Center for AI in Protein Dynamics

VANDERBILT UNIVERSITY

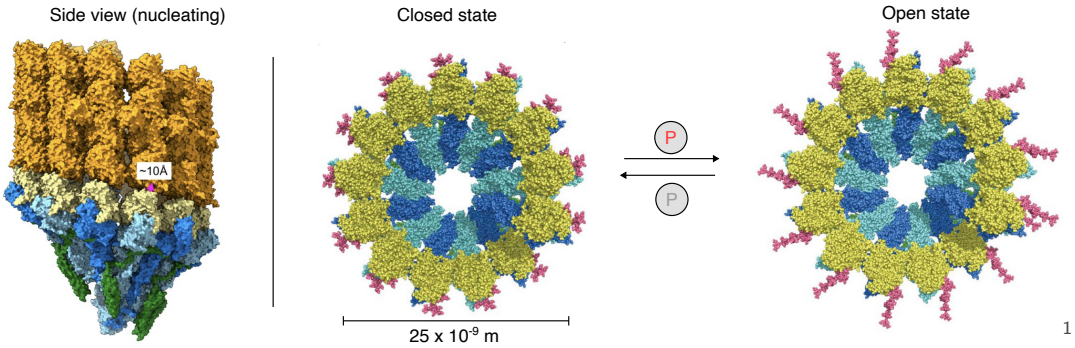
The Post-AlphaFold world: a new algorithmic landscape for structure-function modeling

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

AI Days | March 6th, 2025

Slides: `carlosoliver.co/2025/03/05/aidays.html`

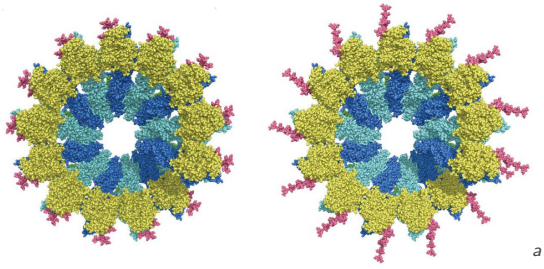
Structure is the language of biology



[1]

[1] [Harris et al., 2018]

Driving Questions



[a] [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

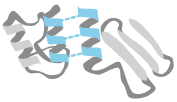
e.g. Binding Pocket Prediction

Self-Supervision

e.g. Pretraining with AlphaFold

Pairwise Residue

e.g. Binding Interface Prediction

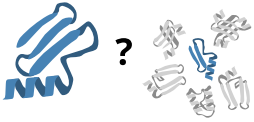


Pairwise Protein

e.g. Structure Alignment or Protein-Protein Interaction

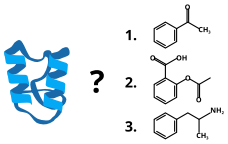
Retrieval

e.g. Similar Structure Search



Ranking

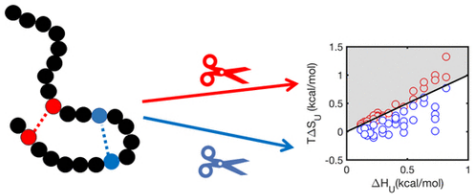
e.g. Drug Screening



2

[2] [Kucera et al., 2023]

