

# The Post-AlphaFold world: a new algorithmic landscape for structure-function modeling

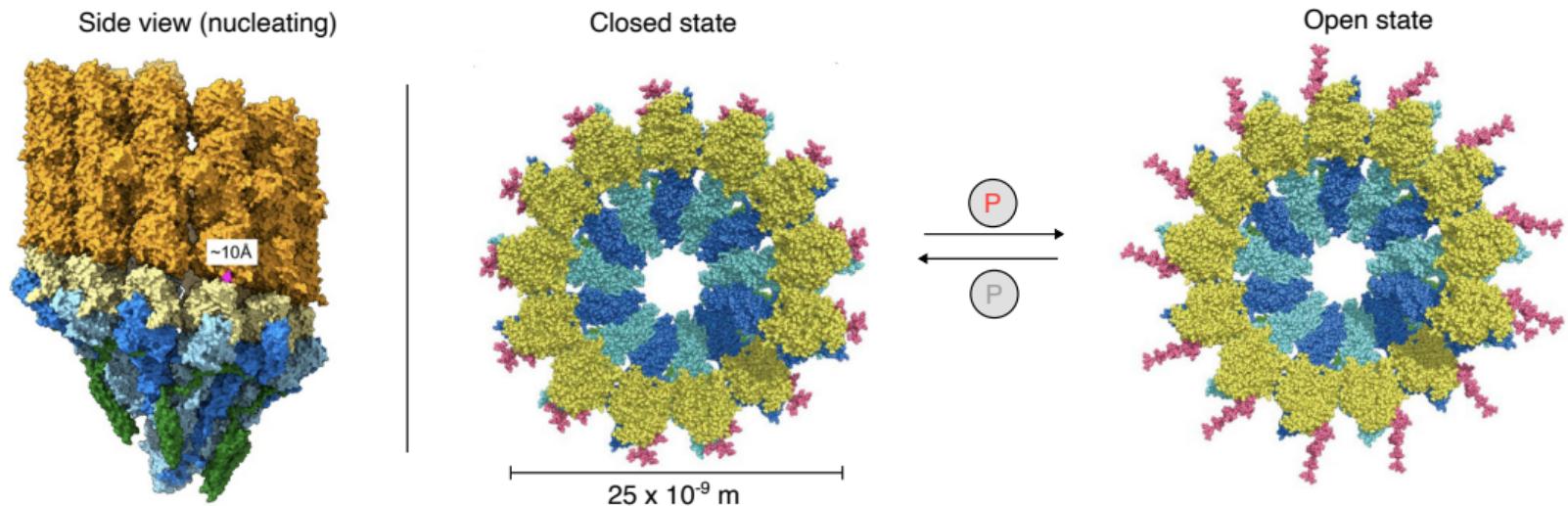
---

Carlos Oliver | Assistant Professor, Center for AI in Protein Dynamics

AI Days | March 6<sup>th</sup>, 2025

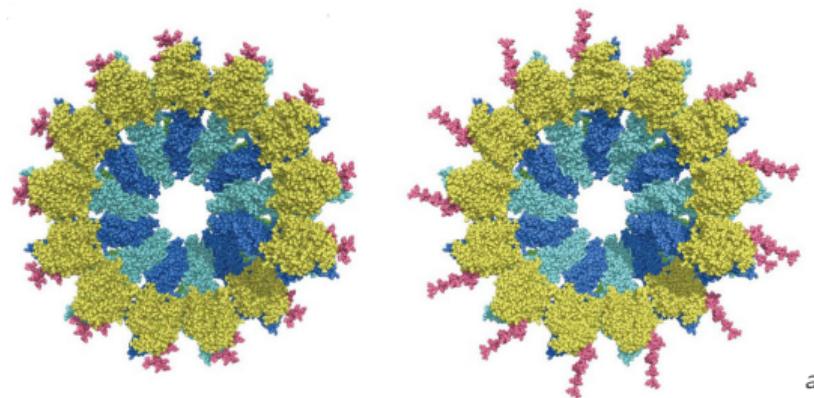
Slides: [carlosoliver.co/2025/03/05/aidays.html](http://carlosoliver.co/2025/03/05/aidays.html)

# Structure is the language of biology



<sup>1</sup>[Harris et al., 2018]

## Driving Questions



---

<sup>a</sup>[Harris et al., 2018]

- What are the **functional** components of the proteins?
- Can we detect **design** principles in these machines?
- How do we optimally **perturb** the machines?

