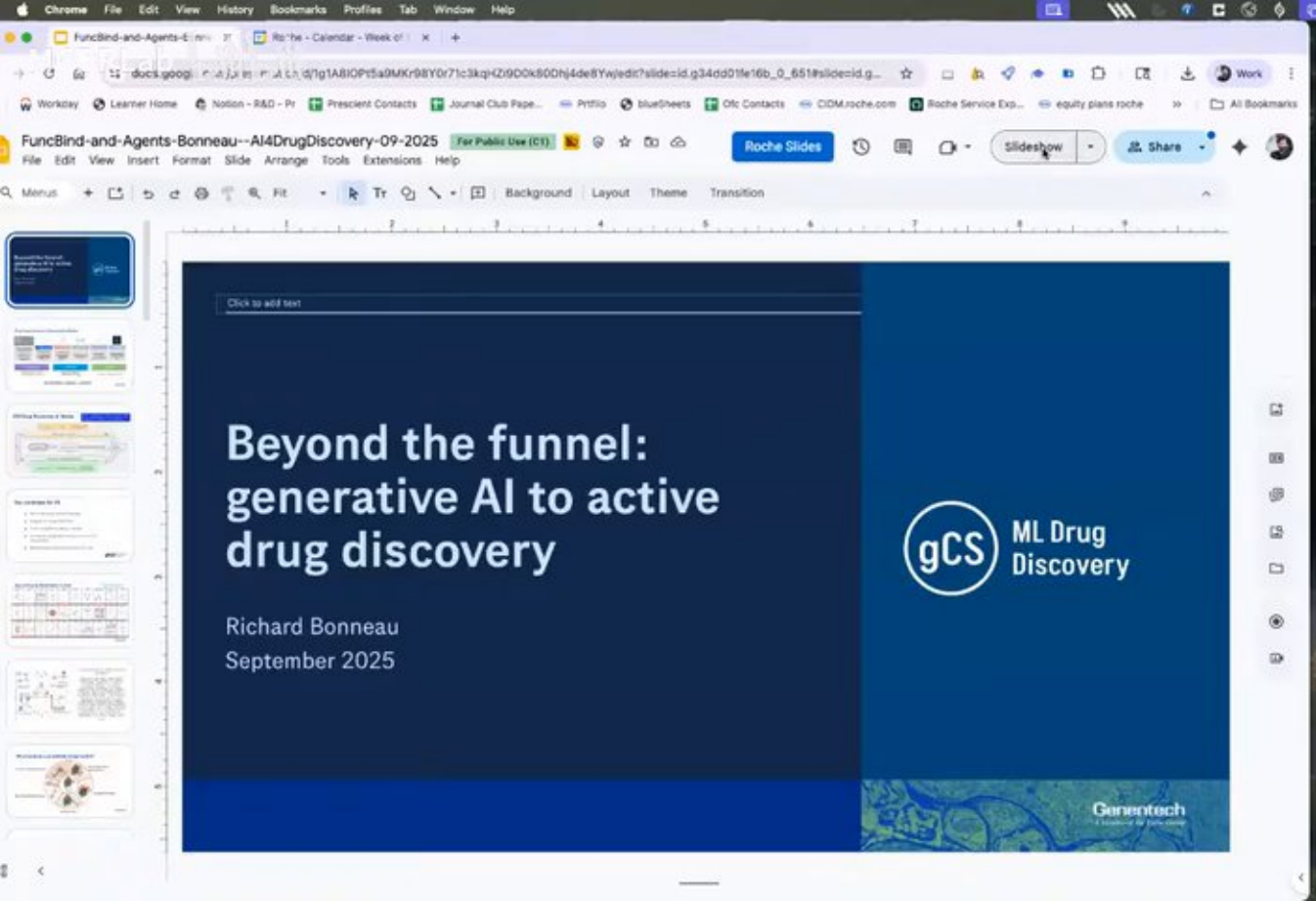
zoom


Zoom Workplace
Meeting
View
Edit
Window
Help
Thu Sep 11 13:30
MCBRLab
bilibili
Macintosh HD
Screenshot
2025-0...6.49.png
"zoom.us" is requesting to bypass the system private window picker and directly access your screen and audio.
This will allow zoom.us to record your screen and system audio, including personal or sensitive information that may be visible or audible.
Allow
Open System Settings
Deny
zoom


MCBRLab
Chrome File Edit View History Bookmarks Profiles Tab Window Help
Thu Sep 11 13:30
Join from Zoom Workplace
nygenome.zoom.us/wc/84781090971?fk=sxP9k3klArhrndR9HZ-zPLqwOP9twTYLBy_CUUZnJkE.DQgAAAATvVjlmxYzZGlmMXlRR1NDdUc...
Work
Workday
Learner Home
Notion – R&D – Pr
Prescient Contacts
Journal Club Pape...
Prtflio
blueSheets
Ofc Contacts
CIDM.roche.com
Roche Service Exp...
equity plans roche
All Bookmarks
zoom
Support
English
Click Open zoom.us on the dialog shown by your browser
If you don't see a dialog, click Join from Zoom Workplace app below
By joining a meeting, you agree to our Terms of Service and Privacy Statement
Join from Zoom Workplace app
Did not open Zoom Workplace app?
Please download and install the app and click Join from Zoom Workplace app again.
Don't have the Zoom Workplace app installed? Download Now
©2025 Zoom Communications, Inc. All rights reserved.
Trust Center | Acceptable Use Guidelines | Legal & Compliance | Do Not Sell My Personal Information | Cookie Preferences



Chrome File Edit View History Bookmarks Profiles Tab Window Help
Thu Sep 11 13:30
Calendar
Today
September 2025
Week
SUN 7
THU 11
FRI 12
SAT 13
Create
September 2025
Meet with...
Search for people
Time Insights
SEP 7 – 13, 2025
More insights
My calendars
Richard Bonneau
Birthdays
Prescient Events
rb133@nyu.edu
Tasks
Other calendars
Holidays in United States
LMDD-Prescient
Phases of the Moon
gCS Calendar
Twin Lights Bike ride
Machine Learning in Computational Biology 2025
Thursday, September 11 · 9:30am – 6:30pm
Daily, until Sep 11, 2025
1 guest
1 yes in a meeting room
Richard Bonneau
Office · Edit
Topic: Machine Learning in Computational Biology 2025
Description: https://www.mlcb.org/
Join from PC, Mac, iPad, or Android:
Passcode:409763
Phone one-tap:
Join via audio:
+1 929 205 6099 US (New York)
Going?
Yes
No
Maybe
Calendar could not load the data. Please try reloading later.
Drive upstate
7:30 – 10am
Marcy Peak
Macintosh HD
zoom


Chrome File Edit View History Bookmarks Profiles Tab Window Help
FuncBind-and-Agents-Bonneau--AI4DrugDiscovery-09-2025
For Public Use (C1)
File Edit View Insert Format Slide Arrange Tools Extensions Help
Roche Slides
Slideshow
Share
Menus
Background
Layout
Theme
Transition
Click to add text
Beyond the funnel: generative AI to active drug discovery
Richard Bonneau
September 2025
gCS
ML Drug Discovery
Genentech

MCBRLab bilibili

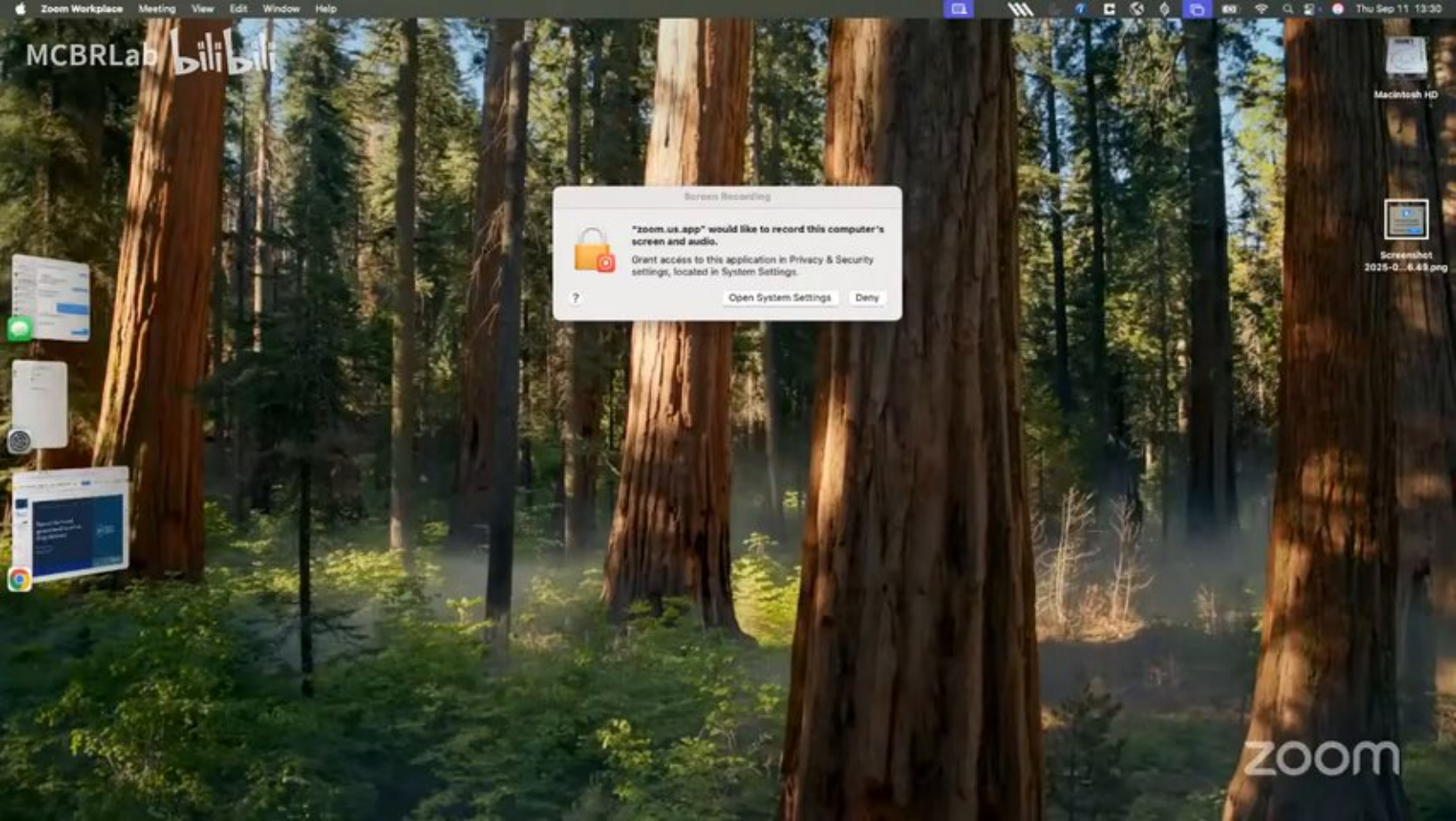
Beyond the funnel: generative AI to active drug discovery

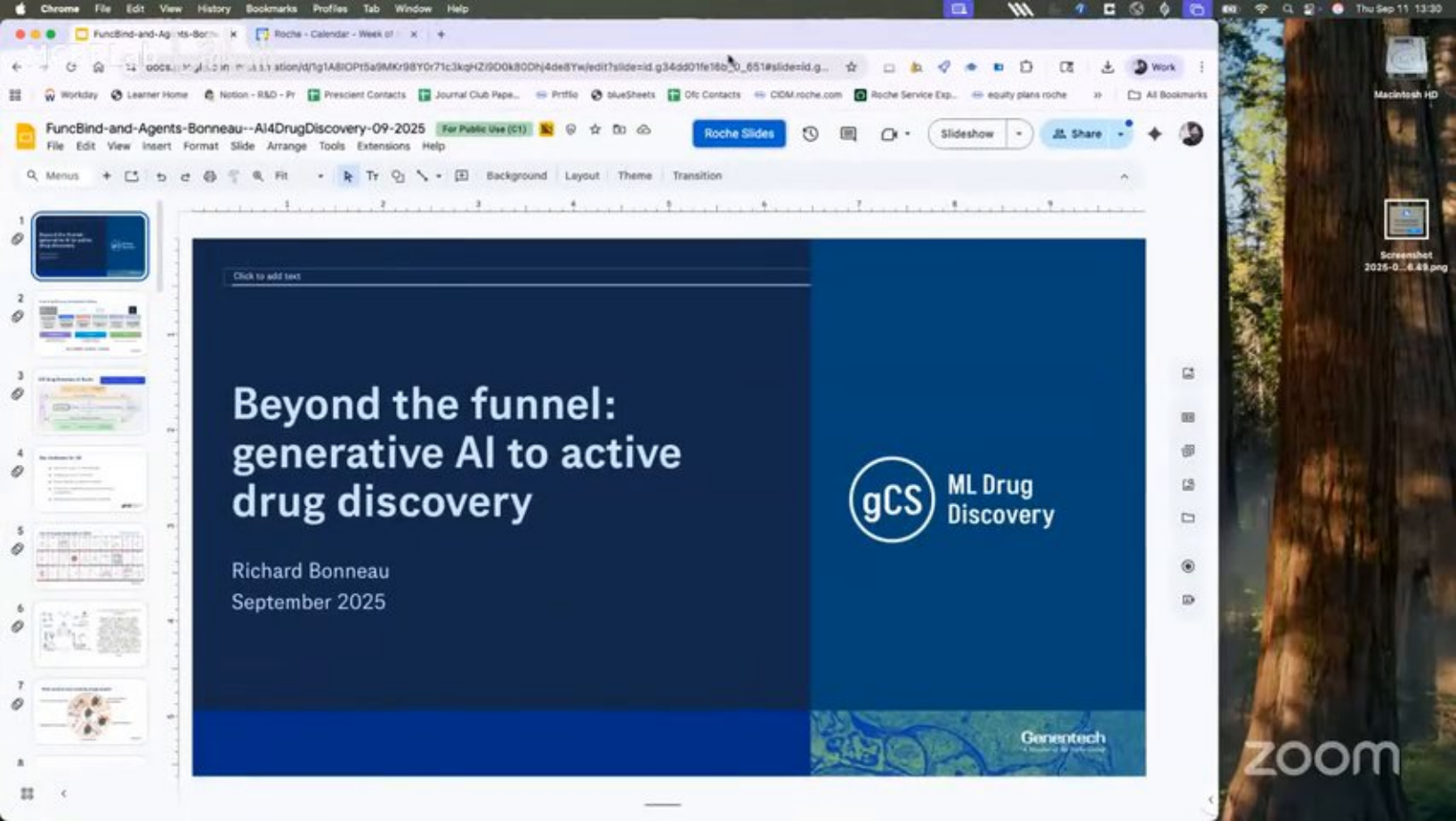
Richard Bonneau

September 2025

gCS ML Drug Discovery

Zoom Genentech
A Member of the Roche Group


Zoom Workplace
Meeting
View
Edit
Window
Help
Thu Sep 11 13:30
MCBRLab
bilibili
Macintosh HD
Screenshot
2025-0...6.48.png
Screen Recording
"zoom.us.app" would like to record this computer's screen and audio.
Grant access to this application in Privacy & Security settings, located in System Settings.
?
Open System Settings
Deny
zoom


FuncBind-and-Agents-Bonneau--AI4DrugDiscovery-09-2025
Roche Slides
Slideshow
Share
Beyond the funnel: generative AI to active drug discovery
Richard Bonneau
September 2025
gCS
ML Drug Discovery
Genentech
zoom

MCBRLab bilibili

Loading...

zoom

MCBRLab bilibili

Beyond the funnel: generative AI to active drug discovery

Richard Bonneau

September 2025

gCS ML Drug Discovery

Genentech
A Member of the Roche Group

AI for Drug Discovery @ Genentech & Roche

Richard Bonneau
ML Drug Discovery

Computational drug discovery, driving new approaches and synergies between small and large molecule drug discovery

Small Mol & Peptide Drug Discovery

computational and machine learning tools for **small molecule and peptide design** and optimization

Large Molecule Drug Discovery

computational and machine learning tools for **large molecule design** and optimization

Frontier Research

foundational machine learning methods and approaches

ML Engineering

machine learning engineering for enabling efforts across all MLDD pillars

Prescient LM

internal development of **large language models** across gRED & Roche

Technology Dev

new capabilities for unmet needs and exploratory work

Platforms

engineering, integration, extensibility, and scaling

Portfolio

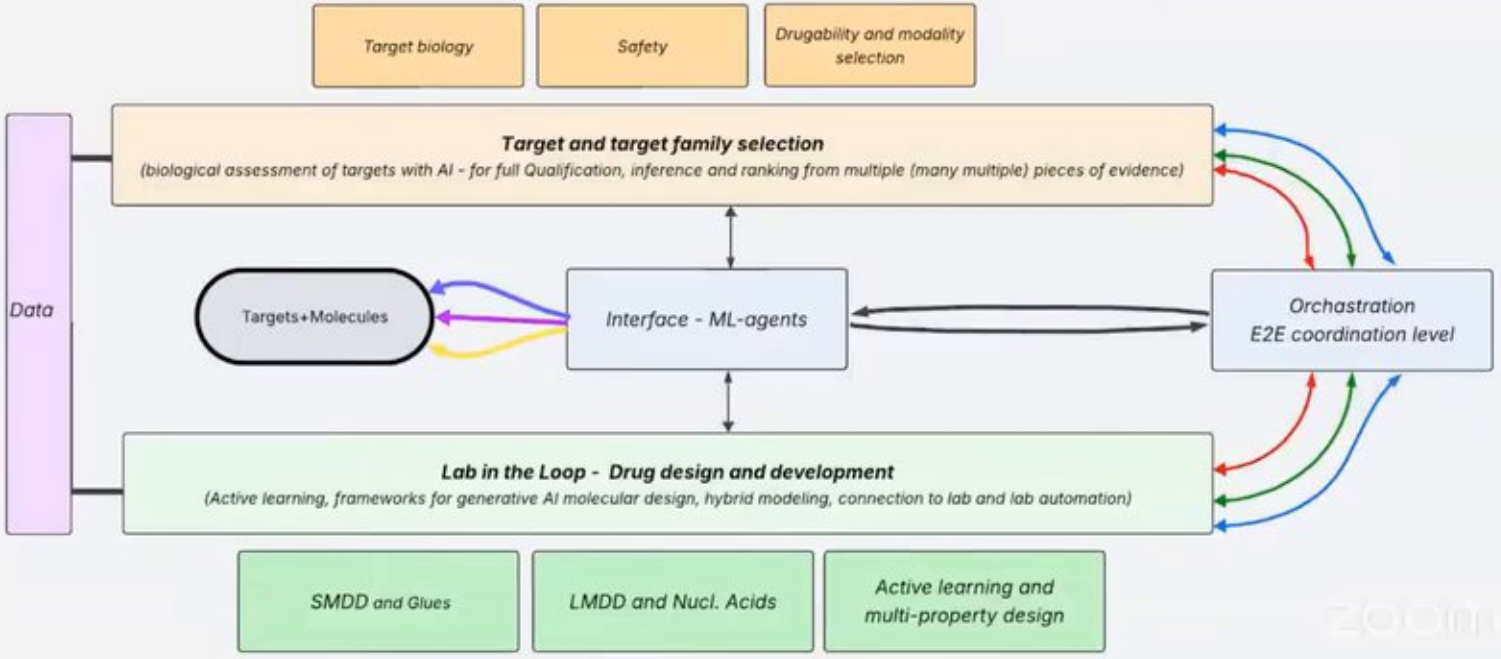
application and program impact

new capabilities, platforms, & portfolio

Genentech
A Member of the Roche Group

E2E Drug Discovery @ Roche

w/ Sara Mostafavi, Fabian Birzele, Anna Craig, Ryan Copping, John Marioni

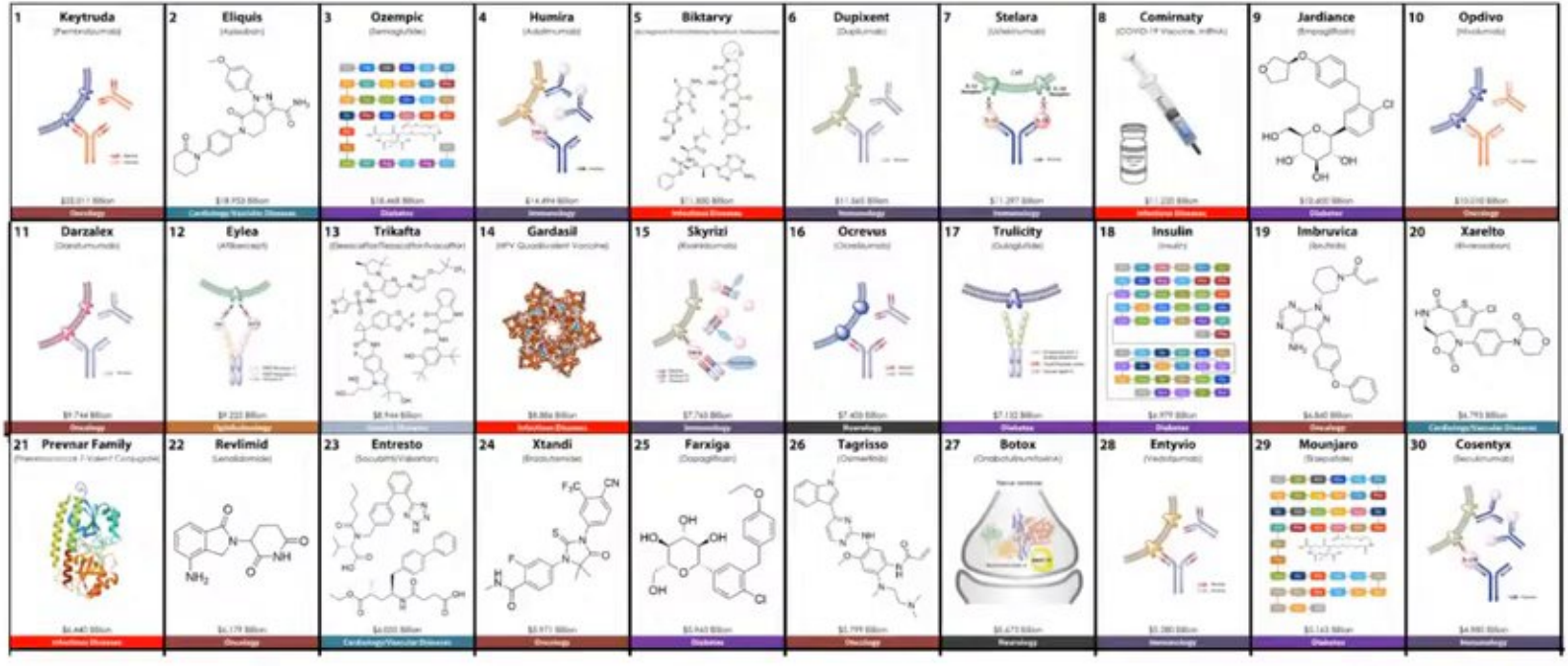


Key challenges for DD

- Lab in the Loop / Active learning
- Integration across timelines
- Cross modality workflows needed
- Contrast in depth/size/runtime of workflow components
- Motivating process optimization with ML

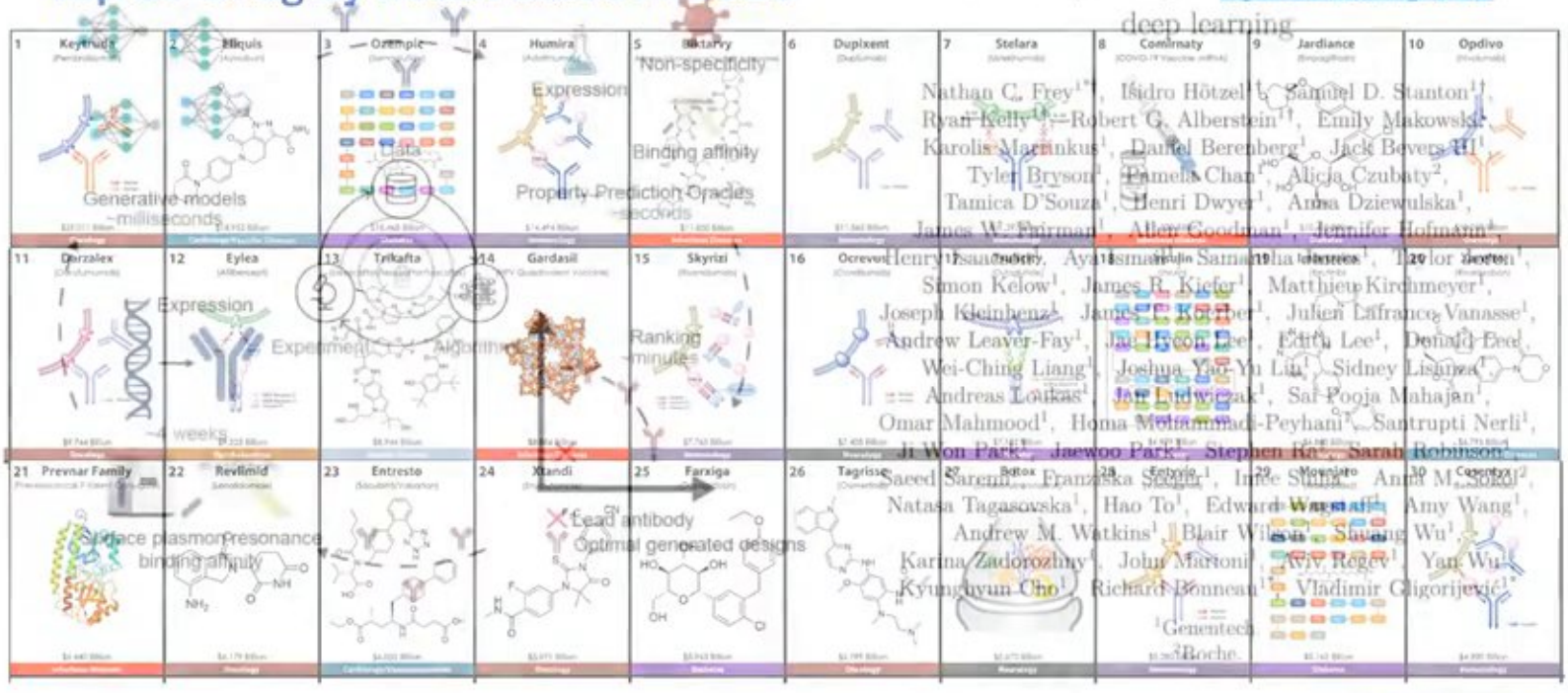
Top 30 Drug by Retail Sales in 2023

Njardarson group



Genentech
A Member of the Roche Group

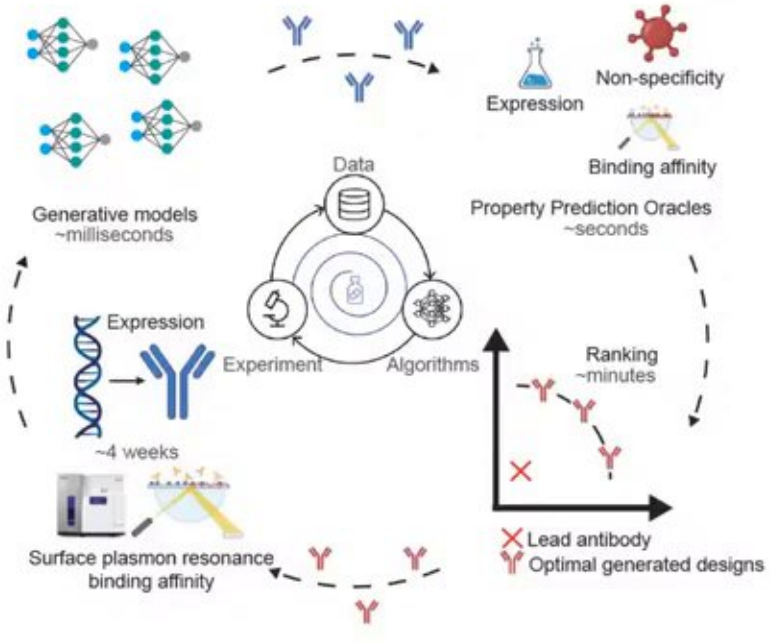
Top 30 Drug by Retail Sales in 2023



Lab-in-the-loop therapeutic antibody design with deep learning

Nathan C. Frey[1*], Isidro Hötzel[1], Samuel D. Stanton[1†], Ryan Kelly[1†], Robert G. Alberstein[1†], Emily Makowski[1], Karolis Martinkus[1], Daniel Berenberg[1], Jack Bevers III[1], Tyler Bryson[1], Pamela Chan[1], Alicia Czubaty[2], Tamica D'Souza[1], Henri Dwyer[1], Anna Dziewulska[1], James W. Fairman[1], Allen Goodman[1], Jennifer Hofmann[1], Henry Isaacson[1], Ayya Jasmin Samantha Jimenez[1], Taylor Joren[1], Simon Kelow[1], James R. Kiefer[1], Matthieu Kirchmeyer[1], Joseph Kleinhenz[1], James C. Kocher[1], Julien Laffrance Vanasse[1], Andrew Leaver-Fay[1], Jae Hyeon Lee[1], Faith Lee[1], Donald Lee[1], Wei-Ching Liang[1], Joshua Yao-Yu Lin[1], Sidney Lisanza[1], Andreas Loukas[1], Jan Ludwiczak[1], Sai Pooja Mahajan[1], Omar Mahmood[1], Homa Mohammadi-Peyhani[1], Santrupti Nerli[1], Ji Won Park[1], Jaewoo Park[1], Stephen Ra[1], Sarah Robinson[1], Saeed Saremi[1], Franziska Seeger[1], Imee Sinha[1], Anna M. Sokol[1], Natasa Tagasovska[1], Hao To[1], Edward Wagstaff[1], Amy Wang[1], Andrew M. Watkins[1], Blair Wilkinson[1], Shuang Wu[1], Karina Zadorozhny[1], John Marioni[1], Aviv Regev[1], Yan Wu[1], Kyunghyun Cho[1], Richard Bonneau[1†], Vladimir Gligorijević[1*]

[1]Genentech [2]Roche.

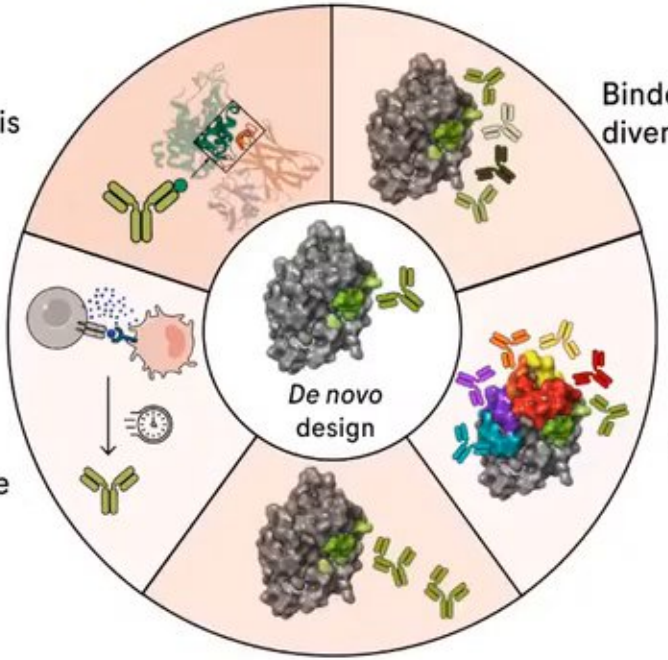


Lab-in-the-loop therapeutic antibody design with deep learning

Nathan C. Frey[1*†], Isidro Hötzel[1†], Samuel D. Stanton[1†], Ryan Kelly[1†], Robert G. Alberstein[1†], Emily Makowski[1], Karolis Martinkus[1], Daniel Berenberg[1], Jack Bevers III[1], Tyler Bryson[1], Pamela Chan[1], Alicja Czubaty[2], Tamica D'Souza[1], Henri Dwyer[1], Anna Dziewulska[1], James W. Fairman[1], Allen Goodman[1], Jennifer Hofmann[1], Henry Isaacson[1], Aya Ismail[1], Samantha James[1], Taylor Joren[1], Simon Kelow[1], James R. Kiefer[1], Matthieu Kirchmeyer[1], Joseph Kleinhenz[1], James T. Koerber[1], Julien Lafrance-Vanasse[1], Andrew Leaver-Fay[1], Jae Hyeon Lee[1], Edith Lee[1], Donald Lee[1], Wei-Ching Liang[1], Joshua Yao-Yu Lin[1], Sidney Lisanza[1], Andreas Loukas[1], Jan Ludwiczak[1], Sai Pooja Mahajan[1], Omar Mahmood[1], Homa Mohammadi-Peyhani[1], Santrupti Nerli[1], Ji Won Park[1], Jaewoo Park[1], Stephen Ra[1], Sarah Robinson[1], Saeed Saremi[1], Franziska Seeger[1], Imee Sinha[1], Anna M. Sokol[2], Natasa Tagasovska[1], Hao To[1], Edward Wagstaff[1], Amy Wang[1], Andrew M. Watkins[1], Blair Wilson[1], Shuang Wu[1], Karina Zadorozhny[1], John Marioni[1], Aviv Regev[1], Yan Wu[1], Kyunghyun Cho[1], Richard Bonneau[1*], Vladimir Gligorijević[1*]

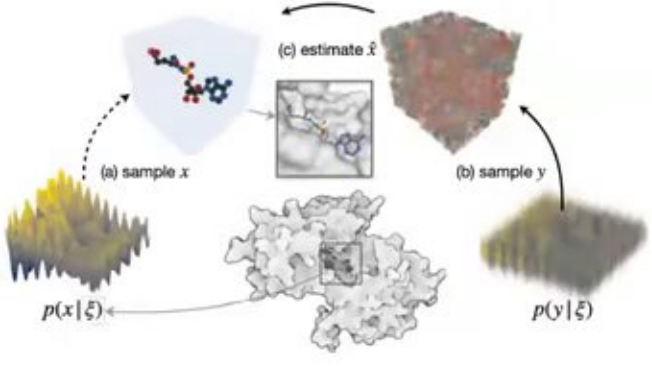
[1]Genentech.
[2]Roche.

What would *de novo* antibody design enable?



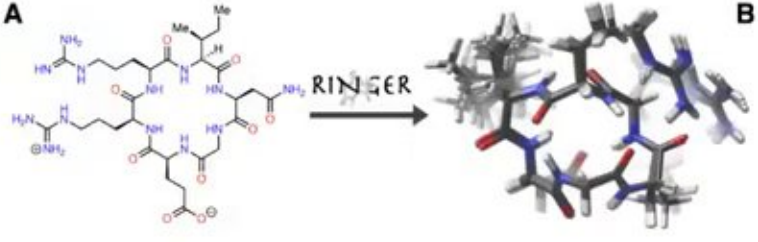
Genentech
A Member of the Roche Group

Molecular Generation and Conformational Sampling



Oliveira Pinheiro et al. "Structure-based drug design by denoising voxel grids." ***ICML 2024***

pocket-conditioned 3D molecule generation with **walk jump sampling**

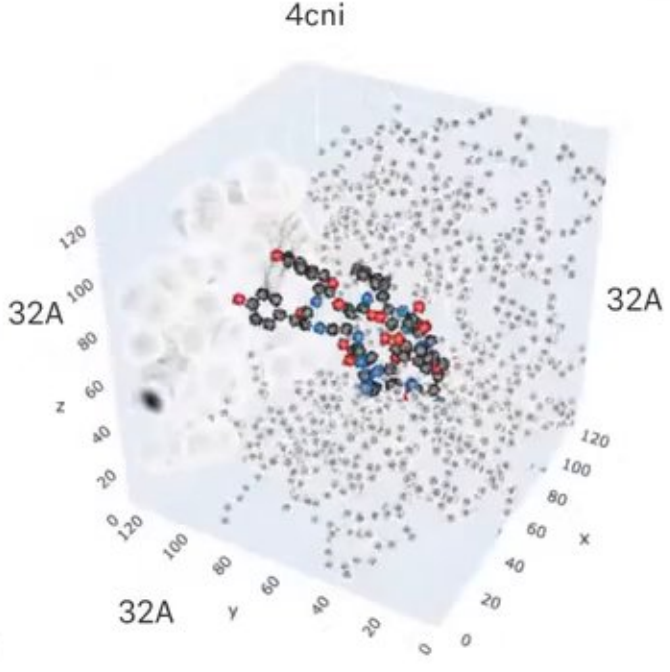


Grambow et al. "Accurate and Efficient Structural Ensemble Generation of Macrocyclic Peptides using Internal Coordinate Diffusion". arXiv:2305.19800v2 2024

efficient conformer generation of macrocyclic peptides and proteins with **walk-jump sampling** and **diffusion** models

Genentech
A Member of the Roche Group

Representing molecules as atomic density fields



Prescient Design
A Genentech Accelerator

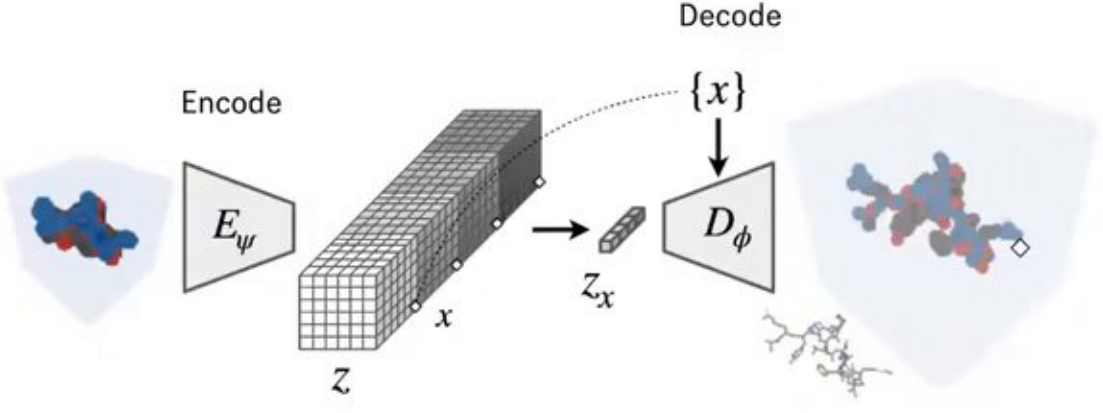
De novo & X-modality : FuncBind

	Unconditional generation	Target conditioned generation
Voxels	VoxMol Pinheiro et al., NeurIPS 2023 Modality: small molecule	VoxBind Pinheiro et al., ICML 2024 Modality: small molecule
Neural Fields	FuncMol Kirchmeyer et al., NeurIPS 2024 Modality: small molecule, macrocyclic peptide	FuncBind (under review) Modality: unified model for small molecules, macrocyclic peptides, antibody CDR loops

Improved scalability and speed going from voxels to neural fields

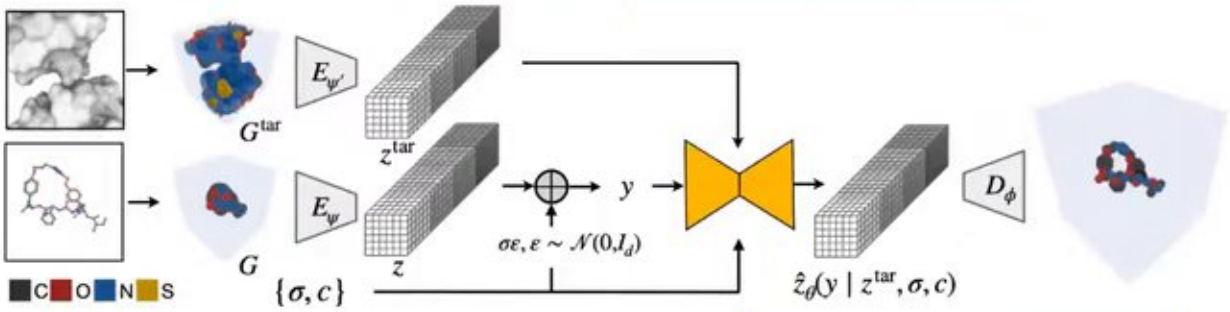
Several technical improvements from FuncMol to FuncBind

Embedding molecules into a latent space using neural fields



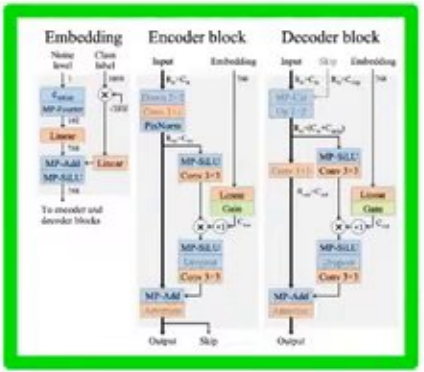
Prescient Design
A Genentech Accelerator

FuncBind's workhorse: a powerful denoiser architecture



Large model leveraging all relevant interface data

Scale via neural field representation



Genentech
A Member of the Roche Group

Sampling via score-based approaches

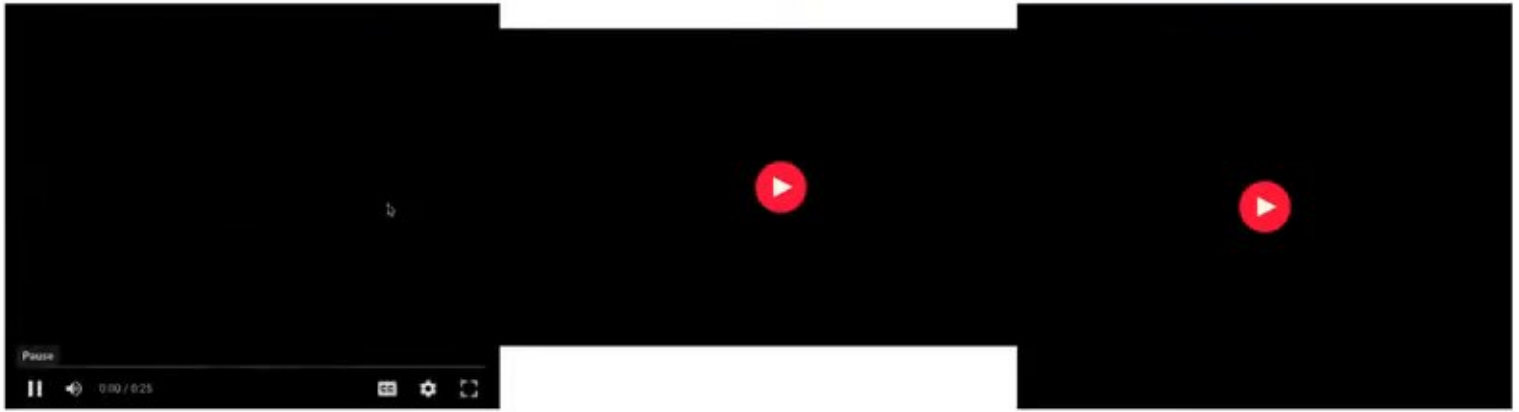
e.g. Walk-Jump Sampling [Saremi & Hyvärinen, 2019 JMLR]

Sample

Prescient Design
A Genentech Accelerator

Pocket conditioned de-novo x-modal generation

small molecule | antibody | peptide

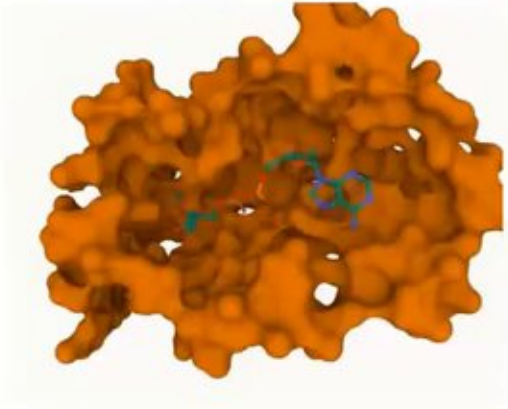


Genentech
A Member of the Roche Group

Pocket conditioned de-novo x-modal generation

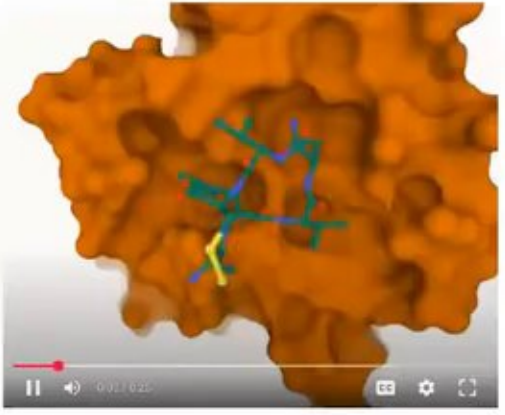
Prescient Design
A Genentech Accelerator

small molecule



antibody

peptide



Genentech
A Member of the Roche Group

FuncBind generates new non canonical amino acids

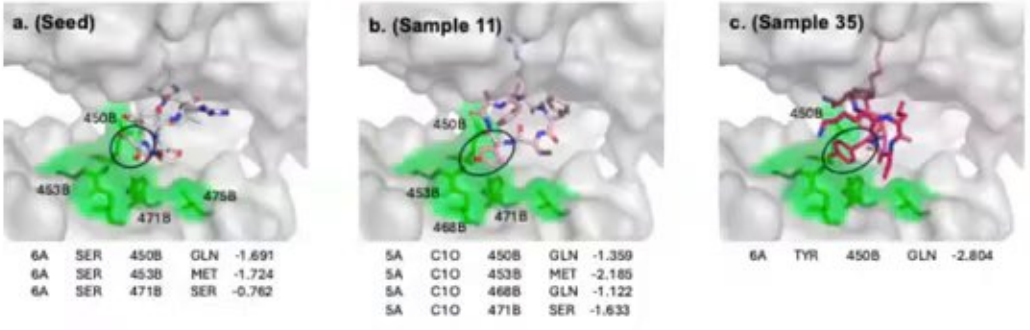
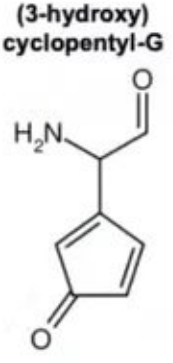


Figure 12: Per-residue energy scores at the same position were calculated using Rosetta's residue energy breakdown for a seed MCP and two samples. The residues analyzed are: (a) serine from the seed, (b) 3-hydroxycyclopentyl-alanine (C1O) from sample 11 (see Figure 11), and (c) tyrosine from sample 35.



C1O

C1O (b) is a new non canonical amino acid that interacts with pocket residues that neither the seed (a) nor a chemically similar AA at the same position (c) engage

Evaluation

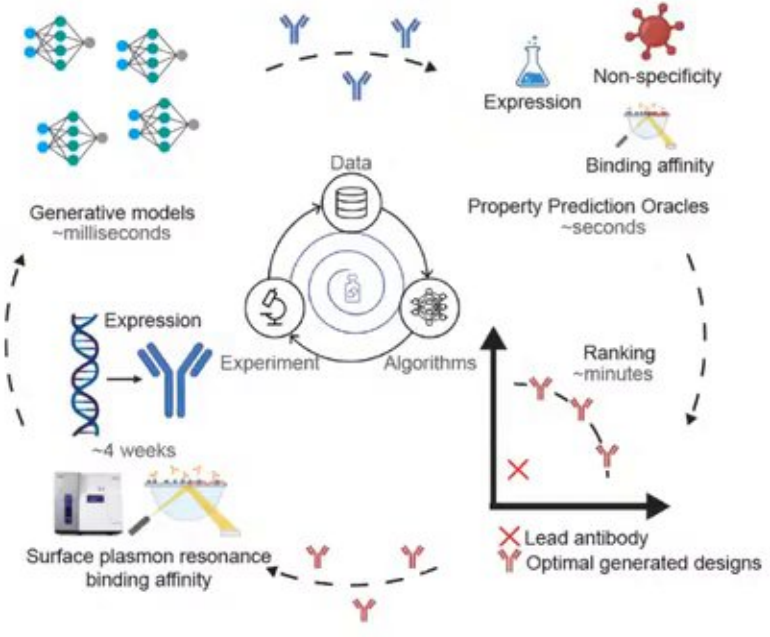
In-silico evaluation on public benchmarks

- state-of-the-art on pocket conditioned small molecule generation
- outperforms the state-of-art on epitope conditioned CDR redesign
- promising results on our new (to be released) benchmark for pocket conditioned macro cyclic peptide generation

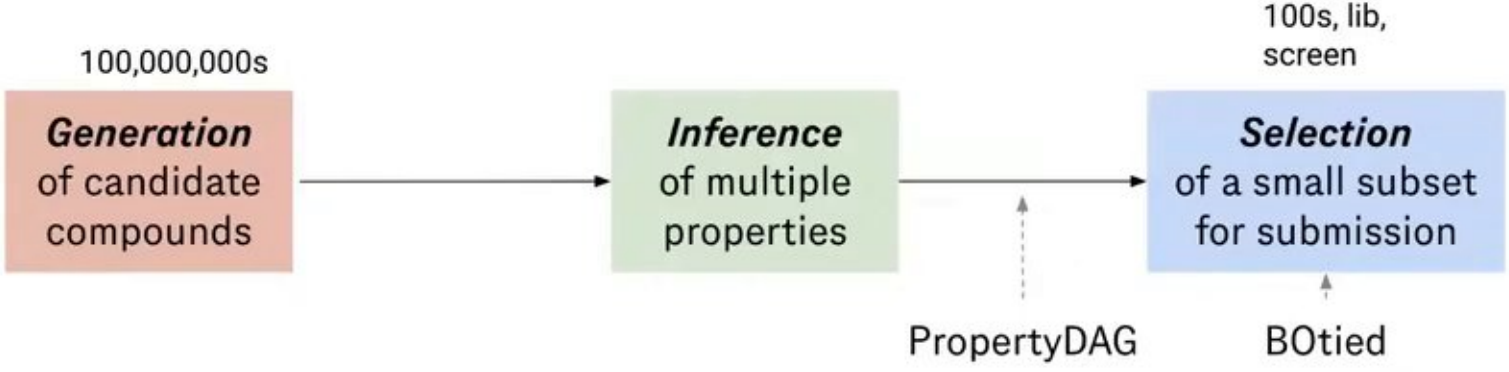
Wet lab validation

- For CDR H3 redesign, we achieved a 42% binding rate on our confident samples and found a 5X better binder than approved drugs for targets.