An approach: direct perturbation



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

a

[a] Bigman, Lavi S., and Yaakov Levy., 2018

Pros

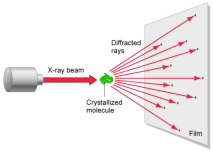
- Directly explainable

Cons

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

The data-driven approach

Structure Data



AlphaFold

Function Data

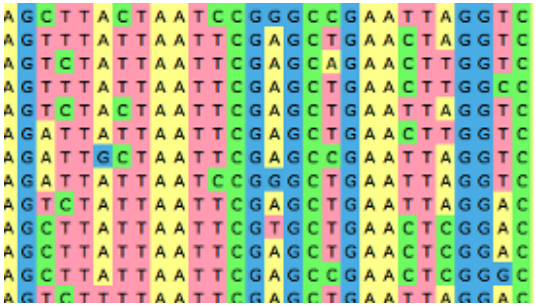
Expasy

PDBbind+

Pfam

Another approach: bioinformatics

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

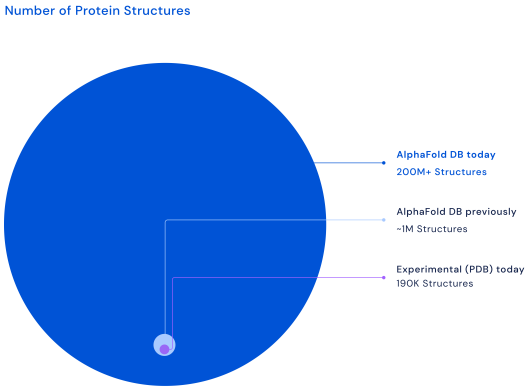


[3]

Main focus of bioinformatics development for the past decades.

[3] http://ugene.net/multiple-sequence-alignment-with-muscle/

And then came AlphaFold...



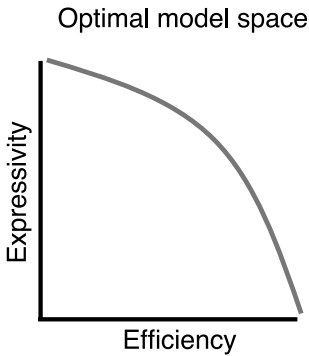
[4]

Massive increase in complexity and scale!

[4] Source: DeepMind blog

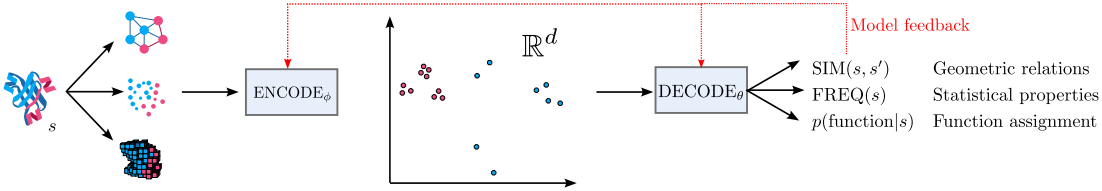
The expressivity vs efficiency tradeoff

- 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?

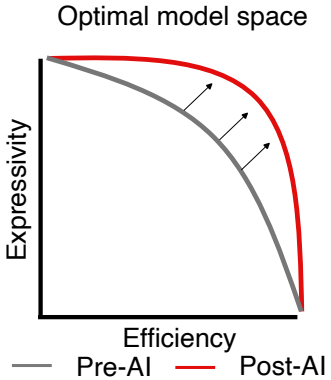
1. Neural network with parameters ϕ encodes proteins from the structure domain to intermediate vectorial space.
2. Neural network with parameters θ decodes proteins to the function domain.



Key: all steps are matrix multiplication-baesd (↑ efficiency), and neural networks can capture complex patterns (↑ expressivity).

The case for AI in bioinformatics

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

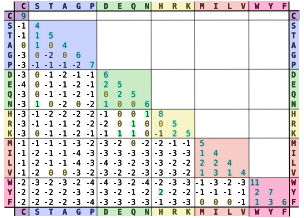


Imperative: discover the next generation of **high-capacity and scalable** bioinformatics tools.

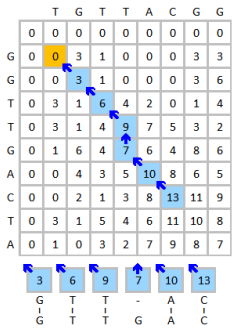
Case study: Protein Alignment

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

Substitution Matrix
(parameters)



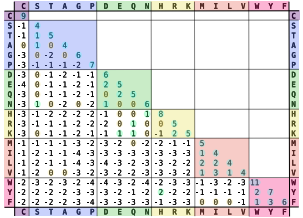
Alignment



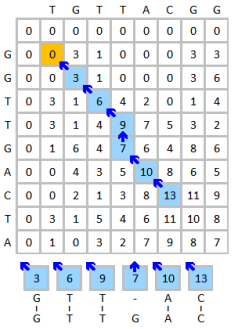
Case study: Protein Alignment

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

Substitution Matrix
(parameters)