# Tasks: mapping from structure to function

## Protein Classification

e.g. Gene Ontology or Enzyme Class



## Residue Classification

e.g. Binding Pocket Prediction



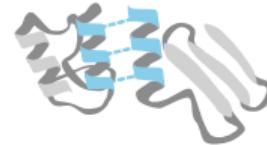
## Self-Supervision

e.g. Pretraining with AlphaFold



## Pairwise Residue

e.g. Binding Interface Prediction



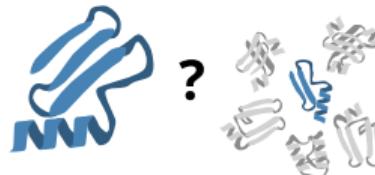
## Pairwise Protein

e.g. Structure Alignment or Protein-Protein Interaction



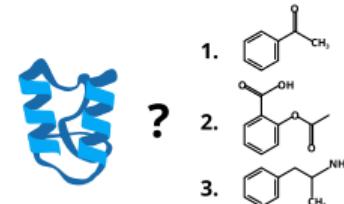
## Retrieval

e.g. Similar Structure Search



## Ranking

e.g. Drug Screening

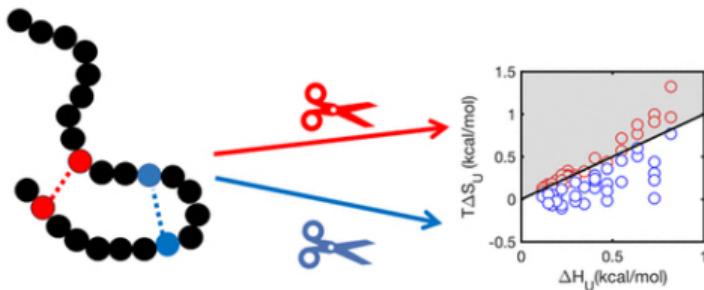


2

<sup>2</sup>[Kucera et al., 2023]

## An approach: direct perturbation or simulation

Physically measure functional properties (MD or wetlab) of a system under perturbations.



$$\Delta G^{\text{Mutant}} > \Delta G^{\text{Mutant}} > \Delta G^{\text{Wild-type}}$$

a

<sup>a</sup>Bigman, Lavi S., and Yaakov Levy., 2018

### Pros

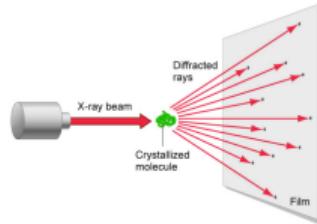
- Directly explainable

### Cons

- Slow & costly
- Lack of generalization (one experiment, one story)

# The data-driven approach

Structure Data



AlphaFold

Function Data



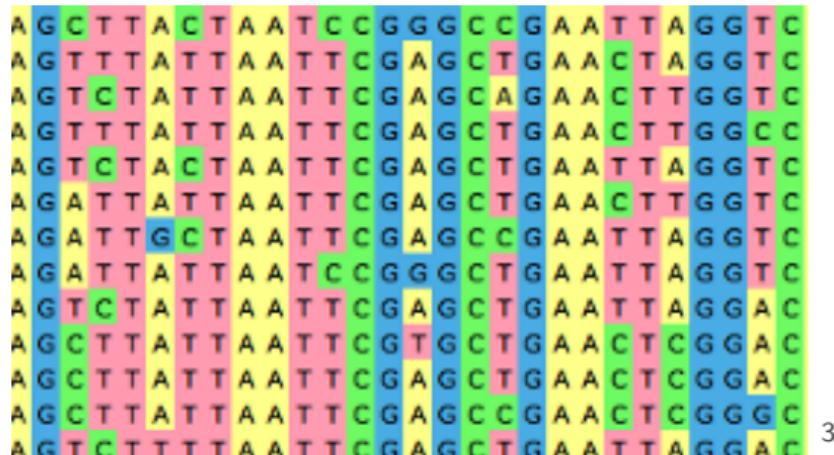
Expasy 

 PDBbind+

Pfam

# The data-driven approach

We can learn more by studying relationships between many different proteins in connection to functional knowledge.