Lab-in-the-loop for DD



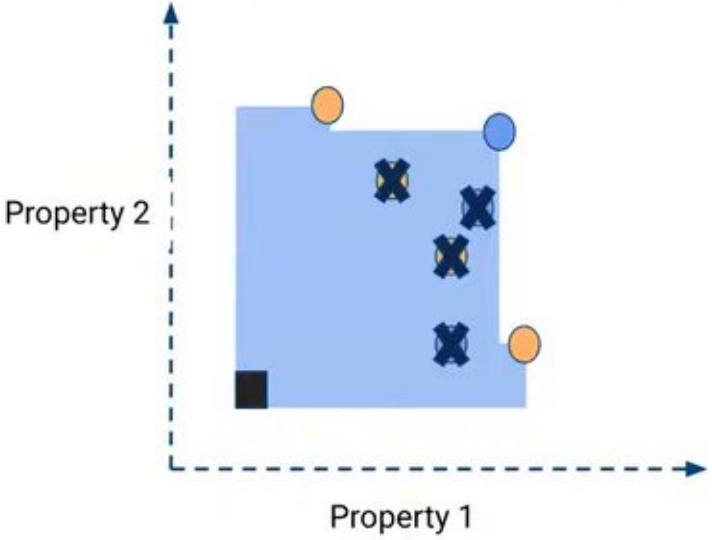
Bayesian optimization for design selection: PropertyDAG and BOtied

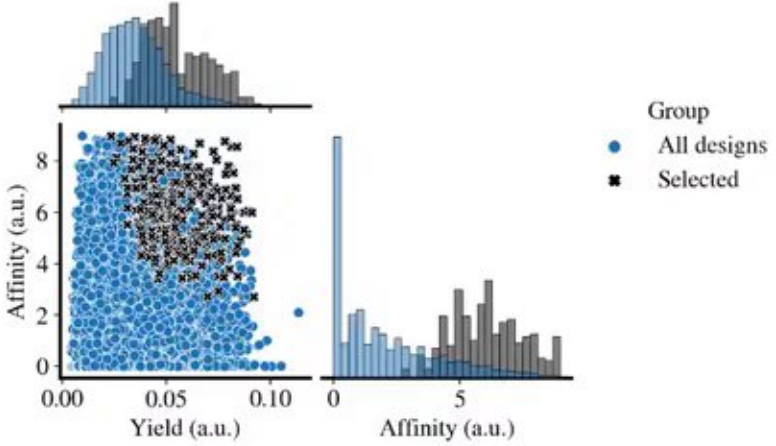


Park, Ji Won, et al. "PropertyDAG: Multi-objective Bayesian optimization of partially ordered, mixed-variable properties for biological sequence design" (2022) NeurIPS, *AI4Science*

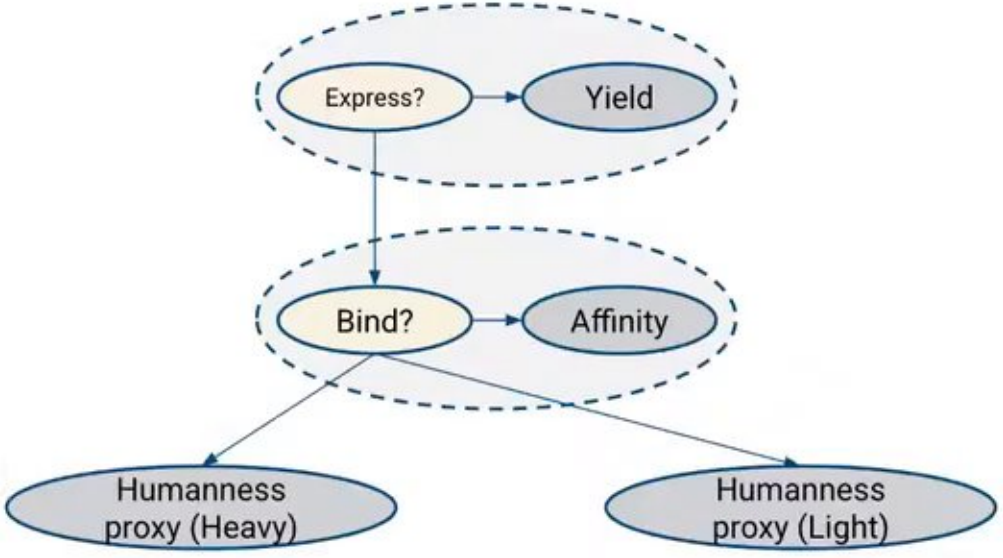
Active learning: multi-objective sample selection through expanding the Pareto frontier

PropertyDAG: Expected hypervolume improvement (EHVI) guides exploration vs. exploitation balance towards expanding the Pareto frontier over multiple developability properties.



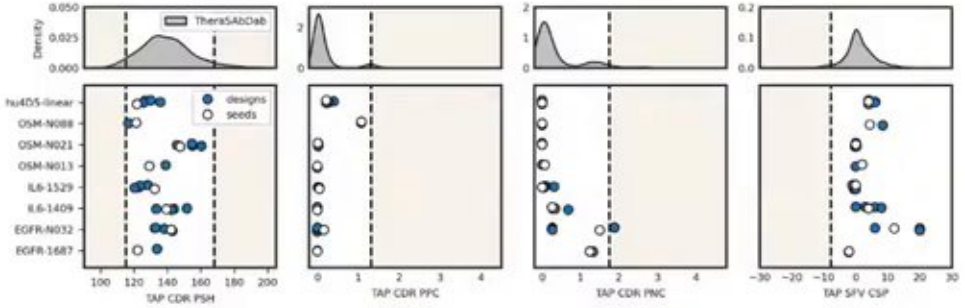


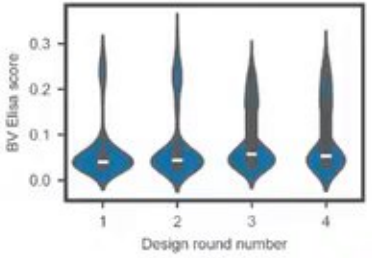
PropertyDAG: antibodies

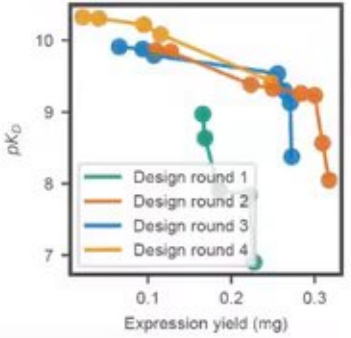


Multi-property optimization enables therapeutic antibody design

Improving expression yield and in-silico developability properties of lead candidate antibodies







gRED Computational Sciences

Generates large amounts of experimental data that is currently underutilized

Large scale screens

- **HTS Screens** (High throughput screening)
- **DEL Screens** (DNA Encoded Libraries)
- **Fragment Screens**

...

Biochemical and cell assays

- **SAR potency assay data** (Structure Activity Relationships)
- **ADME assays** (Absorption, Distribution, Metabolism, Excretion)
- **Toxicology assay data**

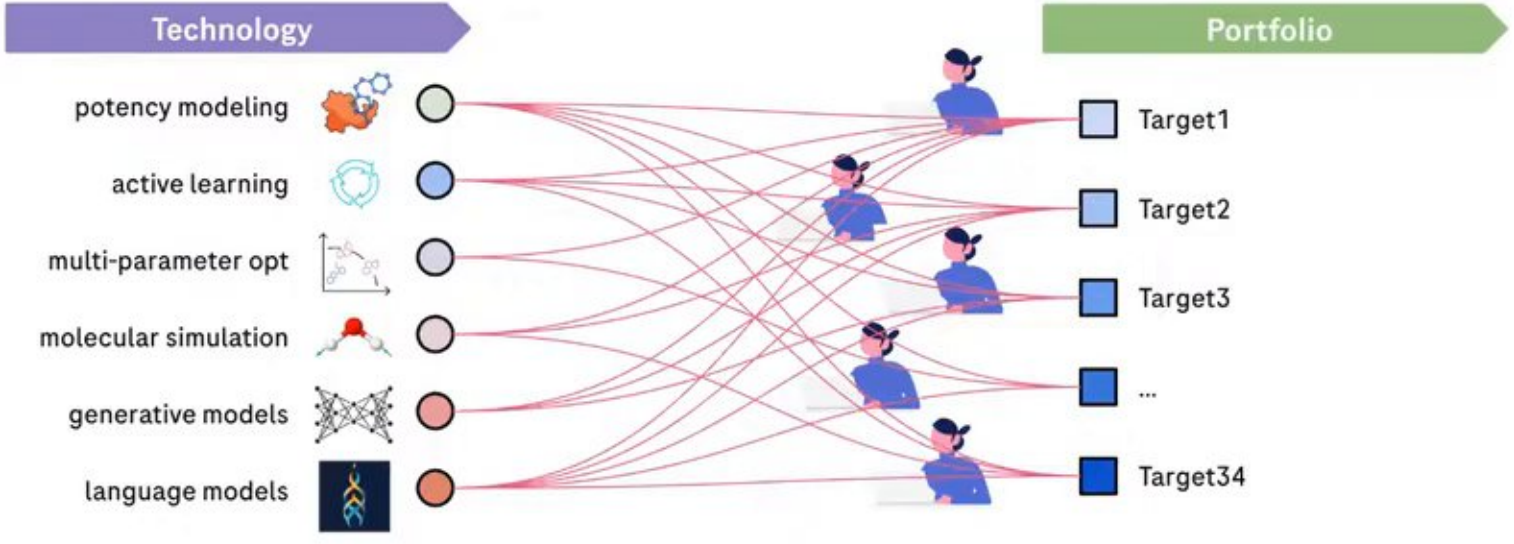
...

Animal studies

- **Toxicology**
- **PK/PD** (Pharmacodynamics, Pharmacokinetics
- **ADME** (Absorption, Distribution, Metabolism, Excretion)

...

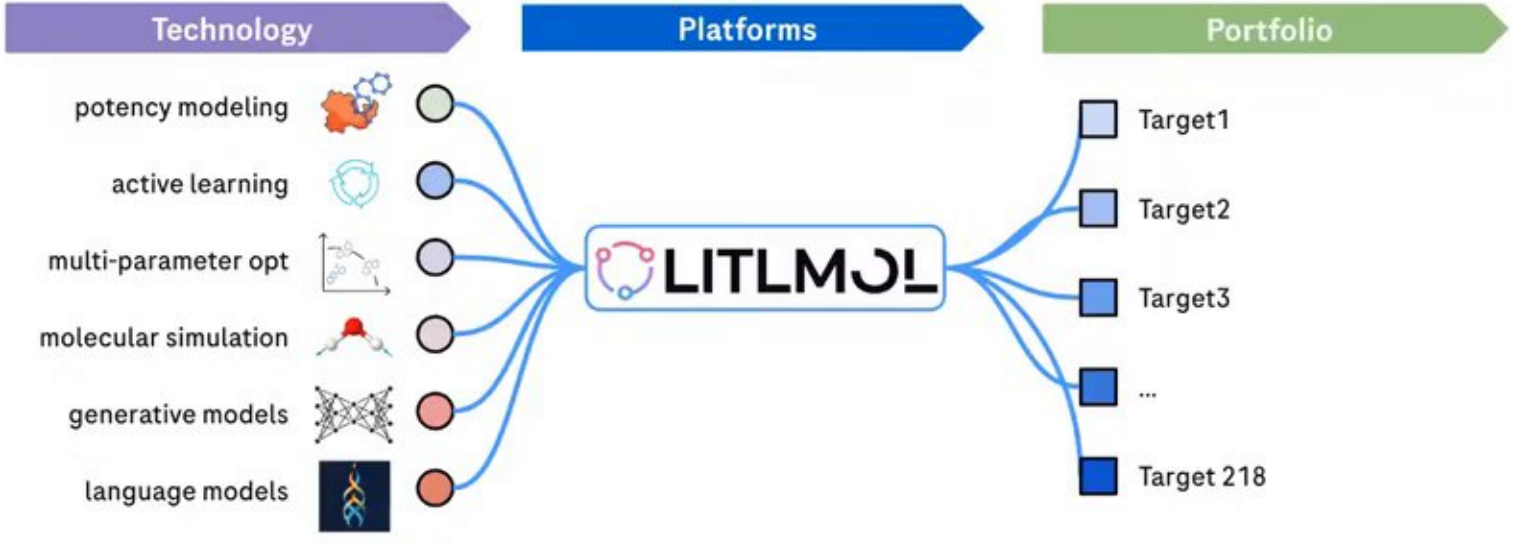
Small Molecule Drug Discovery without a common platform



Ad hoc adoption requires **significant** human **resources** and manual implementation

Does not scale - ***O(N)***

Small Molecule Drug Discovery at Prescient



Instead, we invest in **building systems and platforms** to deliver on the pipeline

We are Clueless about Pretrained Models

- Publicly accessible commercial and open models were pretrained with a massive amount of data of unknown type, origin, composition and content.
- Biases, noise and intentional poisoning in such unknown data will inevitably sway these models' predictions, and we will never know why these models make such predictions.
- For mission-critical scenarios, such as in healthcare, security and others, we must ask ourselves whether we should and will allow such opacity.

Source: Kyunghyun Cho, CHIL '24 Keynote

State-of-the-art LLMs for drug discovery

Foundational models: Train and maintain **Genie-base** and **Genie-large**

Genie-base: 5-15B parameters, supports managed FT, RAG; used for prototyping and running locally

Genie-large: 30-70B parameters, used for complex applications

Improve **drug discovery-related** model capabilities

Tool calling to facilitate the lab-in-the-loop discovery

Reasoning: enable complex QA, multi-hop and better CoT

Multi-modality: train the model on protein sequences, small molecule structures, images

Chat: improve chat interface, instruction following for everyday tasks

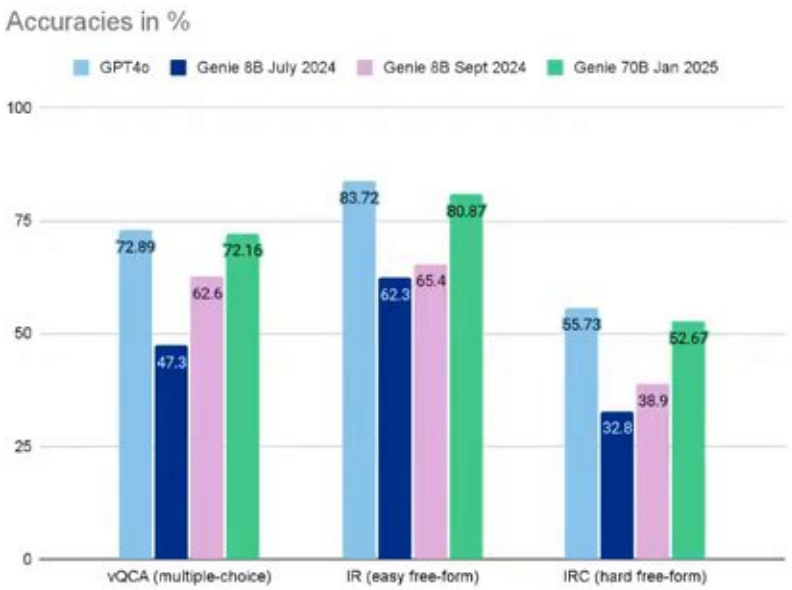
Achieve **state-of-the-art** and make a **public version**

Public version: smaller version trained on a public dataset with agentic and reasoning improvements

Private version: our best state-of-the-art LLM for drug discovery

Genentech
A Member of the Roche Group

Early results: Genie for Target Safety Assessment



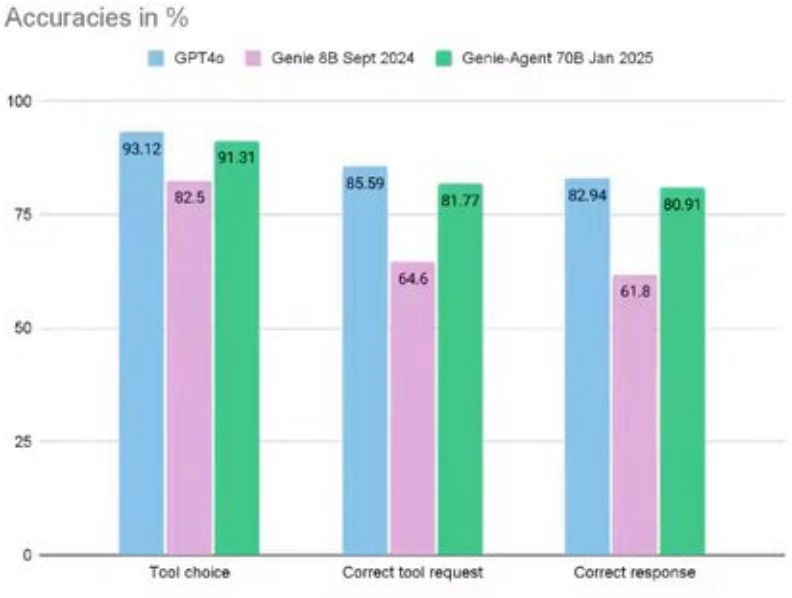
Internal models for core inference

- **Safety**
- **Target assessment and data integration**
- **Digital twins**
- **...**

In collaboration with **gRED DevSci Translational Safety**, **gRED Human Genetics**, **gCS Catalysts**

Genentech
A Member of the Roche Group

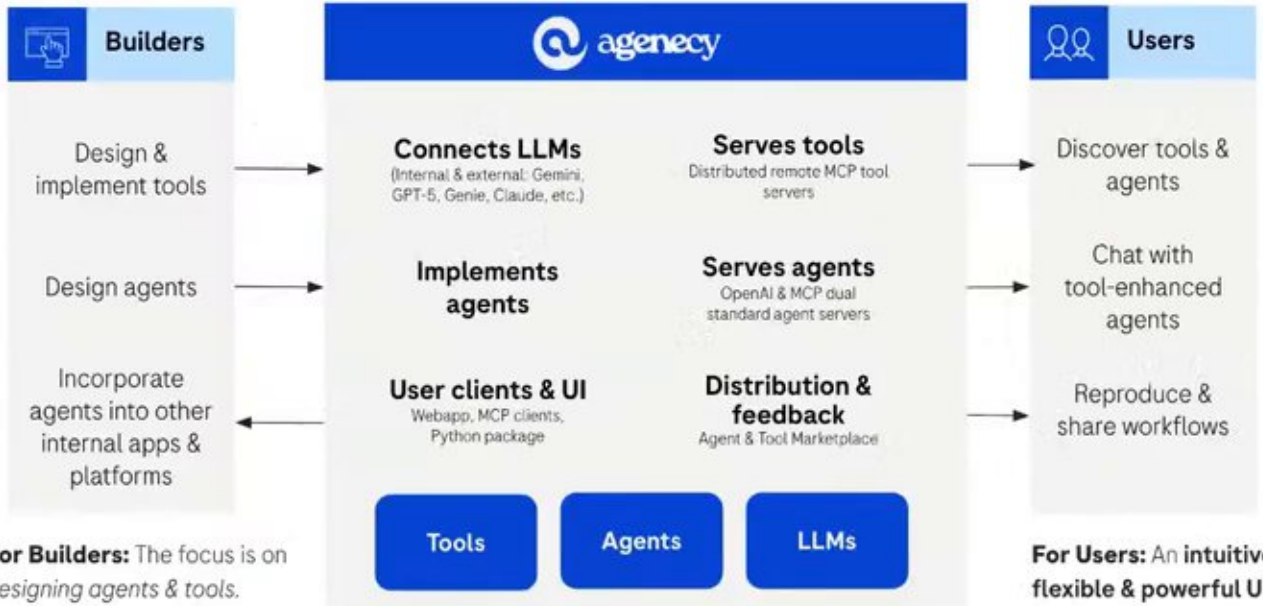
Early results: Genie for Tool Calling



Models for Agents and Orchestration

- Tool calling
- Methods encapsulation and interface
- Automated and semi-automated prototyping and design
- Automated experimental design
- ...

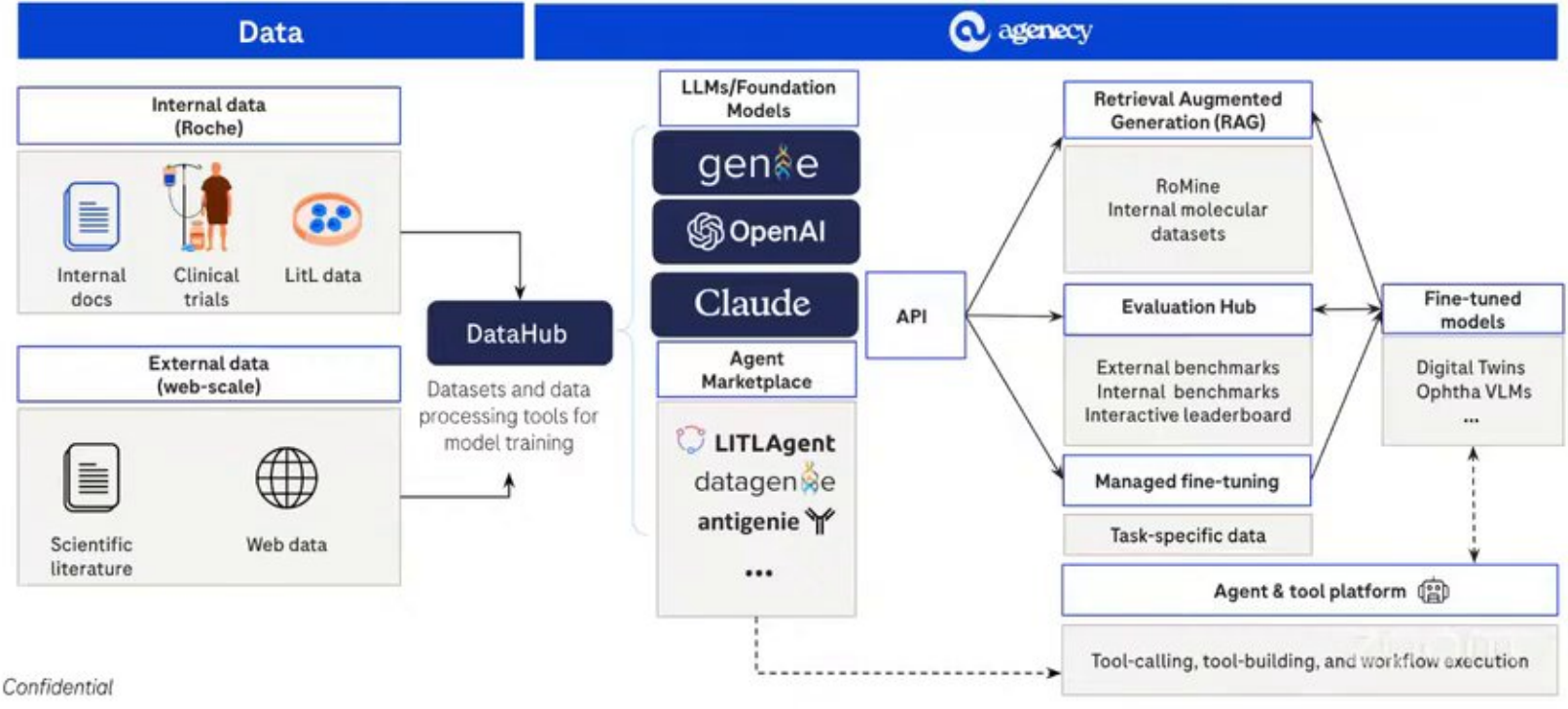
Agenecy: A unified platform for building & using agents



For Builders: The focus is on *designing agents & tools.* Complex engineering is **seamlessly abstracted away.**

For Users: An **intuitive, flexible & powerful UI** to use and chat with different LLMs, agents, and tools.

Agenecy: the Roche agentic ecosystem




Users can
Query LLMs
Select and chat with external and internal LLMs on the webapp
Use `agenecy` package to call a LLM
Using Tools
Browse and find tools
Tool and tool server marketplace on webapp
Use `agenecy` package to list available tools
Trigger tools with natural language
Chat with tool collections on webapp
Call tools in conversations in MCP clients like Cursor etc.
Trigger tools manually
Run manually through webapp UI
Use `agenecy` package to execute tools
Using AI Agents
Browse and find agents
Agent marketplace on webapp
Use `agenecy` package to list available agents
Get agent responses
Select and chat with agents on webapp
Chat with agents in MCP clients like Cursor
Use `agenecy` package to invoke agents

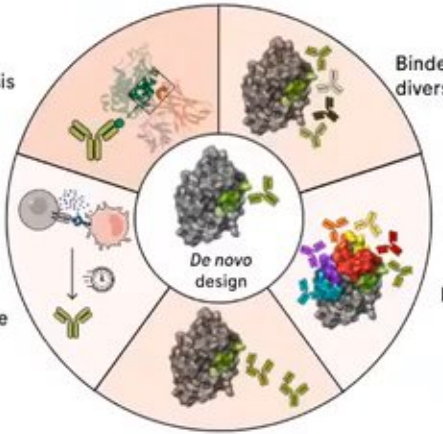

agenecy

Open questions

- Core models and training
- Data layer, model registry, tool/method encapsulation
- GenAI+tools encapsulation and testing
- Integration, interoperation and uptime
- Embrace the Long Tail **? OR**
  Focus on **main workflows** that power value chain?

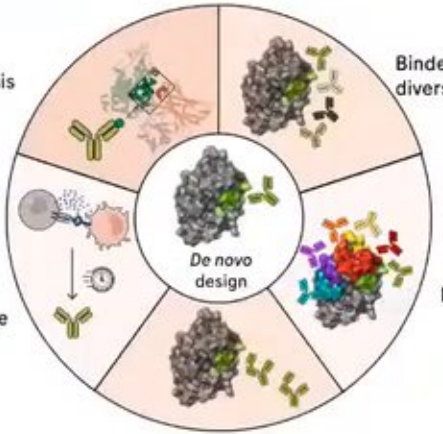
Thanks!

What would *de novo* antibody design enable?



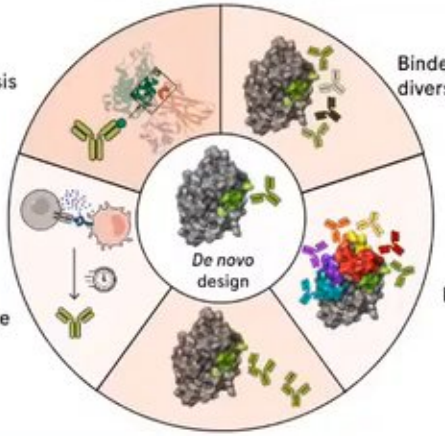
Genentech

What would *de novo* antibody design enable?



Genentech

What would *de novo* antibody design enable?



Genentech

FuncBind generates new non canonical amino acids

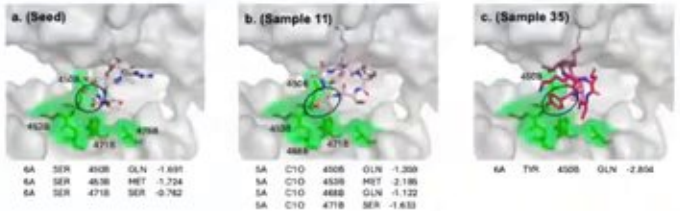
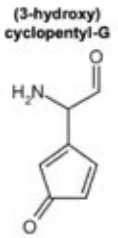


Figure 12: Per-residue energy scores at the same position were calculated using Rosetta's residue energy breakdown for a seed MCP and two samples. The residues analyzed are: (a) serine from the seed, (b) 3-hydroxycyclopentyl-alanine (C1O) from sample 11 (see Figure 11), and (c) tyrosine from sample 35.



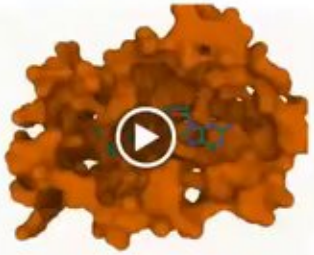
C1O (b) is a new non canonical amino acid that interacts with pocket residues that neither the seed (a) nor a chemically similar AA at the same position (c) engage

Click to add text

Prescient Design
A Genentech Accelerator

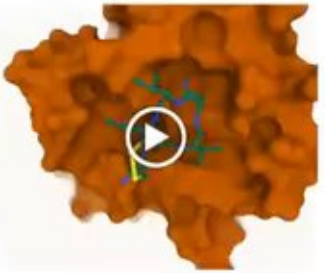
Pocket conditioned de-novo x-modal generation

small molecule



antibody

peptide



Genentech
A Member of the Roche Group

MCBRLab bilibili

Rich Bonneau

zoom

End Show Tips Swap Displays Use Slide Show

00:00 11:21 AM

Slide 1 of 23

bfksdjb

End Show Tips Swap Displays Use Slide Show

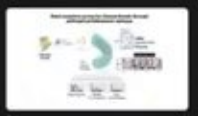
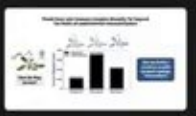
00:08 11:21 AM

Next slide

Slide 1 of 23

bfksdjb





1 2 3 4 5 6 7 8

MCBRLab bilibili

zoom

MCBRLab bilibili

NYGC Events

zoom

MCBRLab bilibili

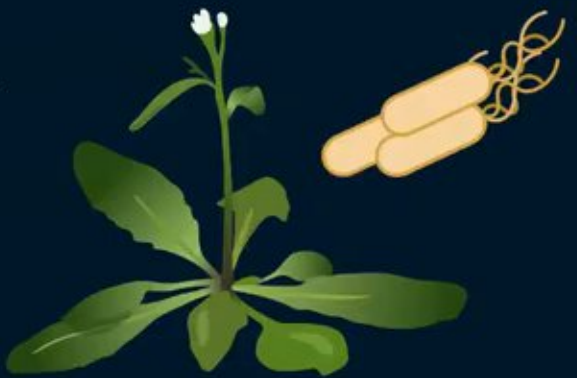
Berkeley
UNIVERSITY OF CALIFORNIA

mamp-ml:

A deep learning approach to predicting epitope immunogenicity in plants

Danielle M. Stevens[1], David Yang[2], Tatiana J. Liang[1], Tianrun Li[3], Brandon Vega[1], Gitta L. Coaker[3], Ksenia Krasileva[1,2]

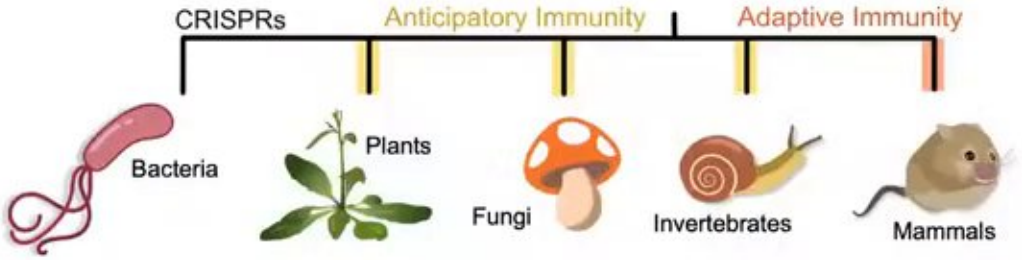
[1] Plant & Microbial Biology, UC Berkeley
[2] Center for Computational Biology, UC Berkeley
[3] Plant Pathology, UC Davis



USDA NIH

zoom
MLCB 2025

Most organisms encoded an innate immune system



Adapted from Sutherland, **Stevens**, et al. 2025. *The Plant Cell.*

Plant receptors survey for diverse threats through pathogen proteinaceous epitopes

Pathogen Protein

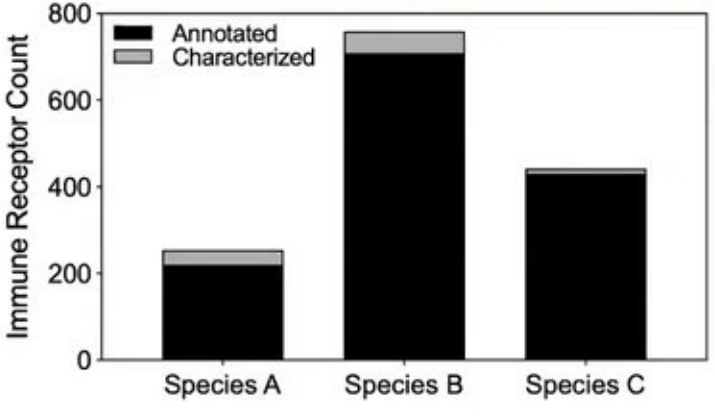
Epitope

Plants have vast immune receptor diversity far beyond the limits of experimental characterization

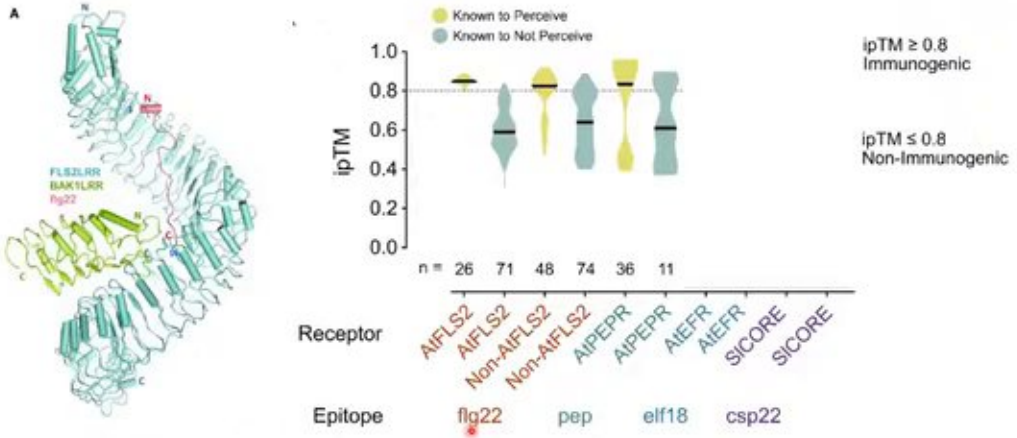
How do they survive?

Plants have vast immune receptor diversity far beyond the limits of experimental characterization

How do they survive?



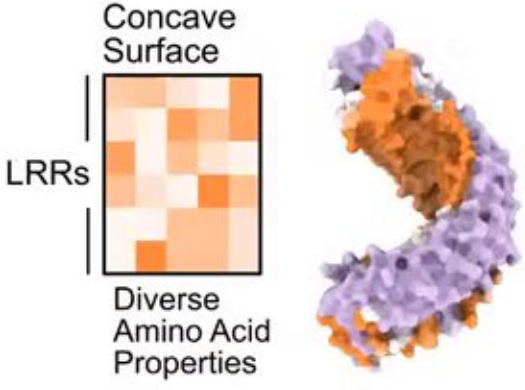
AlphaFold3 fails to predict outcomes when an underlying structure is not solved



AtFLS2-flg22

Sun et al., 2013. *Science;* **Li et al., 2025. *Nature Plants*; Stevens DM** et al., 2025. *BioRxiv.*

Goal: Fine-tune pLM for epitope-receptor immune outcomes



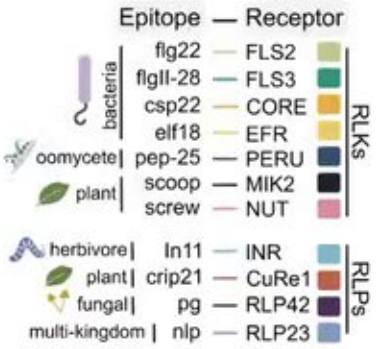
Generating the data needed for fine-tunning a language model

Scraping Data

74 previous published studies from 1999 - 2025

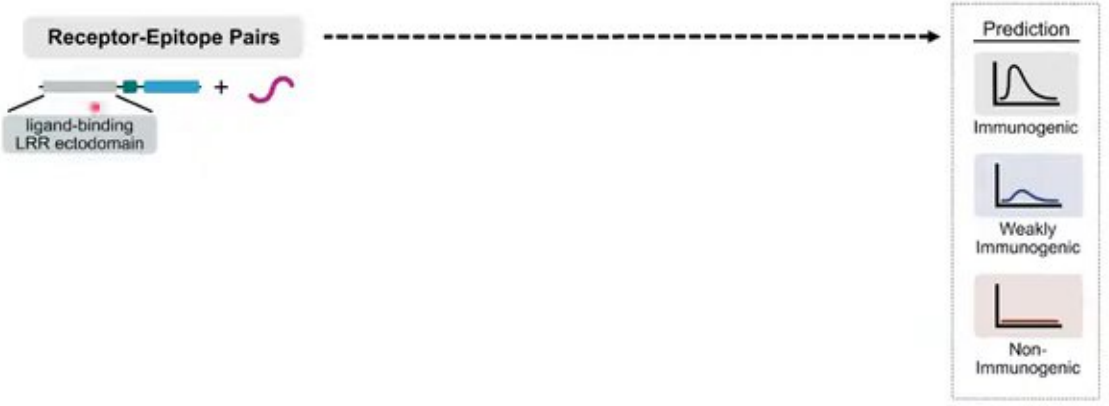
91 plant species

n = 1337 interactions



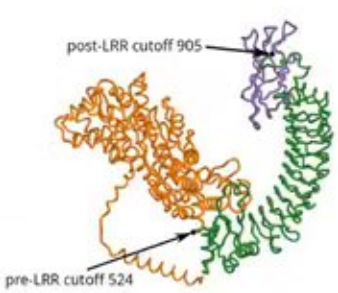
Stevens DM et al., 2025. *BioRxiv.*

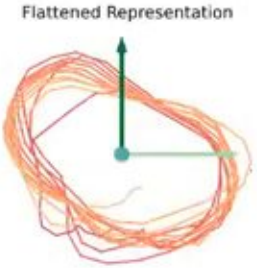
Building a pipeline to extract the epitope-binding LRR Ectodomain



Stevens DM et al., 2025. *BioRxiv.*

Using LRR-Annotation to extract LRR ectodomain and track surface residues

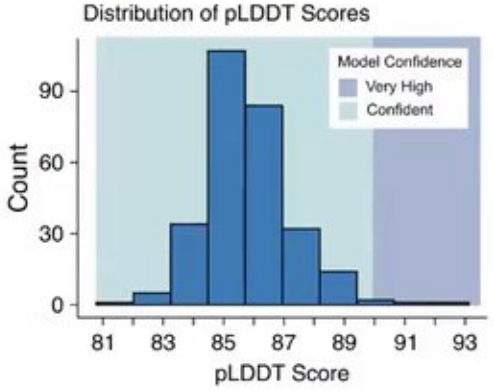


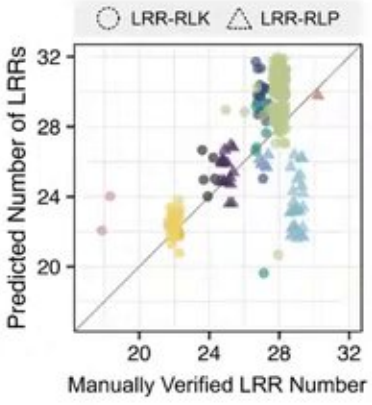


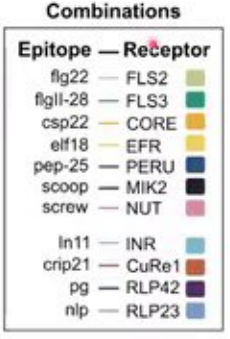
Xu et al. 2024. *PLoS Computational Biology.*

AlphaFold and LRR-Annotation extracts epitope-binding LRR Domain

AlphaFold Receptor → LRR-Annotation → Ectodomain Extraction

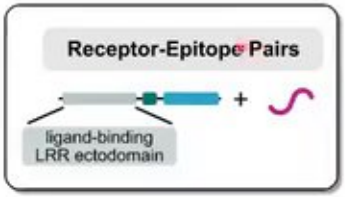






Stevens DM et al., 2025. *BioRxiv.*

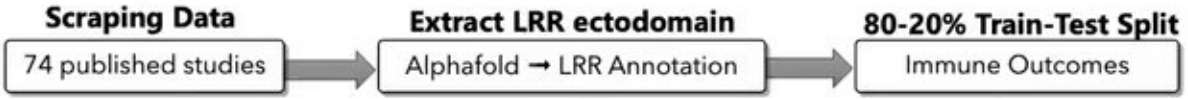
mamp-ml, a fine-tuned pLM enables prediction of immunogenic outcomes

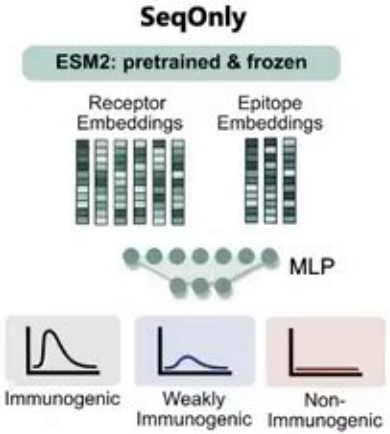


In Feed-Forward NN:
Loss Function = Cross-Entropy
Dropout – 0.2, L2 Regularization to limit overfitting

Stevens DM et al., 2025. *BioRxiv.*

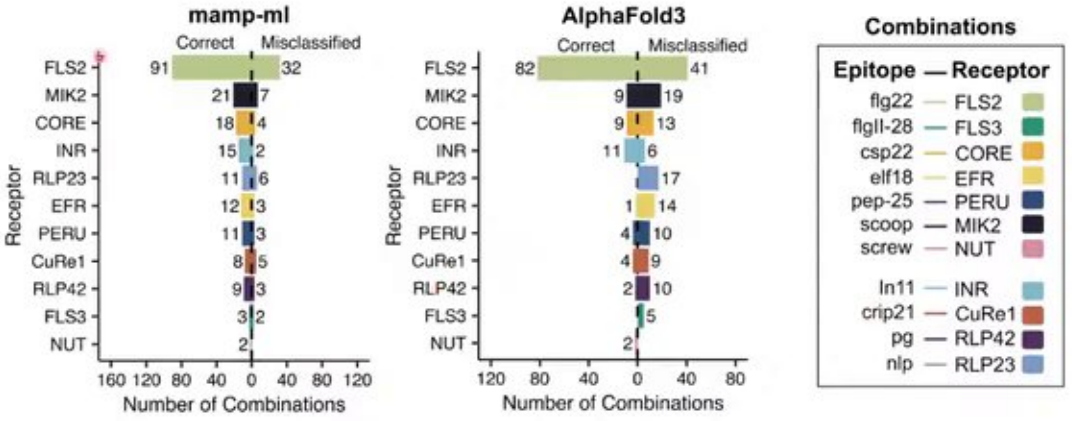
mamp-ml outperforms embeddings only model





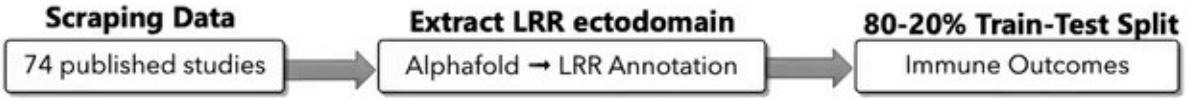
Stevens DM et al., 2025. *BioRxiv.*

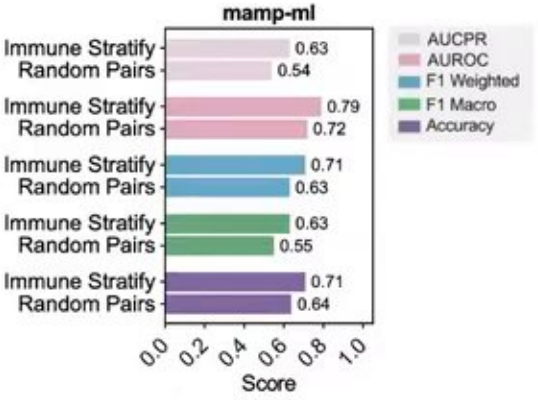
Mamp-ml predicts receptor-epitope immunogenicity without structural context



Stevens DM et al., 2025. *BioRxiv*.

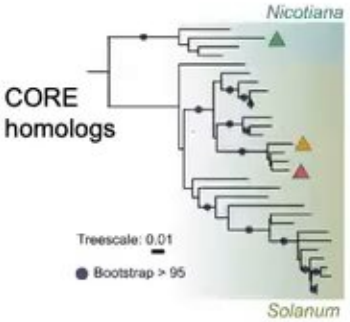
Validating mamp-ml for receptor-epitope interactions

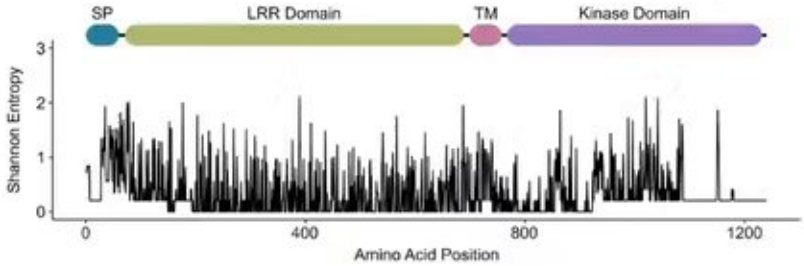




Stevens DM et al., 2025. *BioRxiv.*

Using CORE-csp22 as a receptor-epitope validation dataset

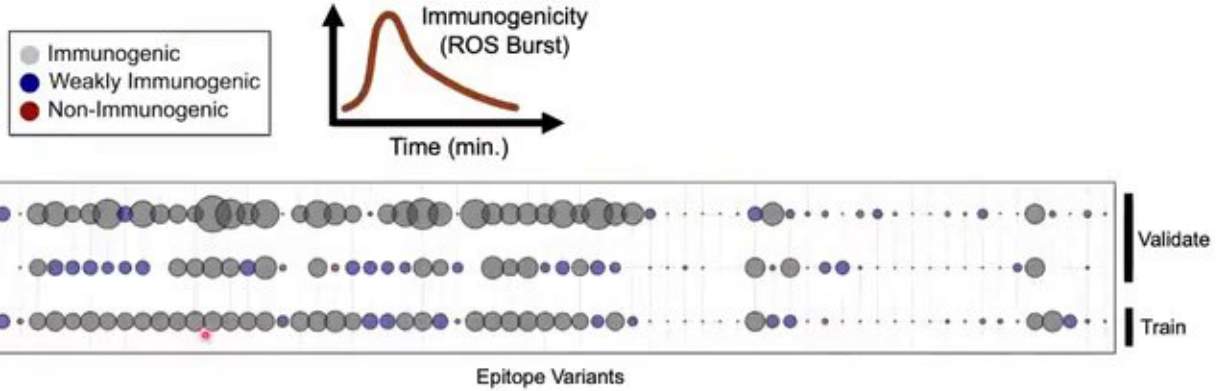




Stevens DM et al., 2025. *BioRxiv.*

Collecting experimental validation data

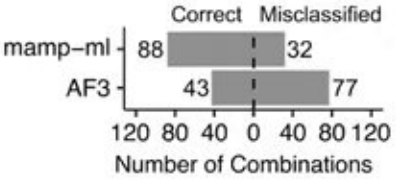
Tatiana Liang,
UC Berkeley Undergraduate



3 plants per plate, Repeated 2 times.