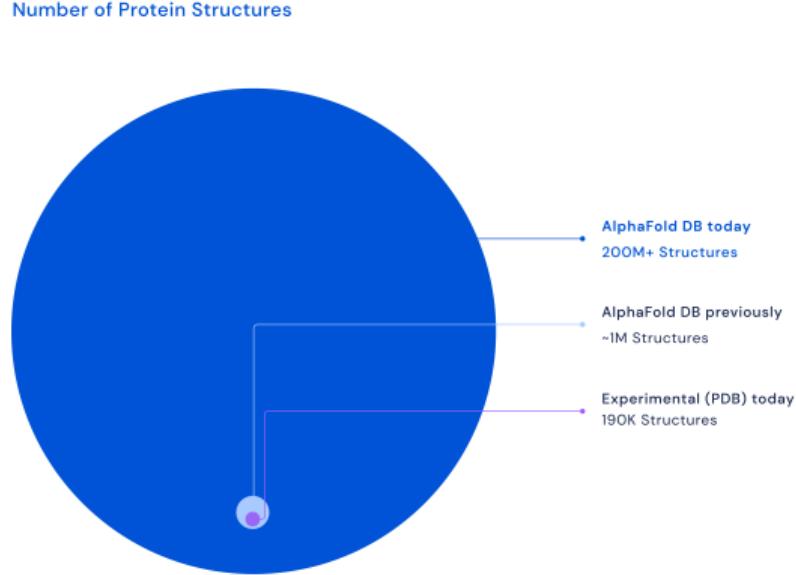
3

Main focus of bioinformatics development for the past decades.

---

<sup>3</sup><http://ugene.net/multiple-sequence-alignment-with-muscle/>

# And then came AlphaFold...

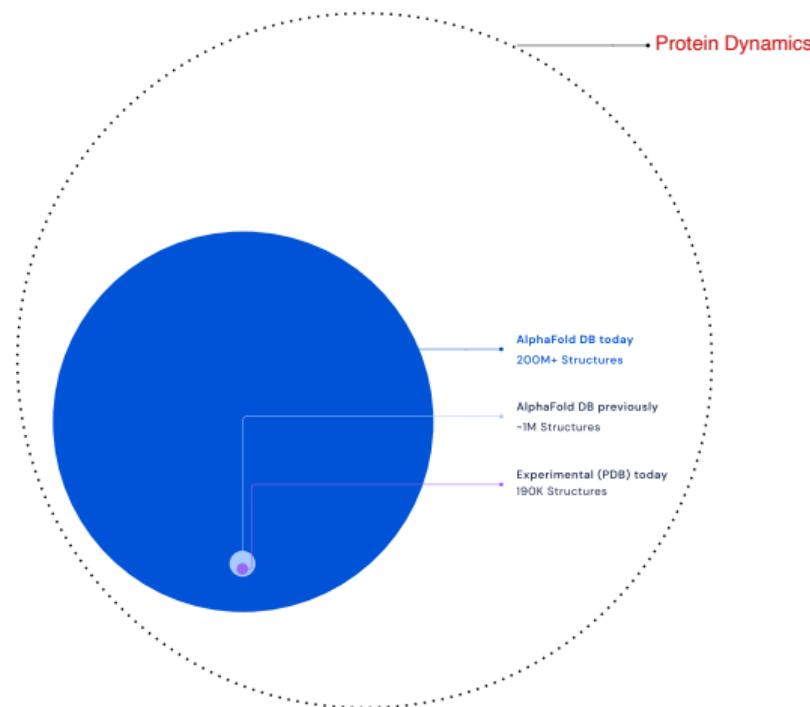


4

Massive increase in complexity and scale!

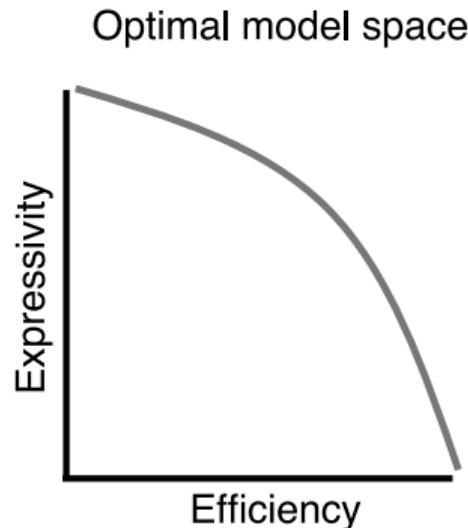
<sup>4</sup>Source: DeepMind blog

# It's going to get worse



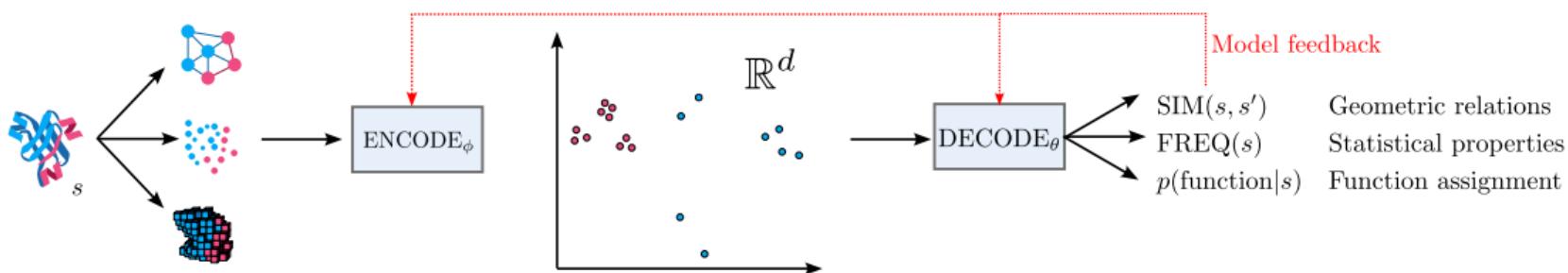
## The expressivity vs efficiency tradeoff

- Models have to balance the degree biological complexity they can capture (expressivity) with the computational speed of running them (efficiency).
- Classical tools (pre-AI) tend to lie in low-expressivity regions.



## How does AI make this shift possible?

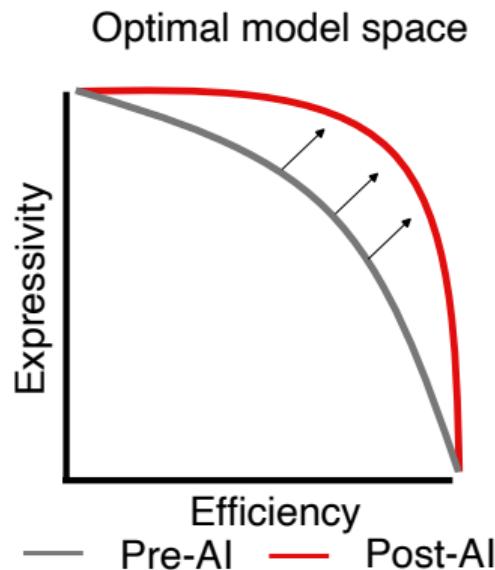
1. Neural network with parameters  $\phi$  encodes proteins from the **structure domain** to intermediate vectorial space.
2. Neural network with parameters  $\theta$  decodes proteins to the **function domain**.



**Key:** all steps are matrix multiplication-based ( $\uparrow$  efficiency), and neural networks can capture complex patterns ( $\uparrow$  expressivity).

# The case for AI in bioinformatics

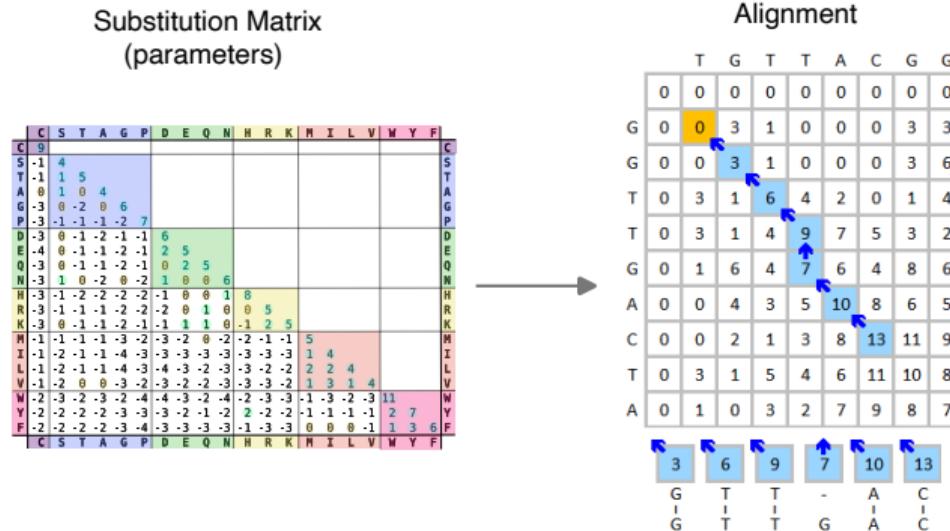
- AI models **efficiently** capture **complex** relationships that **connect** domains.