Stevens DM et al., 2025. *BioRxiv.*

Mamp-ml has improved prediction power over AlphaFold3



Stevens DM et al., 2025. *BioRxiv.*

Conclusions and Future Directions

Fine-tunied models are more accurate than generalized approaches

Mamp-ml performs poorly to completely new receptor-epitope interactions and thus, needs to be generalized in the future

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

Available in GitHub + Google Colab

MCBRLab bilibili

Acknowledgements

Krasileva Lab

- **Ksenia Krasileva**
- China Lunde Shaw
- Wei Wei
- Grace Stark
- Jude Edwards
- Nicole Dubs
- Kyungyong Seong
- Chandler Sutherland
- **Tatiana Liang**
- **Brandon Vega**

Collaborators

- **Gitta Coaker, UC Davis**
- **Jerry Li , UC Davis**
- Georg Felix, University of Tubingen
- **David Yang (Now at UCSF in Kortemme Lab)**
- Zack Lippman, Cold Spring Harbor
- Patrick Shih, UC Berkeley

UC Berkeley, Center for Computational Biology

USDA

Berkeley UNIVERSITY OF CALIFORNIA

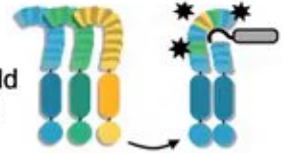
Innovative Genomics Institute

NIH

Conclusions and Future Directions

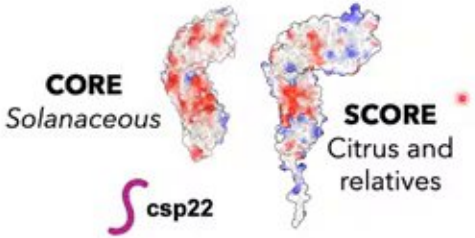
Fine-tunied models are more accurate than generalized approaches → Mamp-ml performs poorly to completely new receptor-epitope interactions and thus, needs to be generalized in the future

Mamp-ml accelerates *in silico* screening of new receptor homologs & epitope variants → Integration with generative module could accelerate engineering improved receptors



Engineer *de novo* receptors

Mamp-ml may enable discovery of convergent evolved receptors to a shared target epitope



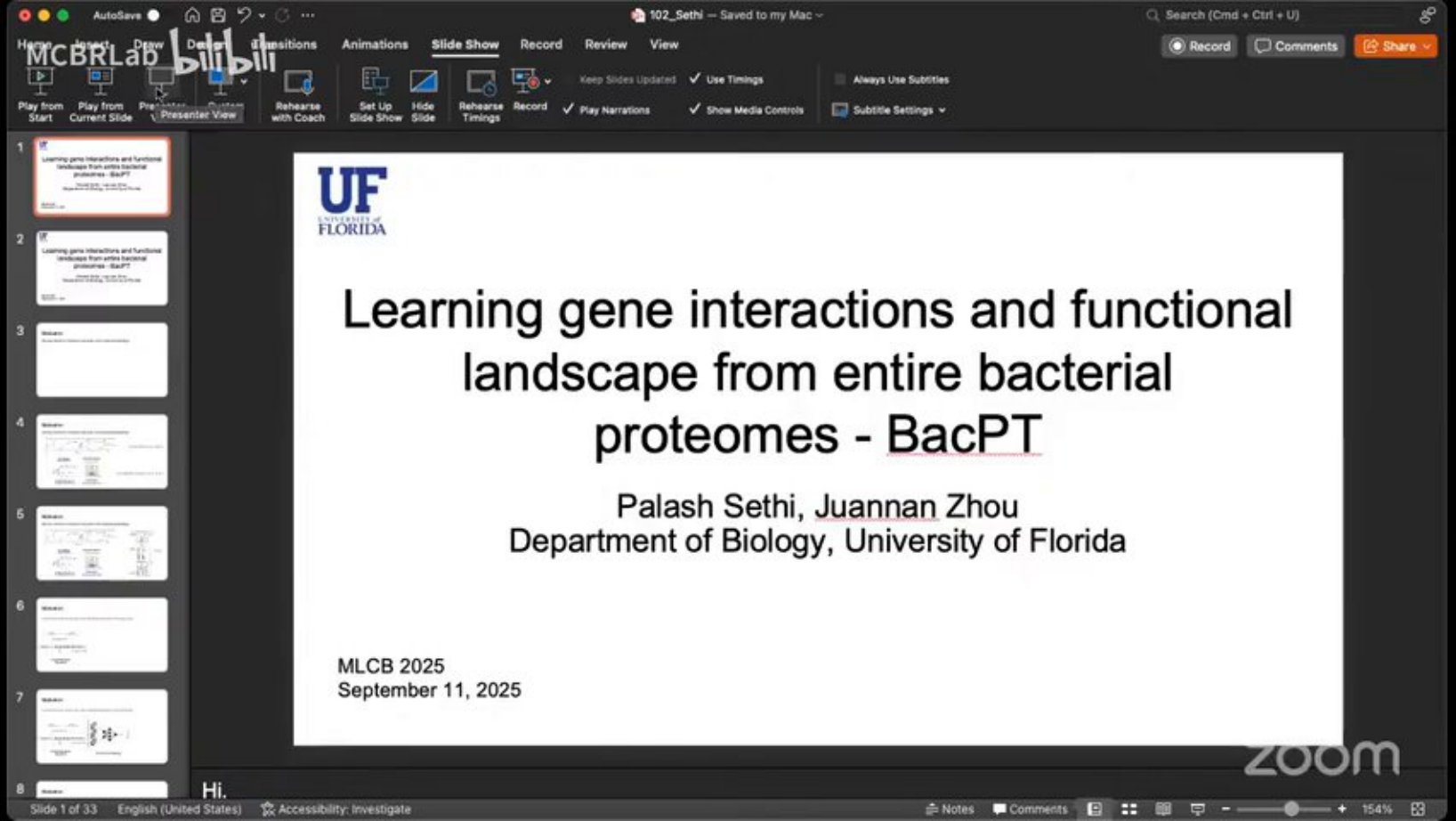
bioRxiv

Available in GitHub + Google Colab

MCBRLab bilibili

NYGC Events

zoom



End Show | Use Slide Show

00:04

2:44 PM

Learning gene interactions and functional landscape from entire bacterial proteomes - BacPT

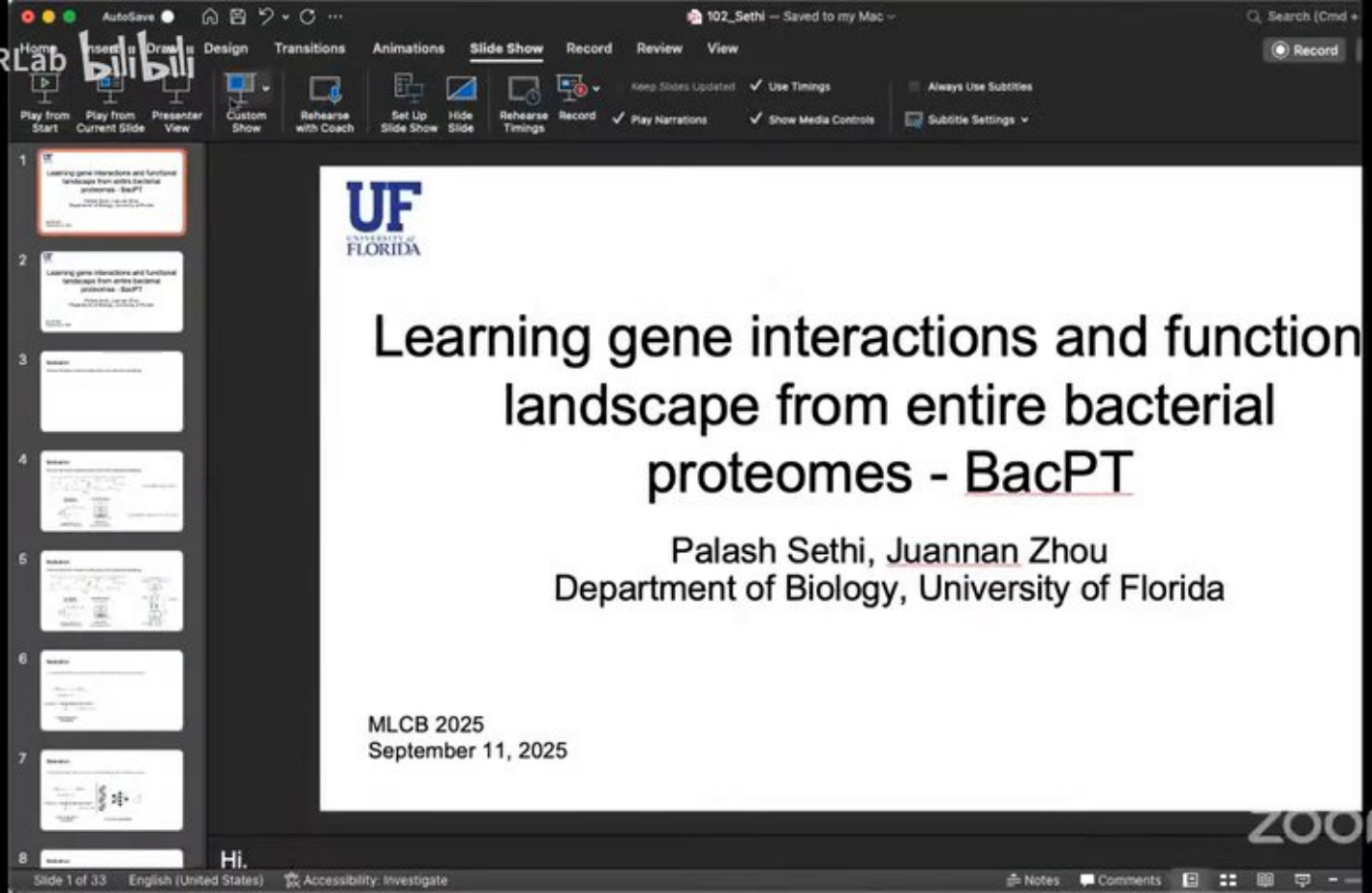
Palash Sethi, Juannan Zhou
Department of Biology, University of Florida

MLCB 2025
September 11, 2025

Slide 1 of 33

Next slide

Hi.


UF
UNIVERSITY of FLORIDA
Learning gene interactions and function landscape from entire bacterial proteomes - BacPT
Palash Sethi, Juannan Zhou
Department of Biology, University of Florida
MLCB 2025
September 11, 2025
Hi.

MCBRLab bilibili

NYGC Events

zoom

Learning gene interactions and functional landscape from entire bacterial proteomes - BacPT

Palash Sethi, Juannan Zhou
Department of Biology, University of Florida

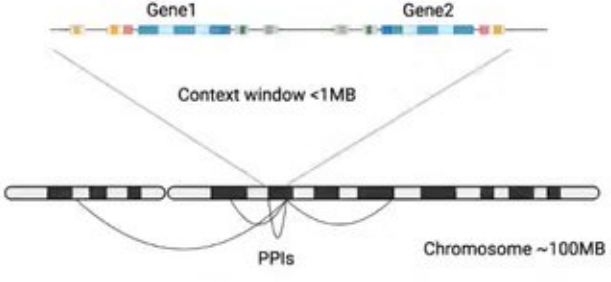
MLCB 2025
September 11, 2025

Motivation

Genome foundation models provide useful local contextual embeddings

Motivation

Contextualized whole-genome embeddings are underexplored



Long-range gene interaction

Whole proteome model for bacteria

Bacterial genomes

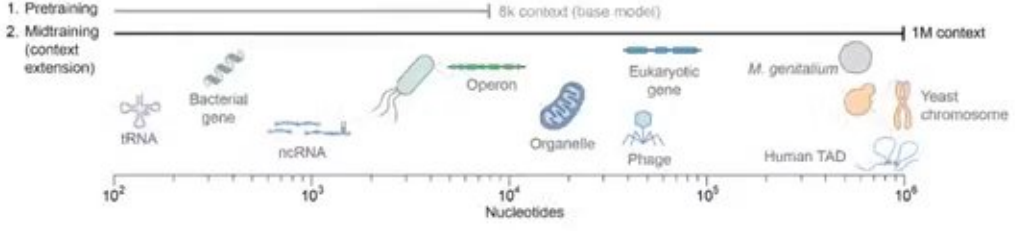
- Small (a few megabases, thousands of genes)
- Minimal intergenic regions (6-14%)
- Large sequence data availability (Refseq, MGnify)
- High-throughput phenotyping (e.g. BacDive database)

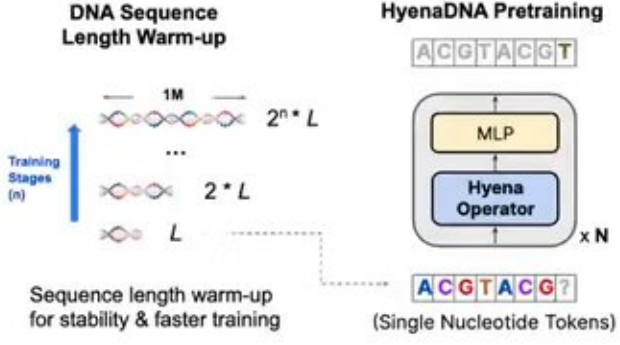
Applications of bacterial proteome foundation models

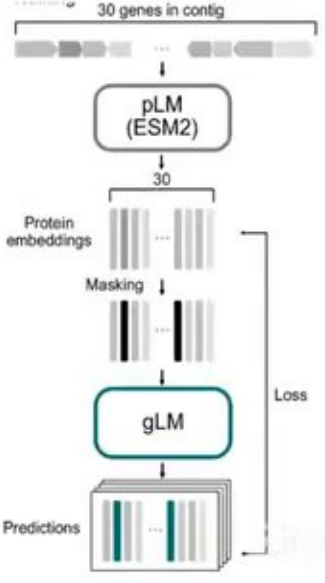
- Genome annotation
- Antibiotic resistance prediction
- Natural product discovery
- Microbiome studies

Motivation

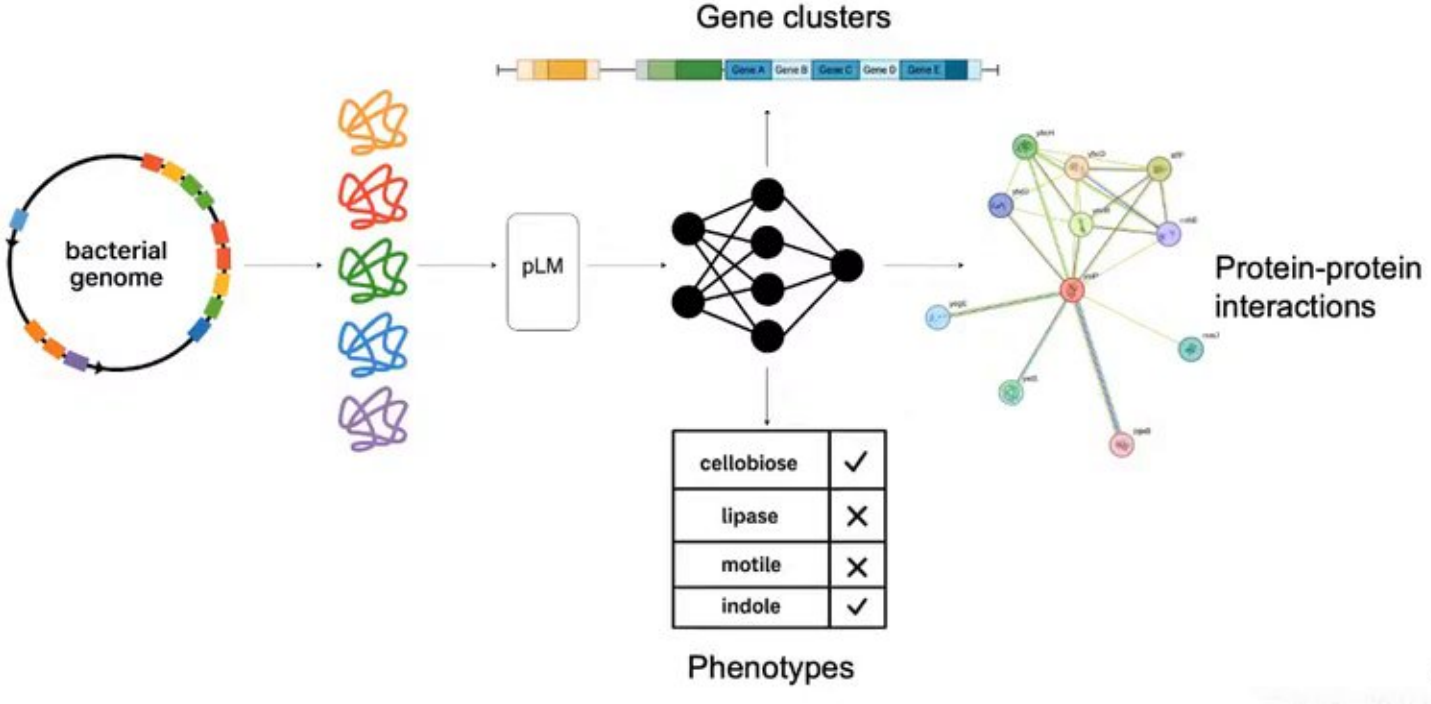
Currency models are limited in their context lengths





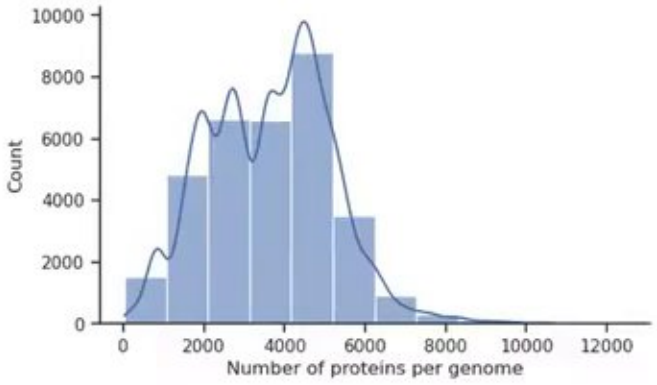


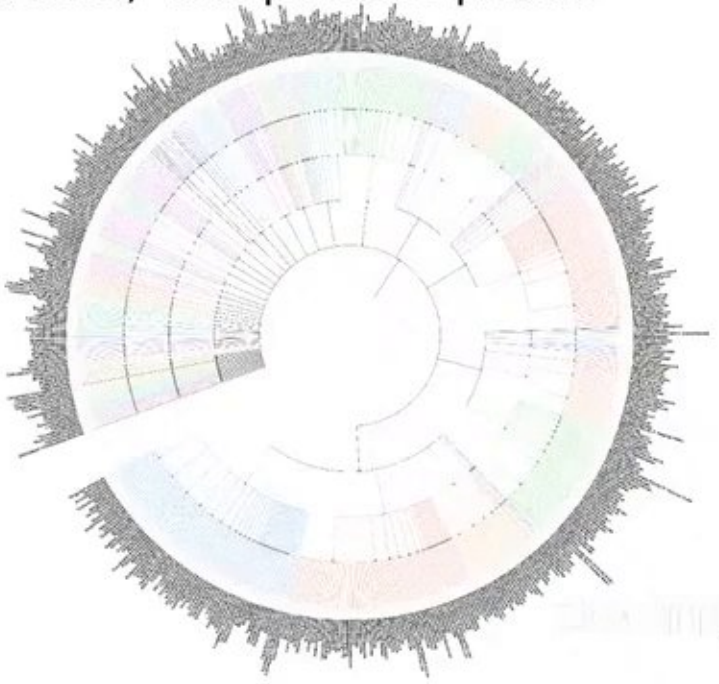
BacPT


Gene clusters
Gene A
Gene B
Gene C
Gene D
Gene E
bacterial genome
pLM
Protein-protein interactions
cellobiose
lipase
motile
indole
Phenotypes

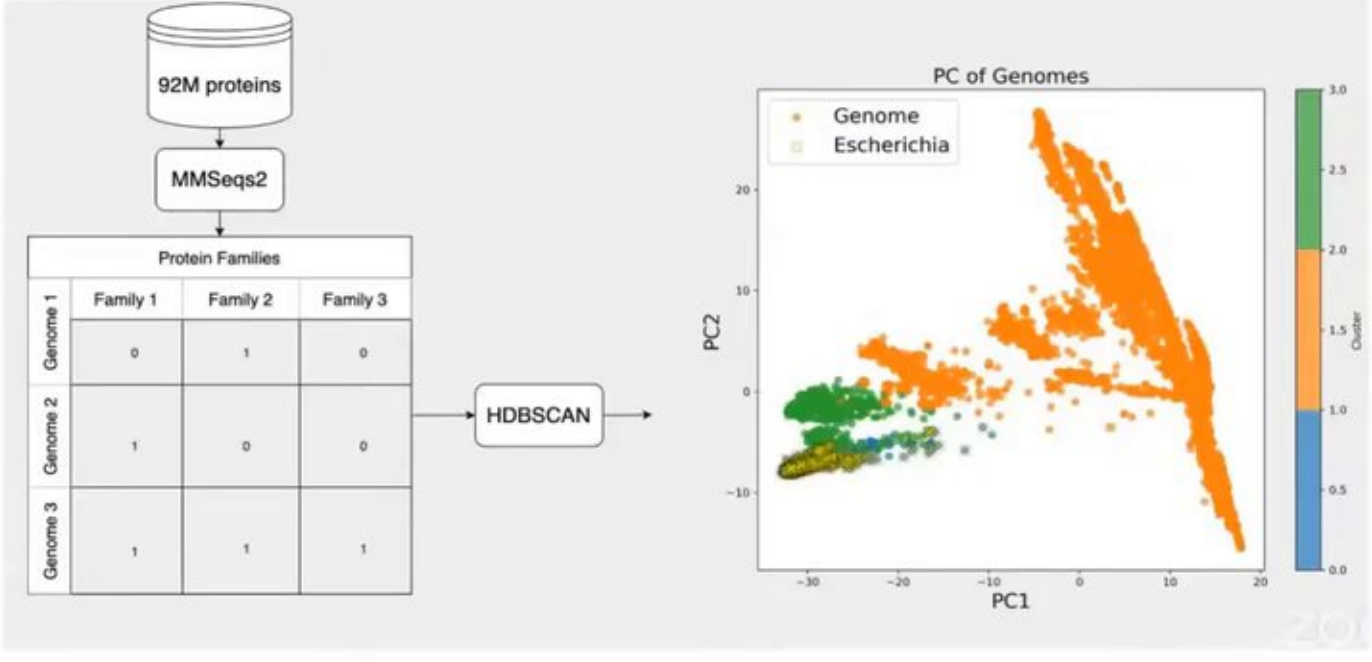
Methods - data

Trained on 33,140 bacterial proteomes from NCBI, ~92M protein sequences



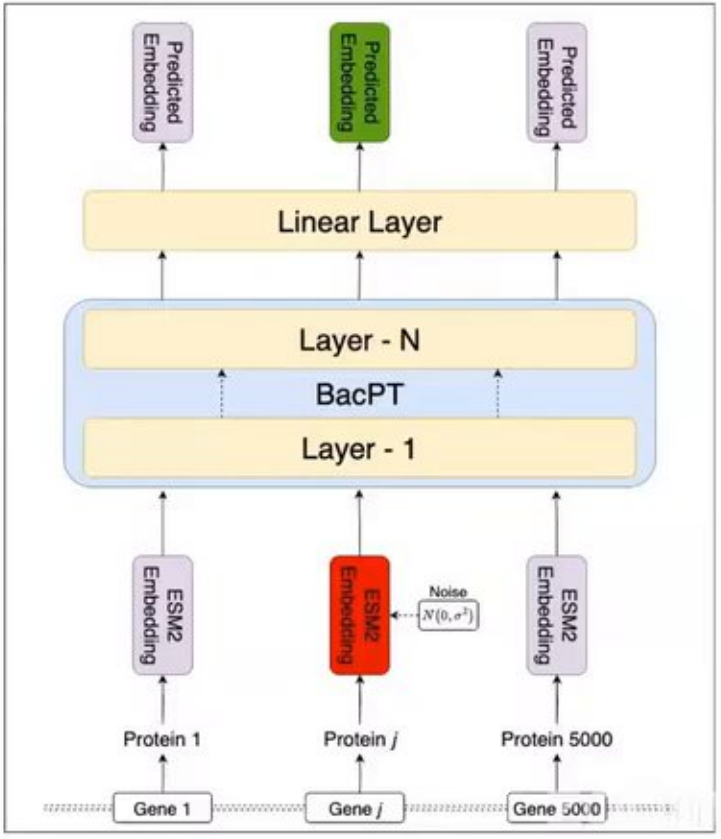


Methods - train/test split



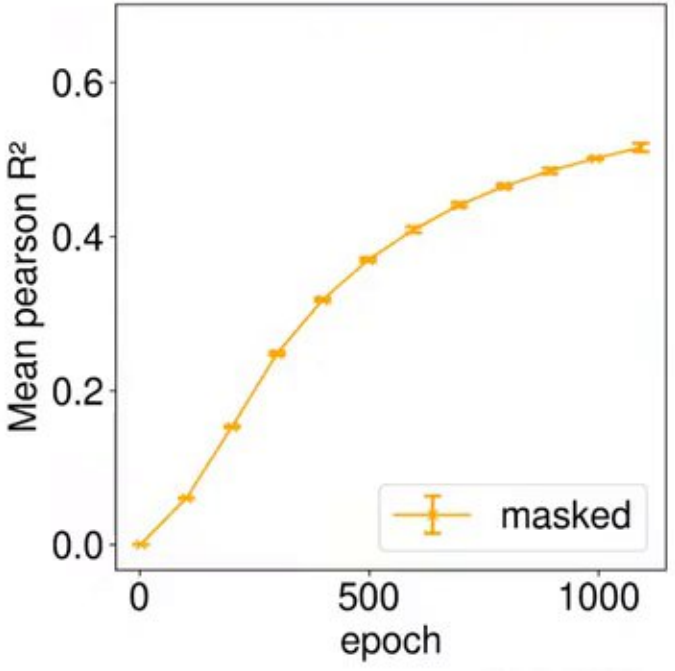
Methods - model architecture

- RoBERTa with rotary position embeddings
- 10 layers, 5 attention heads
- ESM2 - 35M model for input
- 5000 context length
- Trained via progressive noising with **MSE** loss to reconstruct input ESM embeddings
- 1100 epochs over 12 days on 16 NVIDIA A100 GPUs



Performance metrics - 1

- Correlation between predicted protein embeddings and ESM embeddings on test genomes

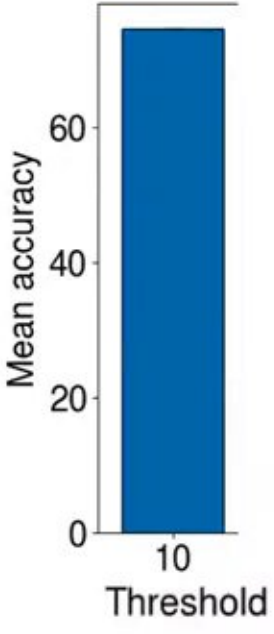


Performance metrics - 2

Metrics	BacPT	gLM
Pseudo Accuracy (%)	63.4	59.2
Absolute Accuracy (%)	78	71.9

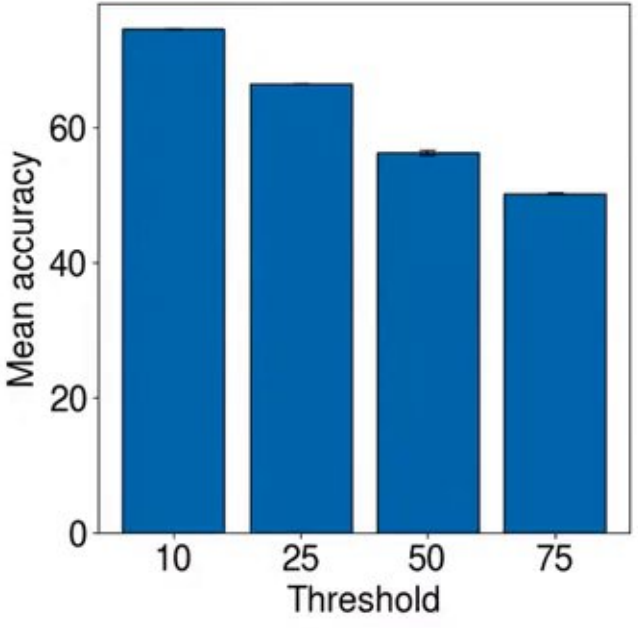
Performance metrics - 3

- Mean accuracy of BacPT at predicting the correct gene cluster for masked proteins, based on 10,000 gene cluster identified in training data



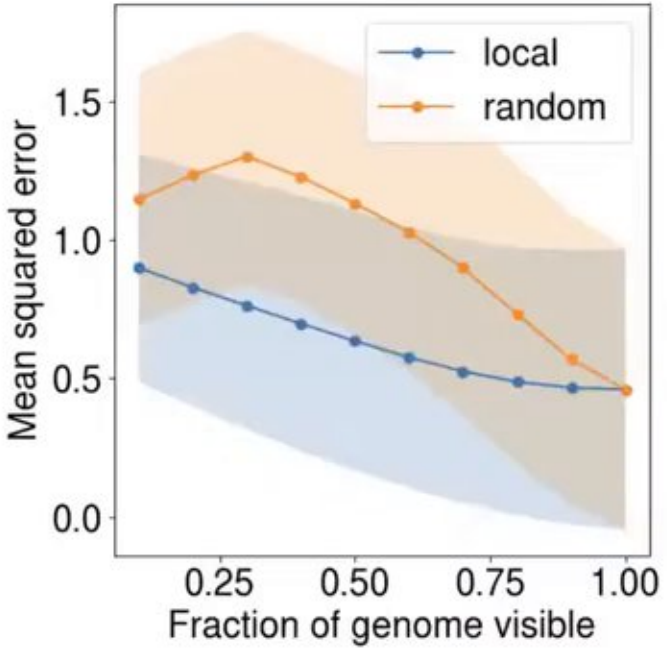
Performance metrics - 3

- Mean accuracy of BacPT at predicting the correct gene cluster for masked proteins, based on 10,000 gene cluster identified in training data



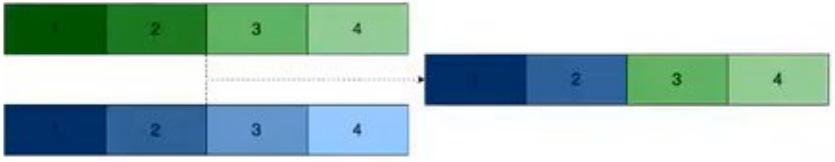
Performance metrics - 4

- Predictions for masked proteins improve when larger proportions of genomes are visible



Performance metric - 5

Can BacPT capture genome integrity?
Yes, linear probes can classify natural versus chimeric genomes

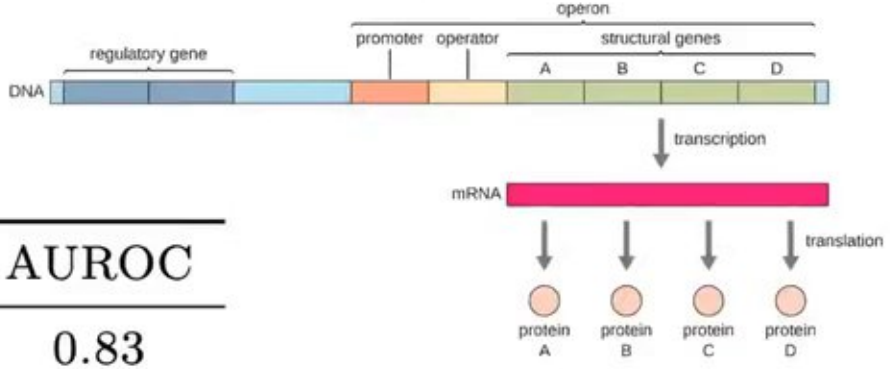


Model	Accuracy	AUROC
BacPT	0.77	0.82
ESM2	0.47	0.53

Applications

1. Operons
2. Protein-protein interactions
3. Whole-proteome representations

BacPT can identify operons

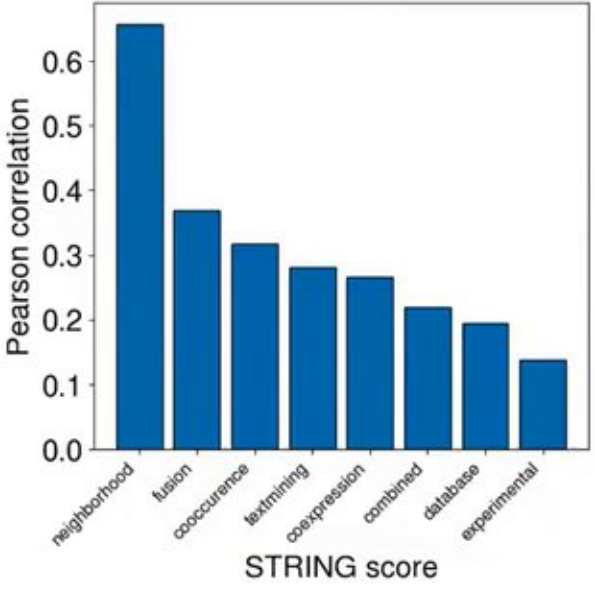


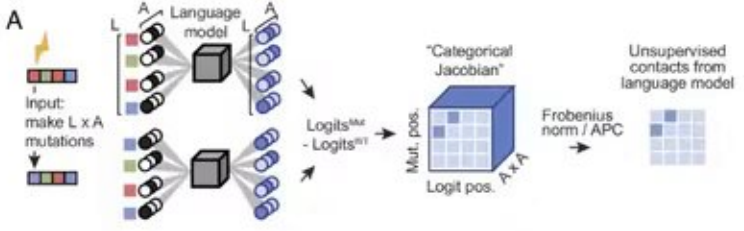
Model	Accuracy	AUROC
BacPT	0.75	0.83
ESM2	0.66	0.75

Linear probing of operon embeddings from BacPT outperform ESM2

BacPT learns protein-protein and other gene interactions

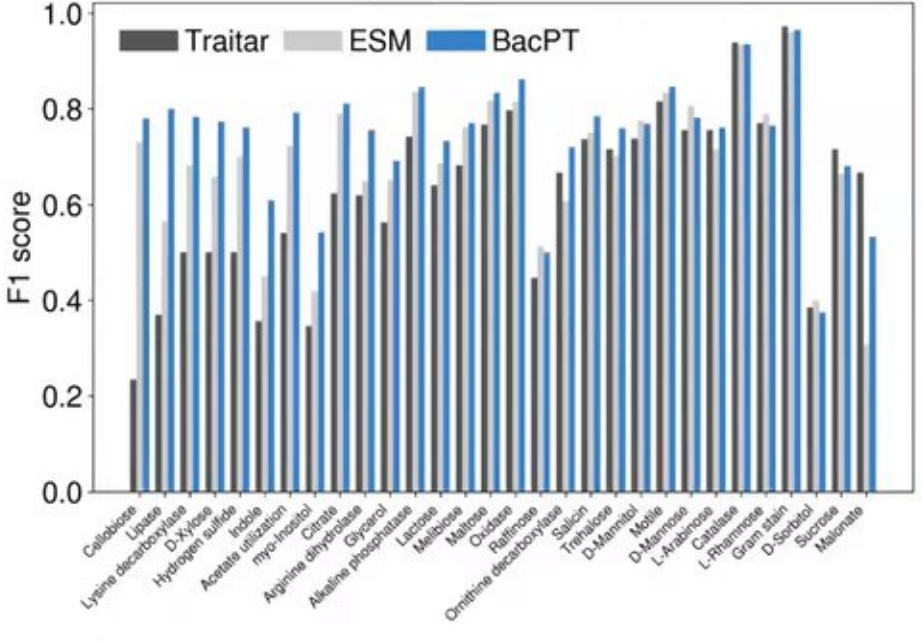
Using jacobian

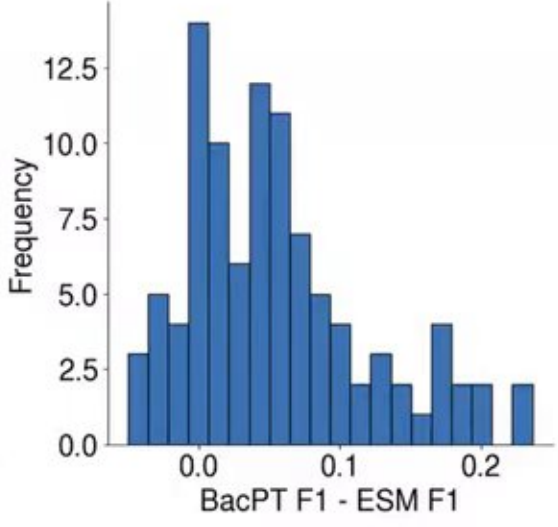




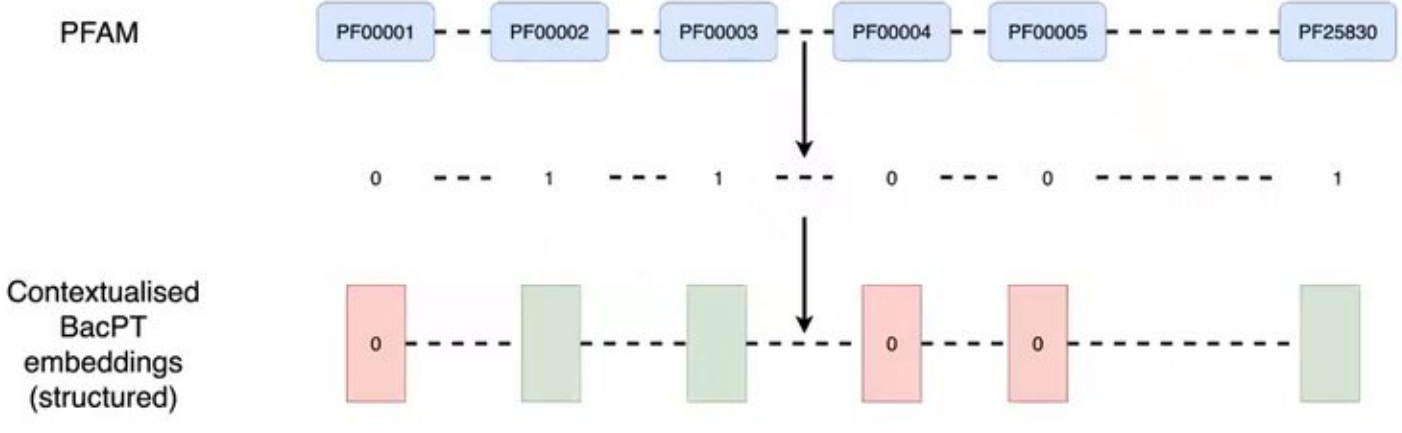
(Zhang et al, 2024)

BacPT predicts organism-level traits

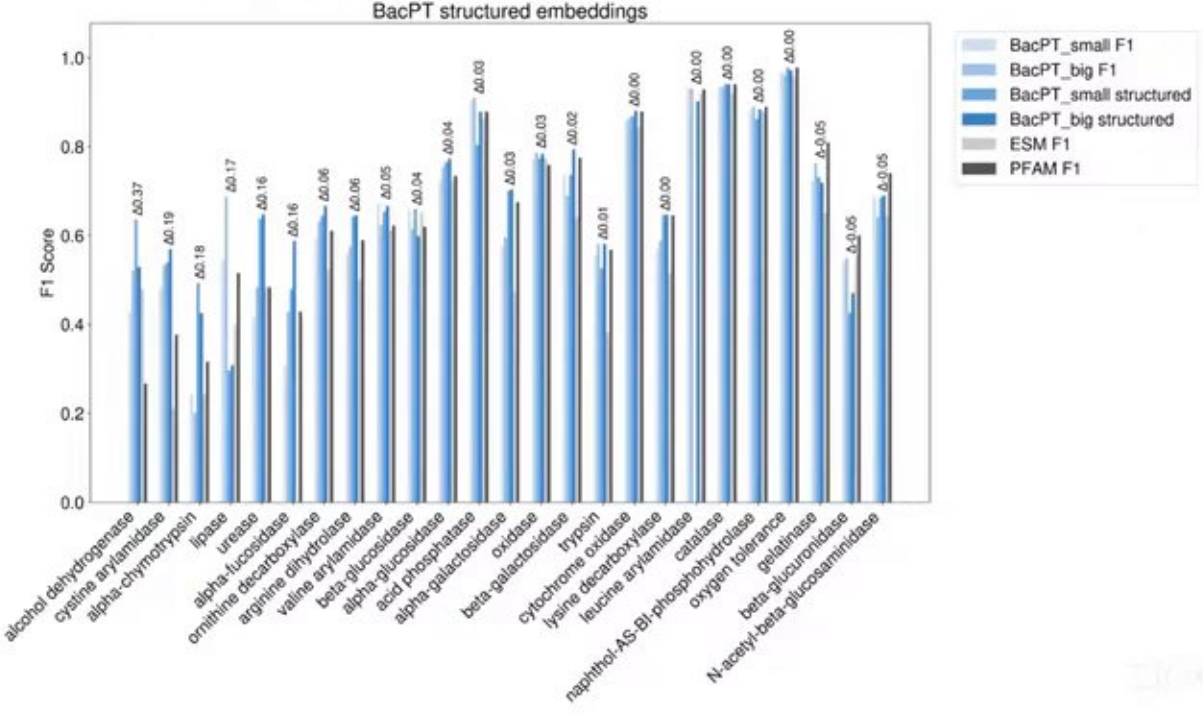




BacPT structural embeddings



BacPT predicts organism-level traits



Conclusion

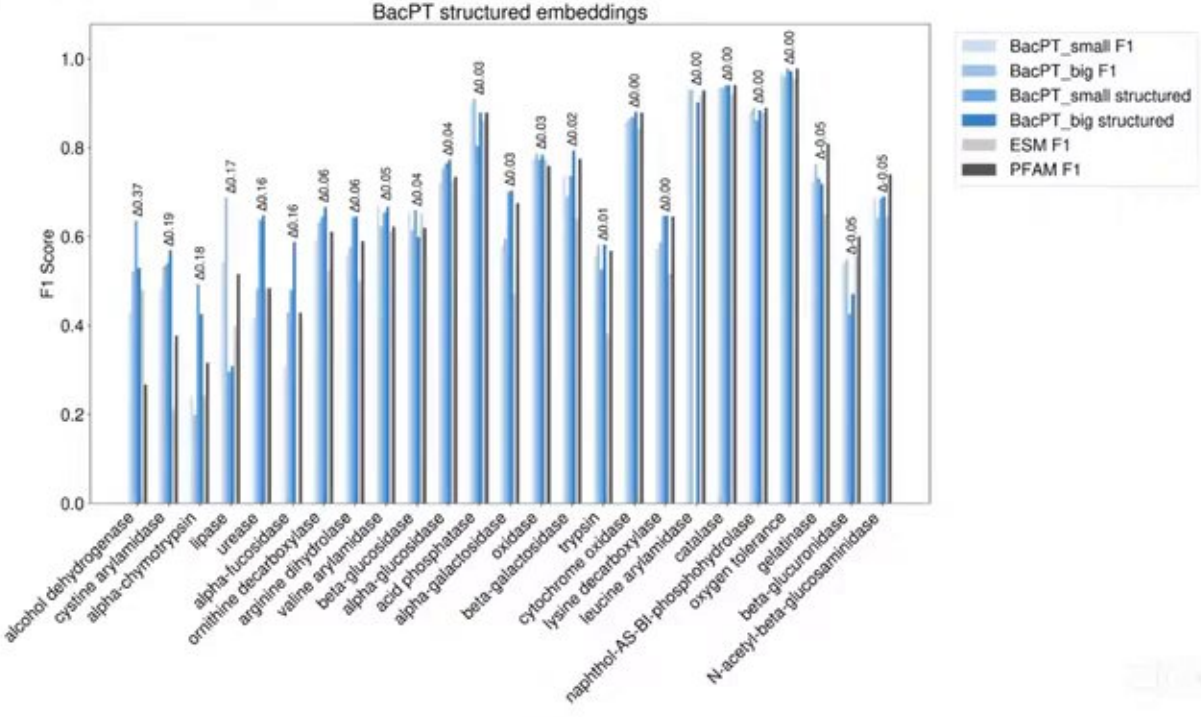
Whole proteome bacterial embeddings via BacPT

- Captures whole-proteome context length
- Identifies higher order genomic structures such as operons
- Learns protein-protein interaction
- Predicts organism-level complex traits

Future Work

- Gene clusters
- Microbial interaction
- Attribution maps for complex traits

BacPT predicts organism-level traits



Conclusion

Whole proteome bacterial embeddings via BacPT

- Captures whole-proteome context length
- Identifies higher order genomic structures such as operons
- Learns protein-protein interaction
- Predicts organism-level complex traits

Future Work

- Gene clusters
- Microbial interaction
- Attribution maps for complex traits

Acknowledgements

Zhou lab @ UF

National Institutes of Health

National Institute of General Medical Sciences

UF | Biodiversity Institute
UNIVERSITY of FLORIDA

- Marc G Chevrette @ UW Madison

Questions?