**Imperative:** discover the next generation of **high-capacity and scalable** bioinformatics tools.

# Case study: Protein Alignment

- How do we measure the similarity of two proteins?
- Prior: fundamental unit is single amino acids ( $\uparrow$  Efficiency)
- Prior: substitution cost for all pairs of amino acids is fixed ( $\downarrow$  Expressivity)



# Case study: Protein Alignment

- How do we measure the similarity of two proteins?
- Prior: fundamental unit is single amino acids ( $\uparrow$  Efficiency)
- Prior: substitution cost for all pairs of amino acids is fixed ( $\downarrow$  Expressivity)

Substitution Matrix  
(parameters)

C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F
C 9																			C
S -1 4																			S
T -1 1 5																			T
A 0 1 0 4																			A
G -3 0 -2 0 6																			G
P -3 -1 -1 -1 -2 7																			P
D -3 0 -1 -2 -1 6																			D
E -4 0 -1 -1 -2 -1 2 5																			E
Q -3 0 -1 -1 -2 -1 0 2 5																			Q
N -3 1 0 -2 0 -2 1 0 0 6																			N
H -3 -1 -2 -2 -2 -2 -1 0 0 1 8																			H
R -3 -1 -1 -1 -2 -2 -1 0 1 0 5																			R
K -3 0 -1 -1 -2 -1 -1 1 1 0 -1 2 5																			K
M -1 -1 -1 -1 -3 -2 -3 -2 0 -2 -2 -1 -1 5																			M
I -1 -2 -1 -1 -4 -3 -3 -3 -3 -3 -3 -3 -3 1 4																			I
L -1 -2 -1 -1 -4 -3 -4 -3 -2 -3 -3 -2 -2 2 2 4																			L
V -1 -2 0 -3 -3 -2 -3 -2 -2 -3 -3 -3 -2 1 3 1 4																			V
W -2 -3 -2 -3 -2 -4 -4 -3 -2 -4 -2 -3 -3 -1 -3 -2 -3 11																			W
Y -2 -2 -2 -2 -3 -3 -3 -2 -1 -2 2 -2 -2 -1 -1 -1 2 7																			Y
F -2 -2 -2 -2 -3 -4 -3 -3 -3 -1 -3 -3 0 0 0 1 3 6																			F
C S T A G P D E Q N H R K M I L V W Y F																			C S T A G P D E Q N H R K M I L V W Y F



Alignment

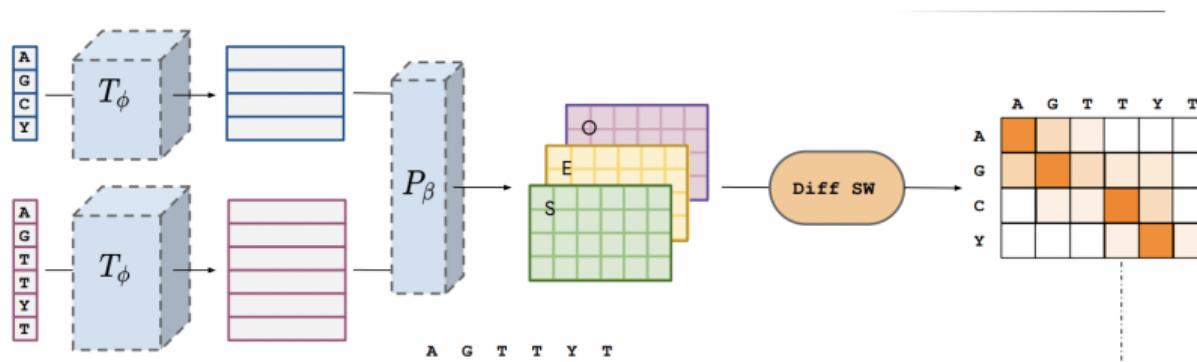
T	G	T	T	A	C	G	G
0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	3
G	0	0	3	1	0	0	3
T	0	3	1	6	4	2	0
T	0	3	1	4	9	7	5
G	0	1	6	4	7	6	4
A	0	0	4	3	5	10	8
C	0	0	2	1	3	8	13
T	0	3	1	5	4	6	11
A	0	1	0	3	2	7	9

Below the alignment grid:

3	6	9	7	10	13
G	T	T	-	A	C
G	T	T	G	A	C

# Beyond fixed substitution costs [Llinares-López et al., 2023]

- Allow flexible substitution costs ( $\uparrow$  Expressivity)
- Substitution costs become a model parameter.
- Parameter is tuned by backpropagation using function data.



# Case study: Protein Alignment

- How do we measure the similarity of two proteins?
- Prior: substitution cost for all pairs of amino acids is fixed ( $\downarrow$  Expressivity)
- **Prior: fundamental unit is single amino acids ( $\uparrow$  Efficiency)**
- Solution: dynamic programming

Substitution Matrix  
(parameters)

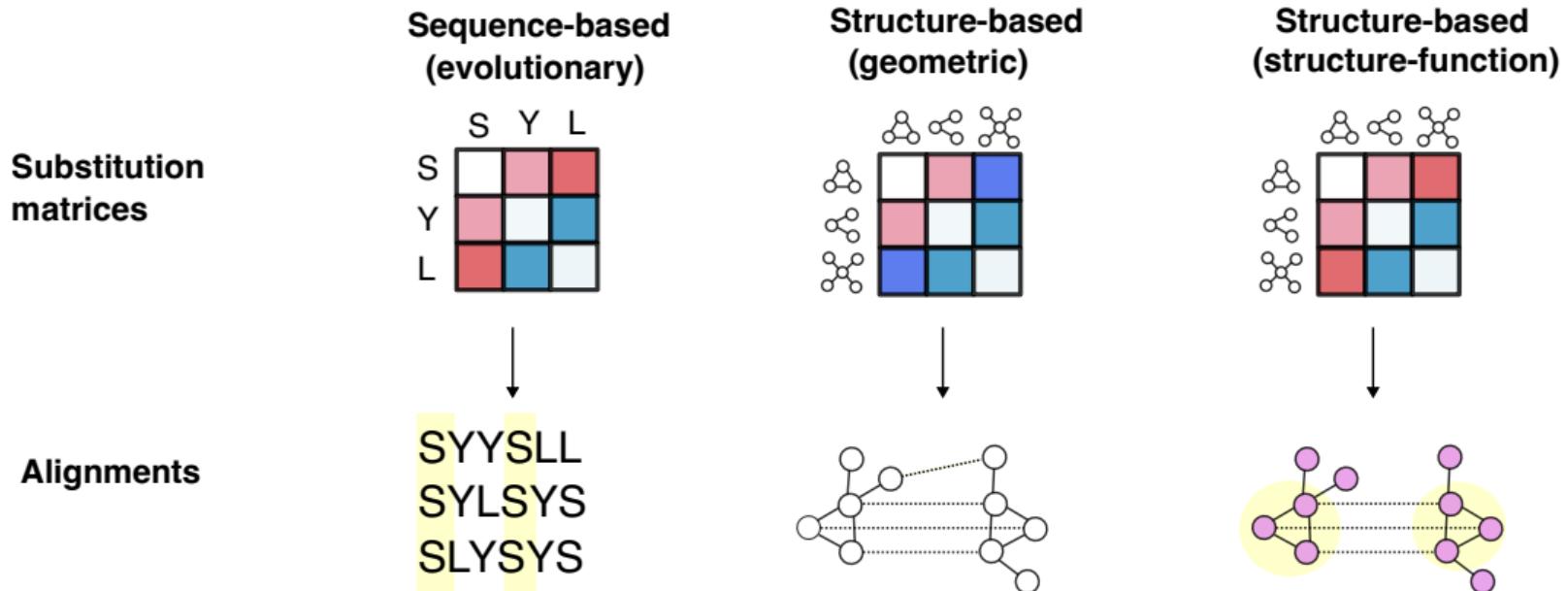
C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F
C	0	-1	4	1	5	0	1	0	4	0	1	0	0	1	0	0	1	0	0
S	-1	0	1	5	0	1	0	4	0	1	0	0	0	1	0	0	1	0	0
T	0	1	5	0	1	0	4	0	1	0	0	0	0	1	0	0	1	0	0
A	0	1	0	4	0	1	0	5	0	1	0	0	0	1	0	0	1	0	0
G	-3	0	-2	0	6	0	2	5	0	1	0	0	0	1	0	0	1	0	0
P	0	3	-1	-1	-2	7	1	0	0	6	0	1	0	0	1	0	0	1	0
D	-3	0	-1	-2	-1	6	2	5	0	1	0	0	0	1	0	0	1	0	0
E	-4	0	-1	-1	-2	1	2	5	0	1	0	0	0	1	0	0	1	0	0
Q	-3	0	-1	-1	-2	1	2	5	0	1	0	0	0	1	0	0	1	0	0
M	-3	1	0	-2	0	2	1	0	0	6	0	1	0	0	1	0	0	1	0
H	-3	-1	-2	-2	-2	-1	0	0	1	8	0	1	0	0	1	0	0	1	0
R	-3	-1	-1	-1	-2	-2	0	1	0	0	5	0	1	0	0	1	0	0	1
K	-3	0	-1	-1	-2	-1	1	1	0	-1	2	5	0	1	0	0	1	0	0
M	-1	-1	-1	-1	-3	-2	-3	0	-2	-2	-1	-1	5	0	1	0	0	1	0
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	0	1	0	0	1
L	-1	-1	-2	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4	0	1	0
V	-1	-2	0	0	-3	-2	-3	-2	-3	-3	-2	1	3	1	4	0	1	0	0
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11	0
Y	-2	-2	-2	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	1	2	7	0	1	0
F	-2	-2	-2	-3	-4	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	0
C	3	6	9	7	10	13													