Learning gene interactions and functional landscapes from entire bacterial proteomes

Palash Sethi[1], Marc G Chevrette[2], and Juannan Zhou*[1]

[1]Department of Biology, University of Florida, Gainesville, FL, 32611
[2]Department of Microbiology & Cell Science, University of Florida, Gainesville, FL, USA

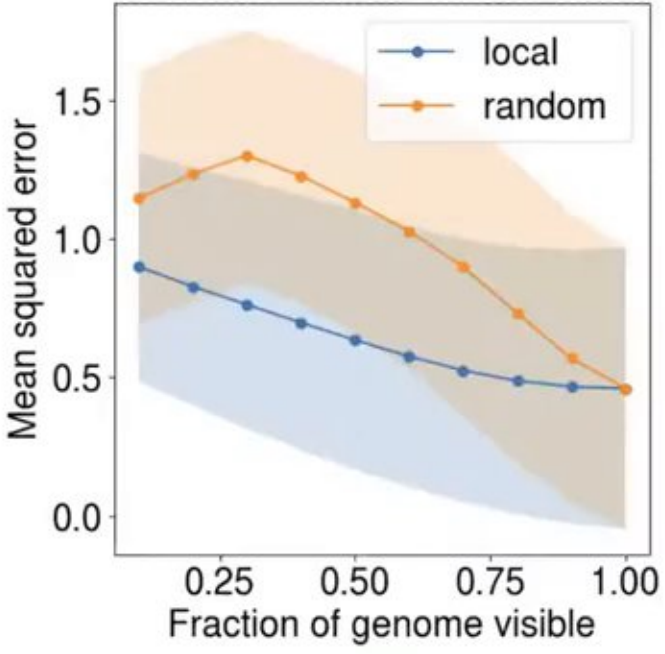
palash.sethi@ufl.edu

@PalashSethi6

SCAN ME

Performance metrics - 4

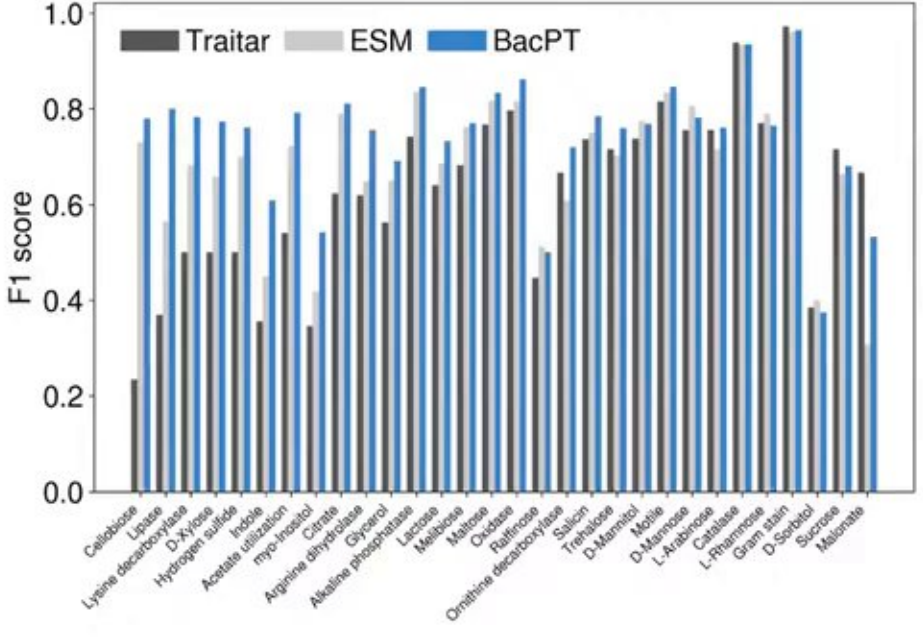
- Predictions for masked proteins improve when larger proportions of genomes are visible

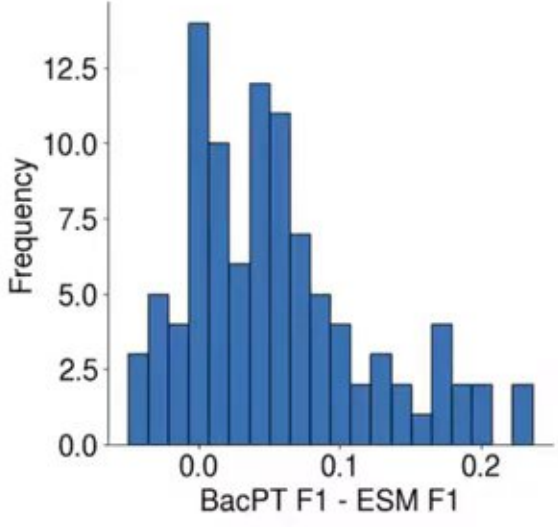


Applications

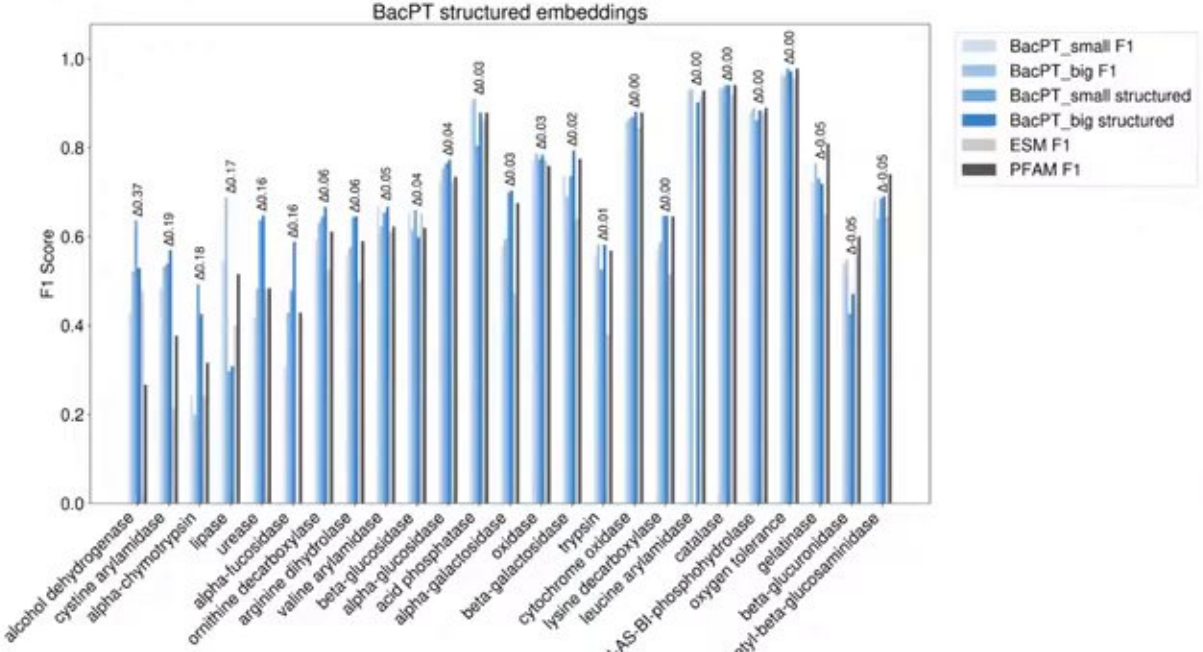
1. Operons
2. Protein-protein interactions
3. Whole-proteome representations

BacPT predicts organism-level traits

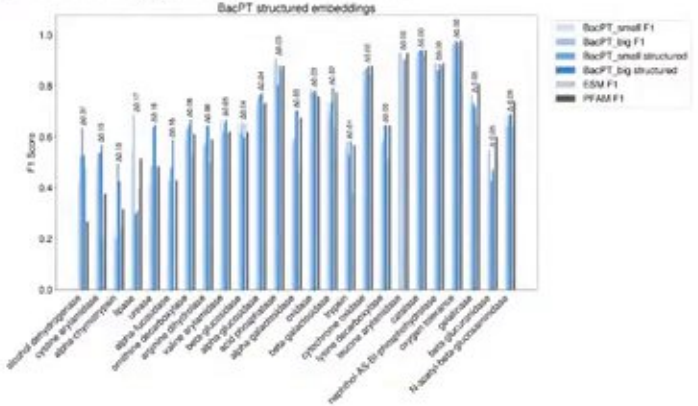




BacPT predicts organism-level traits



BacPT predicts organism-level traits



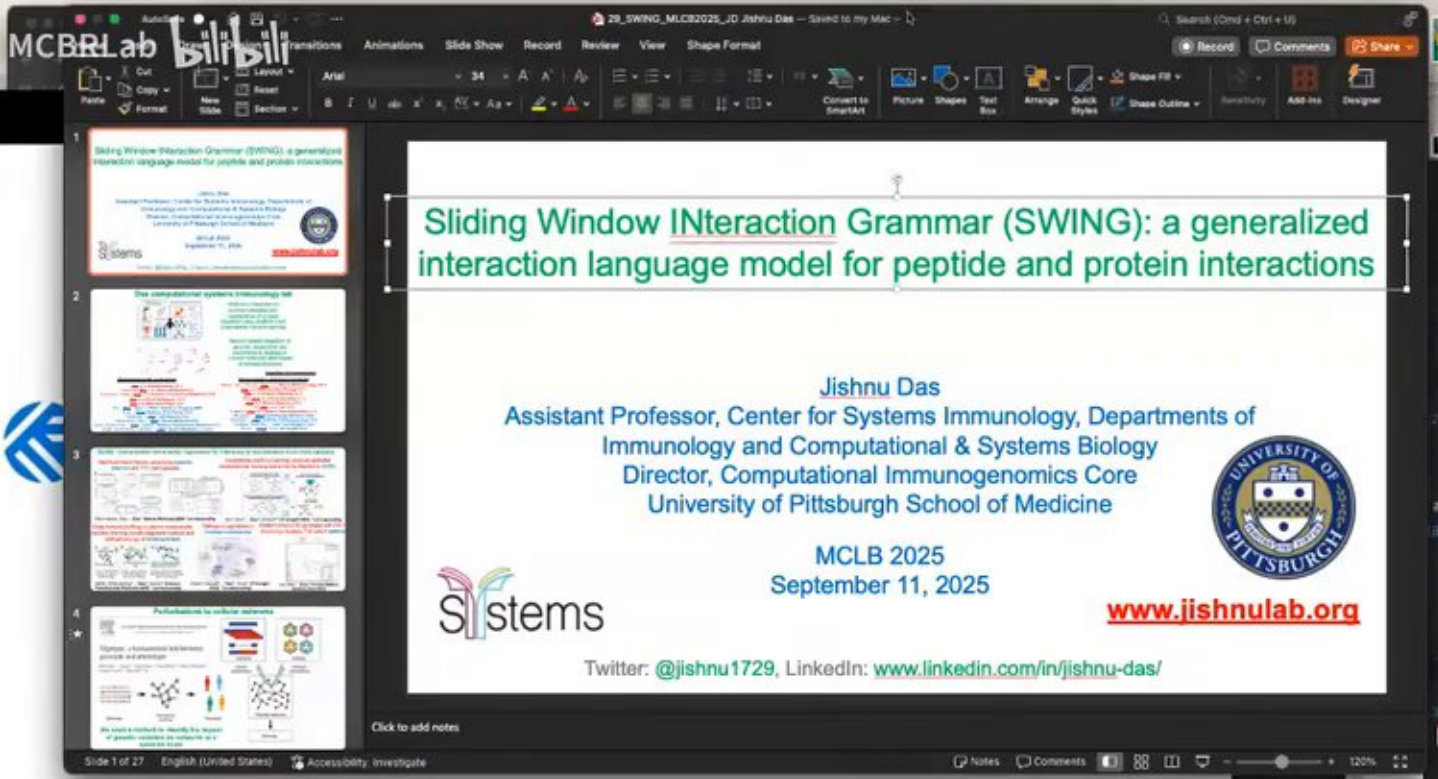
28

- We observe that BacPT's structured embedding outperform the pfam presence-absence method on almost all the traits.

MCBRLab bilibili

NYGC Events

zoom



Restarting at 1.30pm ET

mlcb.org for schedule

Sliding Window INteraction Grammar (SWING): a generalized interaction language model for peptide and protein interactions

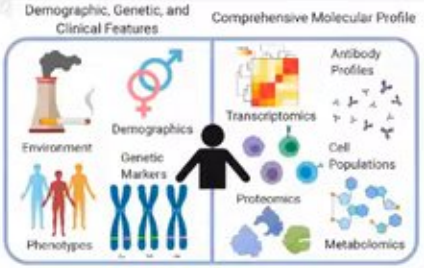
Jishnu Das
Assistant Professor, Center for Systems Immunology, Departments of Immunology and Computational & Systems Biology
Director, Computational Immunogenomics Core
University of Pittsburgh School of Medicine

MCLB 2025
September 11, 2025

www.jishnulab.org

Twitter: @jishnu1729, LinkedIn: www.linkedin.com/in/jishnu-das/

Das computational systems immunology lab



Multi-omic integration to uncover correlates and mechanisms of immune regulation using predictive and interpretable machine learning

Network-based integration of genomic, epigenomic and transcriptomic datasets to uncover molecular phenotypes of immune disorders

*=co-first, ^=corresponding

Representative ML publications

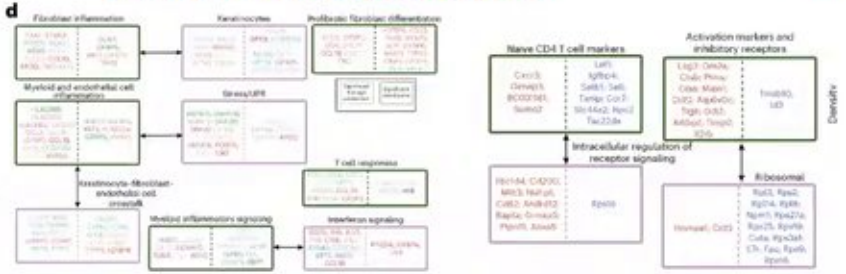
Das et al ***Bioinformatics*** 2012
Ackerman, Das et al ***Nature Medicine*** 2018
Suscovich*, Fallon*, Das* et al ***Science Translational Medicine*** 2020
Das et al ***PLoS Pathogens*** 2020
Das et al ***Med (Cell Press)*** 2021
Wu,…Das^, Sperry^, Billiar^ ***Annals of Surgery 2022***
Bing, … Das^ ***Patterns (Cell Press)*** 2022
Pedireddy, … Das^ ***Cell Reports*** 2022
Rahimikollu, Xiao, … Das^ ***Nature Methods*** 2024
Saha, Chakraborty, … Das^, Sarkar^, ***Science Translational Medicine*** 2024
Steen, Malchenko, Chhibbar, … Das^ ***Nature Methods*** (in press)

Representative networks publications

Wang*, Wei*, Thijssen*, Das* et al ***Nature Biotechnology*** 2012
Das et al ***BMC Systems Biology*** 2012
Das et al ***Science Signaling*** 2013
Wei*, Das* et al ***PLoS Genetics*** 2014
Das et al ***Human Mutation*** 2014
Vo*, Das* et al ***Cell*** 2016
Fragoza*, Das* et al ***Nature Communications*** 2019
Ningappa, … Das^ ***Cell Reports Medicine*** 2022
Berkowitz, … Das^ ***Arthritis & Rheumatology*** 2023
Chhibbar, Guha Roy … Das^ ***Cell Reports*** 2024
Rosen, … Das^, Torok ***JCI Insight*** 2025

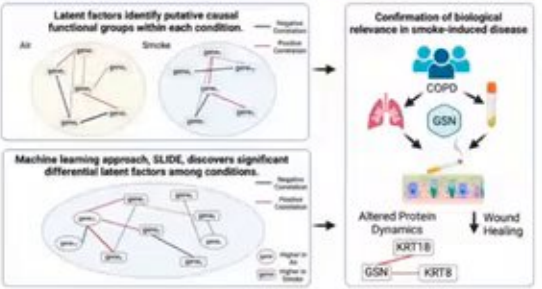
SLIDE - interpretable latent factor regression for inference of mechanisms from omic datasets

Significant latent factors underlying systemic sclerosis and T1D pathogenesis



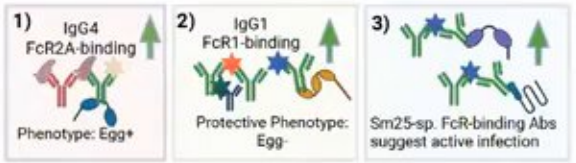
Rahimikollu, Xiao ... **Das^** ***Nature Methods 2024*** ***^corresponding***

Interpretable machine learning uncovers epithelial transcriptional rewiring and a role for Gelsolin in COPD



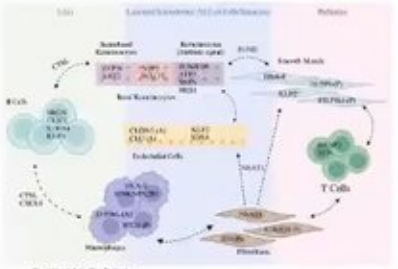
Sui*, Xiao* ... **Das^**, Kliment^ ***JCI Insight 2024; ^corresponding***

Deep humoral profiling coupled to interpretable machine learning unveils diagnostic markers and pathophysiology of schistosomiasis



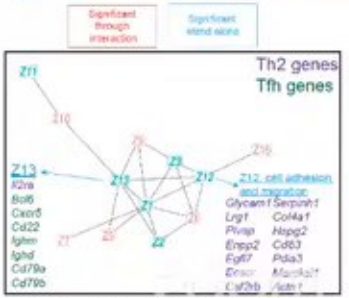
Saha*, Chakraborty* ... **Das^**, Sarkar^ ***Science Translational Medicine 2024; ^corresponding***

Pathogenic signatures in localized scleroderma



Rosen*, Sanyal* ... **Das^**, Torok^ ***JCI Insight 2025; ^corresponding***

Spatial niches of Il2 synergize with Il10 to drive lung migratory Th2 cells in asthma



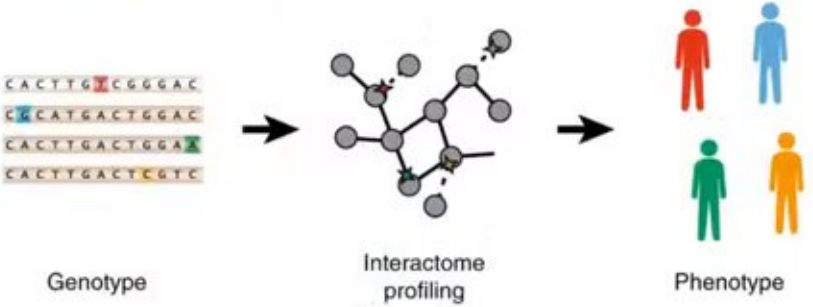
He, Xiao ... **Das,** Poholek ***Nature Immunology 2024***

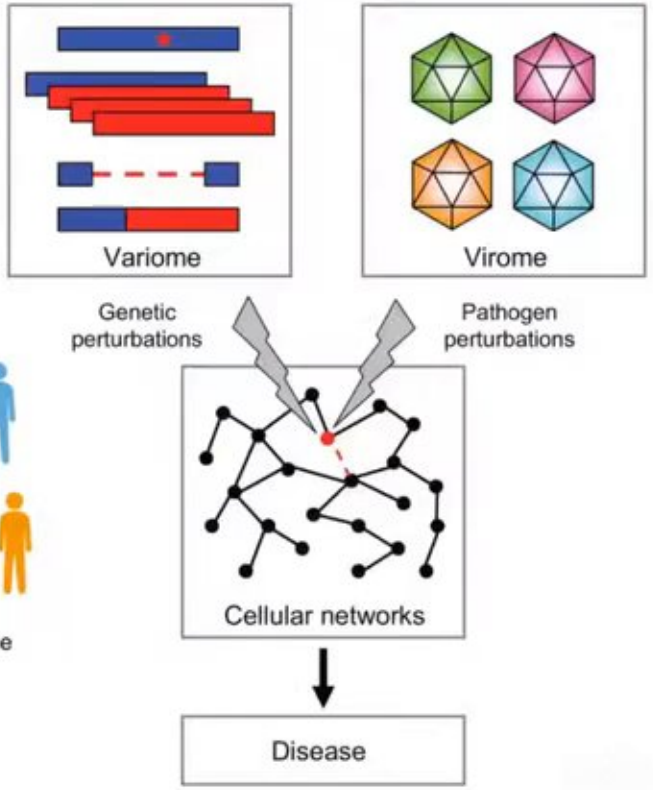
Perturbations to cellular networks

Current Opinion in Genetics & Development
Volume 23, Issue 6, December 2013, Pages 649-657

Edgotype: a fundamental link between genotype and phenotype

Nidhi Sahni [1 2], Song Yi [1 2], Quan Zhong [1 2], Noor Jailkhani [1 2], Benoit Charloteaux [1 2], Michael E Cusick [1 2], Marc Vidal [1 2]





Sliding Window Interaction Grammar (SWING): a generalized interaction language model for peptide and protein interactions

nature methods

Explore content ~ About the journal ~ Publish with us ~

nature > nature methods > articles > article

Article | Open access | Published: 28 July 2025

Sliding Window Interaction Grammar (SWING): a generalized interaction language model for peptide and protein interactions

Jane C. Siwek, Alisa A. Omelchenko, Prabal Chhibbar, Sanya Arshad, AnnaElaine Rosengart, Iliyan Nazarali, Akash Patel, Kiran Nazarali, Javad Rahimikollu, Jeremy S. Tilstra, Mark J. Shlomchik, David R. Koes, Alok V. Joglekar ✉ & Jishnu Das ✉

Nature Methods **22**, 1707–1719 (2025) | Cite this article

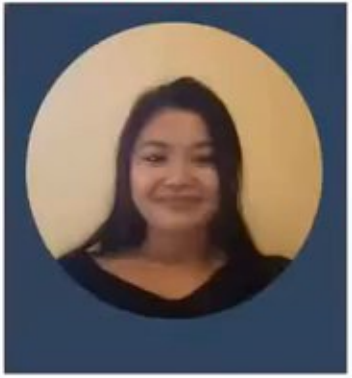
15k Accesses | **1** Citations | **21** Altmetric | Metrics

J Siwek*, A Omelchenko*, P Chhibbar*… **J Das**^ ***Nature Methods*** 2025

Key contributors

Alok Joglekar

Alisa Omelchenko



Jane Siwek

Prabal Chhibbar

Interaction language models to decode peptide & protein interactions

nature biotechnology

Explore content ~ About the journal ~ Publish with us ~

nature > nature biotechnology > articles > article

Article | Published: 26 January 2023

Large language models generate functional protein sequences across diverse families

nature biotechnology

Explore content ~ About the journal ~ Publish with us ~

nature > nature biotechnology > articles > article

Article | Open access | Published: 24 April 2023

Efficient evolution of human antibodies from general protein language models

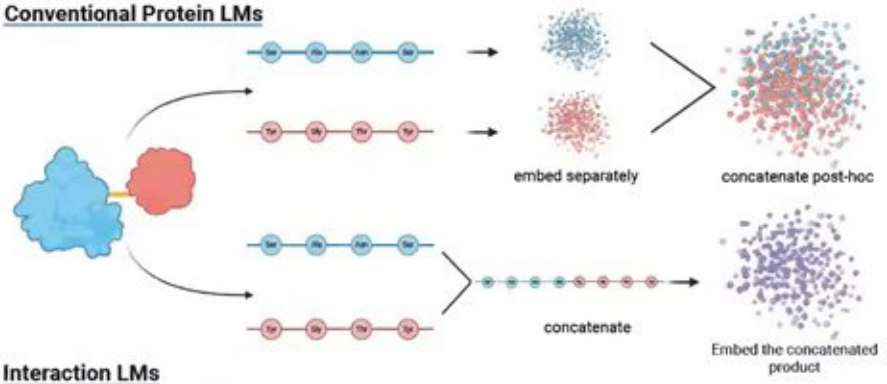
nature machine intelligence

Explore content ~ About the journal ~ Publish with us ~

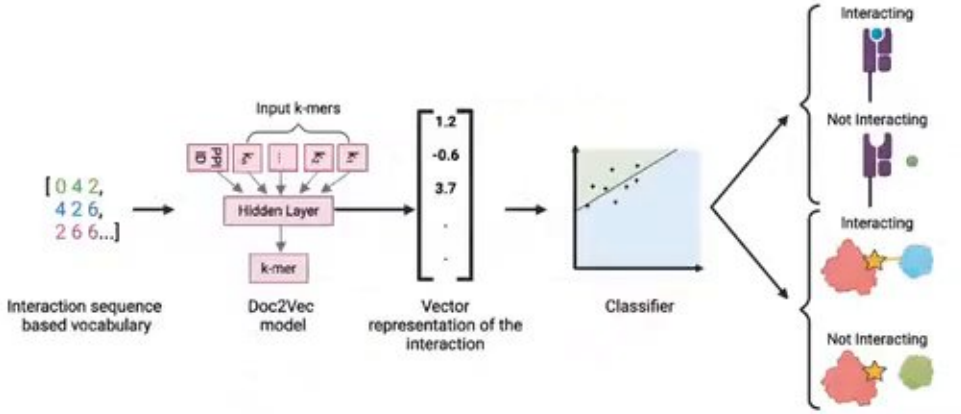
nature > nature machine intelligence > analyses > article

Analysis | Published: 21 March 2022

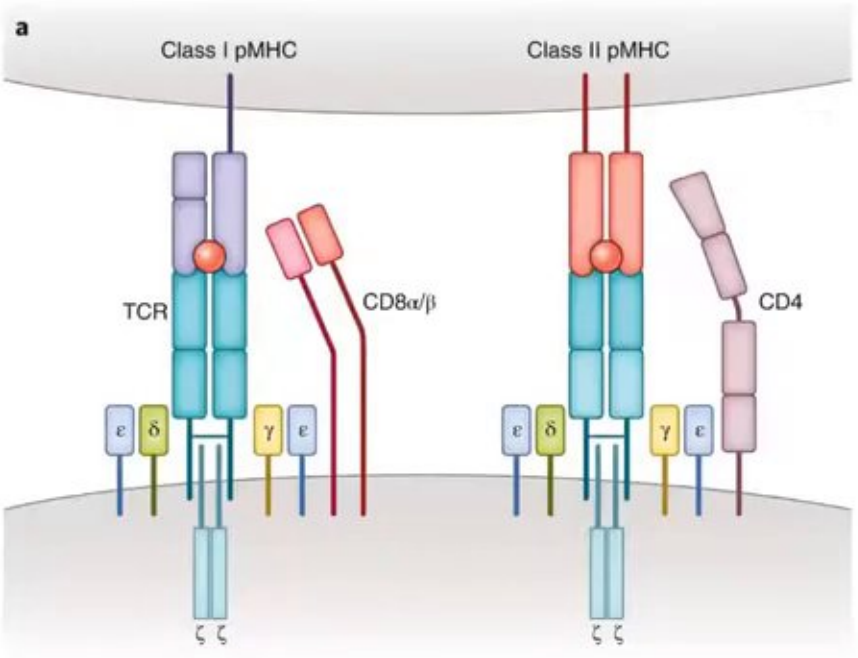
Learning functional properties of proteins with language models



SWING: Sliding Window INteraction Grammar – learning the language of peptide and protein interactions



Predicting pMHC binding is a well-studied and important problem



Joglekar and Li ***Nature Methods*** 2021

Immunogenetics. 2013 Sep; 65(9): 655–665.
Published online 2013 Jun 18. doi: 10.1007/s00251-013-0714-9

MHCcluster, a method for functional clustering of MHC molecules

Martin Thomsen, Claus Lundegaard, Søren Buus, Ole Lund, and Morten Nielsen

JOURNAL ARTICLE

NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data

Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, Morten Nielsen
Author Notes

Nucleic Acids Research, Volume 48, Issue W1, 02 July 2020, Pages W449–W454,
https://doi.org/10.1093/nar/gkaa379

nature biotechnology

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

nature › nature biotechnology › brief communications › article

Brief Communication | Published: 14 October 2019

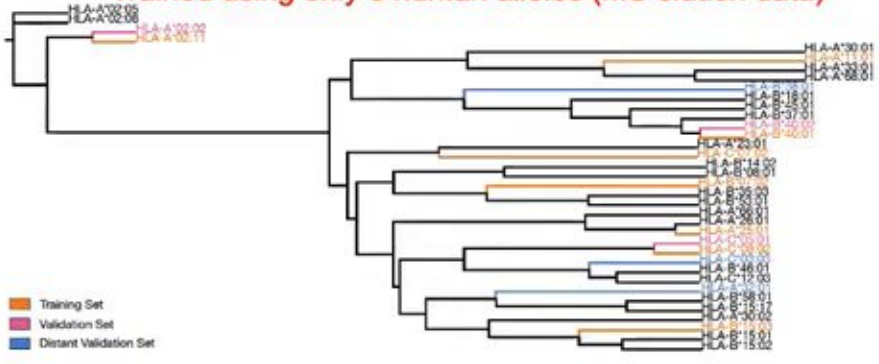
Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes

Julien Racle, Justine Michaux, Georg Alexander Rockinger, Marion Arnaud, Sara Bobisse, Chloé Chong, Philippe Guillaume, George Coukos, Alexandre Harari, Camilla Jandus, Michal Bassani-Sternberg & David Gfeller

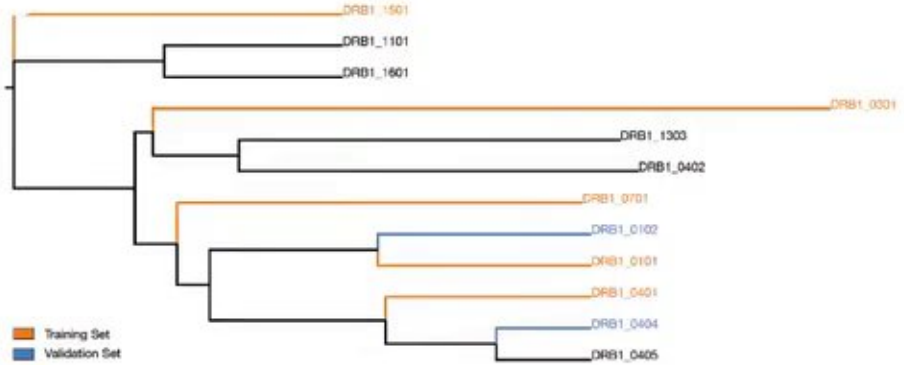
Nature Biotechnology 37, 1283–1286 (2019) | Cite this article

Learning the language of Class I pMHC interactions

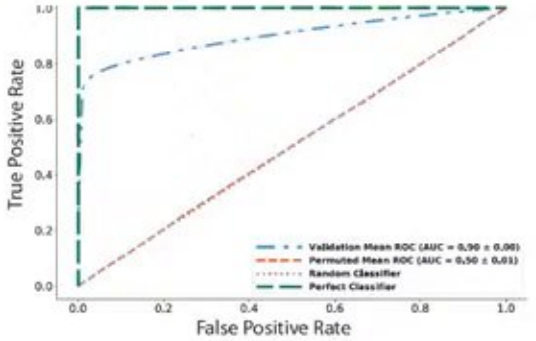
Trained using only 8 human alleles (MS elution data)



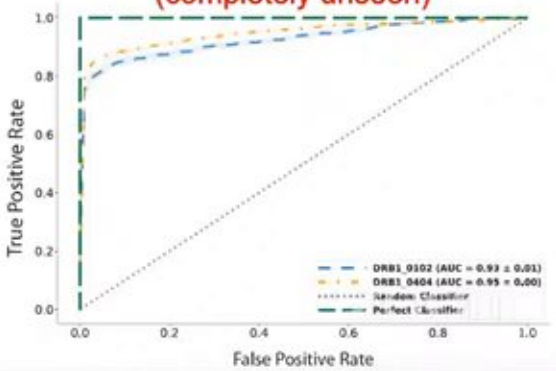
Learning the language of Class II pMHC interactions



K-fold cross-validation on the training set



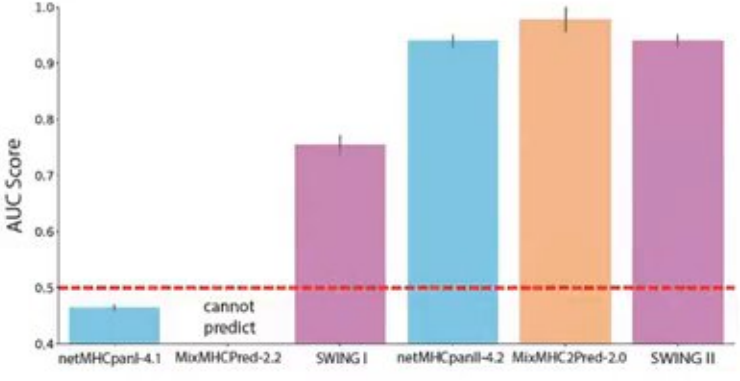
Cross prediction on the validation set (completely unseen)



Generalizing the language of all pMHC interactions

Generalizing the language of all pMHC interactions

No existing method can go from Class I -> Class II

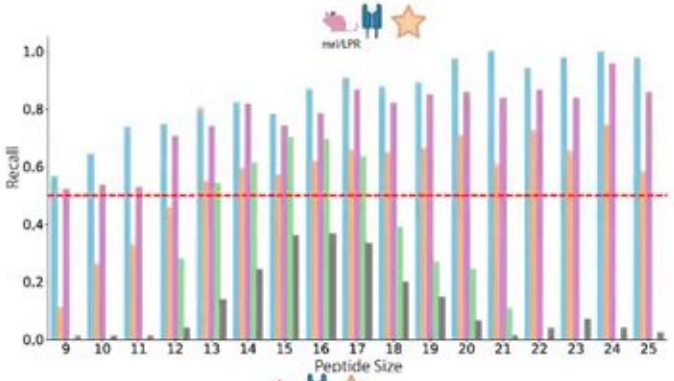


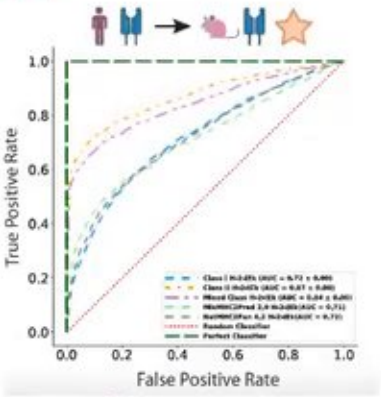
SWING can make de-novo predictions of pMHC interactions in a murine model of SLE

Few-shot learning – prediction of de-novo pMHC binding across organisms

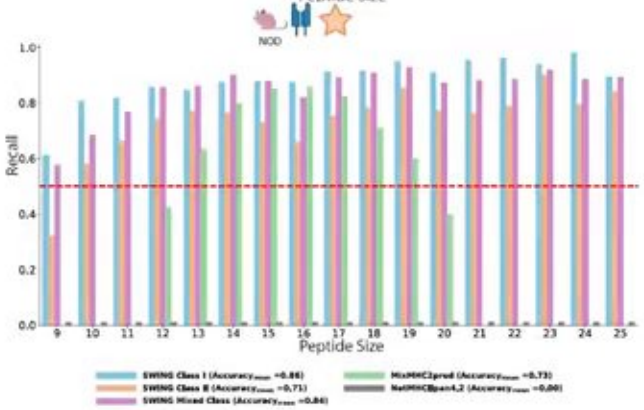
SWING can make de-novo predictions of pMHC interactions in a murine model of SLE

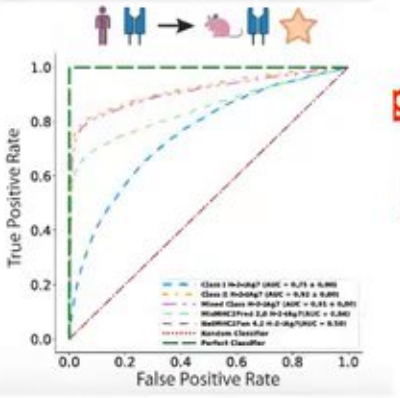
SWING model learnt using human data highly predictive of H-2-IEk binding in a murine model of SLE (Mrl/lpr)





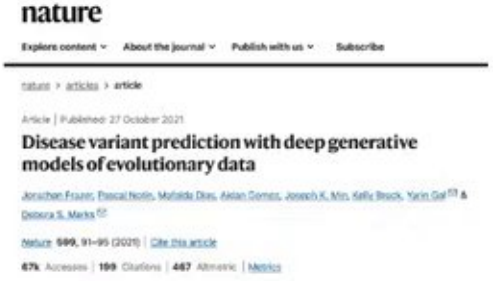
SWING model learnt using human data highly predictive of IAg7 binding in a murine model of T1D (NOD)





Few-shot learning – prediction of de-novo pMHC binding across organisms

Modern variant effect predictors learn complex relationships but not the nuanced language of protein interactions


nature

Explore content ~ About the journal ~ Publish with us ~ Subscribe

nature > articles > article

Article | Published: 27 October 2021

Disease variant prediction with deep generative models of evolutionary data

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal & Debora S. Marks

Nature 599, 91–95 (2021) | Cite this article

67k Accesses | 199 Citations | 467 Altmetric | Metrics


nature genetics

Explore content ~ About the journal ~ Publish with us ~

nature > nature genetics > articles > article

Article | Open access | Published: 10 August 2023

Genome-wide prediction of disease variant effects with a deep protein language model

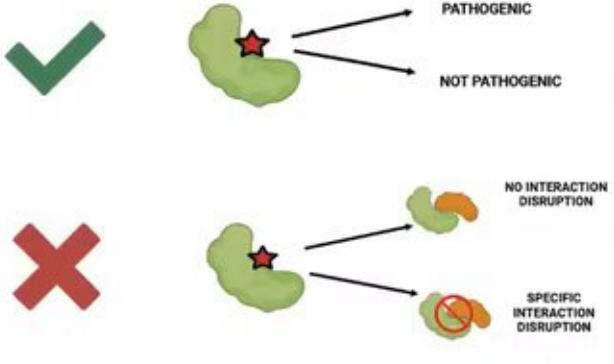
Nadav Brandes, Grant Goldman, Charlotte H. Wang, Chun Jimmie Ye & Vasilis Ntranos

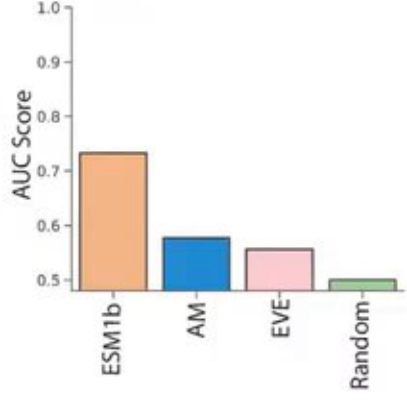
Nature Genetics 55, 1512–1522 (2023) | Cite this article

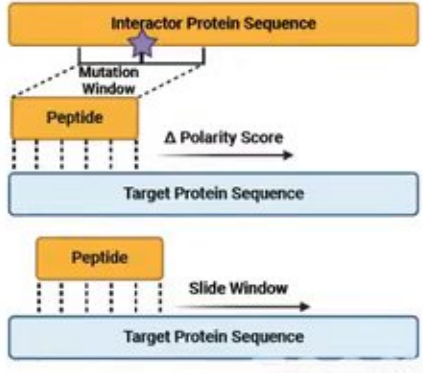
58k Accesses | 28 Citations | 180 Altmetric | Metrics


Science

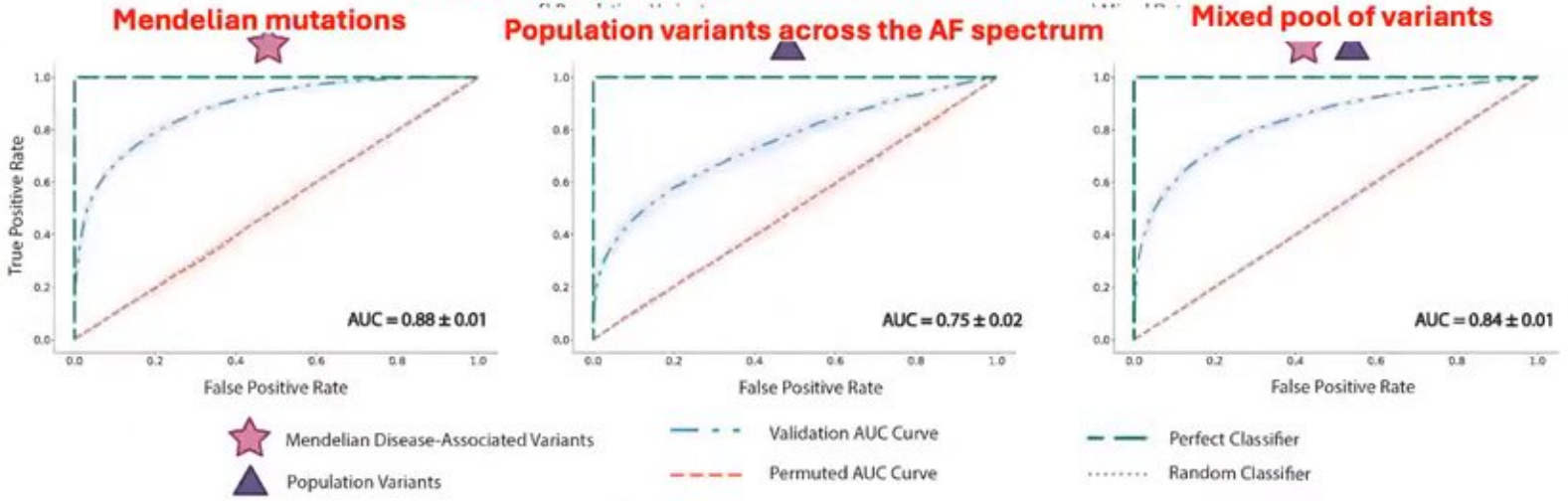
Accurate proteome-wide missense variant effect prediction with AlphaMissense

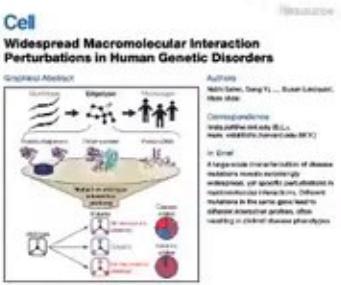






SWING can predict disruptions of protein interactions by different classes of missense mutations



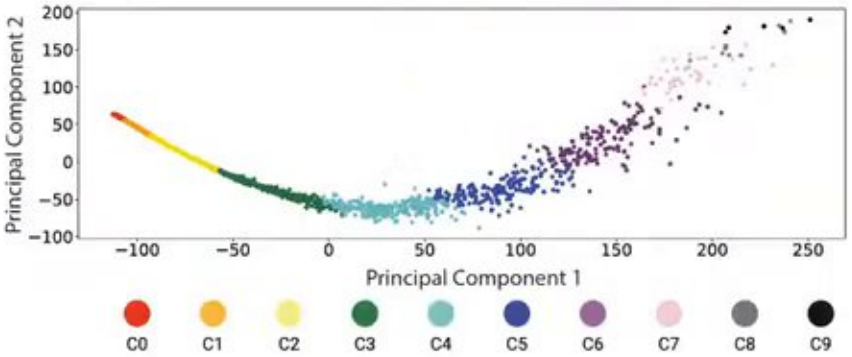

Cell

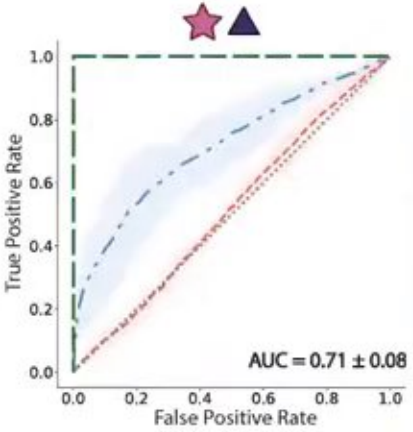
Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders

nature communications

Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations

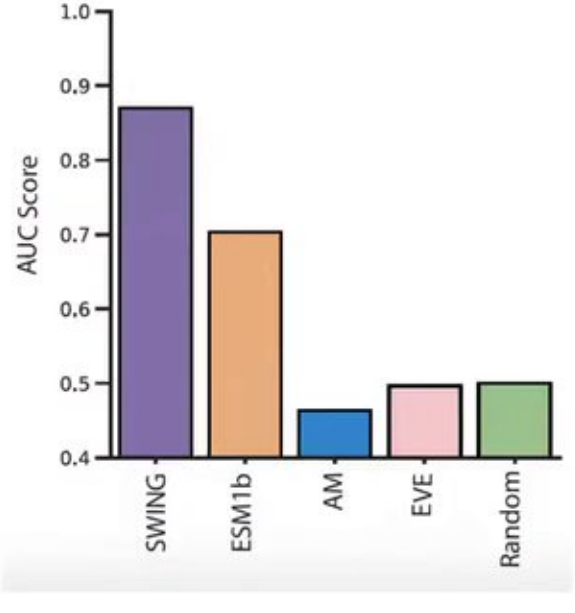
SWING learns relationships well beyond sequence similarity



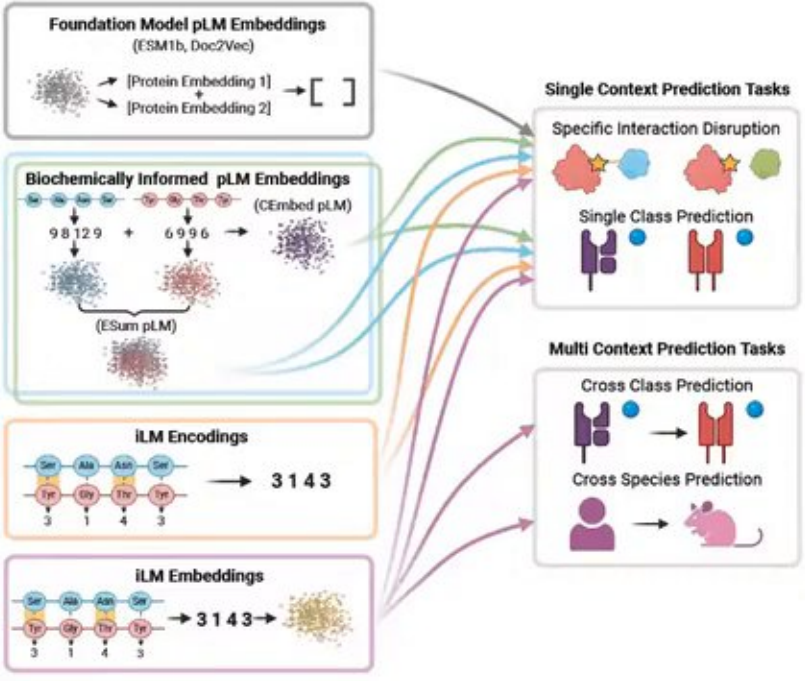


	Validation Clusters	Dropped Clusters	AUC Score
Fold 1	C0, C1	C2, C3	0.67
Fold 2	C2, C3	C1, C4	0.70
Fold 3	C4, C5	C3, C6	0.72
Fold 4	C6, C7	C5, C8	0.78

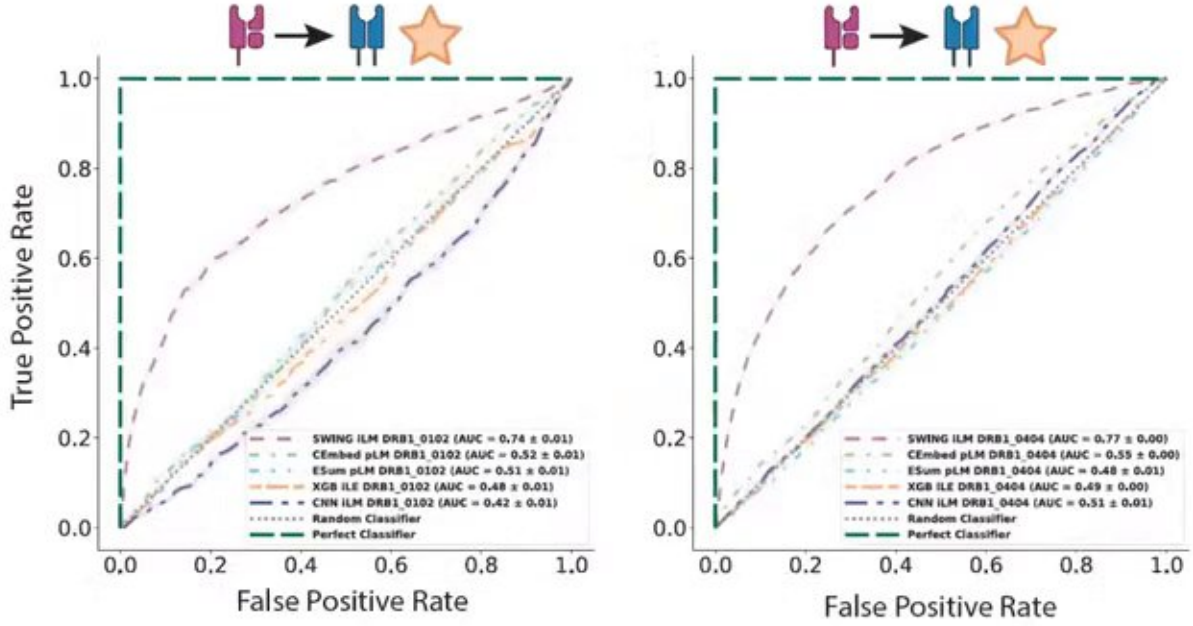
SWING outperforms state-of-the-art VEP methods and uncovers molecular phenotypes missed by these approaches



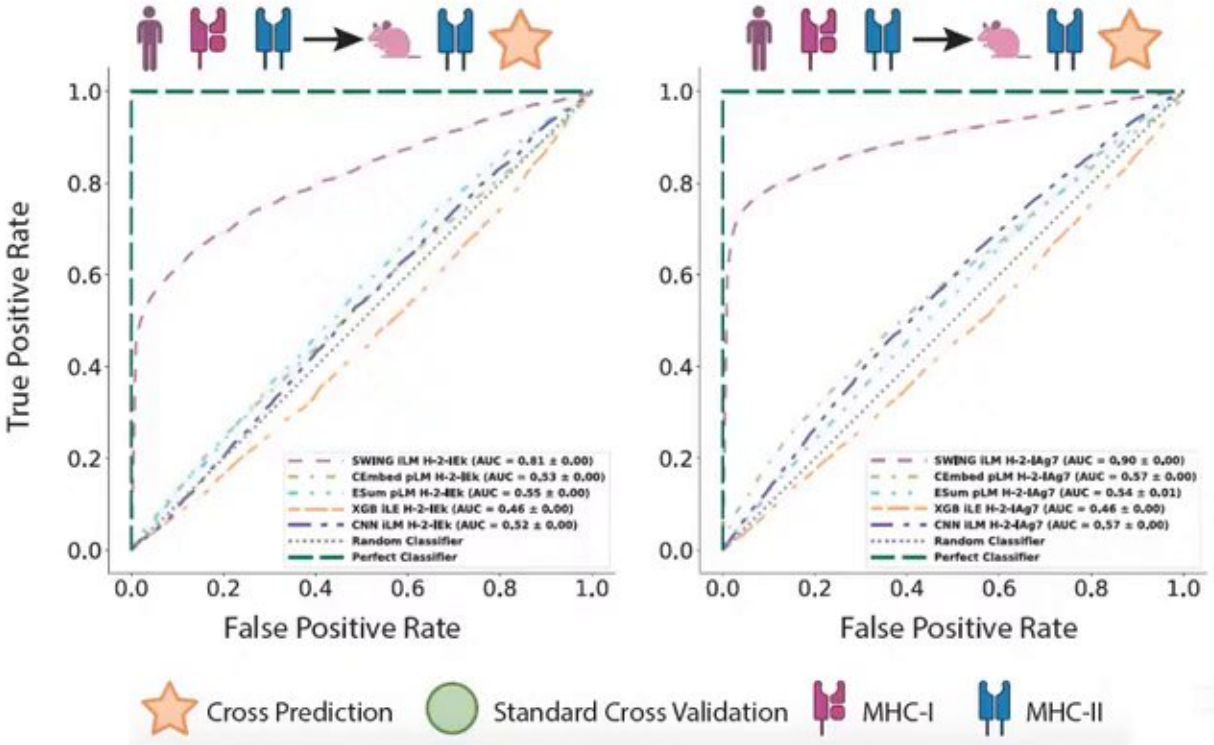
SWING performs better than/as well as related architectures for single context prediction tasks



SWING significantly outperforms related architectures for multi-context prediction tasks and unseen alleles - I



SWING significantly outperforms related architectures for multi-context prediction tasks and unseen alleles - II



Conclusions

	SWING	NetMHCPan	MixMHCPred	AF	ESM1b	EVE
Embeds the interaction	✓	✗	✗	✗	✗	✗
Works across peptide-peptide interaction contexts	✓	✗	✗	✗	✗	✗
Predicts Class I pMHC interactions	✓	✓	✓			
Predicts Class II pMHC interactions	✓	✓	✓			
Predicts Class I and Class II interactions with the same model	✓	✗	✗			
Predicts across MHC Classes	✓	✗	✗			
Generalizes across species while trained on one species	✓	✗	✗			
Predicts variant effects on the phenotype	✓			✓	✗	✓
Predicts edgotype effects on the phenotype	✓			✗	✓	✗

Siwek, Omelchenko, Chhibbar ... **Das *Nature Methods*** 2025

Das Lab

Systems Immunology Faculty Position at Pitt

Interested candidates should send a CV and statement of research interests to Rachel Gottschalk (rachel.gottschalk@pitt.edu) and Jishnu Das (jishnu@pitt.edu).