Alignment

T	G	T	T	T	A	C	G	G											
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	0	0	3	6	0	0	0	0	0	0	0	0
T	0	3	1	6	4	2	0	1	4	0	0	0	0	0	0	0	0	0	0
T	0	3	1	4	9	7	5	3	2	0	0	0	0	0	0	0	0	0	0
G	0	1	6	4	7	6	4	8	6	0	0	0	0	0	0	0	0	0	0
A	0	0	4	3	5	10	8	6	5	0	0	0	0	0	0	0	0	0	0
C	0	0	2	1	3	8	13	11	9	0	0	0	0	0	0	0	0	0	0
T	0	3	1	5	4	6	11	10	8	0	0	0	0	0	0	0	0	0	0
A	0	1	0	3	2	7	9	8	7	0	0	0	0	0	0	0	0	0	0

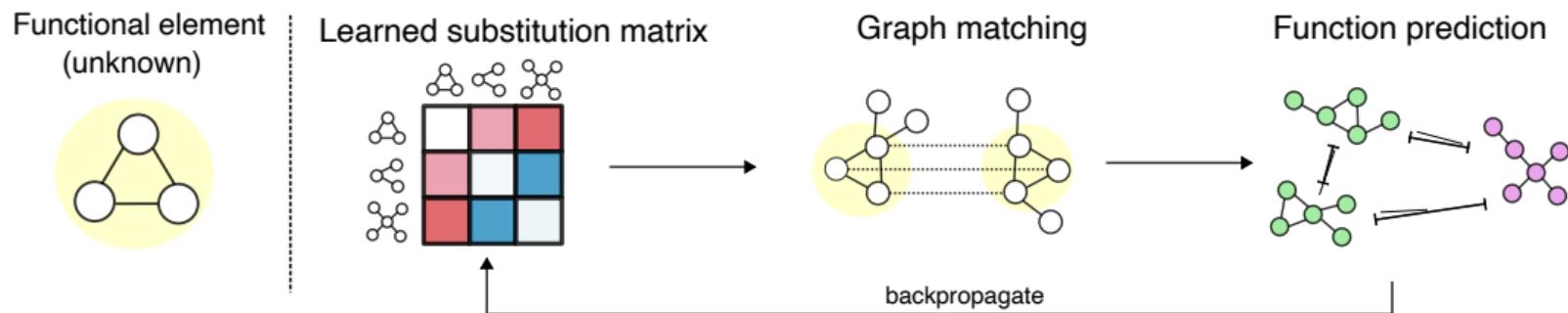
3 6 9 7 10 13  
G T T G A C  
I I I G A C

# Beyond residue-level alphabets [Pellizzoni et al., 2024]

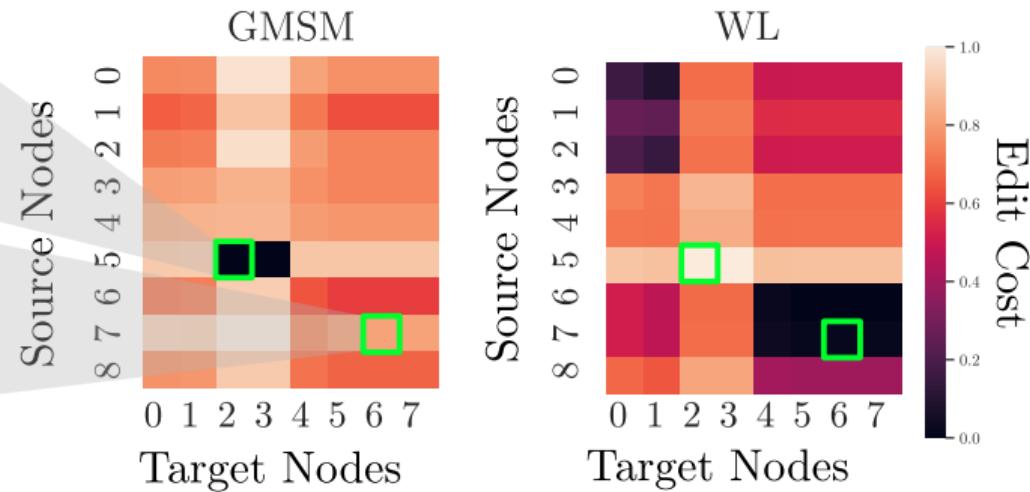
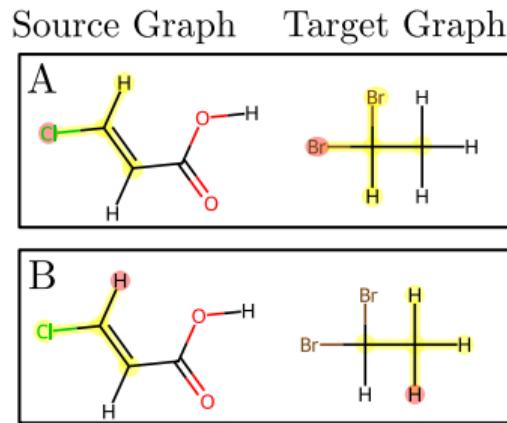