The Center for Systems Immunology (CSI) and the Department of Immunology at the University of Pittsburgh are seeking to appoint a tenure-track Assistant Professor. Candidates with strong research training and accomplishments in systems and computational immunology and/or synthetic biology and immune engineering will be considered. The position offers a competitive start-up package and dedicated space within a new, highly interactive research building on the University campus that houses faculty in the CSI, the Departments of Immunology and the Hillman Cancer Center.

The candidate will be expected to develop a leading externally funded research group; mentor graduate students including those in the CMU-Pitt Computational Biology program and contribute to graduate teaching within such programs affiliated with the School of Medicine.

The University of Pittsburgh has excellence in Immunology and Computational Biology made possible by a highly collaborative faculty, a diverse and strong trainee pool, and world class core facilities and infrastructure.

Interested candidates should send a CV and statement of research interests to Rachel Gottschalk (rachel.gottschalk@pitt.edu) and Jishnu Das (jishnu@pitt.edu).

Our Funders

4 R01/R01-equivalents grants as PI/MPI

NIAID New Innovators DP2 Award

DP2AI164325 (Role: PI, Active)

NHGRI IGVF Consortium U01

U01HG012041 (Role: MPI, Active)

NIAID Flu Systems Vaccinology R01

R01AI170108 (Role: MPI, Active)

NIAID Flu Systems Vaccinology U01

U01AI179514 (Role: MPI, Active)

Rainin Innovator Award

1. DoD Idea Development Award
2. DoD SL20118

Other Grants

1. NIAID R01AI167711
2. NIAID U19AI181984

NIAID R35HL166219

NIMH R01MH136952

NEI U01EY034711

NCI R01CA2774731

NIAMS P50 CORT P50AR080612-01

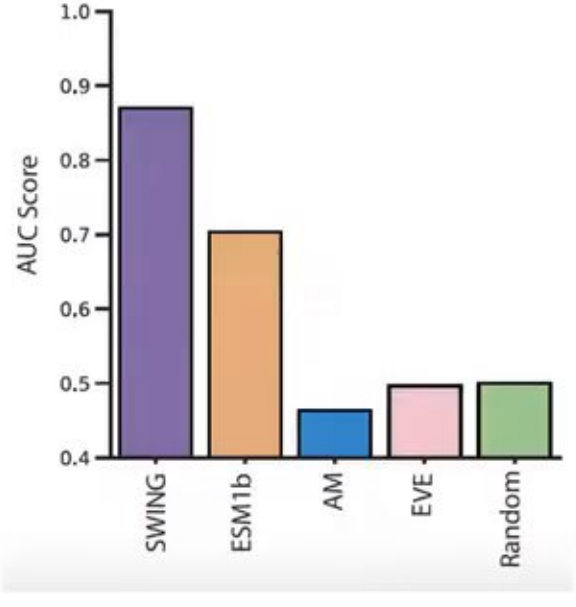
CIHR DCO150GP

jishnu@pitt.edu

www.jishnulab.org

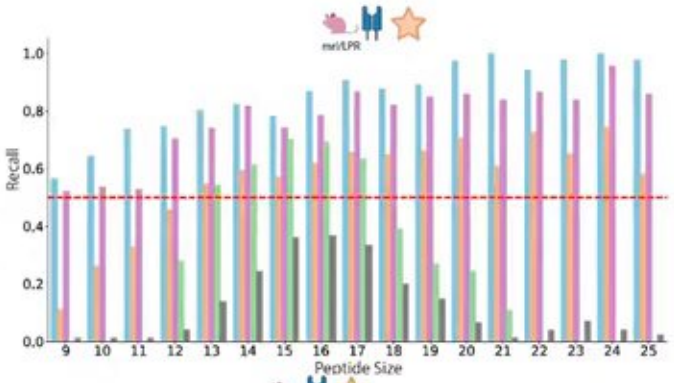
Twitter: @jishnu1729

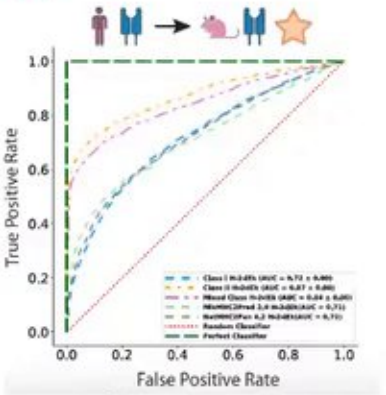
SWING outperforms state-of-the-art VEP methods and uncovers molecular phenotypes missed by these approaches



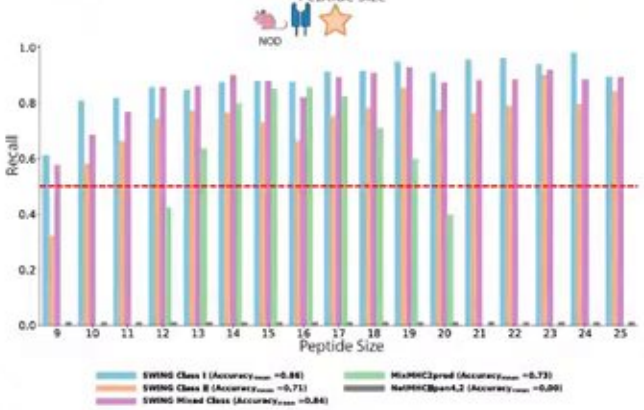
SWING can make de-novo predictions of pMHC interactions in a murine model of SLE

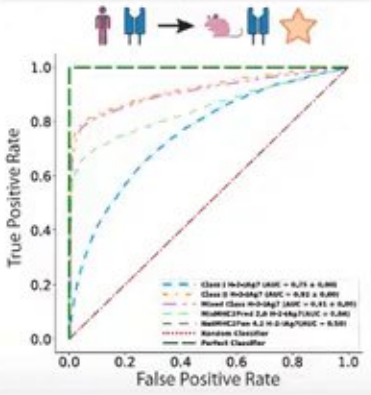
SWING model learnt using human data highly predictive of H-2-IEk binding in a murine model of SLE (Mrl/lpr)





SWING model learnt using human data highly predictive of IAg7 binding in a murine model of T1D (NOD)



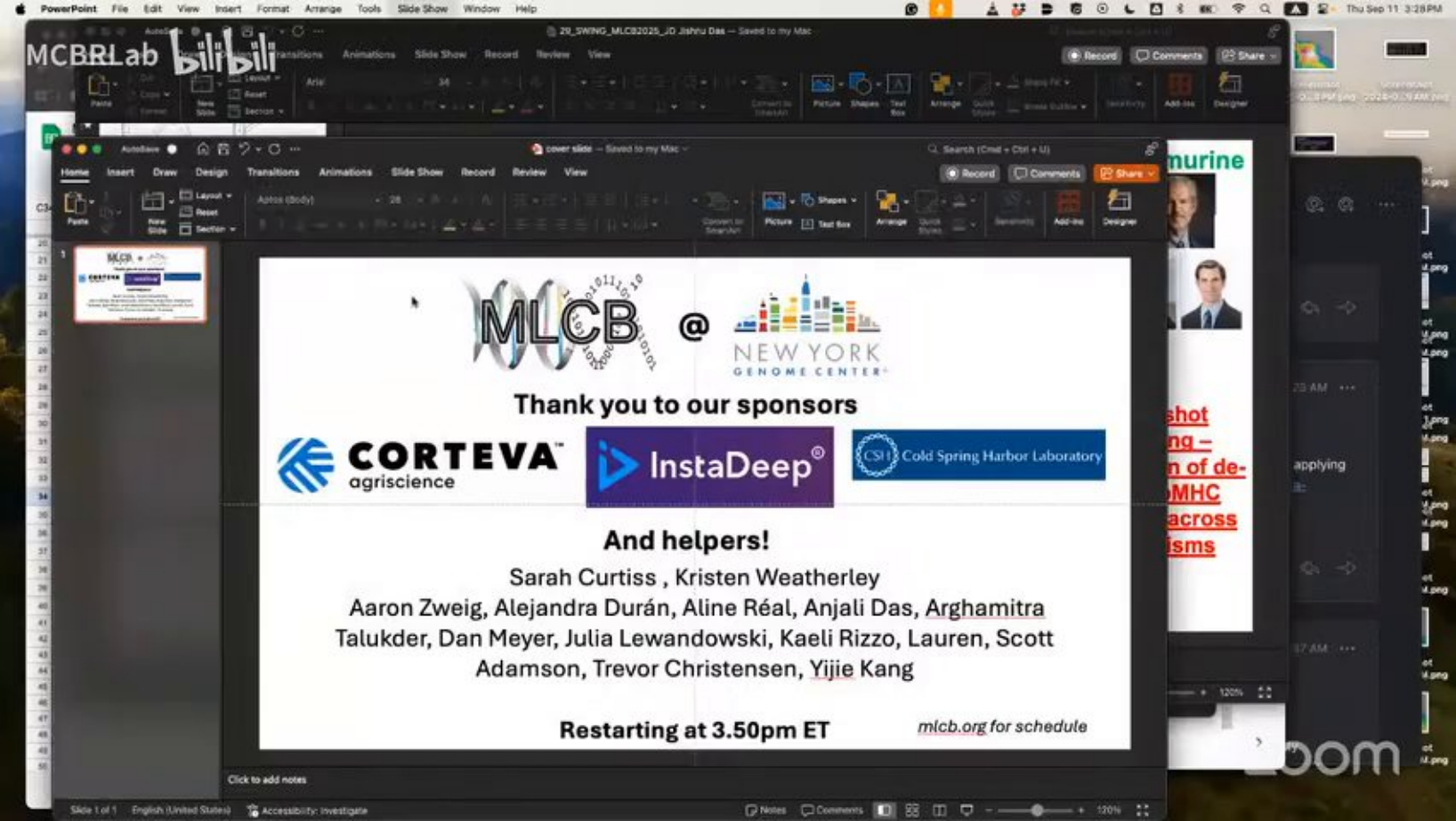


Few-shot learning – prediction of de-novo pMHC binding across organisms

MCBRLab bilibili

NYGC Events

zoom


MCBRLab
bilibili
MLCB @ NEW YORK GENOME CENTER
Thank you to our sponsors
CORTEVA agriscience
InstaDeep
Cold Spring Harbor Laboratory
And helpers!
Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang
Restarting at 3.50pm ET
mlcb.org for schedule

MLCB @ NEW YORK GENOME CENTER

Thank you to our sponsors

CORTEVA™ agriscience

InstaDeep®

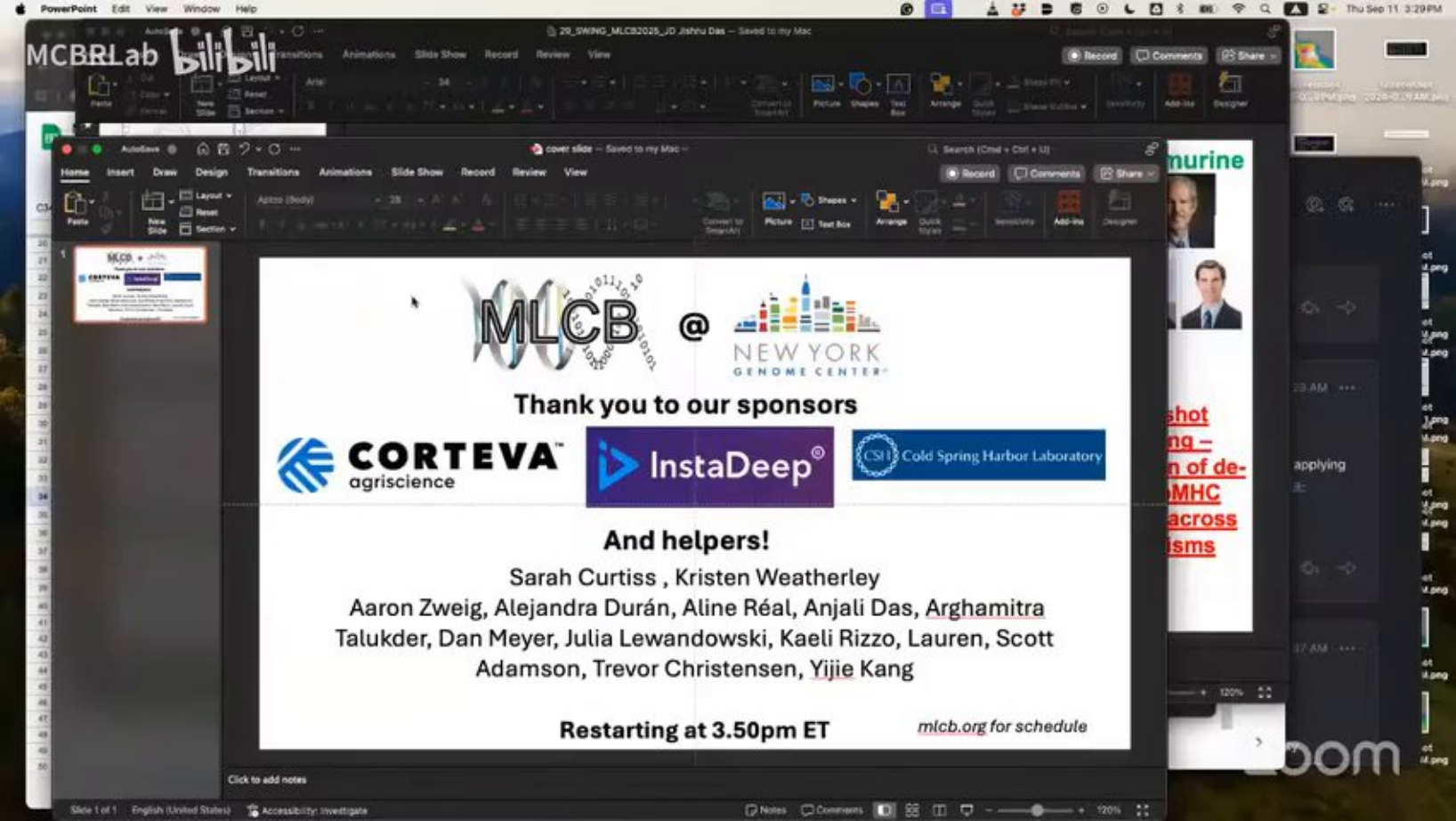
CSH Cold Spring Harbor Laboratory

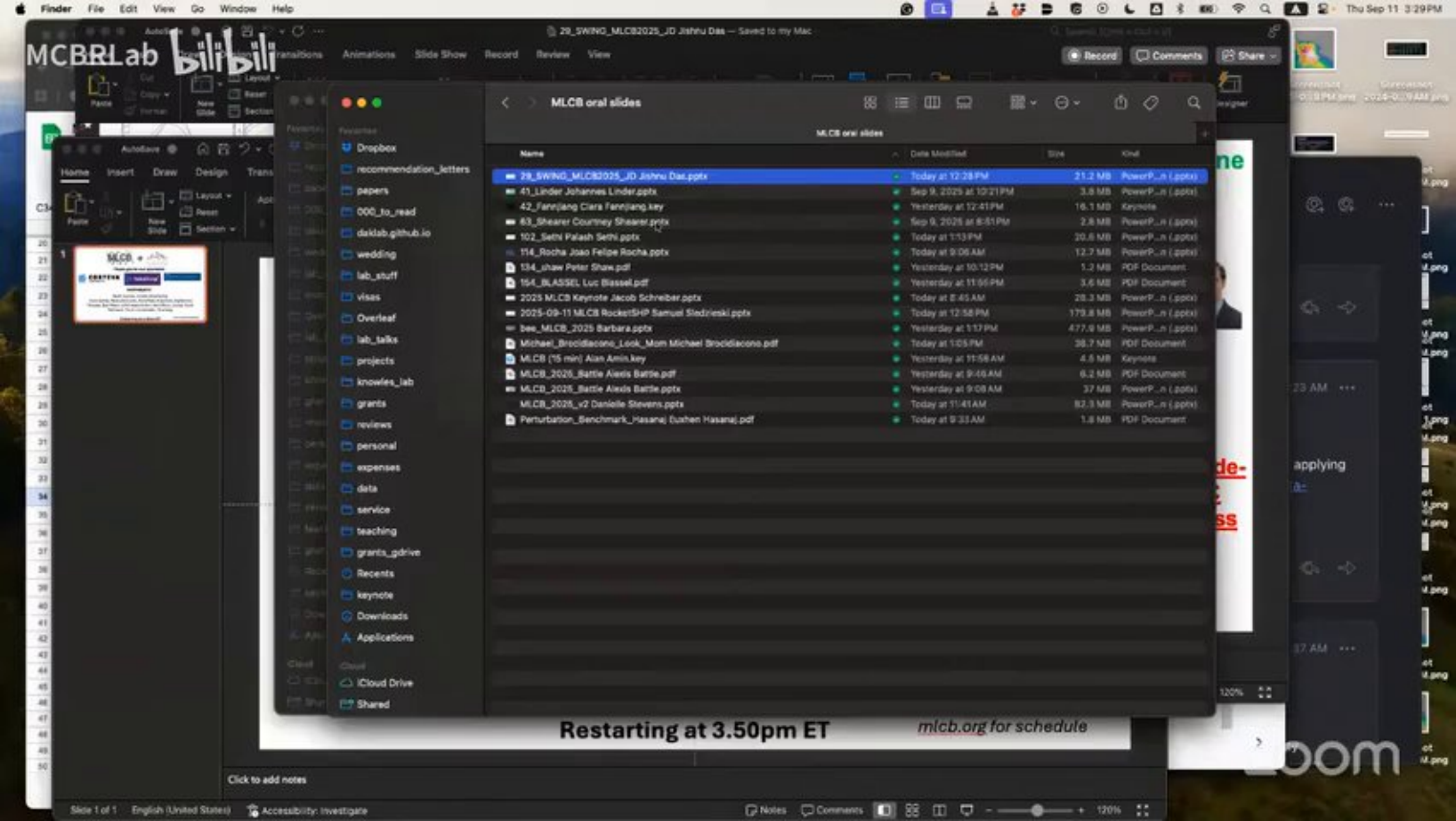
And helpers!

Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang

Restarting at 3.50pm ET

mlcb.org for schedule


MCBRLab
bilibili
MLCB
@
NEW YORK GENOME CENTER
Thank you to our sponsors
CORTEVA agriscience
InstaDeep
Cold Spring Harbor Laboratory
And helpers!
Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang
Restarting at 3.50pm ET
mlcb.org for schedule


MCBRLab
bilibili
MLCB oral slides
Restarting at 3.50pm ET
mlcb.org for schedule

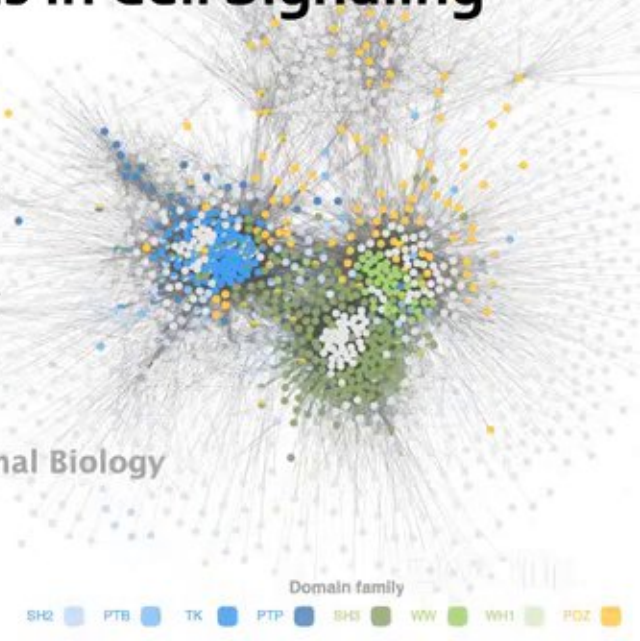
MCBRLab bilibili

NYGC Events

zoom

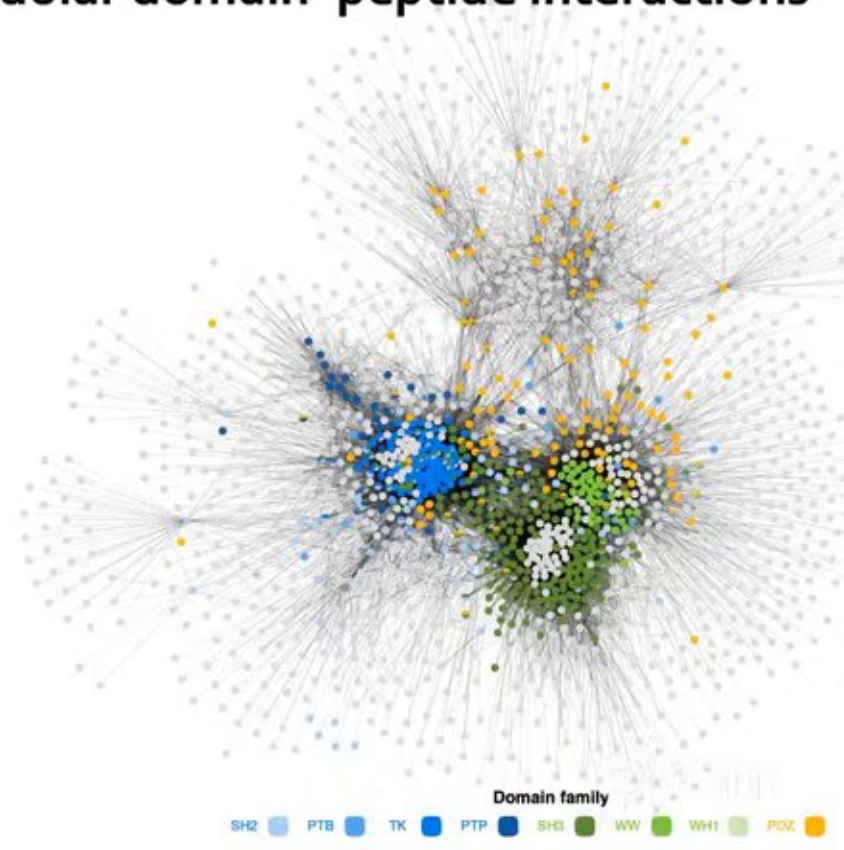
Machine Learning the Binding Affinities of Protein–Peptide Interactions in Cell Signaling

Julia R Rogers
AlQuraishi Lab
Columbia University



Machine Learning in Computational Biology
September 11, 2025

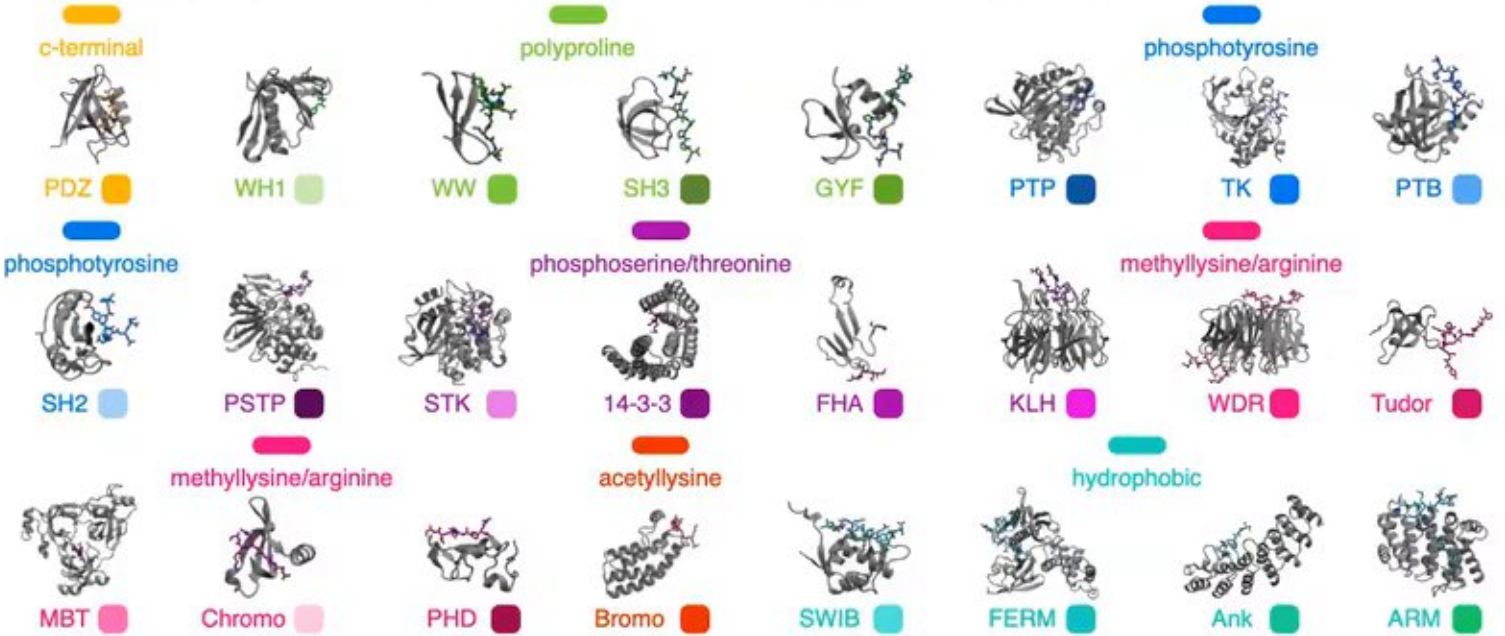
Cellular signals processed by modular domain–peptide interactions



Cunningham, J. M.; et al. *Nat. Methods* **2020**, *17*, 175.

Diversity of modular domain–peptide interactions

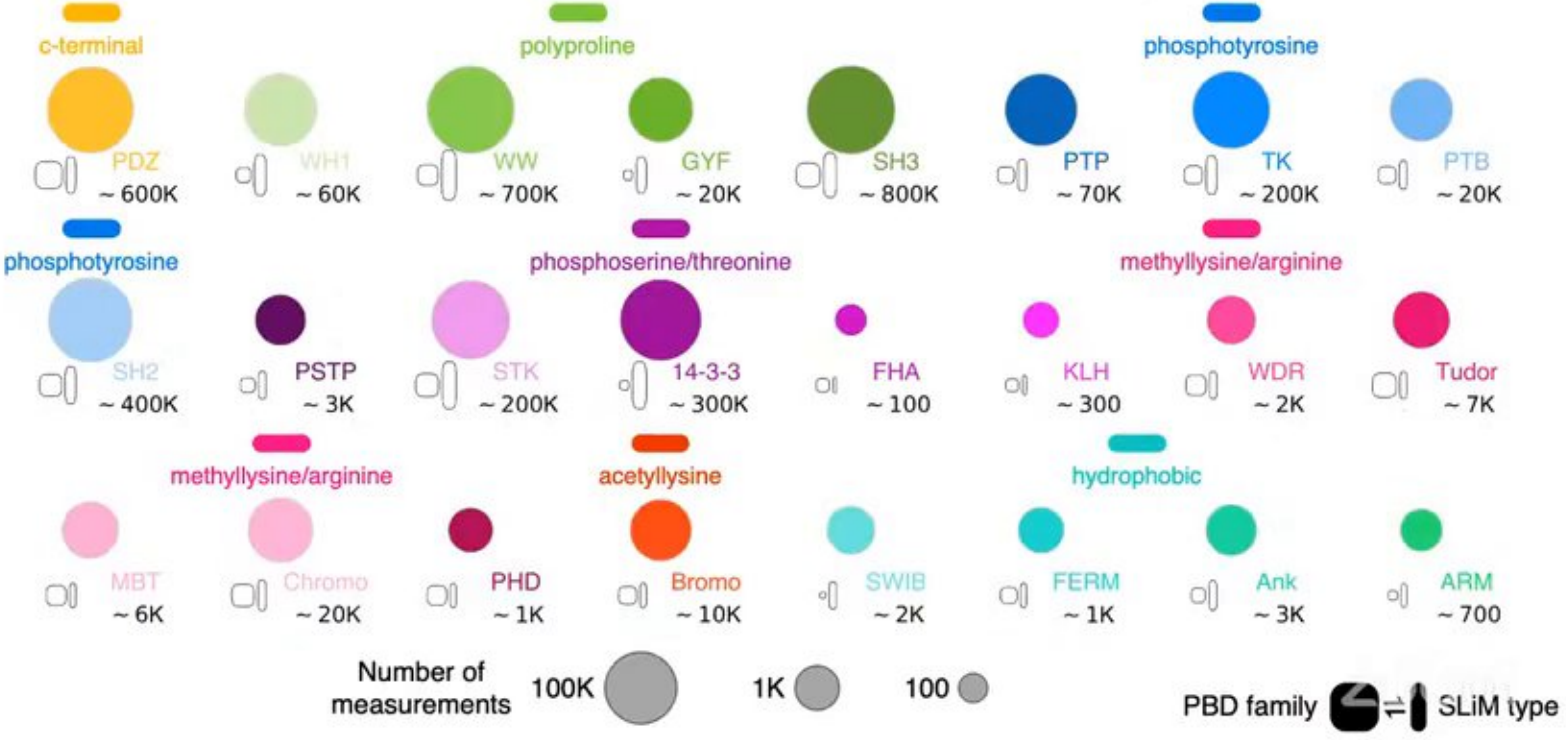
(out of ~100-200 human PBD families and ~7 binding chemistry)



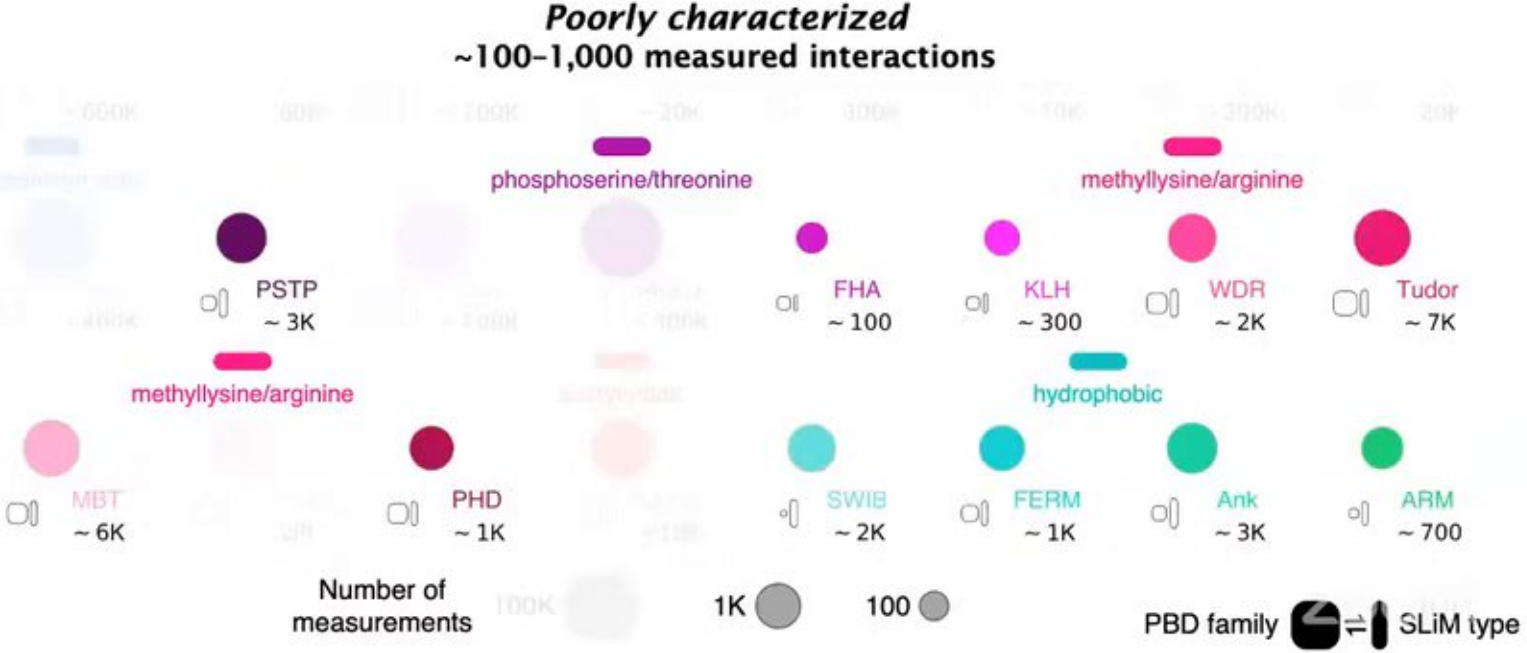
Pawson, T. *Cell* **2004**, *116*, 191.
Mayer, B. J. *Nat. Rev.* **2015**, *16*, 691.

PBD family ⇌ SLiM type

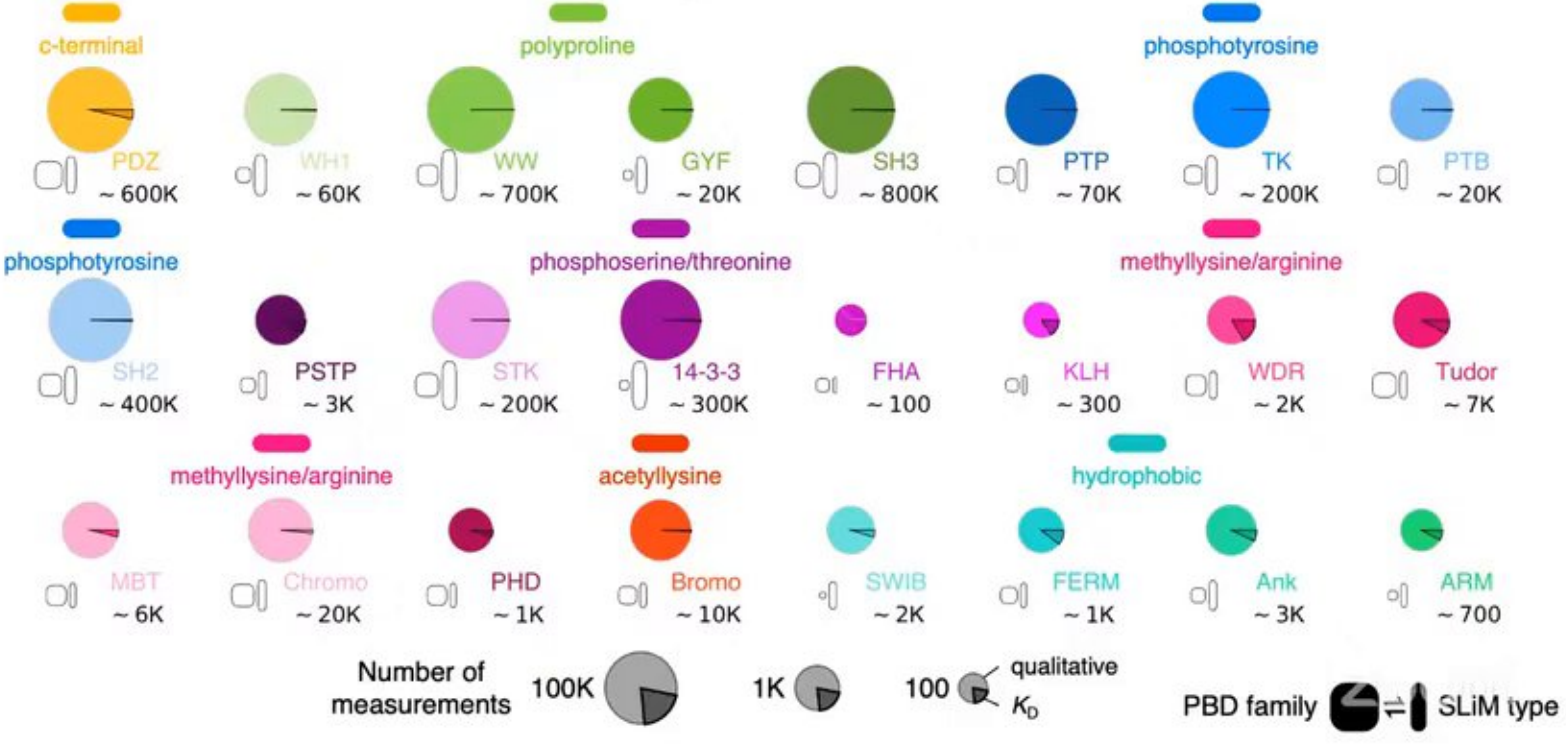
Diversity of modular domain–peptide interactions and disparate experimental coverage

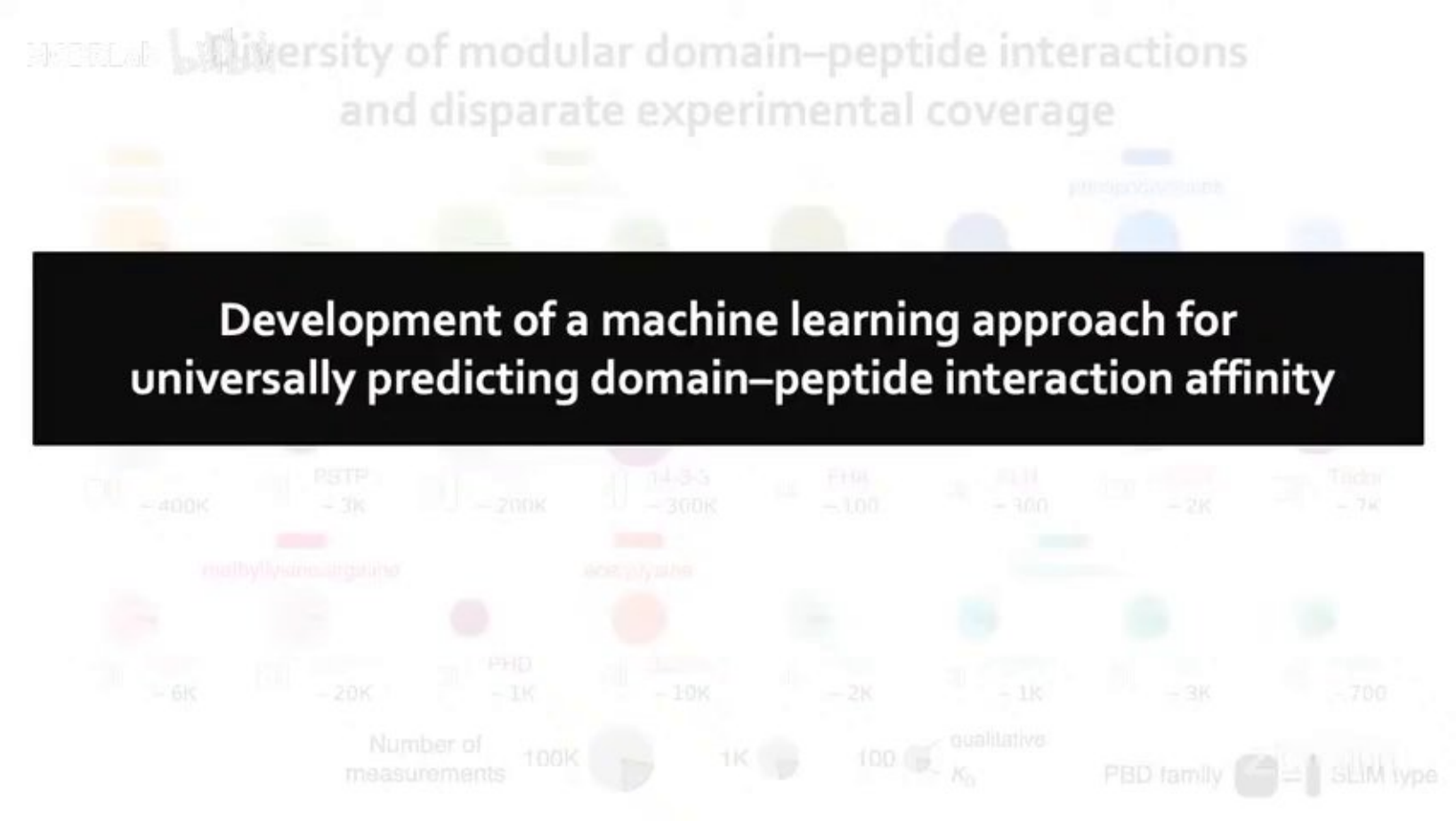


Diversity of modular domain–peptide interactions and disparate experimental coverage

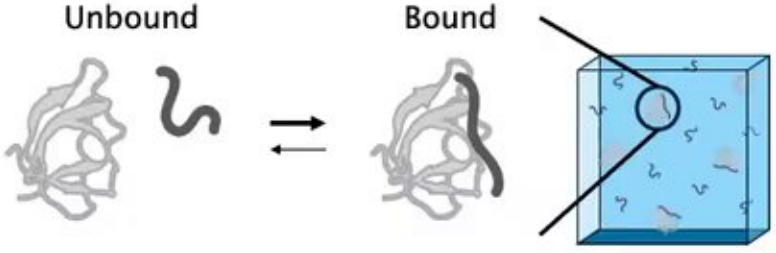


Diversity of modular domain–peptide interactions and limited knowledge of quantitative affinities



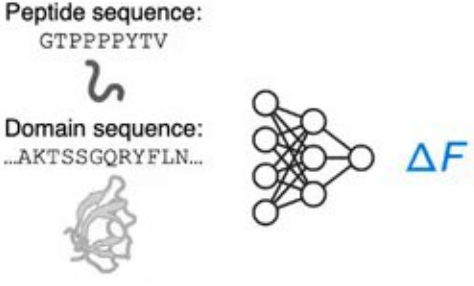

University of modular domain–peptide interactions
and disparate experimental coverage
Development of a machine learning approach for
universally predicting domain–peptide interaction affinity
Number of
measurements
100K
1K
100
qualitative
PBD family
SLiM type

Statistical mechanical description of domain–peptide interactions



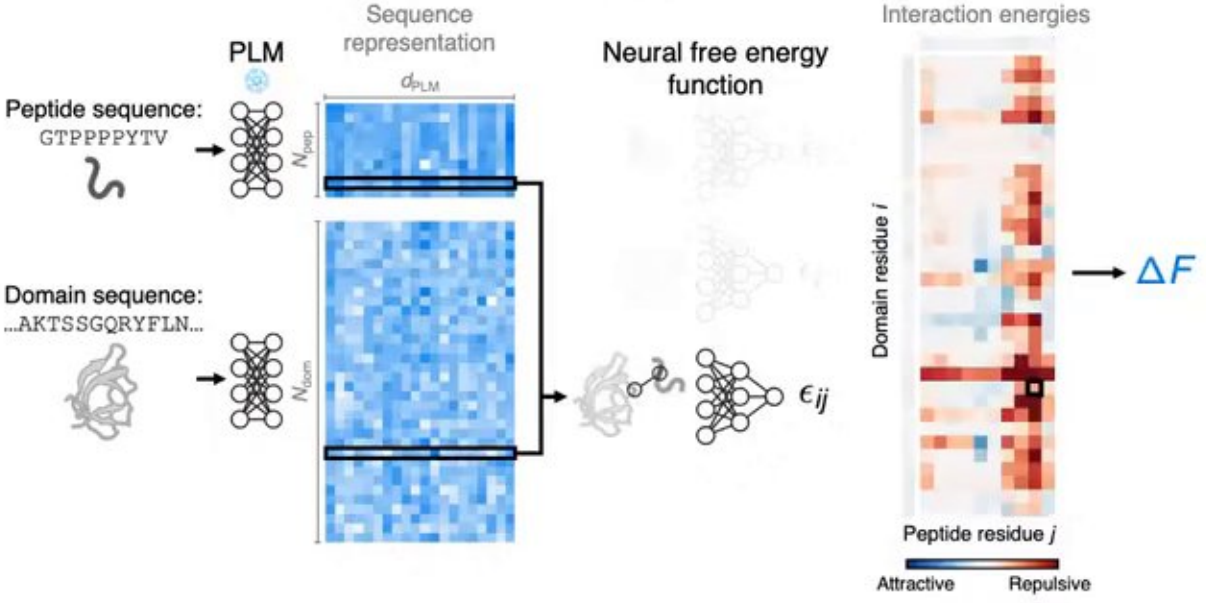
Machine learning a quantitative statistical mechanical model for ΔF

to predict interaction affinity directly from amino acid sequence



Machine learning a quantitative statistical mechanical model for ΔF

parameterized by representations shared across domain families

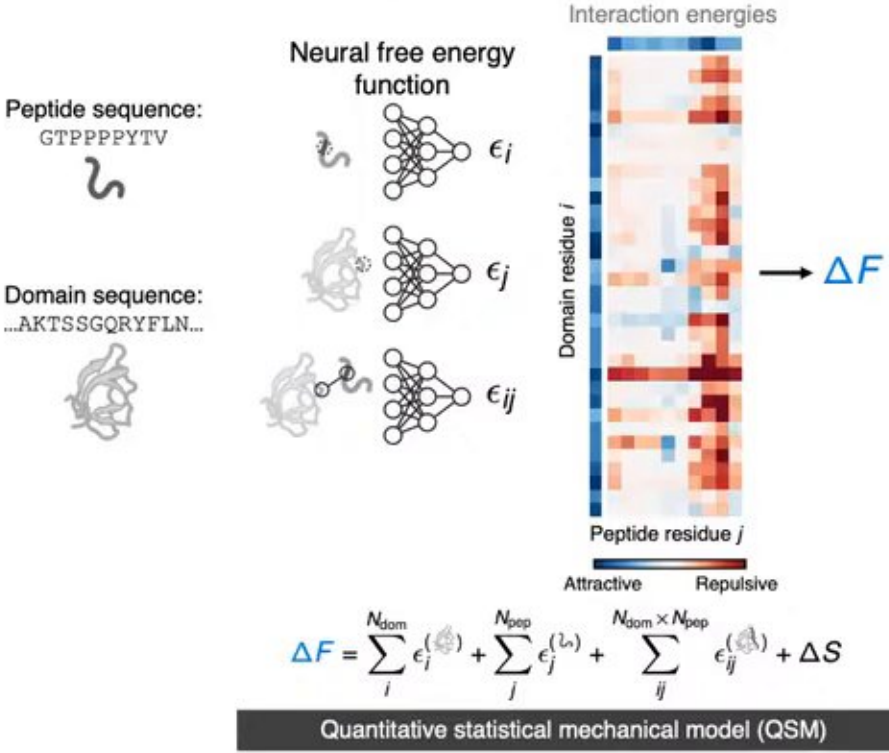


$$\Delta F = \sum_{i}^{N_{dom}} \epsilon_i^{(\text{dom})} + \sum_{j}^{N_{pep}} \epsilon_j^{(\text{pep})} + \sum_{ij}^{N_{dom} \times N_{pep}} \epsilon_{ij}^{(\text{int})} + \Delta S$$

Quantitative statistical mechanical model (QSM)

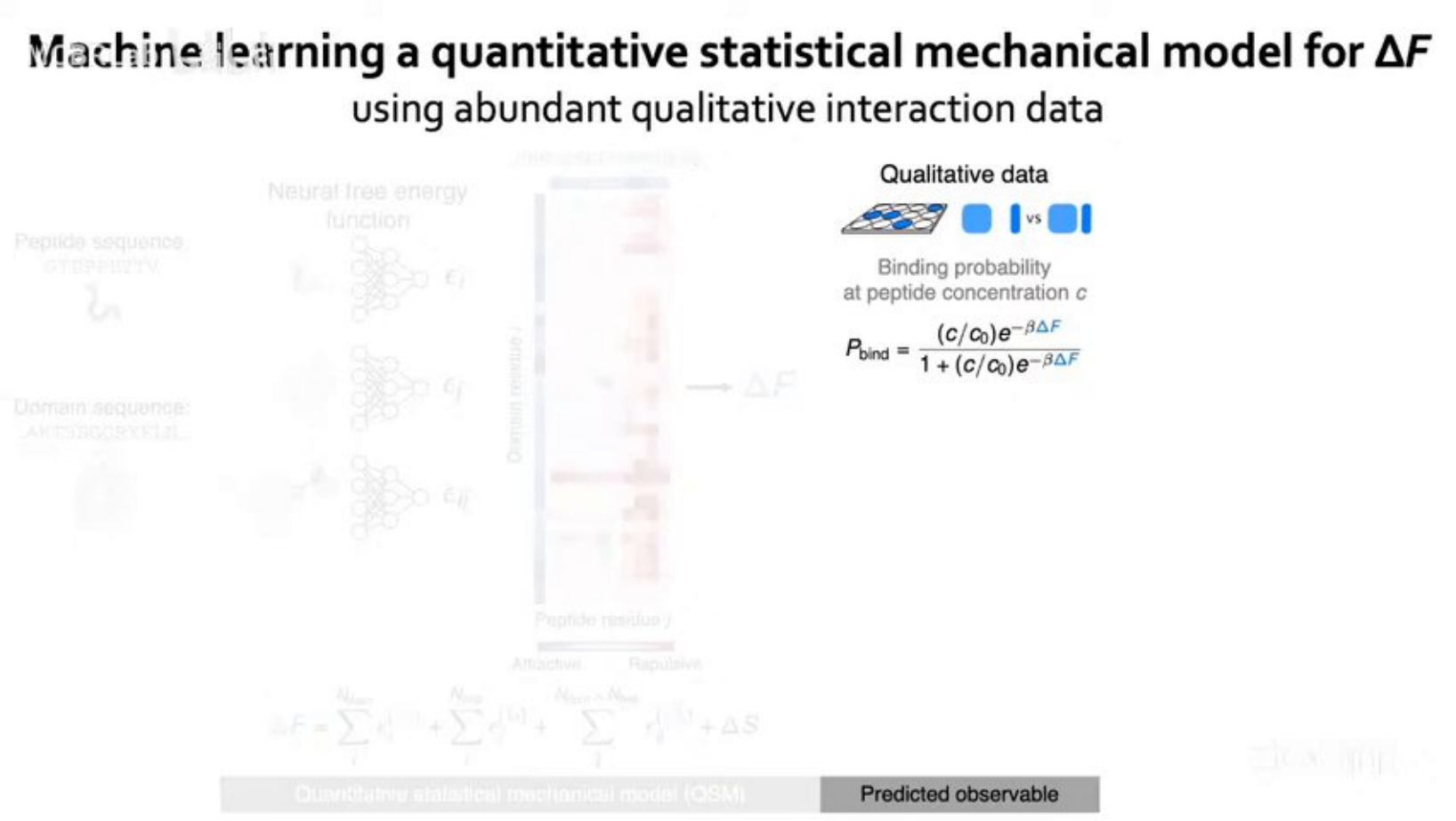
Machine learning a quantitative statistical mechanical model for ΔF

How to infer a free energy function from data?



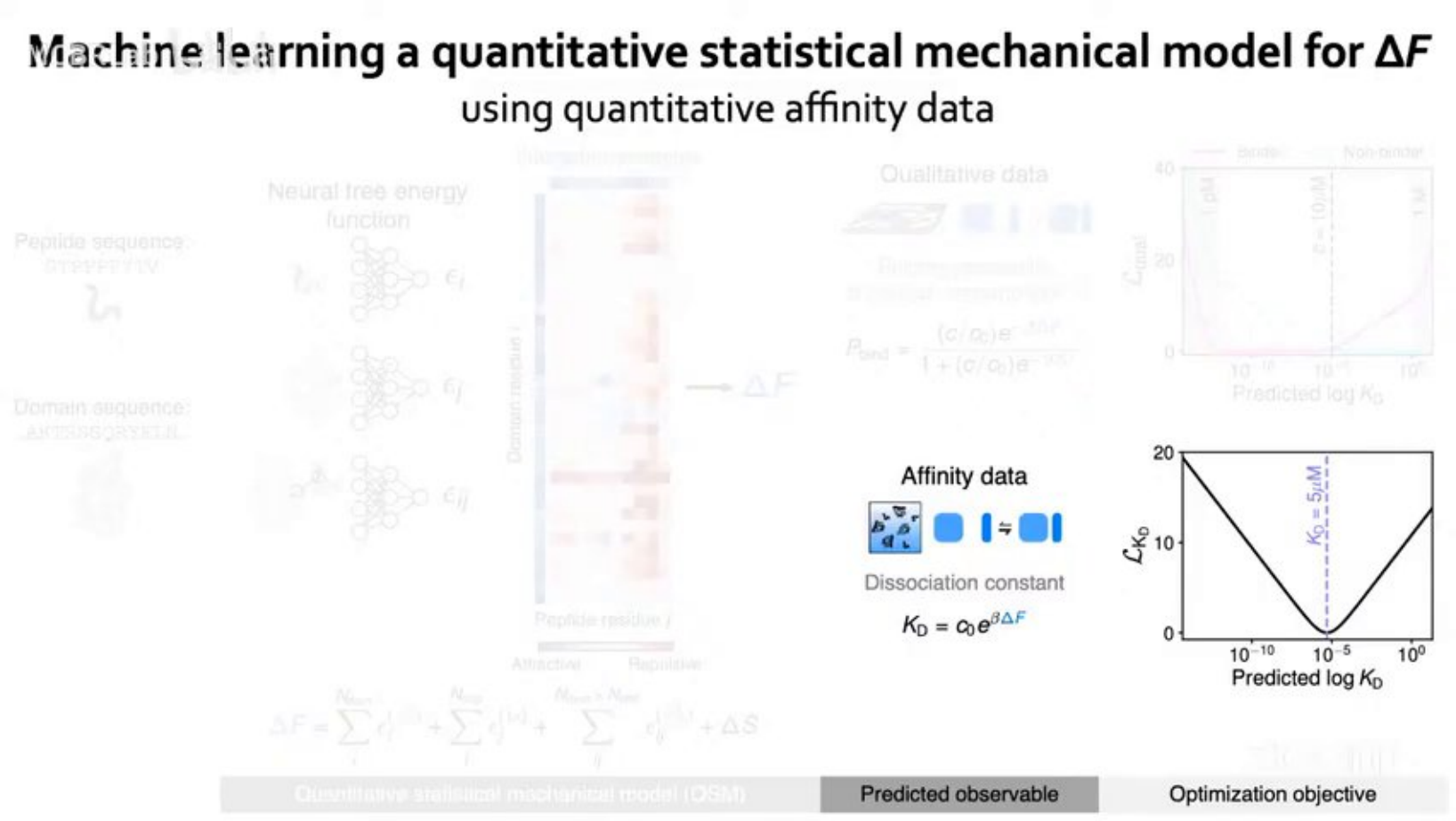
Machine learning a quantitative statistical mechanical model for ΔF

using abundant qualitative interaction data



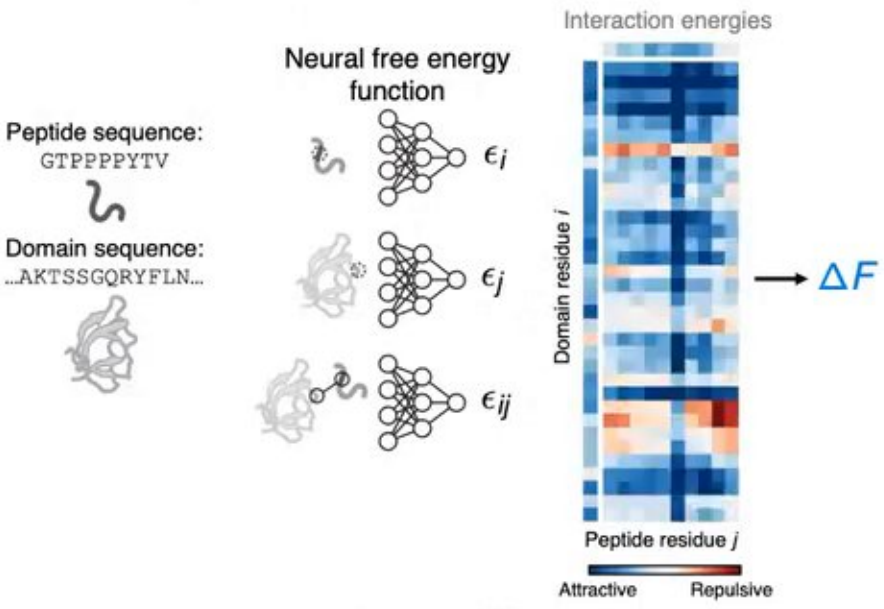
Machine learning a quantitative statistical mechanical model for ΔF

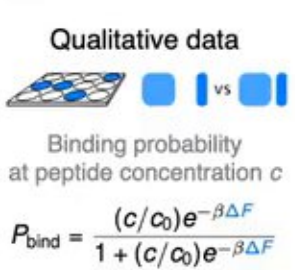
using quantitative affinity data

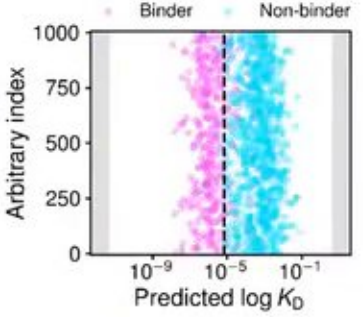


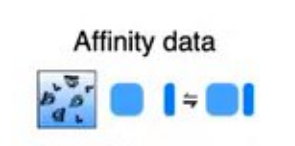
Machine learning a quantitative statistical mechanical model for ΔF

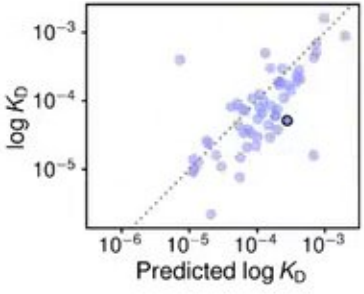
using a multitask learning objective to leverage all available interaction data









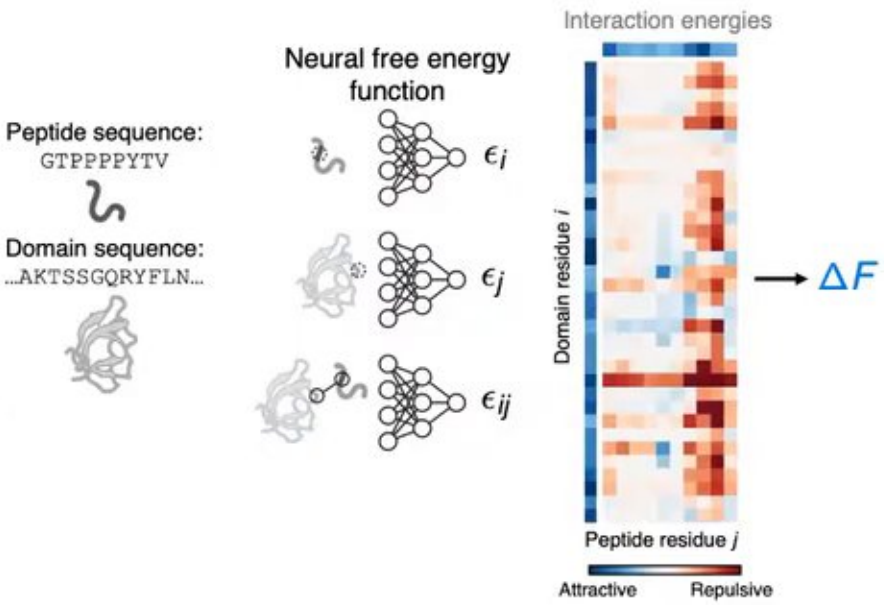


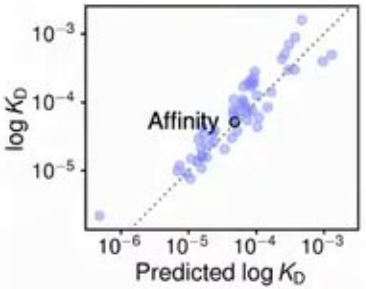
$$\Delta F = \sum_i^{N_{\text{dom}}} \epsilon_i^{(\text{dom})} + \sum_j^{N_{\text{pep}}} \epsilon_j^{(\text{pep})} + \sum_{ij}^{N_{\text{dom}} \times N_{\text{pep}}} \epsilon_{ij}^{(\text{dom-pep})} + \Delta S$$

$$\mathcal{L} = \mathcal{L}_{\text{qual}}(\text{vs}) + \alpha \mathcal{L}_{K_D}(\doteq)$$

Quantitative statistical mechanical model (QSM)	Predicted observable	Optimization objective

Machine learning a quantitative statistical mechanical model for ΔF





Quantitative statistical mechanical model (QSM) → Predicted observable

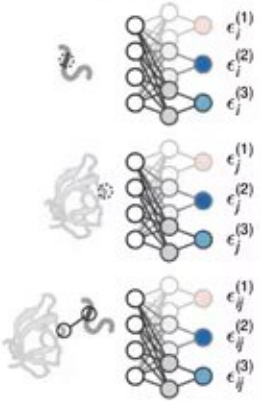
Machine learning a quantitative statistical mechanical model for ΔF

that also estimates predictive error

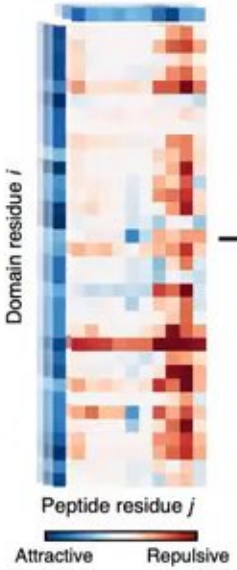
Peptide sequence:
GTPPPPYTV

Domain sequence:
...AKTSSGQRYFLN...

Neural free energy function



Interaction energies



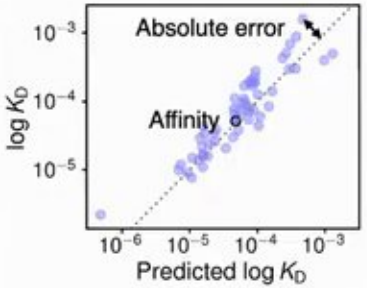
$\langle \Delta F \rangle$

Ensemble neural free energy function

⇒ Estimates affinity + absolute error

Scientific application

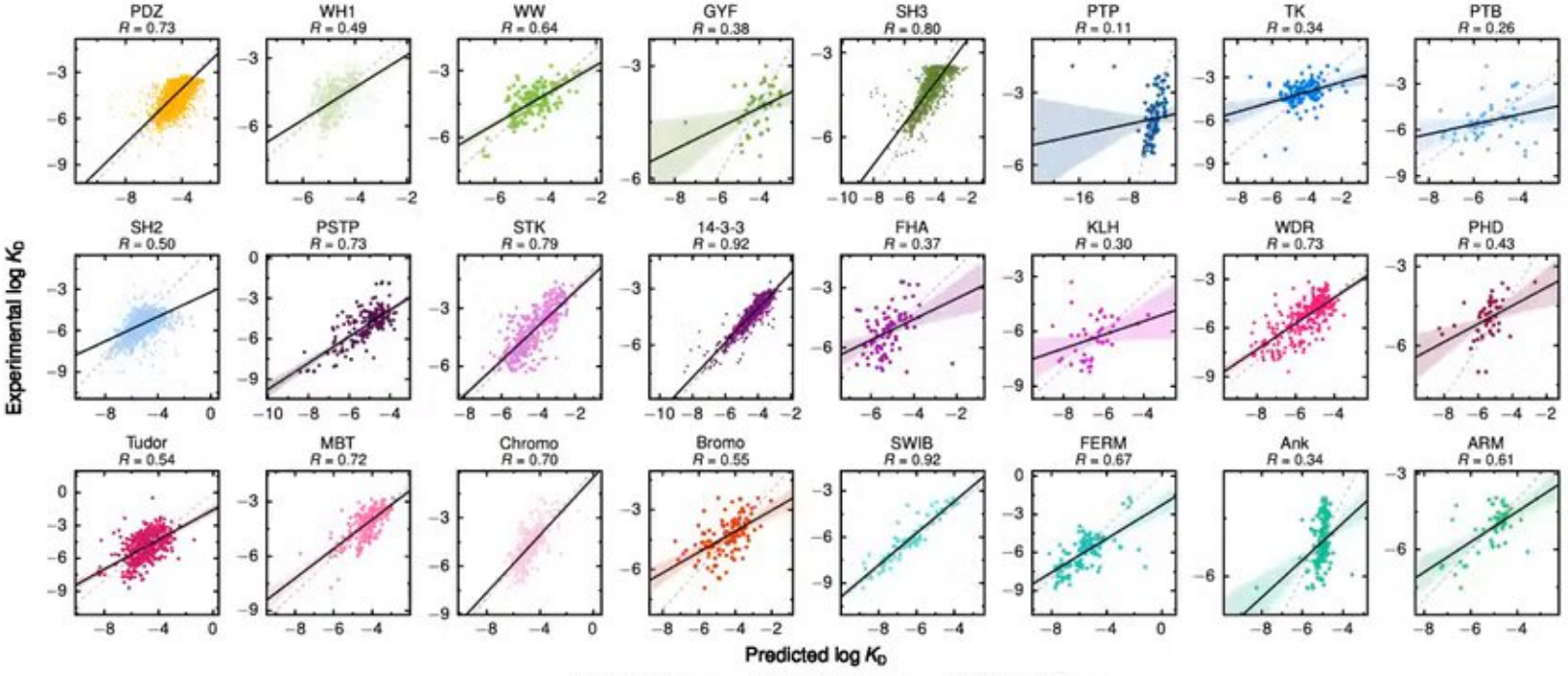
Ensemble mean := Predicted binding free energy



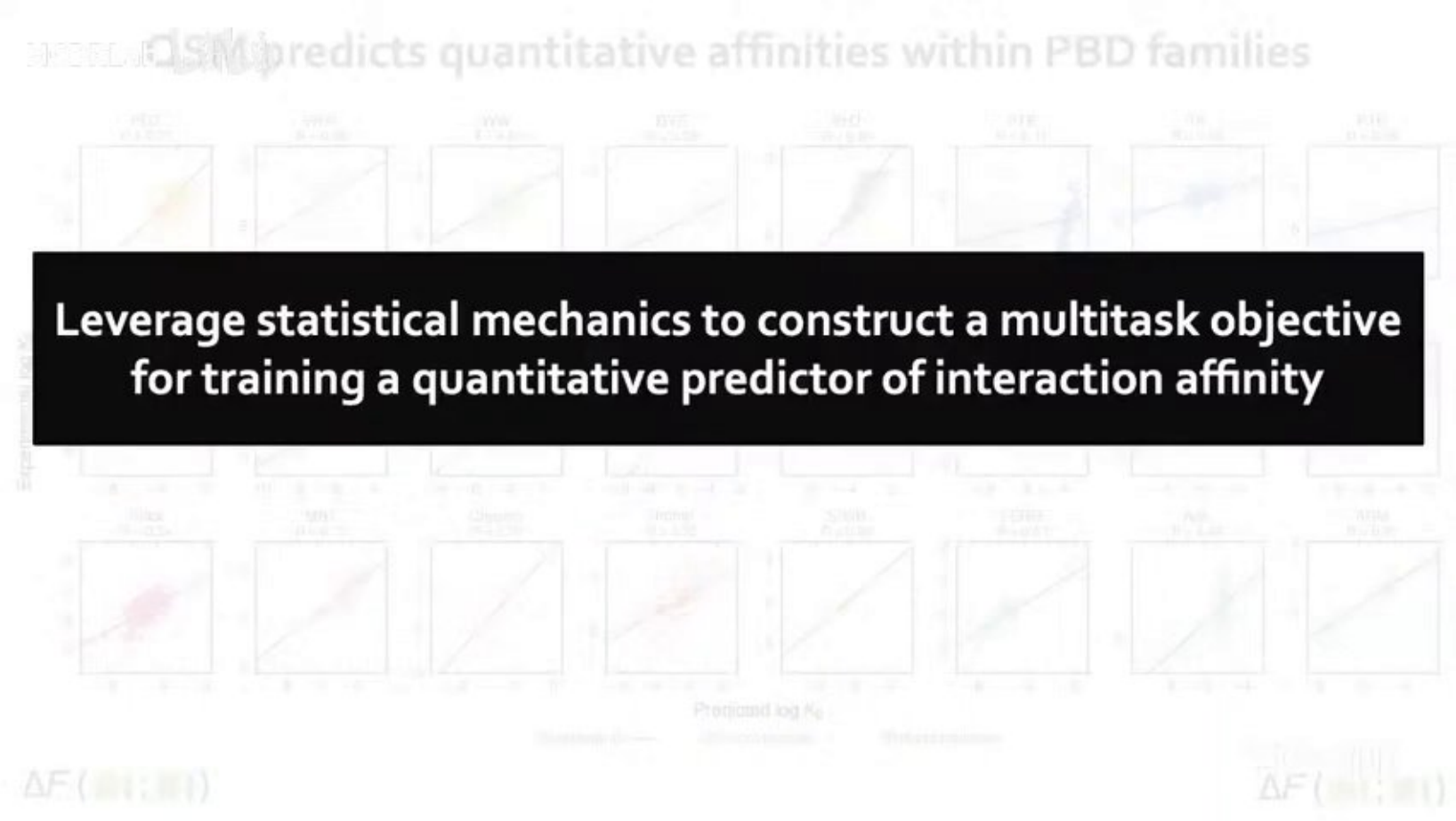
Quantitative statistical mechanical model (QSM) | Predicted observable

Do QSM models predict quantitative affinities within PBD families?

QSM predicts quantitative affinities within PBD families

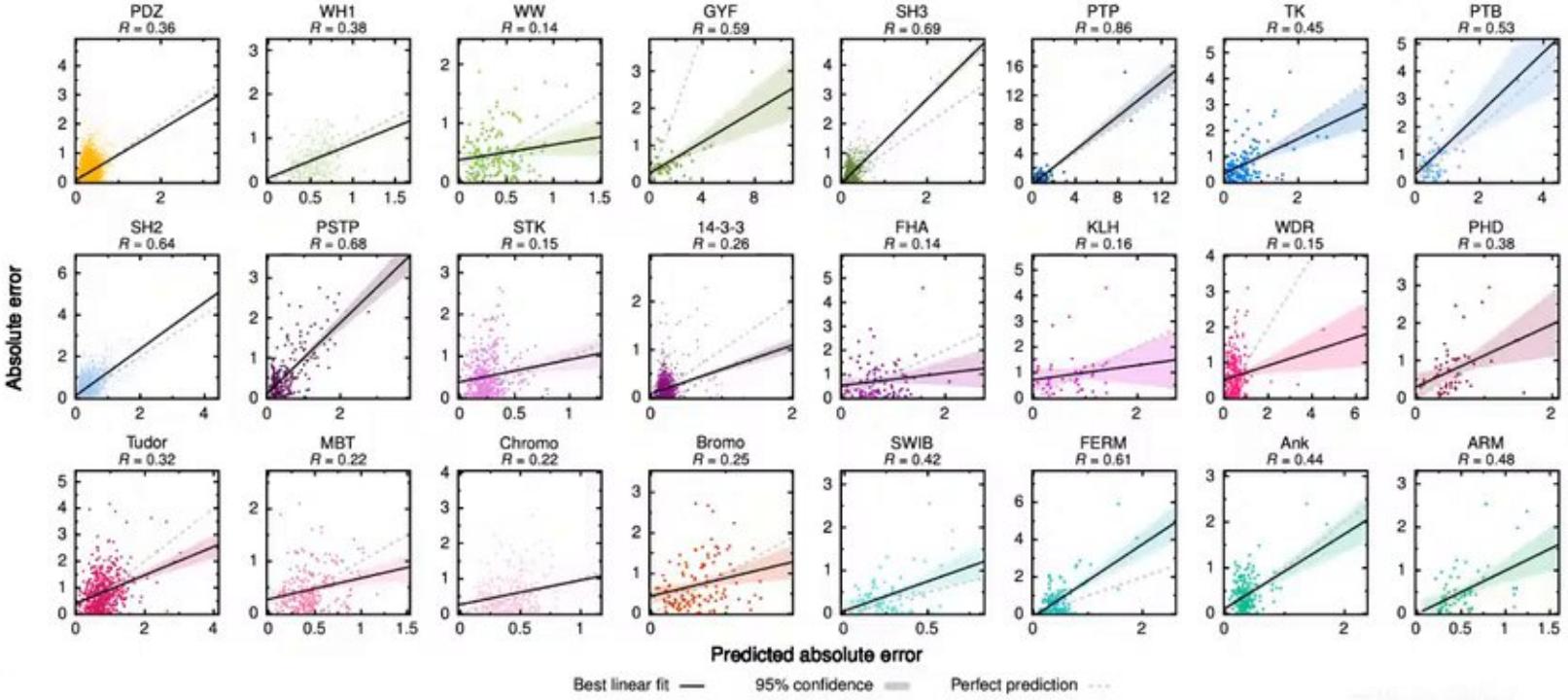


ΔF (;)


predicts quantitative affinities within PBD families
Leverage statistical mechanics to construct a multitask objective for training a quantitative predictor of interaction affinity

How trustworthy are QSM affinity predictions?

QSM estimates uncertainty by predicting its own absolute error

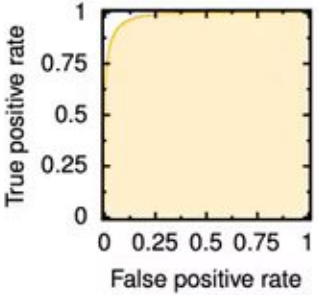


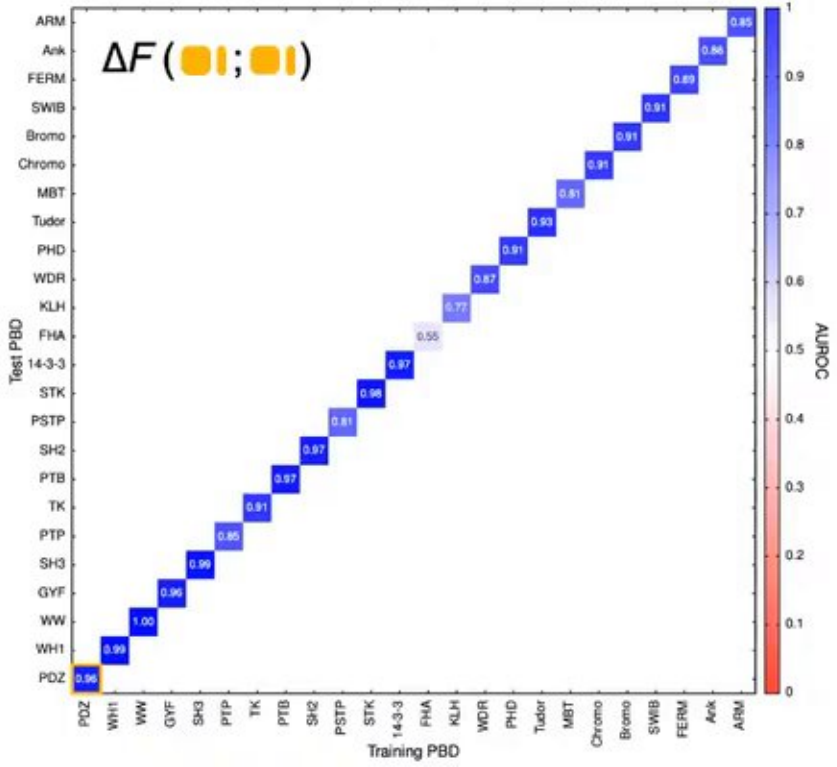
$\Delta F($; $)$

Are QSM free energy functions universal?

QSM free energy functions identify interactions within PBD families

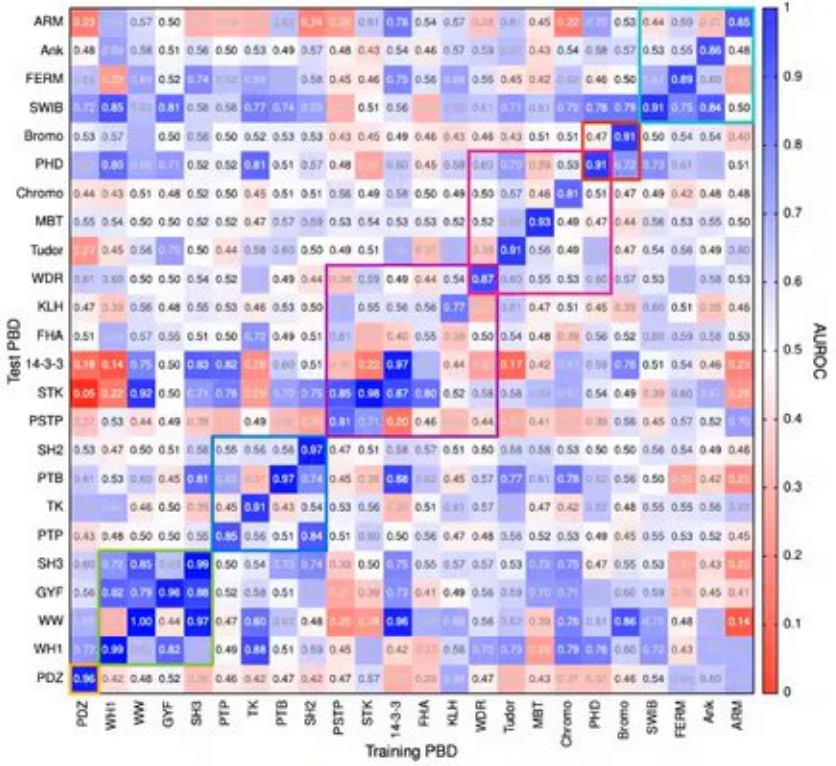
Evaluated for identifying 'interaction' vs 'non-interacting' domain–peptide pairs





QSM free energy functions transfer across PBD families

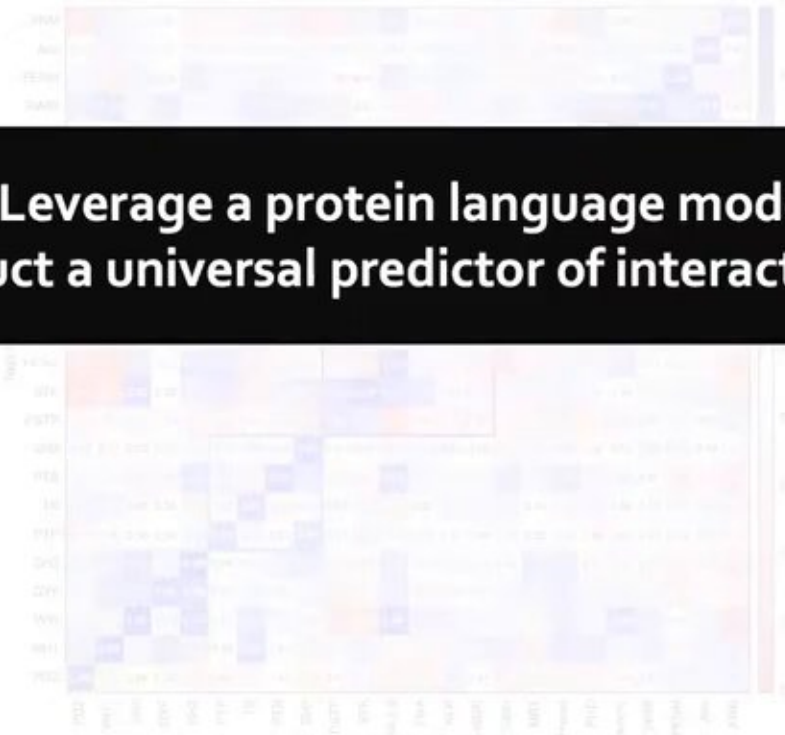
Evaluated as classifier of 'interaction' vs 'non-interacting' domain–peptide pairs



$\Delta F(\bullet;\bullet)$

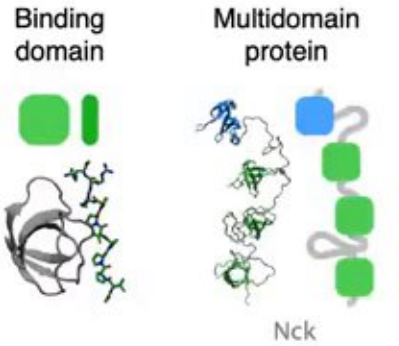
Q5M free energy functions transfer across PBD families

Evaluated as classifier of 'interaction' vs 'non-interacting' domain–peptide pairs



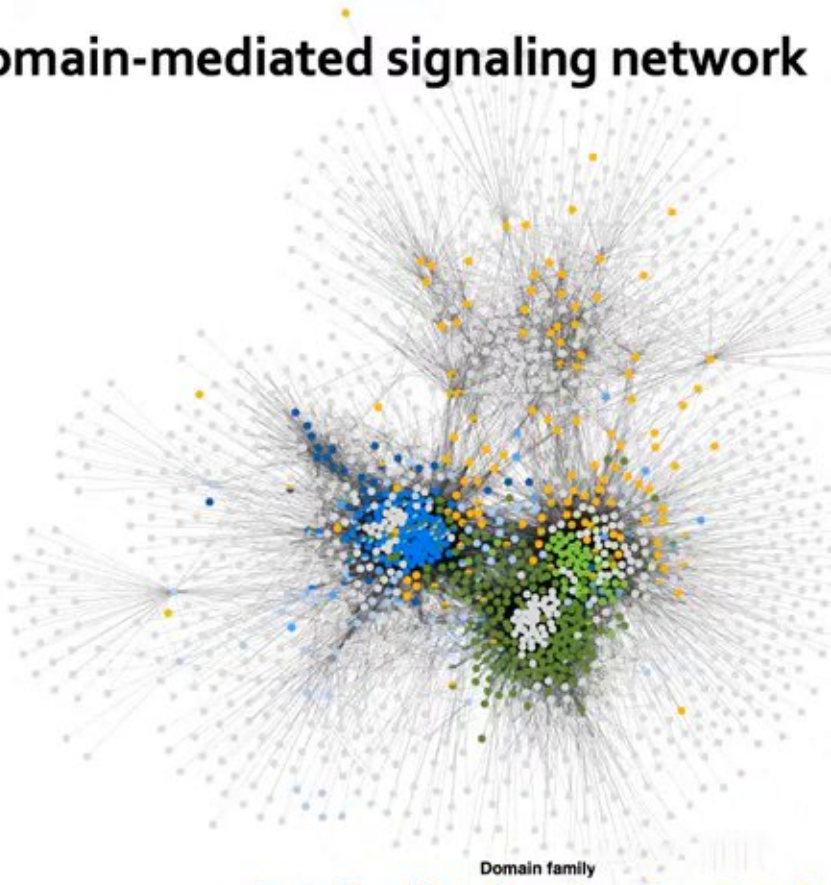
Leverage a protein language model
to construct a universal predictor of interaction affinity

Can QSM discover multivalent protein–protein interactions mediated by domain–peptide interactions?



Okrut, J. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E6436.

Constructing a proteome-scale domain-mediated signaling network

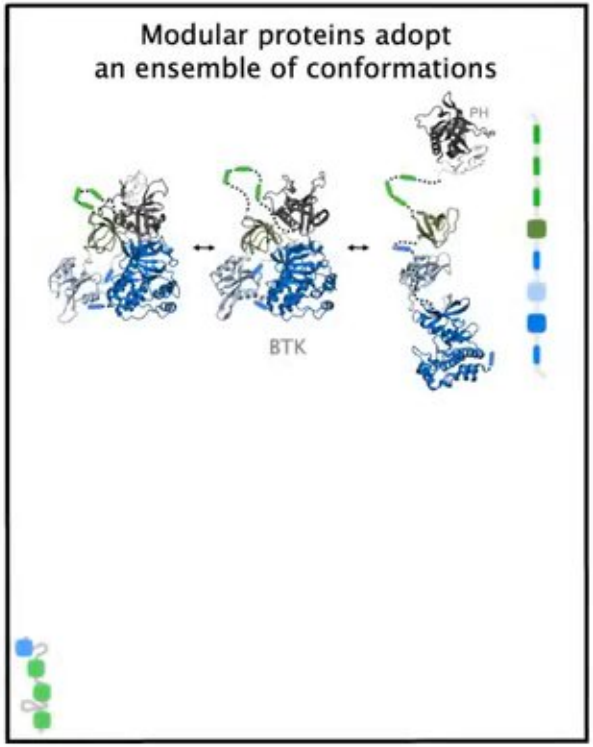


Cunningham, J. M.; et al. *Nat. Methods* **2020**, *17*, 175.

Constructing a proteome-scale domain-mediated signaling network

Quantitative models of cell signaling enabled by machine learning

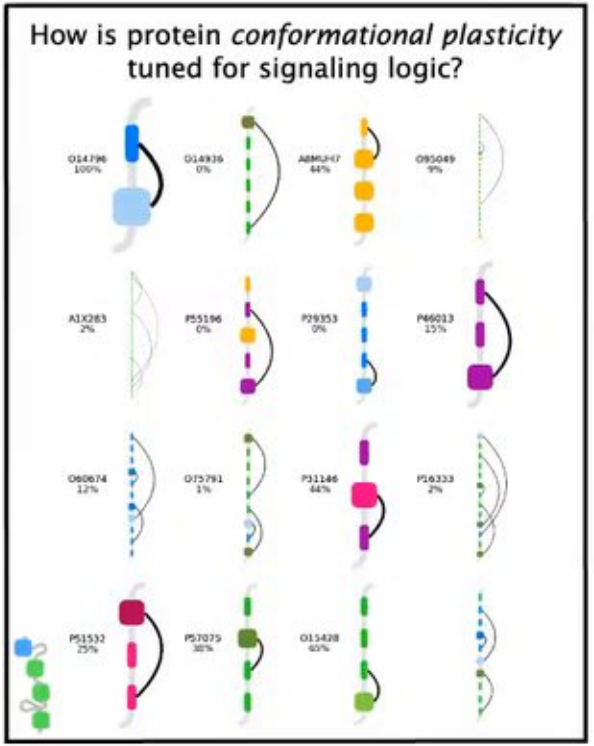
to decode the biophysical logic of cellular information processing



Joseph, R. E., et al. *Structure* **2017**, *25*, 1481.

Quantitative models of cell signaling enabled by machine learning

to decode the *biophysical logic* of cellular information processing

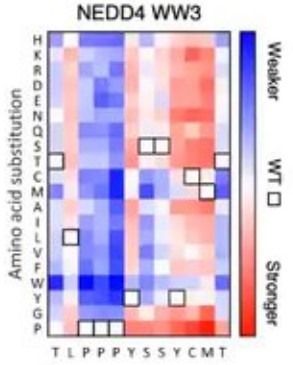


PBD-mediated protein–protein interaction network

Quantitative models of cell signaling enabled by machine learning

to decode the *biophysical logic* of cellular information processing

How does sequence variation tune PBD–SLiM *binding affinity*?

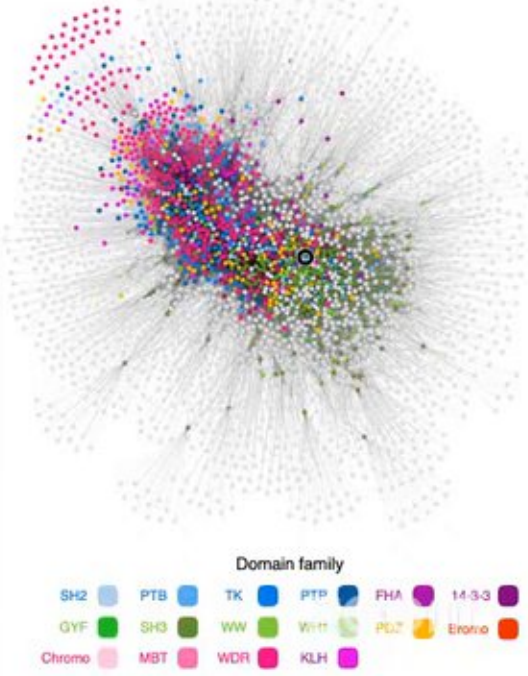


In silico
deep mutational scanning

How is protein *conformational plasticity* tuned for signaling logic?

What *multi-input computations* are performed by modular proteins?

PBD-mediated protein–protein interaction network



Acknowledgements

AlQuraishi Lab

Mohammed AlQuraishi

Data curation

Ethan Eickmann
Undergraduate

Allen Na
Masters

Sev Ihnat
Masters

Bridget Liu
Undergraduate

Ashley Hsu
Undergraduate

Teemu Rönkkö
Visiting PhD

Lindsey Yang
Undergraduate

Henry Low
MD/PhD

Sumaiyah Rahman
Undergraduate

Angelina Yan
Undergraduate

Martin Culka
Postdoc

Model benchmarking

Shriya Mahakala
Undergraduate

Emily Duniec
Masters

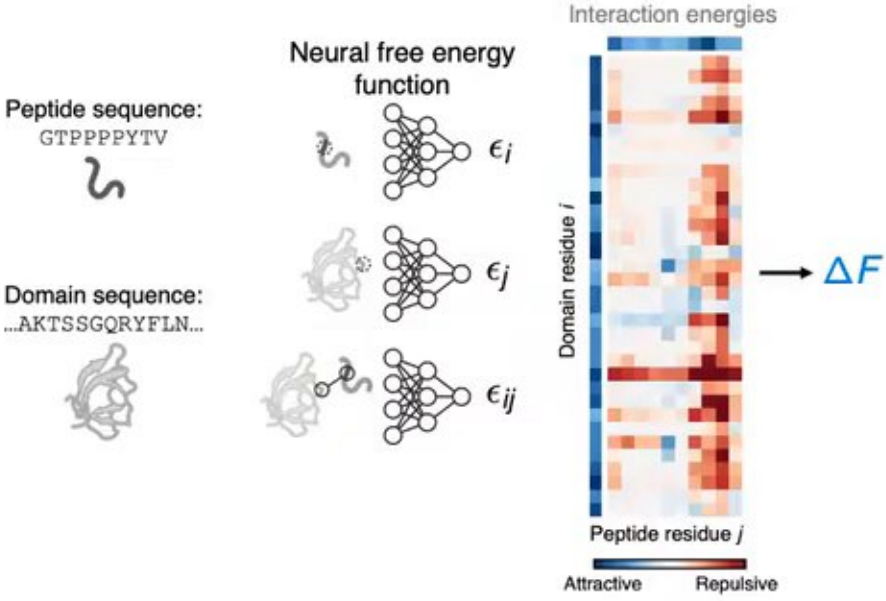
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

JANE COFFIN CHILDS Fund for Medical Research

BURROUGHS WELLCOME FUND
Career Award at the Scientific Interface

Machine learning a quantitative statistical mechanical model for ΔF

How to infer a free energy function from data?

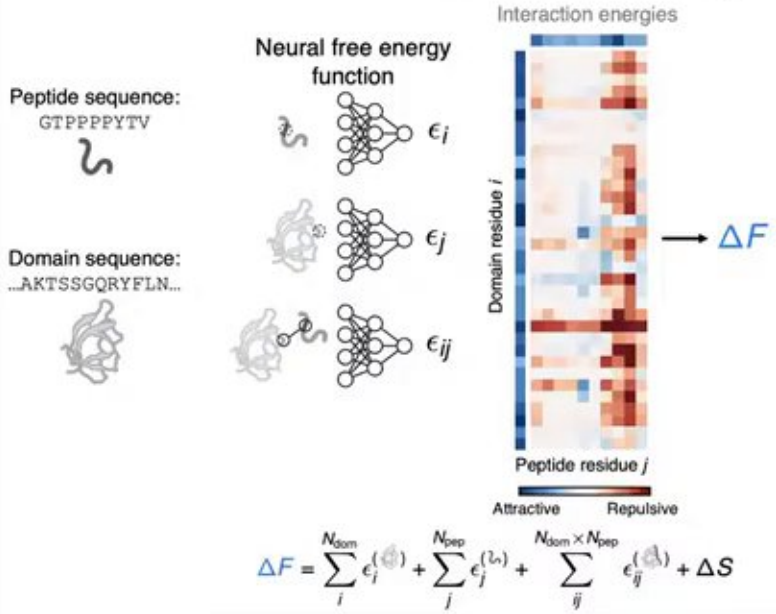


$$\Delta F = \sum_{i}^{N_{\text{dom}}} \epsilon_i^{(\text{dom})} + \sum_{j}^{N_{\text{pep}}} \epsilon_j^{(\text{pep})} + \sum_{ij}^{N_{\text{dom}} \times N_{\text{pep}}} \epsilon_{ij}^{(\text{dom-pep})} + \Delta S$$

Quantitative statistical mechanical model (QSM)

Machine learning a quantitative statistical mechanical model for ΔF

How to infer a free energy function from data?