# Learn SM via graph matching

- We decompose the protein into higher order subunits → local neighbourhoods ( $\uparrow$  expressivity).



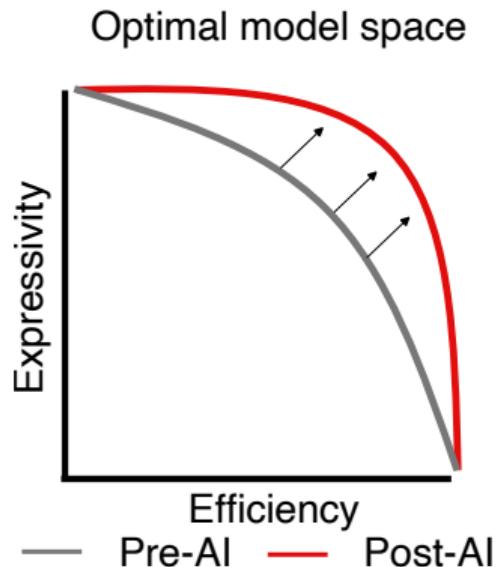
# Learned substitution costs reflect functional substructures



**Figure:** Learned substitution matrices from GMSM vs structure-only WL kernel.

## Perspectives

- Many more algorithms remain to be discovered around the new Pareto front.
- Exploration will unlock insights in more complex modalities (e.g protein ensembles)

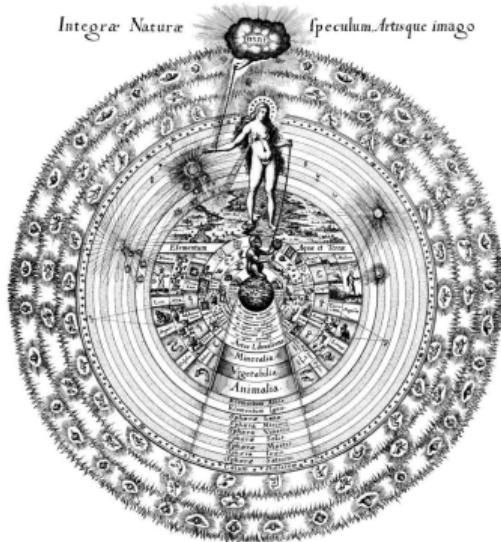


## Acknowledgements

### ETH Zürich & Max Planck Institute

- Paolo Pellizzoni
- Karsten Borgwardt
- Dexiong Chen
- Tim Kucera
- Philip Hartout
- Leslie O'Bray

## Contact



- Email: carlos.oliver@vanderbilt.edu
  - Webpage: carlosoliver.co
  - Lab: oliverlaboratory.com
  - X: @carlosgoliver

## Bibliography i

-  Harris, J., Shadrina, M., Oliver, C., Vogel, J., and Mittermaier, A. (2018).  
**Concerted millisecond timescale dynamics in the intrinsically disordered carboxyl terminus of  $\gamma$ -tubulin induced by mutation of a conserved tyrosine residue.**  
*Protein Science*, 27(2):531–545.
-  Kucera, T., Oliver, C., Chen, D., and Borgwardt, K. (2023).  
**Proteinshake: Building datasets and benchmarks for deep learning on protein structures.**  
In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

## Bibliography ii

-  Llinares-López, F., Berthet, Q., Blondel, M., Teboul, O., and Vert, J.-P. (2023).  
**Deep embedding and alignment of protein sequences.**  
*Nature methods*, 20(1):104–111.
  
-  Pellizzoni, P., Oliver, C., and Borgwardt, K. (2024).  
**Graph-matching-based substitution matrices.**  
In *Research in Computational Molecular Biology*.