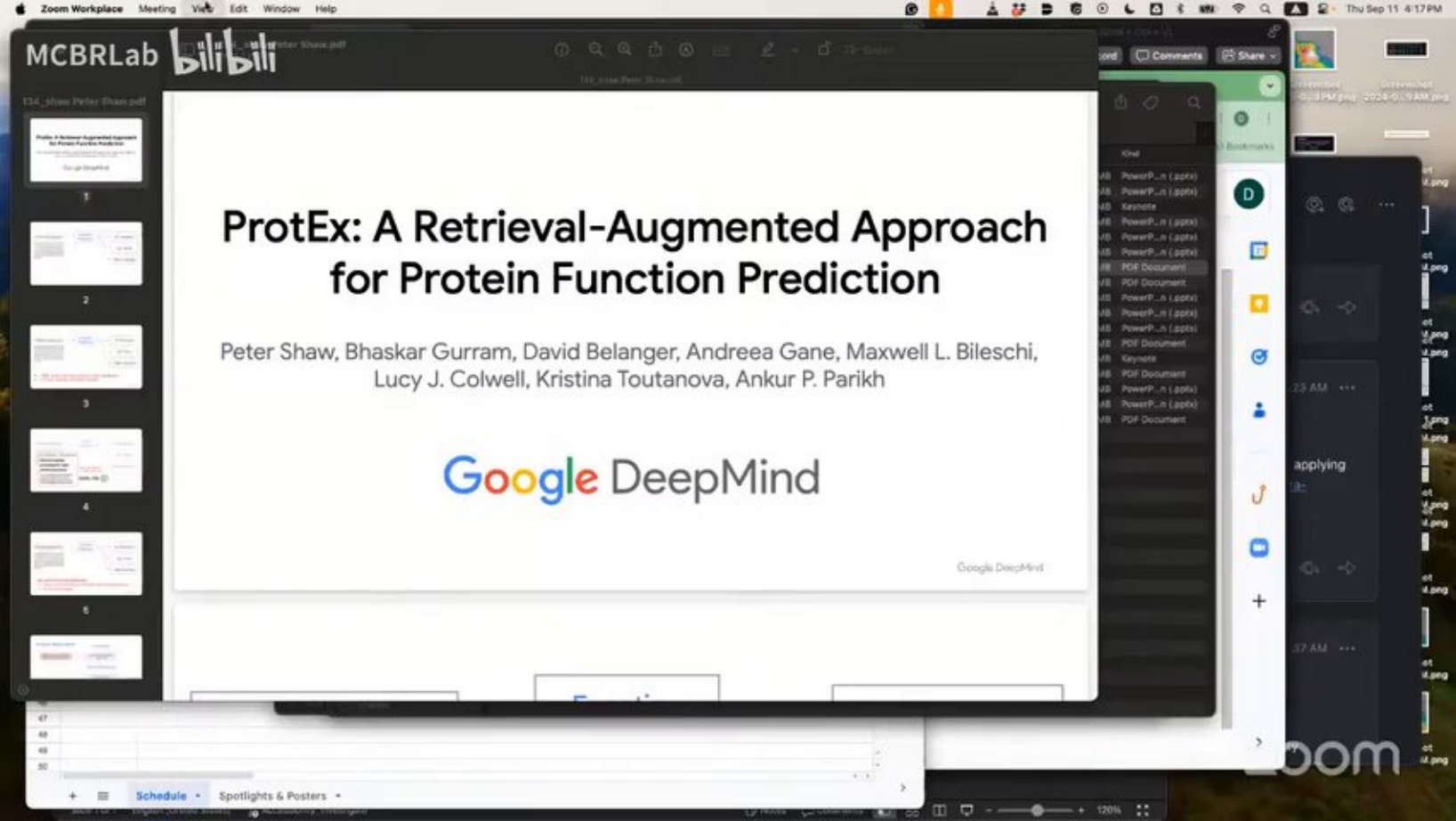


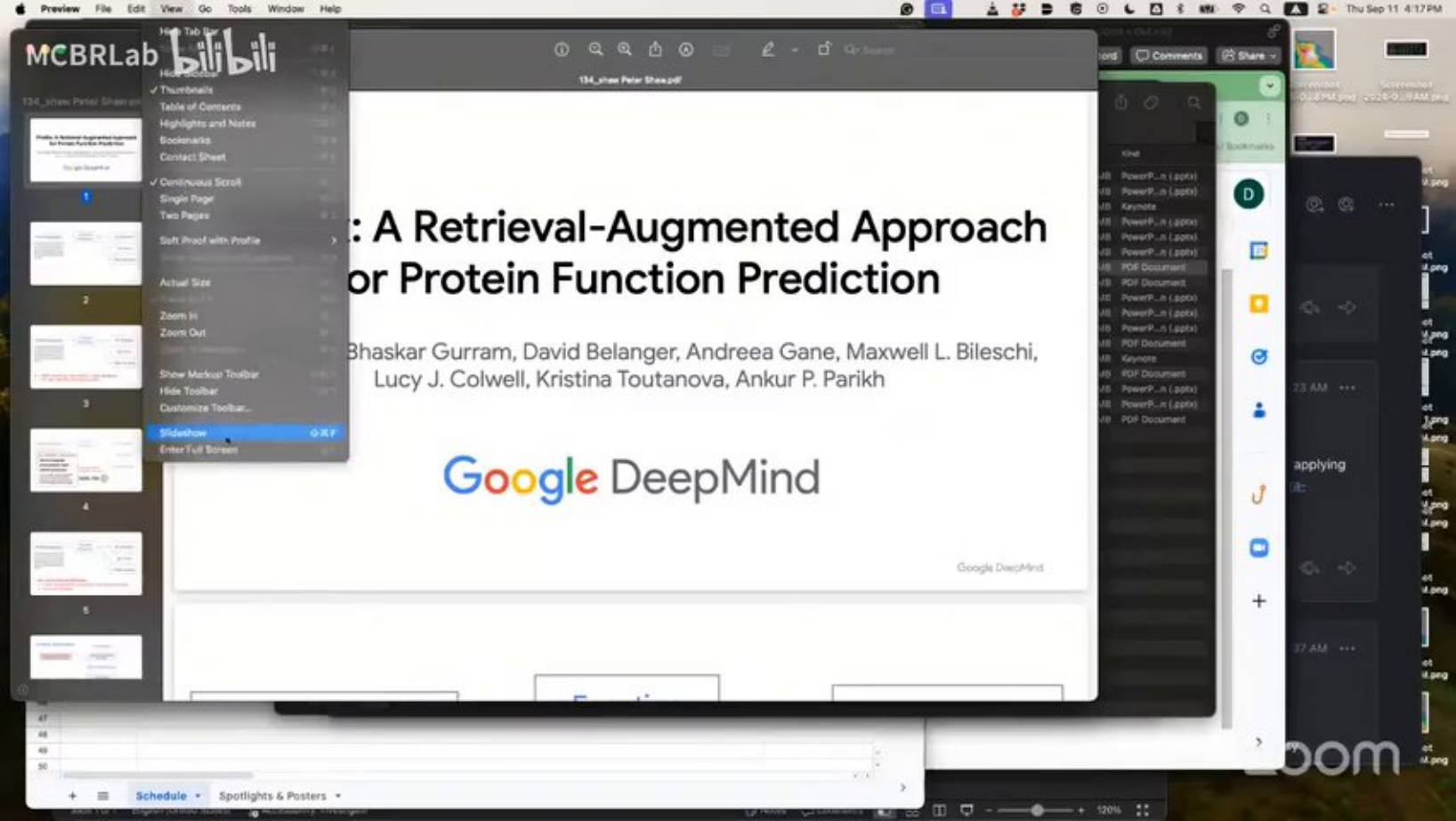
Quantitative statistical mechanical model (QSM)

MCBRLab bilibili

NYGC Events

zoom

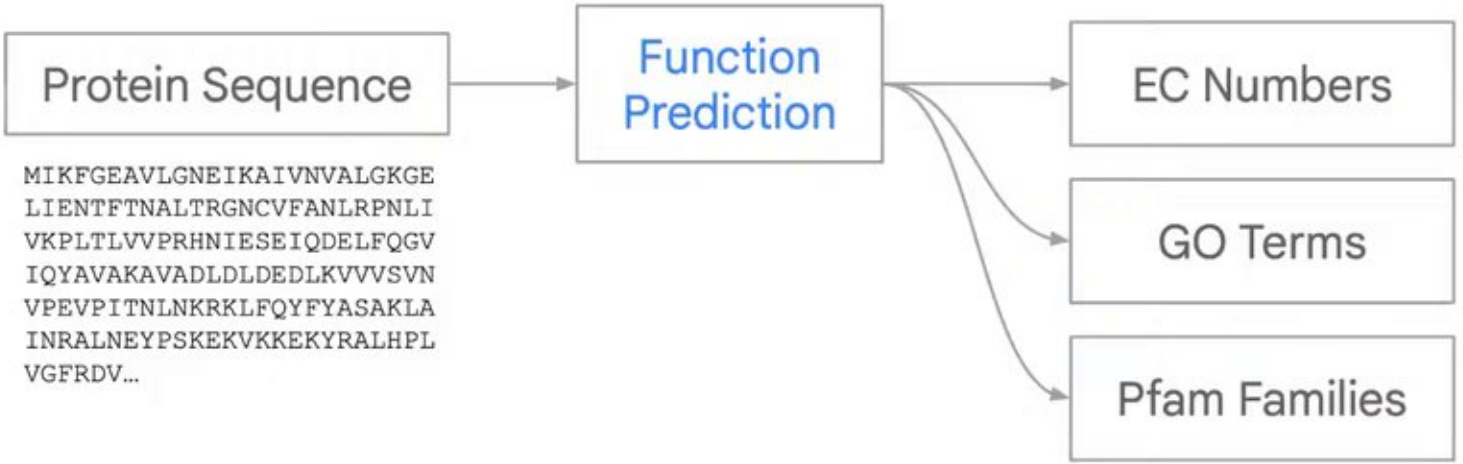

MCBRLab
bilibili
ProtEx: A Retrieval-Augmented Approach for Protein Function Prediction
Peter Shaw, Bhaskar Gurram, David Belanger, Andreea Gane, Maxwell L. Bileschi, Lucy J. Colwell, Kristina Toutanova, Ankur P. Parikh
Google DeepMind
Google DeepMind
Schedule
Spotlights & Posters


MCBRLab
bilibili
: A Retrieval-Augmented Approach
or Protein Function Prediction
Bhaskar Gurram, David Belanger, Andreea Gane, Maxwell L. Bileschi,
Lucy J. Colwell, Kristina Toutanova, Ankur P. Parikh
Google DeepMind
Schedule
Spotlights & Posters

ProtEx: A Retrieval-Augmented Approach for Protein Function Prediction

Peter Shaw, Bhaskar Gurram, David Belanger, Andreea Gane, Maxwell L. Bileschi, Lucy J. Colwell, Kristina Toutanova, Ankur P. Parikh

Google DeepMind



Protein Sequence

MIKFGEAVLGNEIKAIVNVALGKGE
LIENTETNALTRGNCVEANLRPNLI
VKP
IQY
VPE
INR
VGF

Function Prediction

EC Numbers

GO Terms

Pfam Families

News > All EMBL-EBI news > Technology and innovation

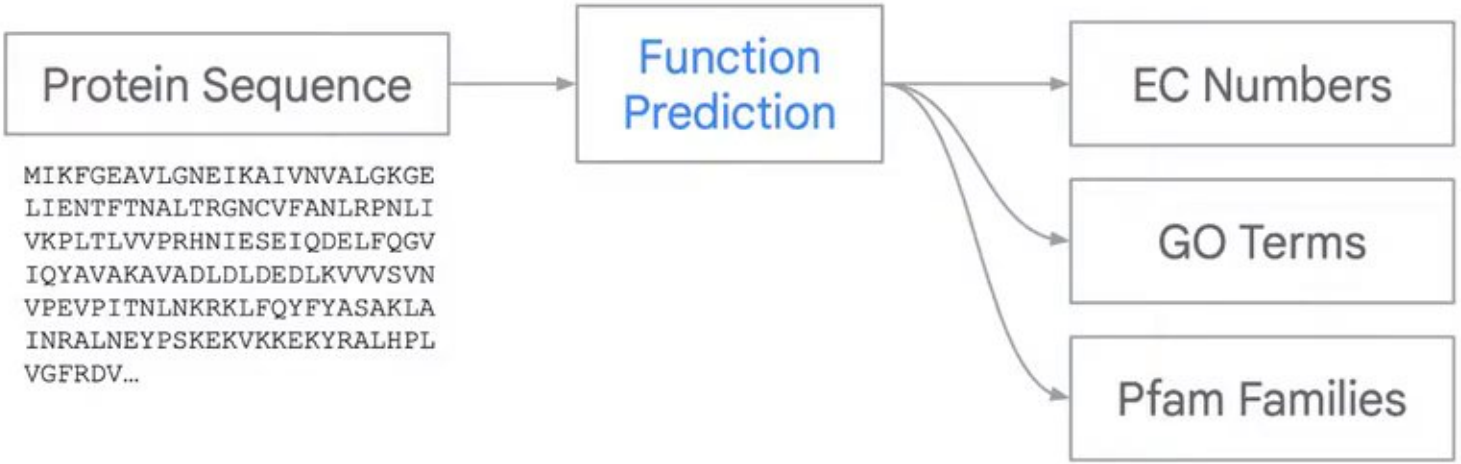
Natural language processing for rapid protein annotation

Over 40 million protein annotations have been added to the UniProt database using a Google Research natural language processing model

Gane et al. (2022) in collaboration with

EMBL-EBI

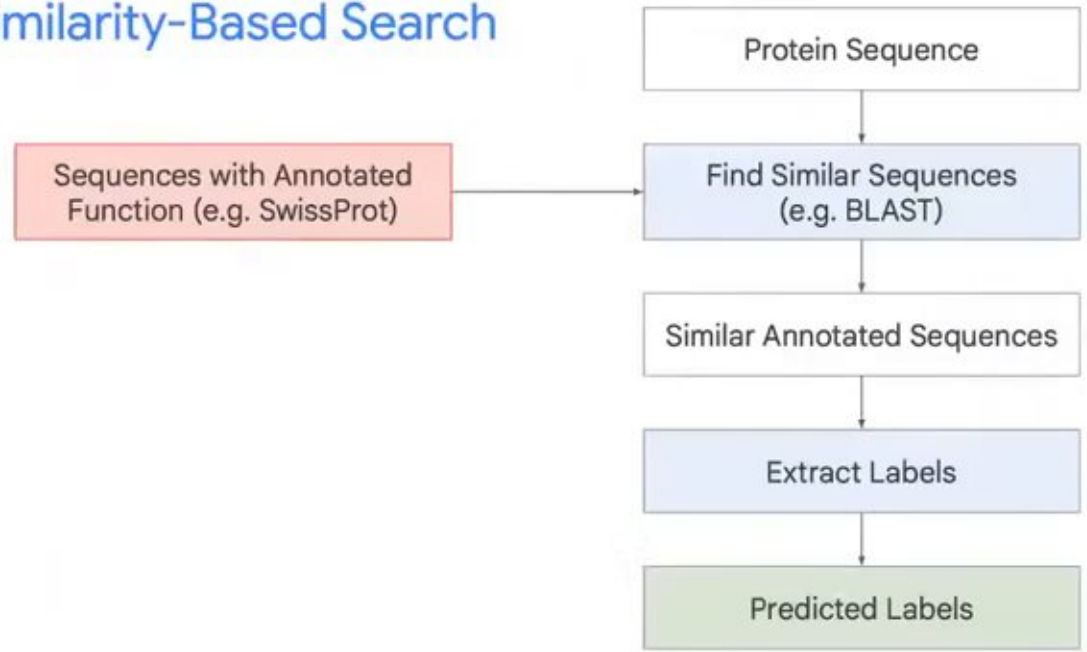
Google DeepMind



Key machine learning challenges:

- *Protein sequences that are dissimilar from annotated proteins*
- *Rare functional labels*

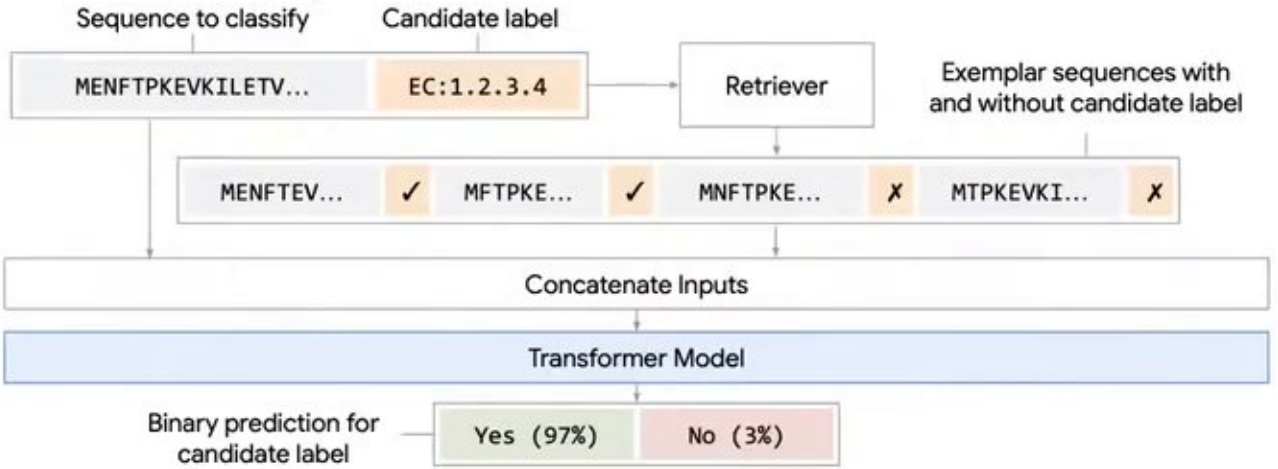
Similarity-Based Search



Work in NLP shows that retrieval-augmented models...

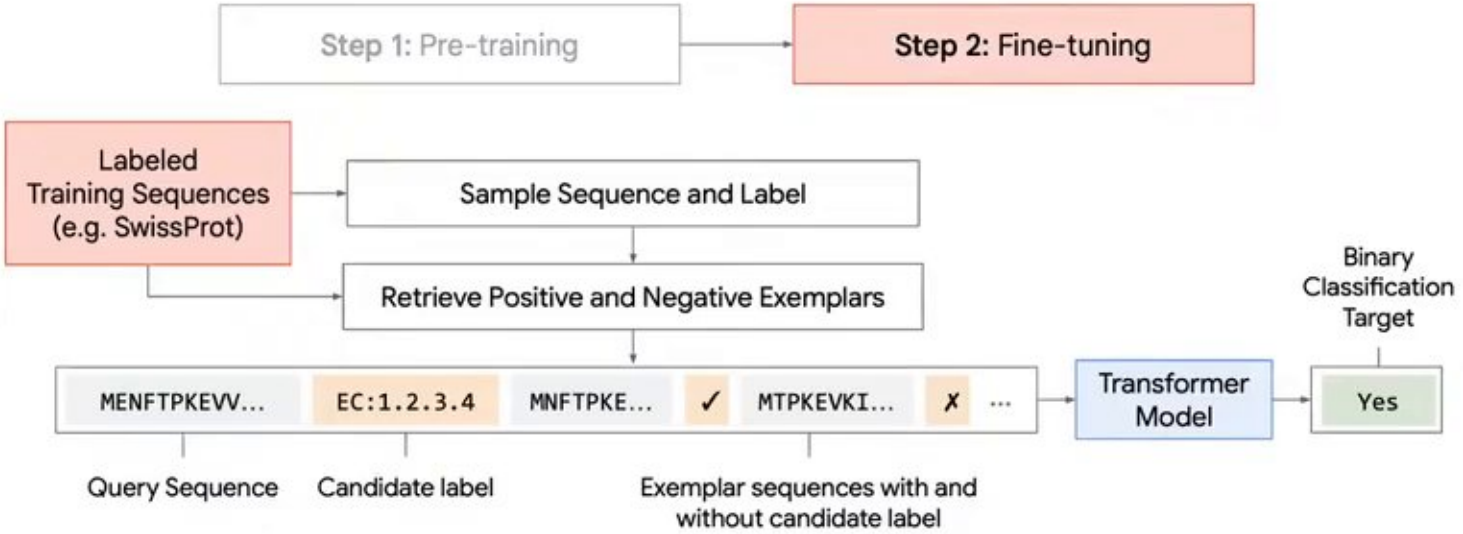
- Better capture **tail knowledge** than fully parametric approaches (Kandpal et al., 2023)
- Can perform **few-shot in-context learning** by conditioning on relevant task examples (Min et al., 2022)

ProtEx Overview



ProtEx Training Pipeline

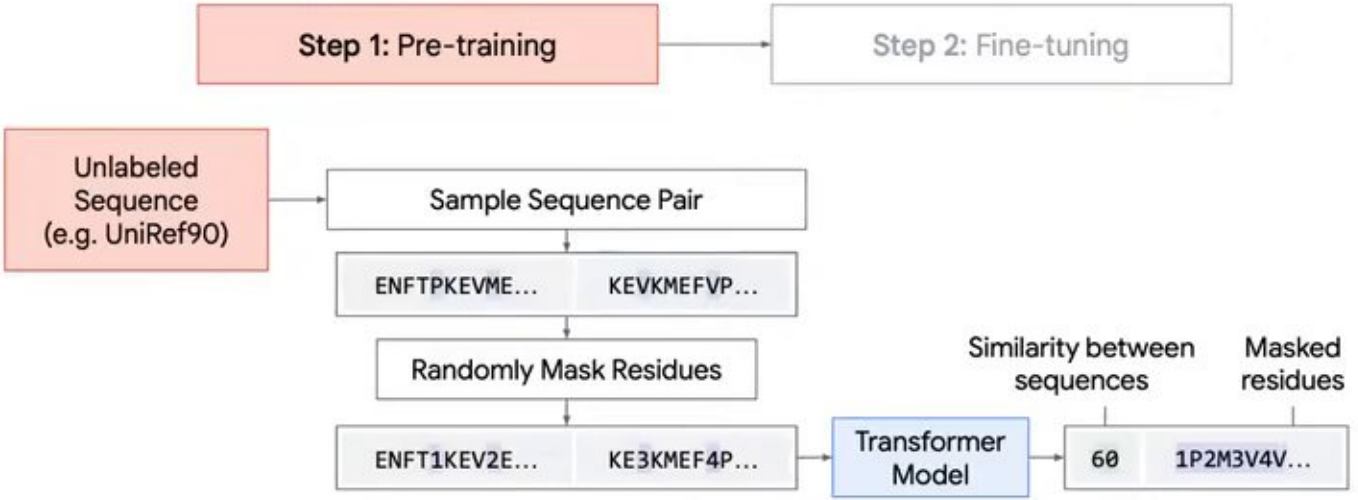
ProtEx Training Pipeline: Fine-tuning



Bias training time sampling towards less similar exemplars to improve generalization

Masked language modeling (MLM; Devlin et al., 2018) implemented as sequence-to-sequence task following T5 (Raffel et al., 2020)

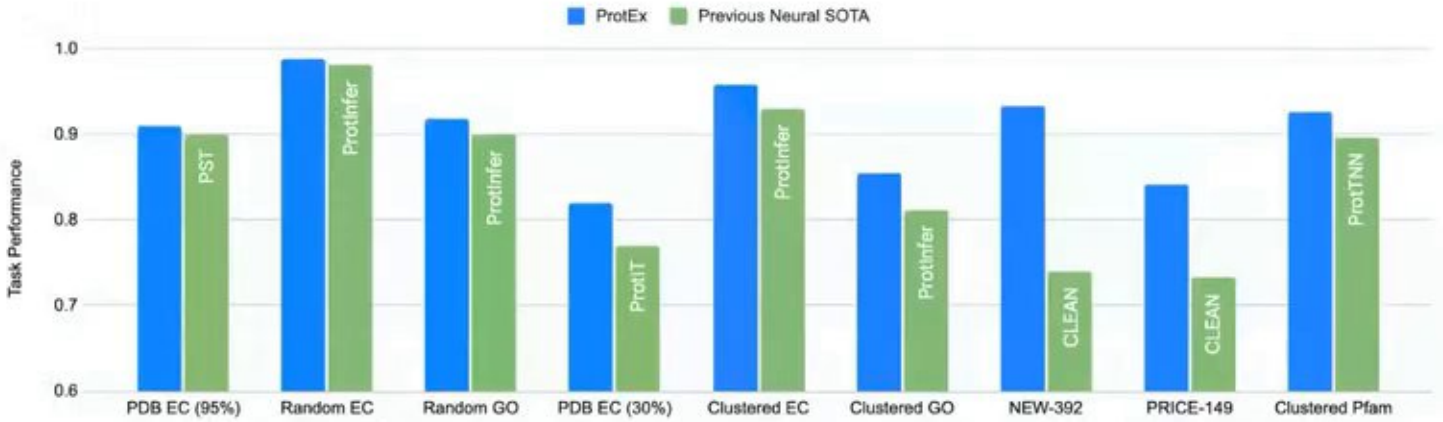
ProtEx Training Pipeline: Pre-training



Pre-training encourages model to implicitly align and compare multiple sequences

Experiments and Results

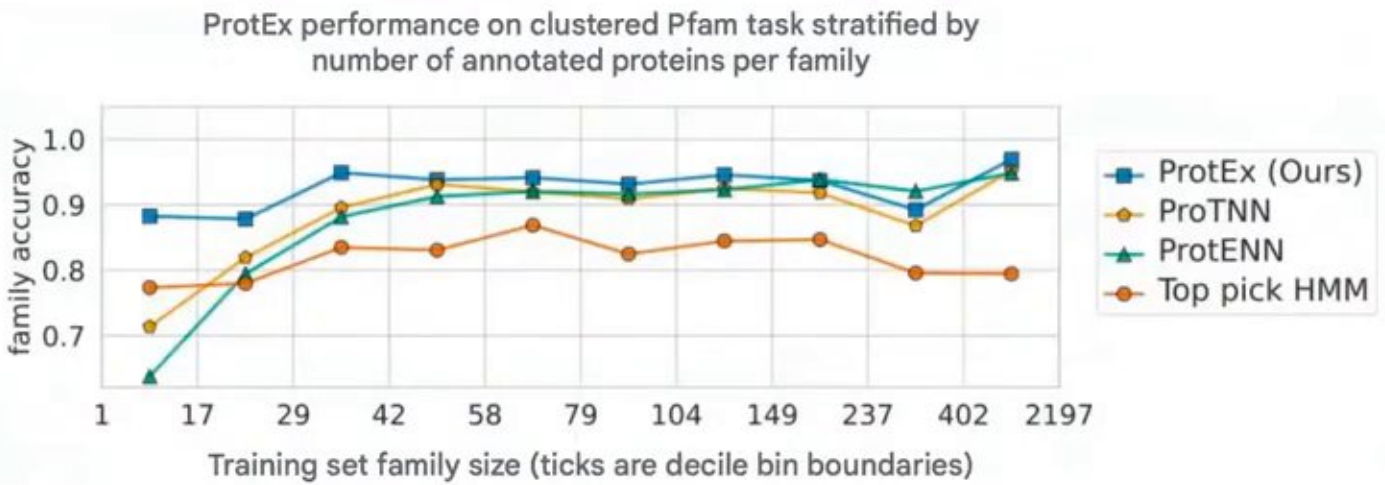
Task Performance on EC, GO, and Pfam Benchmarks



ProtEx outperforms strong neural methods from prior work across EC, GO, and Pfam baseline tasks

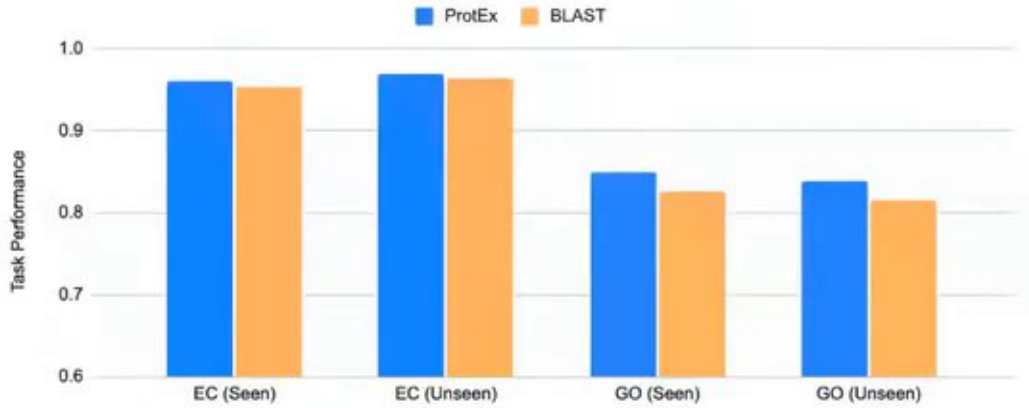
We report Max F1 for the random and clustered EC and GO splits of Sanderson et al. (2023), Weighted AUC for NEW-392 and PRICE-149 splits of Yu et al. (2023), and Family Accuracy for the Pfam clustered split of Bileschi et al. (2022). Previous SOTA neural methods are from PST (Chen et al., 2024), ProtInfer (Sanderson et al., 2023), ProtIR (Zhang et al., 2024), CLEAN (Yu et al., 2023), and ProtTNN (Dohan et al., 2021).

Performance for rare functional labels



ProtEx's improvements are largest when fewer training examples are available for a given functional label

Performance for unseen functional labels



ProtEx generalizes well even if we randomly remove a subset of functional labels during fine-tuning

Conclusions and Future Work

- Retrieving and conditioning on relevant annotated proteins can improve function prediction accuracy
- "Agents" with tools are an increasingly popular instance of inference-time retrieval
- What other types of information could be useful to retrieve at inference time for protein function prediction?

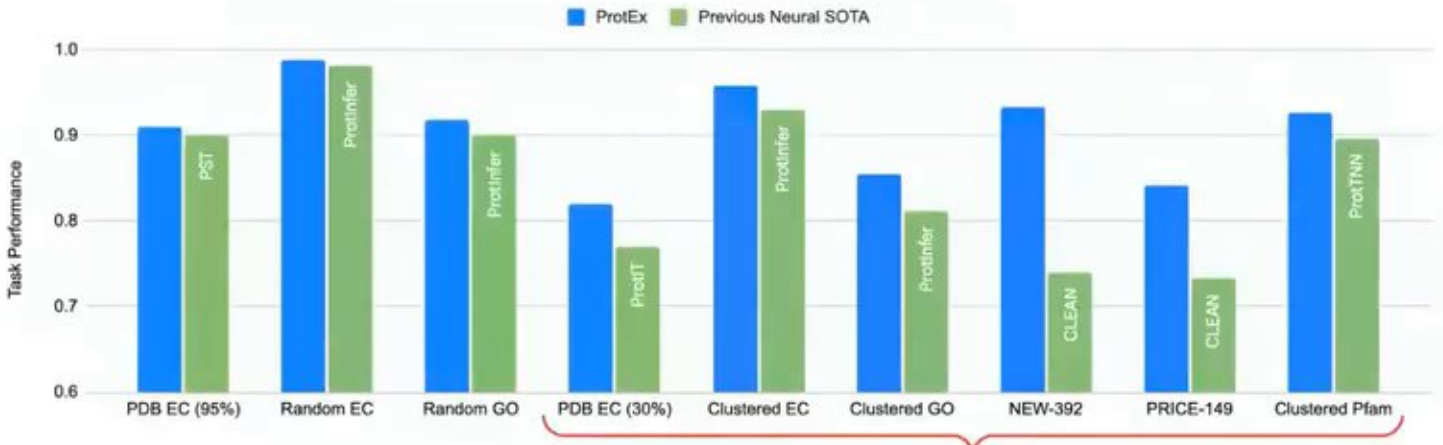
Thank you!

https://github.com/google-deepmind/protex

Peter Shaw, Bhaskar Gurram, David Belanger, Andreea Gane, Maxwell L. Bileschi,
Lucy J. Colwell, Kristina Toutanova, Ankur P. Parikh

Google DeepMind

Task Performance on EC, GO, and Pfam Benchmarks

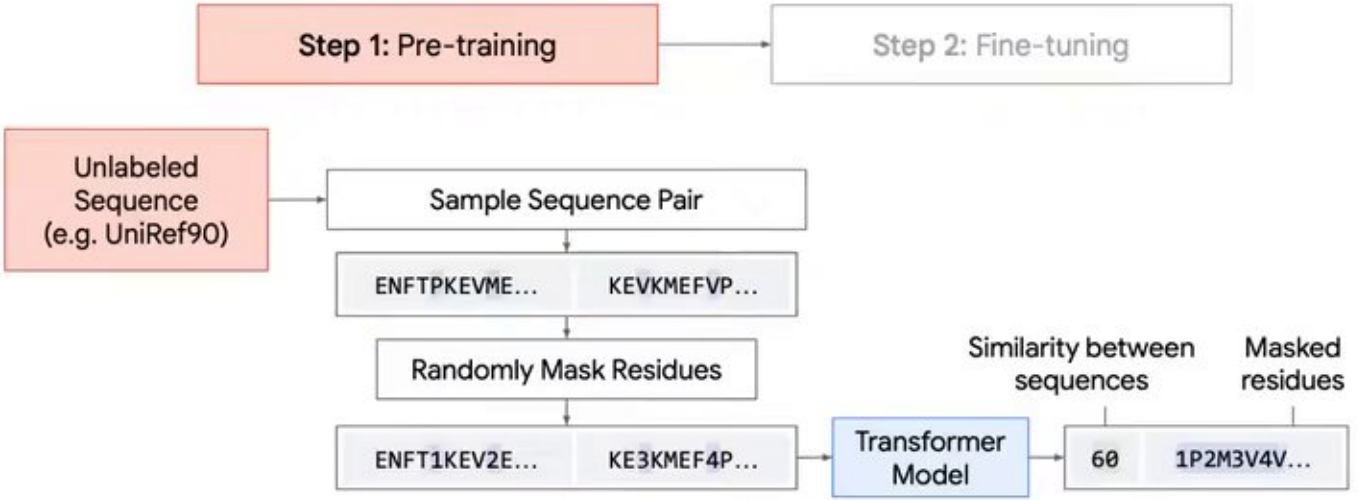


Especially for sequences dissimilar to the training set

We report Max F1 for the random and clustered EC and GO splits of Sanderson et al. (2023), Weighted AUC for NEW-392 and PRICE-149 splits of Yu et al. (2023), and Family Accuracy for the Pfam clustered split of Bileschi et al. (2022). Previous SOTA neural methods are from PST (Chen et al., 2024), ProtInfer (Sanderson et al., 2023), ProtIR (Zhang et al., 2024), CLEAN (Yu et al., 2023), and ProtTNN (Dohan et al., 2021).

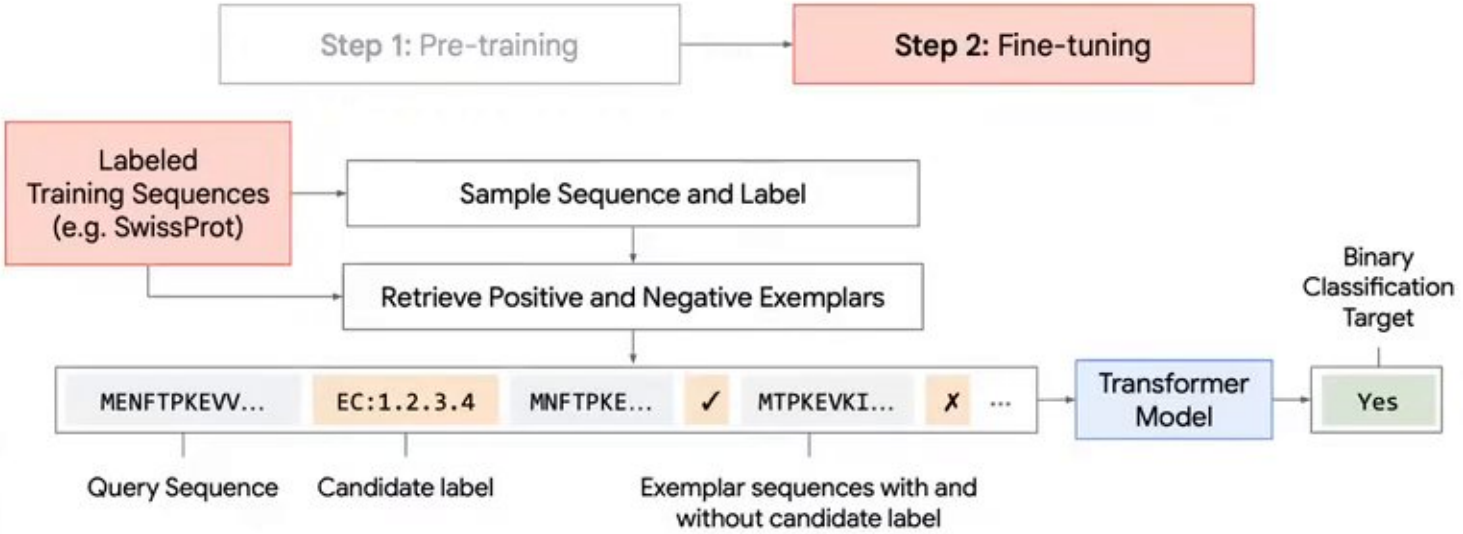
Masked language modeling (MLM; Devlin et al., 2018) implemented as sequence-to-sequence task following T5 (Raffel et al., 2020)

ProtEx Training Pipeline: Pre-training



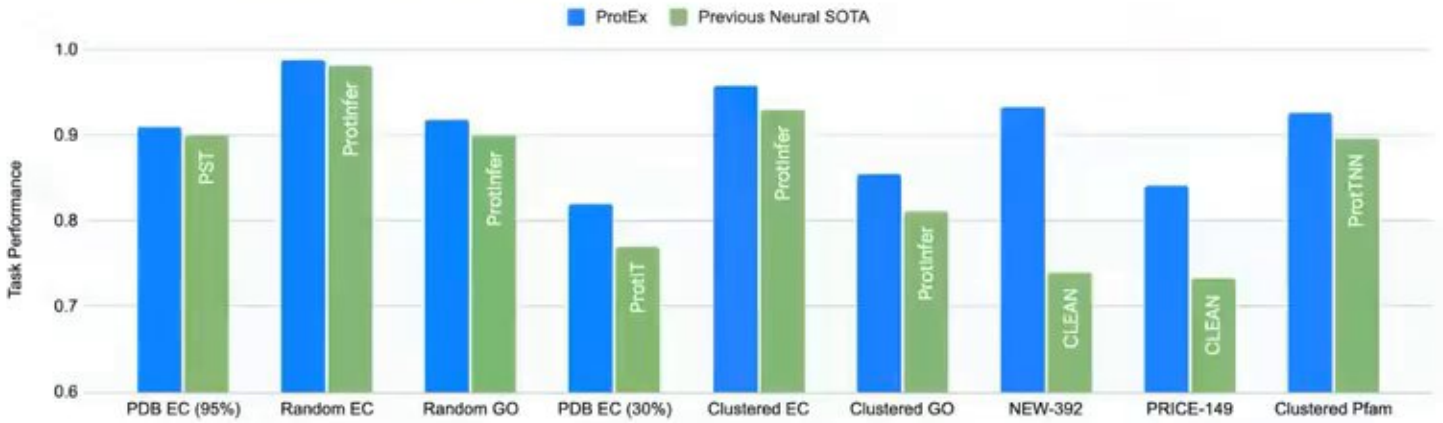
Pre-training encourages model to implicitly align and compare multiple sequences

ProtEx Training Pipeline: Fine-tuning



Bias training time sampling towards less similar exemplars to improve generalization

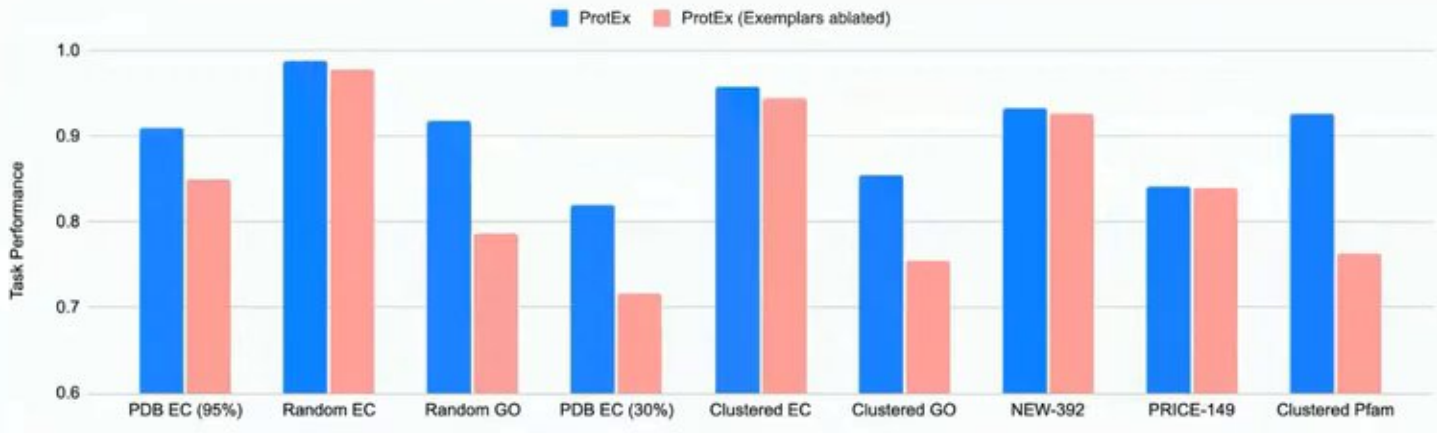
Task Performance on EC, GO, and Pfam Benchmarks



ProtEx outperforms strong neural methods from prior work across EC, GO, and Pfam baseline tasks

We report Max F1 for the random and clustered EC and GO splits of Sanderson et al. (2023), Weighted AUC for NEW-392 and PRICE-149 splits of Yu et al. (2023), and Family Accuracy for the Pfam clustered split of Bileschi et al. (2022). Previous SOTA neural methods are from PST (Chen et al., 2024), ProtInfer (Sanderson et al., 2023), ProtIR (Zhang et al., 2024), CLEAN (Yu et al., 2023), and ProtTNN (Dohan et al., 2021).

Task Performance on EC, GO, and Pfam Benchmarks



Ablations highlight that conditioning on exemplars is key to ProtEx's performance

We report Max F1 for the random and clustered EC and GO splits of Sanderson et al. (2023), Weighted AUC for NEW-392 and PRICE-149 splits of Yu et al. (2023), and Family Accuracy for the Pfam clustered split of Bileschi et al. (2022).

Conclusions and Future Work

- Retrieving and conditioning on relevant annotated proteins can improve function prediction accuracy
- "Agents" with tools are an increasingly popular instance of inference-time retrieval
- What other types of information could be useful to retrieve at inference time for protein function prediction?

Thank you!

https://github.com/google-deepmind/protex

Peter Shaw, Bhaskar Gurram, David Belanger, Andreea Gane, Maxwell L. Bileschi,
Lucy J. Colwell, Kristina Toutanova, Ankur P. Parikh

Google DeepMind

MCBRLab bilibili

NYGC Events

zoom

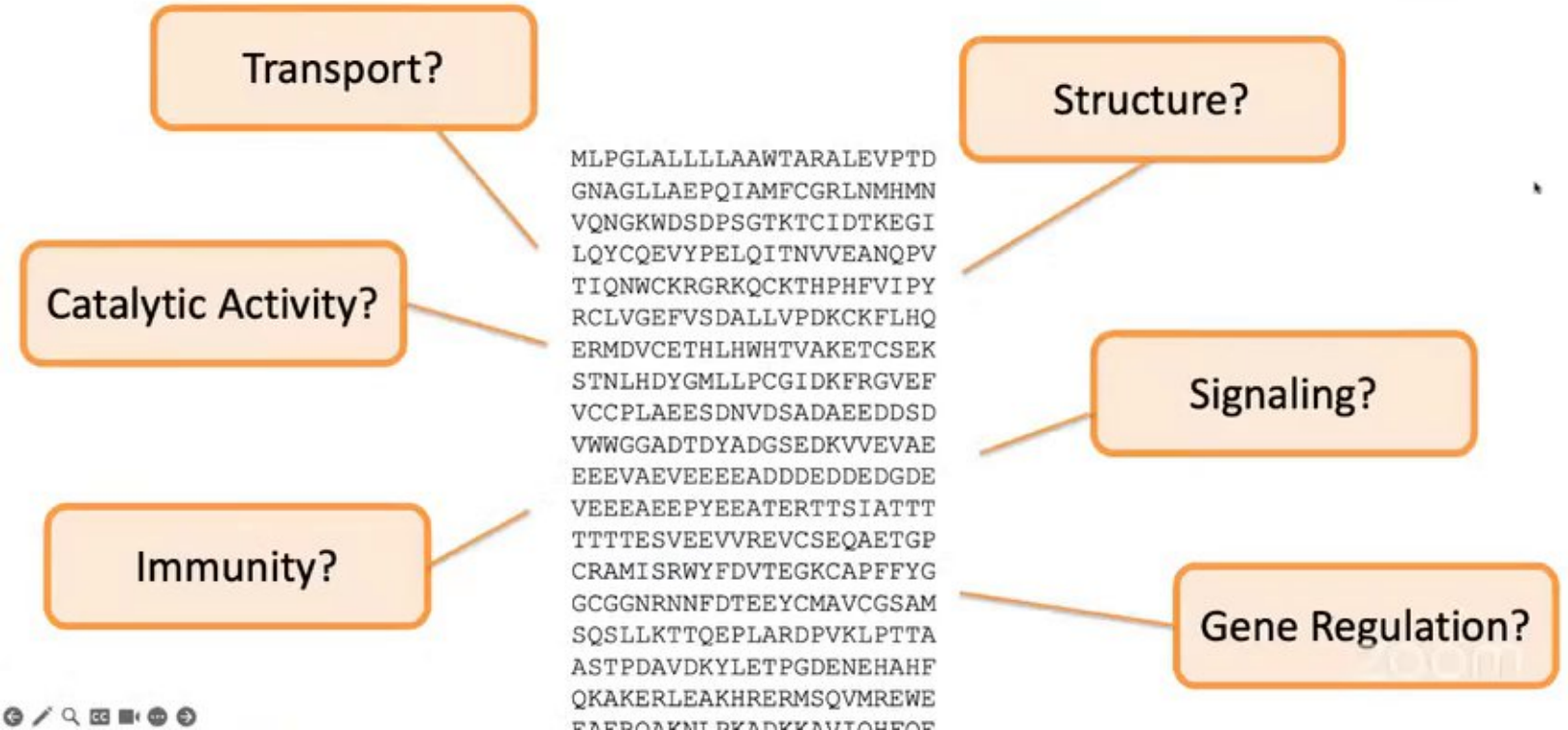
Learning Protein Dynamics at Proteome Scale

Samuel Sledzieski
Machine Learning in Computational Biology
New York Genome Center
11 September 2025

FLATIRON INSTITUTE

How do we learn about protein function?

FLATIRON INSTITUTE



How do we learn about protein function?

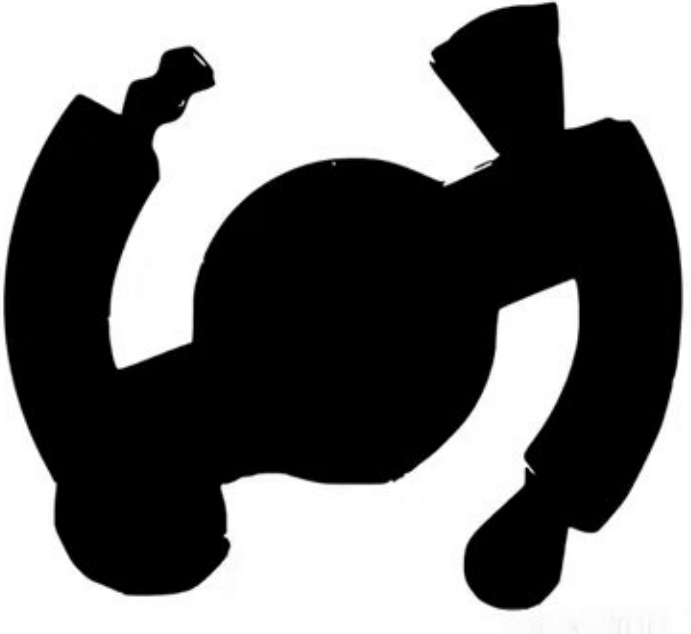
FLATIRON INSTITUTE

```
MLPGLALLLLAAWTARALEVPTDG
NAGLLAEPQIAMFCGRLNMHMNVQ
NGKWDSDPSGTKTCIDTKEGILQY
CQEVYPELQITNVVEANQPVTIQN
WCKRGRKQCKTHPHFVIPYRCLVG
EFVSDALLVPDKCKFLHQERMDVC
ETHLHWHTVAKETCSEKSTNLHDY
GMLLPCGIDKFRGVEFVCCPLAEE
SDNVDSADAEEDDSDVWWGGADTD
YADGSEDKVVEVAEEEEVAEVEEE
EADDDEDDEDGDEVEEEAEEPYEE
ATERTTSIATTTTTTESVEEVVR
EVCSEQAETGPCRAMISRWYFDVT
EGKCAPFFYGGCGGNRNNFDTEEY
CMAVCGSAMSQSLLKTTQEPLARD
PVKLPTTAASTPDAVDKYLETPGD
ENEHAHFQKAKERLEAKHRERMSQ
VMREWEEAERQAKNLPKADKKAVI
QHFQEKVESLEQE
```

How do we learn about protein function?

FLATIRON INSTITUTE

```
MLPGLALLLLAAWTARALEVPTDG
NAGLLAEPQIAMFCGRLNMHMNVQ
NGKWDSDPSGTKTCIDTKEGILQY
CQEVYPELQITNVVEANQPVTIQN
WCKRGRKQCKTHPHFVIPYRCLVG
EFVSDALLVPDKCKFLHQERMDVC
ETHLHWHTVAKETCSEKSTNLHDY
GMLLPCGIDKFRGVEFVCCPLAEE
SDNVDSADAEEDDSDVWWGGADTD
YADGSEDKVVEVAEEEEVAEVEEE
EADDDEDDEDGDEVEEEAEEPYEE
ATERTTSIATTTTTTTESVEEVVR
EVCSEQAETGPCRAMISRWYFDVT
EGKCAPFFYGGCGGNRNNFDTEEY
CMAVCGSAMSQSLLKTTQEPLARD
PVKLPTTAASTPDAVDKYLETPGD
ENEHAHFQKAKERLEAKHRERMSQ
VMREWEEAERQAKNLPKADKKAVI
QHFQEKVESLEQE
```



How do we learn about protein function?

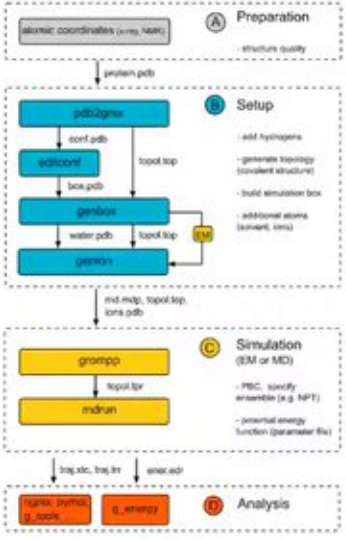
FLATIRON INSTITUTE

- Allosteric regulation requires conformational change.
- Receptor activation and signaling requires helical motion.
- Natively disordered regions are functional on binding.
- Catalytic sites may be dynamically accessible.

Current approaches to modeling dynamics

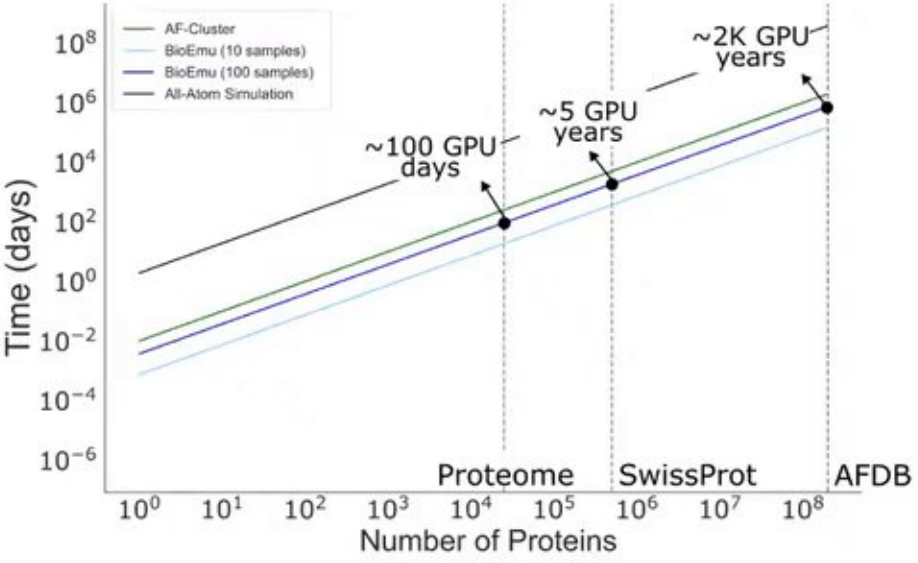
FLATIRON INSTITUTE

Simulation



https://www3.mpibpc.mpg.de/groups/de_groot/compbio/p2/index.html

Limitations – genome scale × variant effect

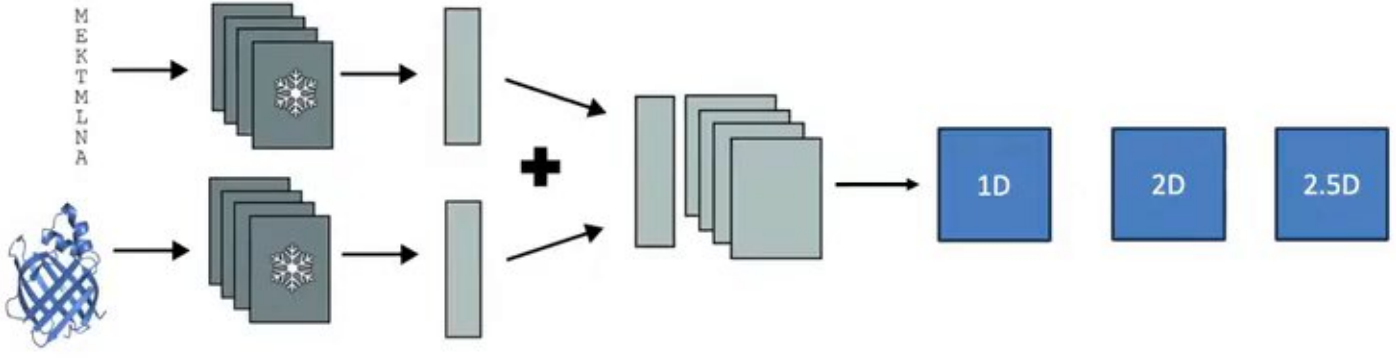


- We aim for high throughput *de novo* prediction of properties of protein dynamics.
- Differential splicing of isoforms, missense variants, somatic mutations all contribute to differential dynamics (>1 million in ClinVar as of 2021)
- Sequence space is vast: we want to functionally characterize non-human and engineered proteins

On single A100-80GB

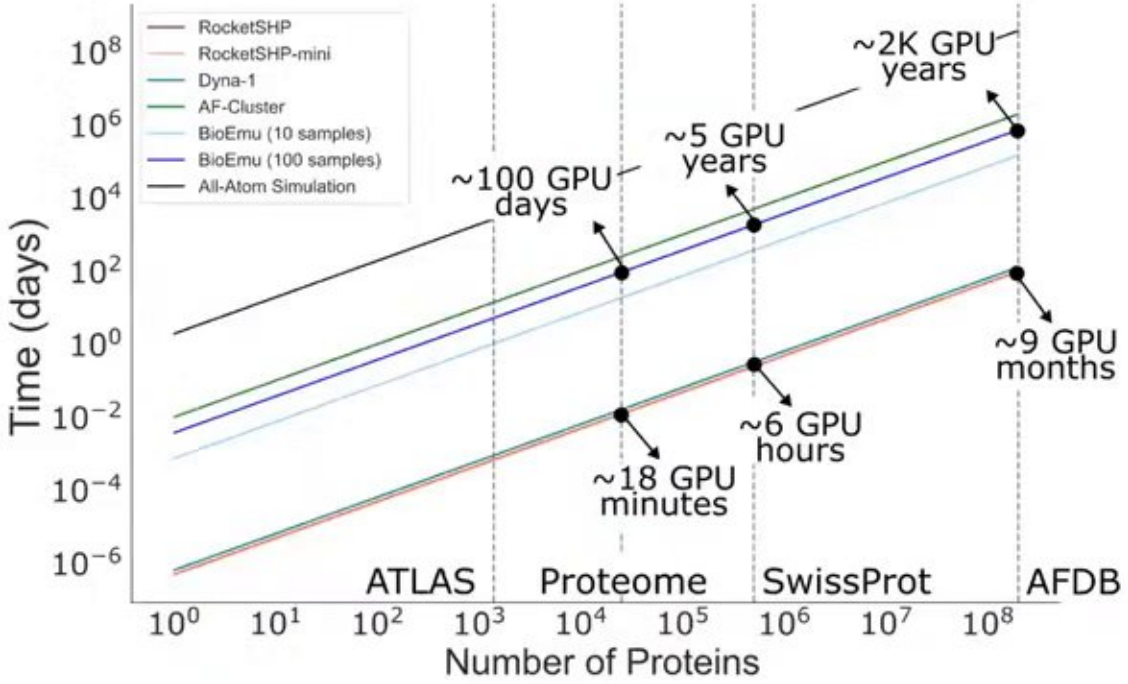
RocketSHP: fast Structure Heterogeneity Profiles

FLATIRON INSTITUTE



RocketSHP is highly scalable

FLATIRON INSTITUTE



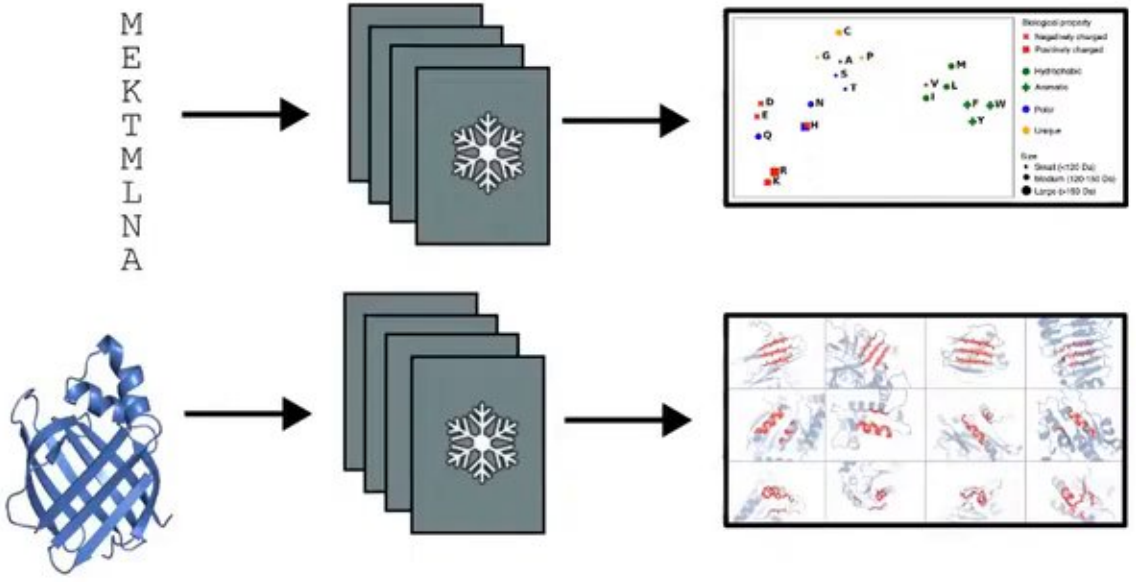
~0.1s per protein

>1,000x speedup

On single A100-80GB

Enabled by rich pre-trained features...

FLATIRON INSTITUTE



Hayes et al. "Simulating 500 million years of evolution with a language model." Science (2025)

...and large-scale MD data sets

FLATIRON INSTITUTE

ATLAS: protein flexibility description from atomistic molecular dynamics simulations

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly* and Tatiana Galochkina*

Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM, BIGR, F-75014 Paris, France

*To whom correspondence should be addressed. Tel: +33 1 81 72 43 30, Email: tatiana.galochkina@u-paris.fr
Correspondence may also be addressed to Jean-Christophe Gelly, Tel: +33 1 81 72 43 22, Email: jean-christophe.gelly@u-paris.fr

- 1390 proteins with diverse ECOD domains
- 3 replicates at 300K
- 2fs time step, coordinates saved every 10ps for 100ns
- GROMACS w/ CHARMM36m
- Clustered into 1,039 Foldseek clusters for train/validation/test splits

mdCATH: A Large-Scale MD Dataset for Data-Driven Computational Biophysics

Antonio Mirarchi[1,†], Toni Giorgino[2,†,*], and Gianni De Fabritiis[1, 3, 4, *]

[1]Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Carrer Dr. Aiguader 88, Barcelona, 08003, Spain.
[2]Biophysics Institute, National Research Council (CNR-IBF), Via Celoria 26, Milan, 20133, Italy
[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, Barcelona, 08010, Spain.
[4]Acellera Labs, Doctor Trueta 183, Barcelona, 08005, Spain.
*corresponding author(s): Gianni De Fabritiis (g.defabritiis@gmail.com), Toni Giorgino (toni.giorgino@cnr.it)
†A. M. and T. G. contributed equally to this work

- mdCATH: 5,398 CATH domains
- 5 replicates at 5 temperatures (320K, 348K, 379K, 413K, 450K)
- 4fs time step, coordinates saved every 1ns for ~464ns
- ACEMD on GPUGRID.net w/ CHARMM22
- Clustered into 3,178 Foldseek clusters for train/validation/test splits

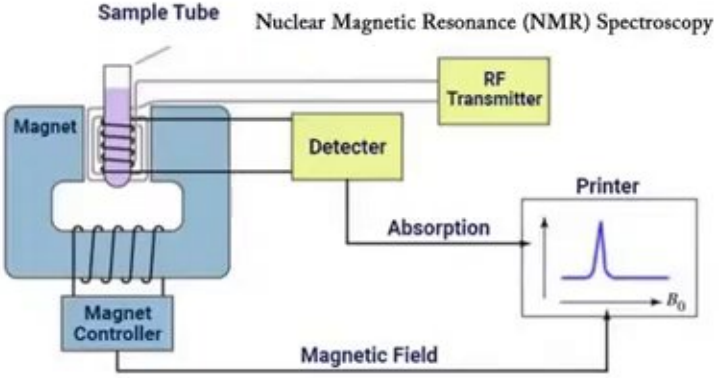
RocketSHP efficiently predicts descriptors of protein dynamics

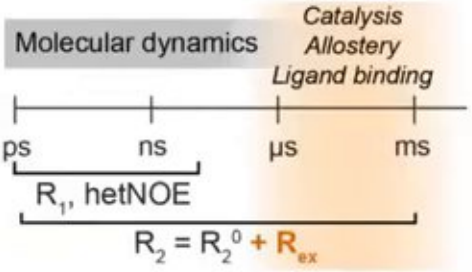
FLATIRON INSTITUTE

Model	RMSF		GCC-LMI		SHP	Model Information		
	RMSE	Spear. ρ (↑)	GDD	IMSD	KL-Div	Seq.	Struct.	Params.
RocketSHP-mini	0.104	0.680	1.023	2.532	**1.089**	✓	✗	**1.5M**
RocketSHP-seq	0.090	0.719	0.944	2.387	1.225	✓	✗	29.7M
RocketSHP	**0.083**	**0.789**	**0.859**	1.898	1.551	✓	✓	30.2M
BioEmu (10 samples)	0.192	0.715	2.550	3.357	1.923	✓	✗	31.2M
BioEmu (100 samples)	0.202	0.748	1.215	**1.871**	1.778	✓	✗	31.2M
Dyna-1	0.361	0.267	-	-	-	✓	✓	188.9M
Dyna-1 (calibrated)	0.191	0.267	-	-	-	✓	✓	188.9M

Predicting NMR hetNOE with RocketSHP

FLATIRON INSTITUTE





Wayment-Steele et al. "Learning millisecond protein dynamics from what is missing in NMR spectra." bioRxiv (2025)

RocketSHP predicts temperature-dependent behavior

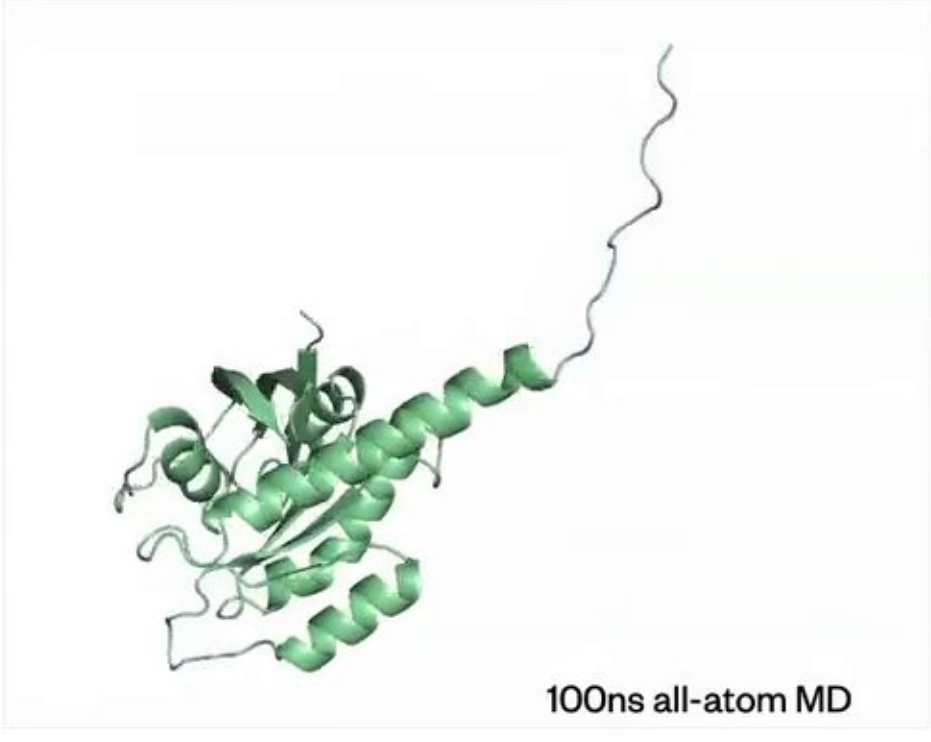
FLATIRON INSTITUTE

$$B_{obs} = B_0 e^{kt} \Longrightarrow B_{T2} = B_{T1} e^{k(T2-T1)}, k = 0.0045K^{-1}$$

de Sa Ribeiro and Lima. "Linking b-factor and temperature-induced conformational transition." Biophysical Chemistry (2023).

Case Study: Allosteric Networks in KRAS

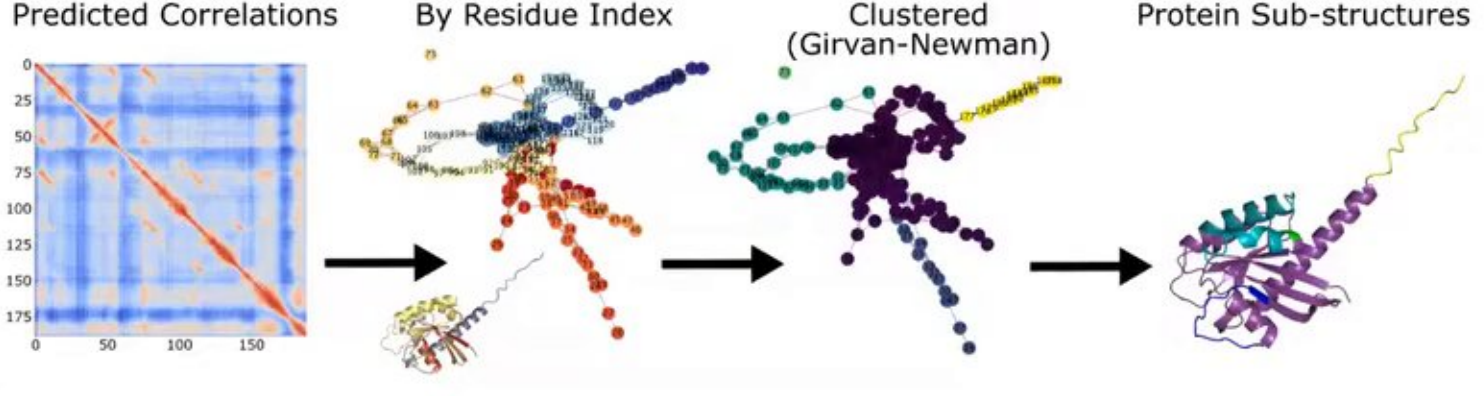
FLATIRON INSTITUTE



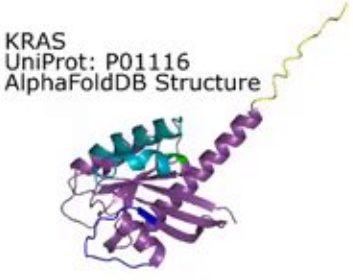
- KRAS is a frequently-mutated oncogene implicated in diverse cancer types
- Popular but challenging therapeutic target due to allosteric behavior
- "Molecular switch" relays signals only on binding by GTP
- Can we model the "allosteric network" of KRAS?

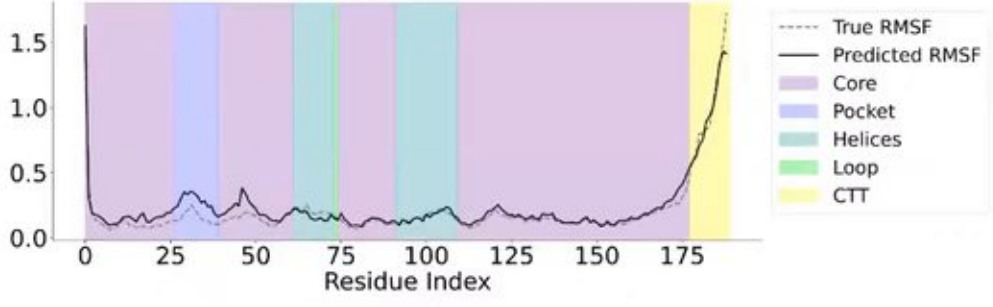
Case Study: Allosteric Networks in KRAS

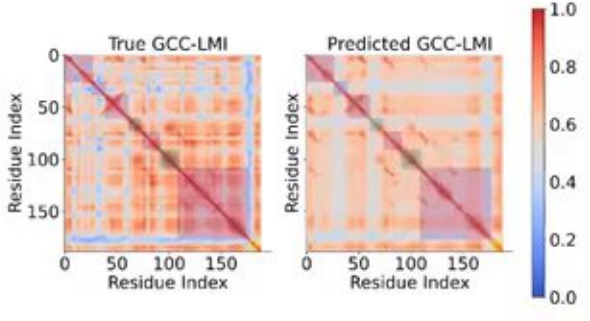
FLATIRON INSTITUTE

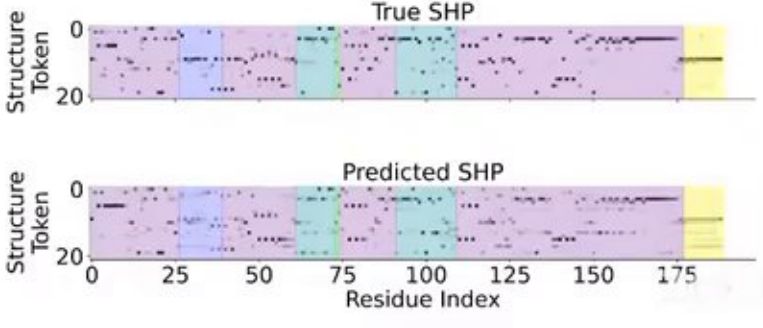


Case Study: Allosteric Networks in KRAS



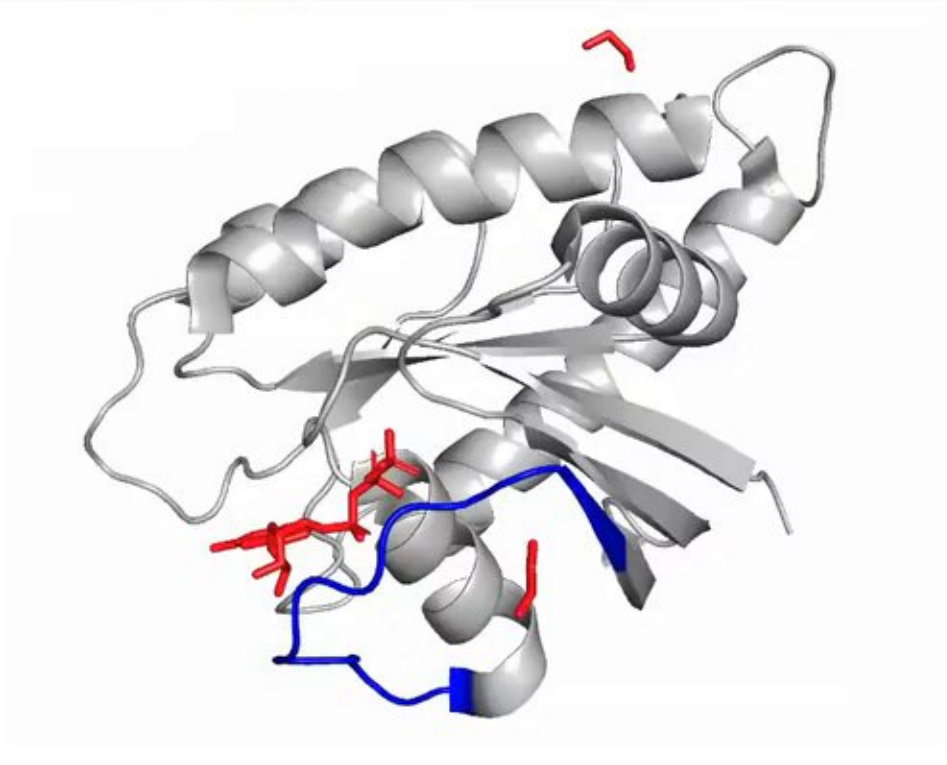






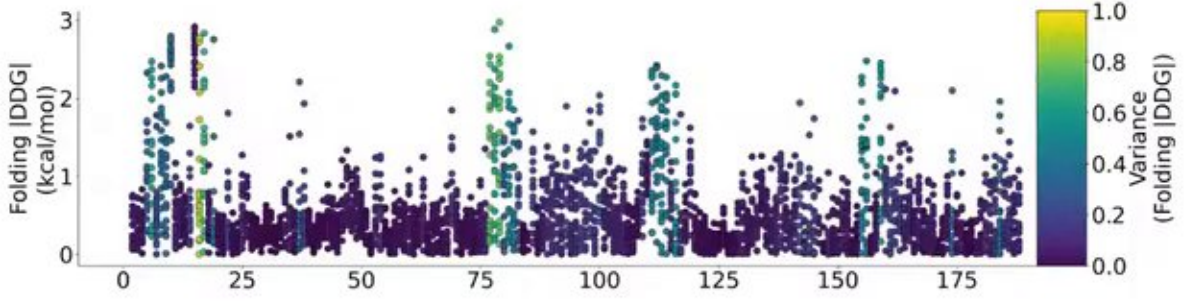
Case Study: Allosteric Networks in KRAS

FLATIRON INSTITUTE



PDB: 4DSN
KRAS w/ bound ligand

Case Study: Allosteric Networks in KRAS

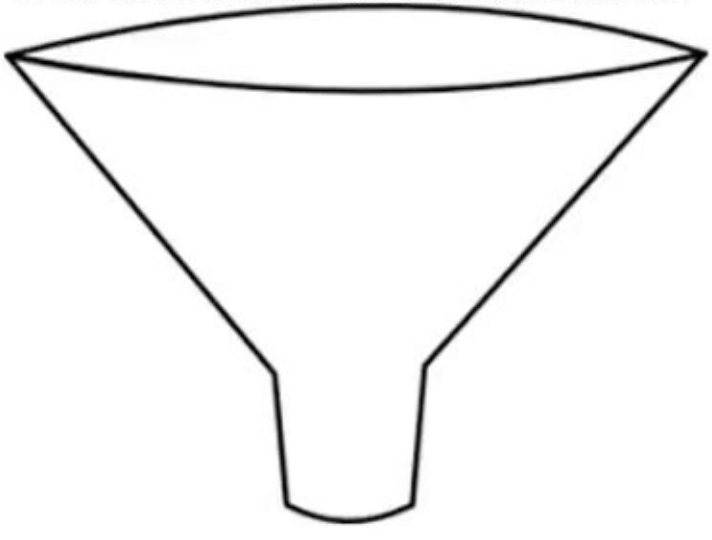


Weng et al. "The energetic and allosteric landscape for KRAS inhibition." Nature (2024)

Outlook and Limitations

FLATIRON INSTITUTE

MLPGLALLLLAAWTARALEVPTDGNAGLLAEPQIAMFCGRL
NMHMNVQNGKWDSDPSGTKTCIDTKEGILQYCQEVYPELQI
TNVVEANQPVTIQNWCKRGRKQCKTHPHFVIPYRCLVGEFV
SDALLVPDKCKFLHQERMDVCETHLHWHTVAKETCSEKSTN



Acknowledgements

FLATIRON INSTITUTE

SIMONS FOUNDATION

Sonya Hanson
Structural and Molecular Biophysics

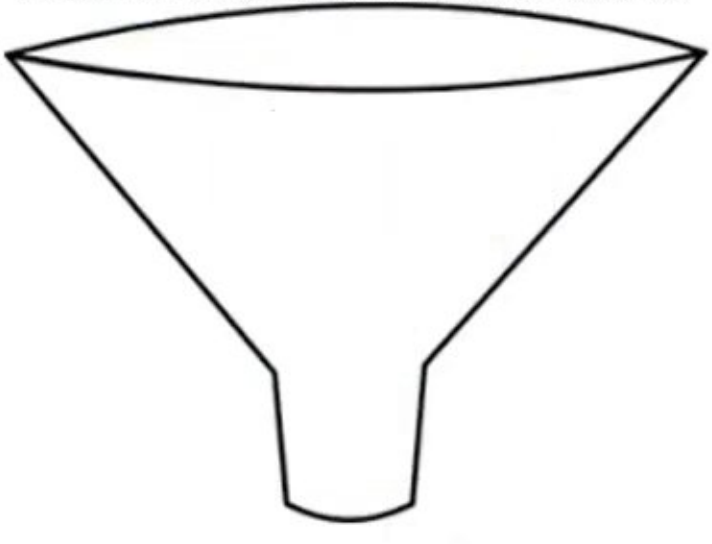
Pilar Cossio, Olga Troyanskaya, Miro Astore, Mihir Bafna, Siavash Golkar, Bowen Jing, Mahsa Mofidi, Anand Ojha, Abhilash Sahoo, Lane Votapka

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

Outlook and Limitations

FLATIRON INSTITUTE

MLPGLALLLLAAWTARALEVPTDGNAGLLAEPQIAMFCGRL
NMHMNVQNGKWDSDPSGTKTCIDTKEGILQYCQEVYPELQI
TNVVEANQPVTIQNWCKRGRKQCKTHPHFVIPYRCLVGEFV
SDALLVPDKCKFLHQERMDVCETHLHWHTVAKETCSEKSTN



RocketSHP efficiently predicts descriptors of protein dynamics

FLATIRON INSTITUTE

Model	RMSF		GCC-LMI		SHP	Model Information		
	RMSE	Spear. ρ (↑)	GDD	IMSD	KL-Div	Seq.	Struct.	Params.
RocketSHP-mini	0.104	0.680	1.023	2.532	**1.089**	✓	✗	**1.5M**
RocketSHP-seq	0.090	0.719	0.944	2.387	1.225	✓	✗	29.7M
RocketSHP	**0.083**	**0.789**	**0.859**	1.898	1.551	✓	✓	30.2M
BioEmu (10 samples)	0.192	0.715	2.550	3.357	1.923	✓	✗	31.2M
BioEmu (100 samples)	0.202	0.748	1.215	**1.871**	1.778	✓	✗	31.2M
Dyna-1	0.361	0.267	-	-	-	✓	✓	188.9M
Dyna-1 (calibrated)	0.191	0.267	-	-	-	✓	✓	188.9M

MCBRLab bilibili

NYGC Events

zoom

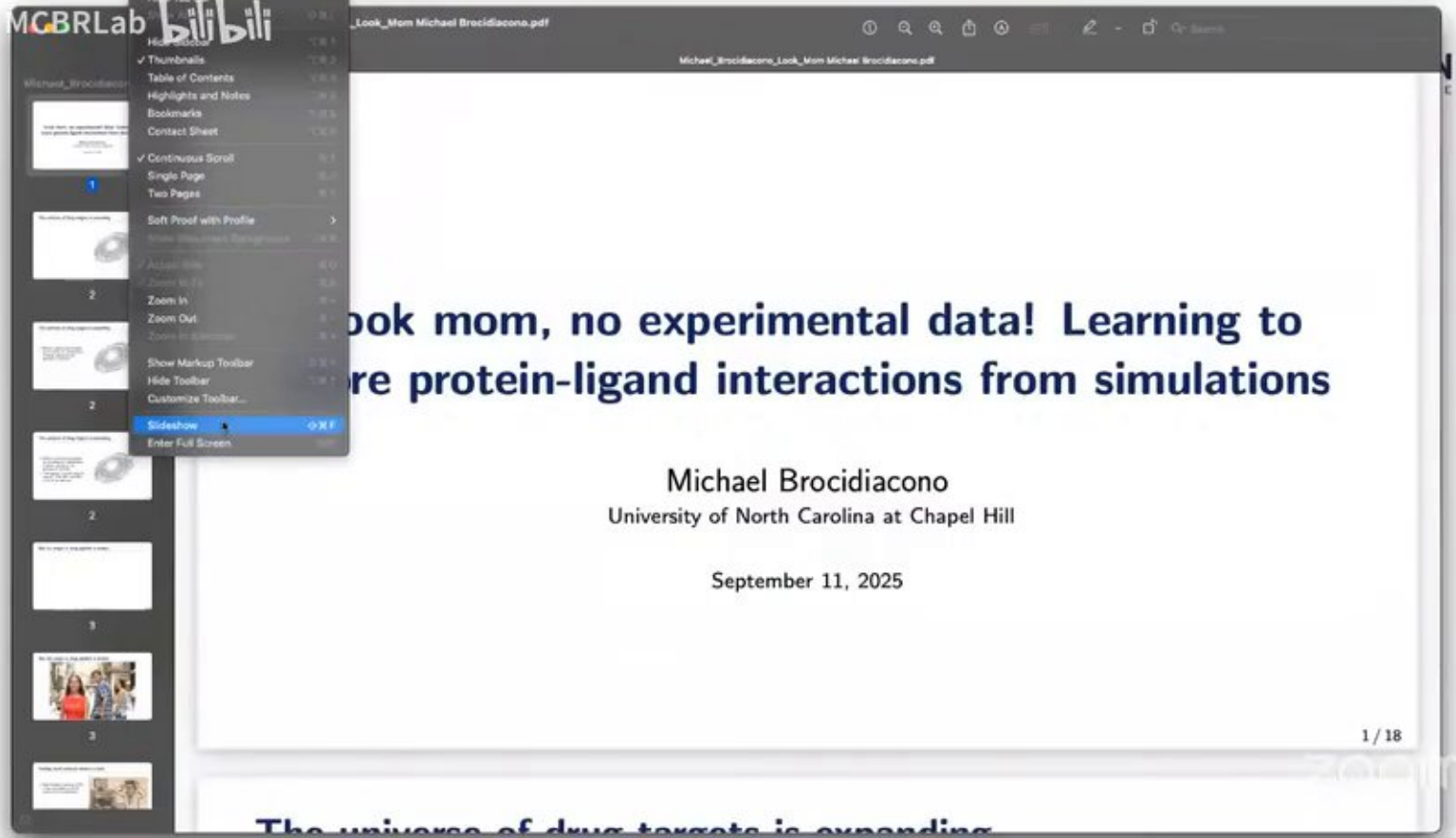
Look mom, no experimental data! Learning to score protein-ligand interactions from simulations

Michael Brocidiacono
University of North Carolina at Chapel Hill

September 11, 2025





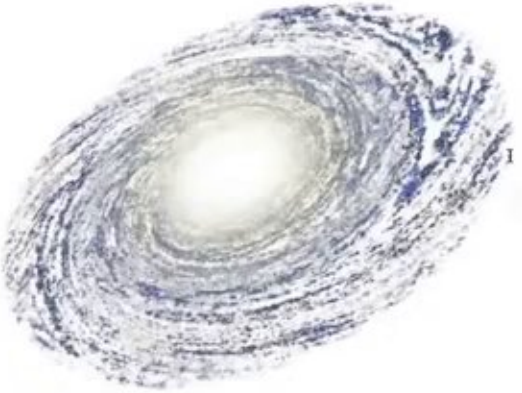


MCBRLab
bilibili
Thumbnails
Table of Contents
Highlights and Notes
Bookmarks
Contact Sheet
Continuous Scroll
Single Page
Two Pages
Soft Proof with Profile
Zoom In
Zoom Out
Show Markup Toolbar
Hide Toolbar
Customize Toolbar...
Slideshow
Enter Full Screen
ook mom, no experimental data! Learning to
re protein-ligand interactions from simulations
Michael Brocidiacono
University of North Carolina at Chapel Hill
September 11, 2025
1 / 18

Look mom, no experimental data! Learning to score protein-ligand interactions from simulations

Michael Brocidiacono
University of North Carolina at Chapel Hill

September 11, 2025

The universe of drug targets is expanding



But the target to drug pipeline is broken

But the target to drug pipeline is broken

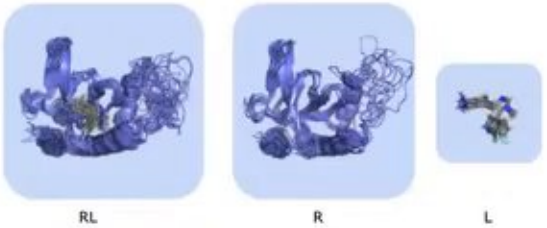
Finding small-molecule binders is hard

- High-throughput screening (HTS) is slow and expensive (and hits require a lot of optimization)

Can we computationally identify binders?

- Can pure AI solve our problems?

What's going on with free energy calculations?



- Free energy is very spicy to calculate (10K dimensional integral).

$$\Delta G_{\text{binding}} = kT \log \frac{Z_{RL}}{Z_R Z_L}$$

$$Z = \int_{-\infty}^{\infty} e^{\frac{-U(x)}{kT}} dx$$

Combining ML and physics

Force matching

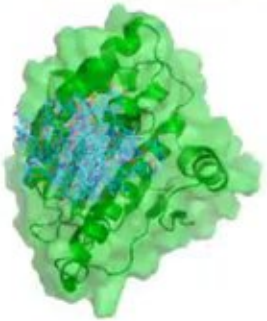
If we have a dataset of forces from MD simulations (pretty cheap), we can use a neural network to remove degrees of freedom from the system.

$$\text{Minimize} \left(-\frac{df_{NN}(x_S)}{dx_S} - \langle \mathbf{F}_{x_S} \rangle \right)^2$$

Ligand force matching (LFM) overview

a)

30-60K random
small molecules
(ZINC database)



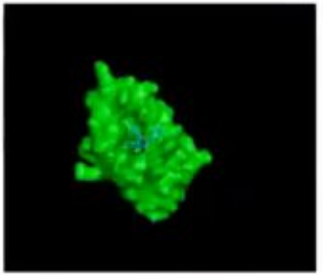
Is it working?

1. Choose targets with many known binders in ChEMBL (w/ and w/o cocrystal poses)

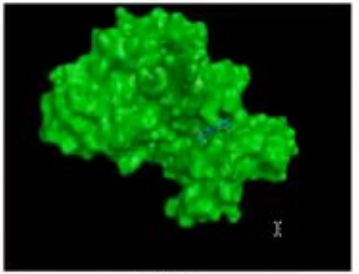
Benchmarking targets

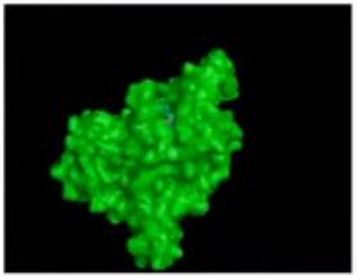
MCL1

48,284 training datapoints

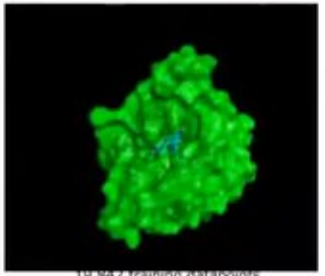


CDK2



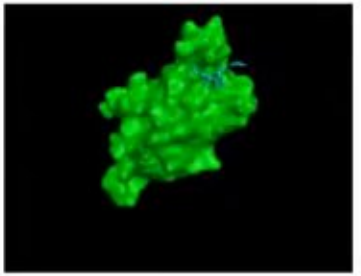


29,275 training datapoints



19,842 training datapoints

BRD4



Results (crystal poses)

Model	EF^{B}_{max}	AUC	% < 2 Å

What's going on?

DD actives + UD decoys

Model	EF^{B}_{max}	AUC
Vina	1.0 [1.0, 2.6]	0.23 ± 0.02
GNINA	7.9 [4.9, 17]	**0.65 ± 0.02**
LFM	**13 [7.9, 31]**	0.55 ± 0.04

Moving forward

- LFM is achieving promising results, though poor docking is holding us back

Acknowledgments

Grants
NIH R01GM140154

Tropsha Lab
Dr. Alexander Tropsha
Dr. Eugene Muratov
Enes Kelestemur
Dr. Holli-Joi Martin
Dr. James Wellnitz
Dr. Jon-Michael Beasley
Kelvin Idanwekhai
Kushal Koirala
Dr. Matthew Hart
Nyssa Tucker
Dr. Marcello DeLuca
Ricardo Teighi

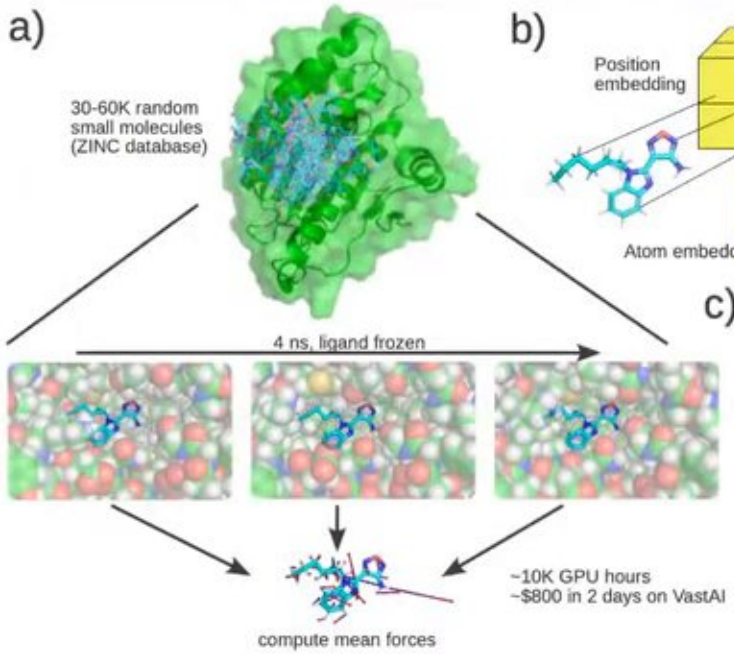
Popov Lab
Dr. Konstantin Popov
Dr. James Wellnitz
Brandon Novoy
Rishabh Dey

Koes Lab
Dr. David Koes
Rishal Aggarwal
Dr. Paul Francoeur

Results (crystal poses)

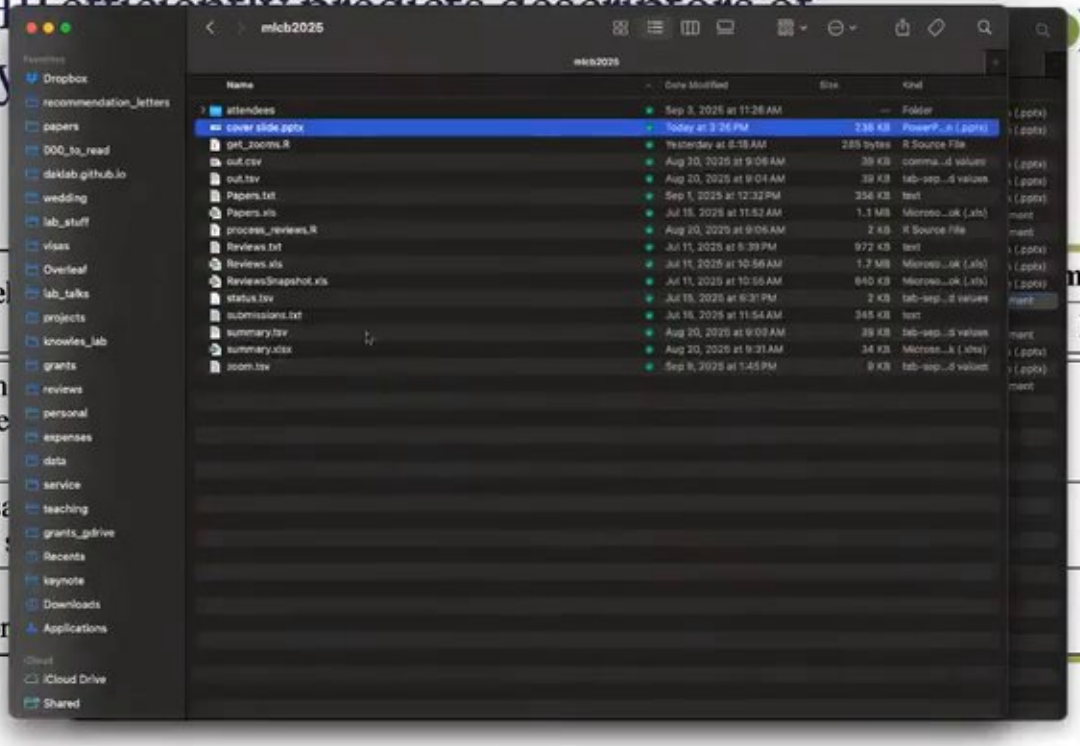
Model	EF^{B}_{max}	AUC	% < 2 Å
Vina (UD)	3.7 [1.0, 13]	0.36 ± 0.09	13 ± 5.0 %

Ligand force matching (LFM) overview



RocketSHP efficiently predicts descriptors of protein dy

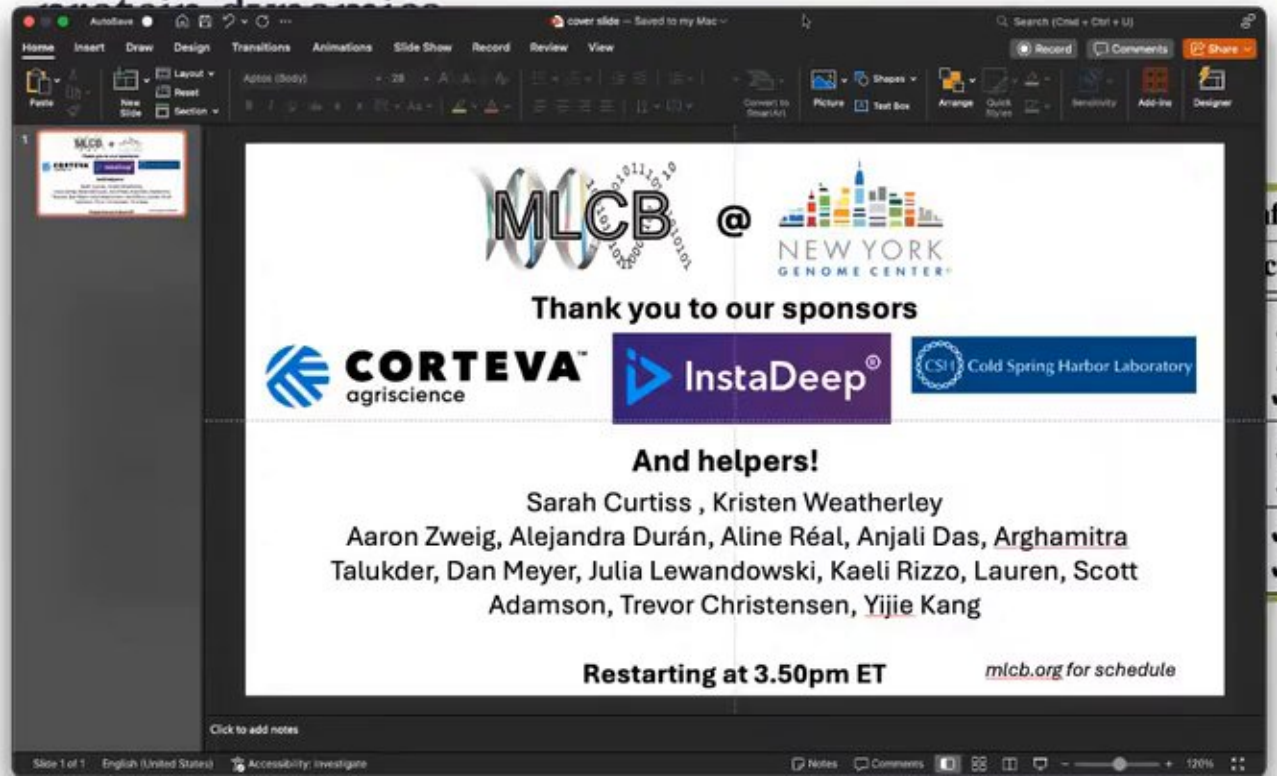
FLATIRON INSTITUTE



Model	nation Params.
RocketSHP-m	**1.5M**
RocketSHP-se	29.7M
RocketSHP	30.2M
BioEmu (10 sa	31.2M
BioEmu (100	31.2M
Dyna-1	188.9M
Dyna-1 (calibr	188.9M

RocketSHP efficiently predicts descriptors of

FLATIRON INSTITUTE



nformation	
ct.	Params.
✗	1.5M
✗	29.7M
✓	30.2M
✗	31.2M
✗	31.2M
✓	188.9M
✓	188.9M

MLCB @ NEW YORK GENOME CENTER

Thank you to our sponsors

CORTEVA agriscience

InstaDeep®

CSH Cold Spring Harbor Laboratory

And helpers!

Sarah Curtiss , Kristen Weatherley
Aaron Zweig, Alejandra Durán, Aline Réal, Anjali Das, Arghamitra Talukder, Dan Meyer, Julia Lewandowski, Kaeli Rizzo, Lauren, Scott Adamson, Trevor Christensen, Yijie Kang

Restarting at 3.50pm ET

mlcb.org for schedule