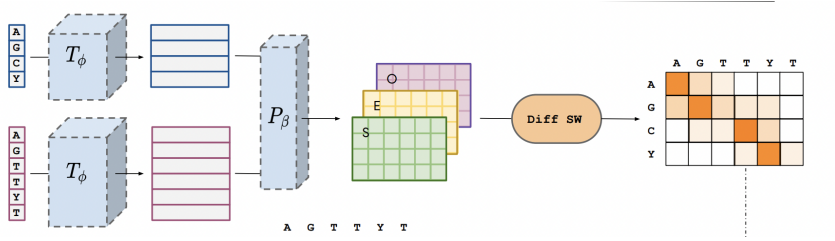


Alignment



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

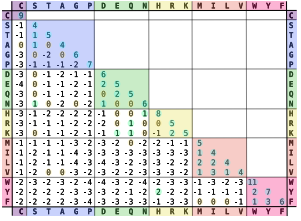
- Allow flexible substitution costs (↑ Expressivity)
- Substitution costs become a model parameter.
- Parameter is tuned by backpropagation using function data.



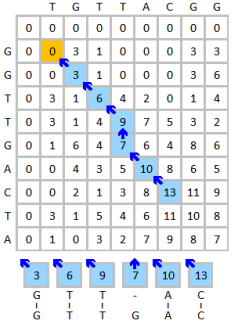
Case study: Protein Alignment

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

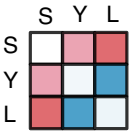
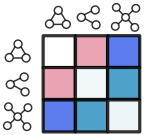
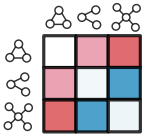
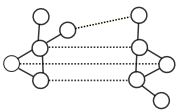
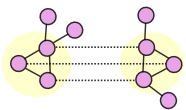
Substitution Matrix
(parameters)



Alignment

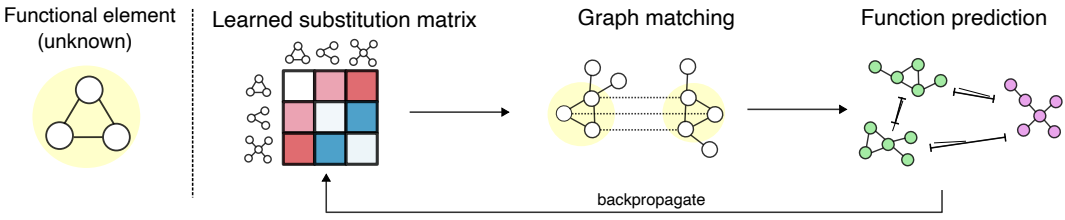


Beyond residue-level alphabets [Pellizzoni et al., 2024]

	Sequence-based (evolutionary)	Structure-based (geometric)	Structure-based (structure-function)
Substitution matrices			
Alignments	SYYSLL SYLSYS SLYSYS		

Learn SM via graph matching

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



Learned substitution costs reflect functional substructures

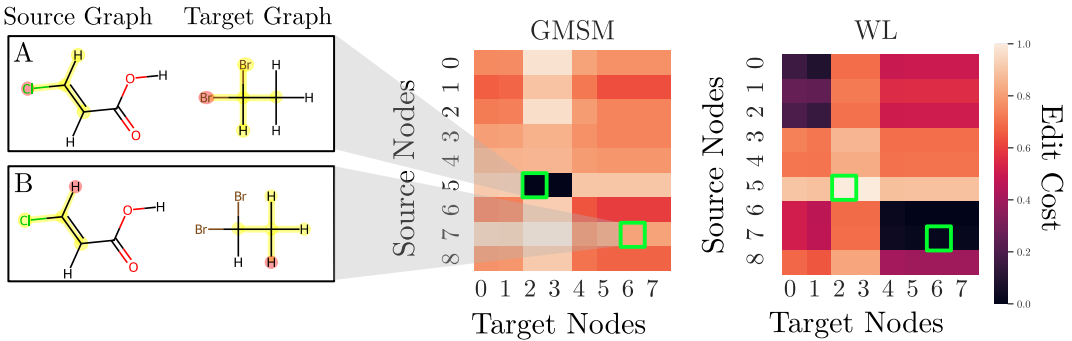
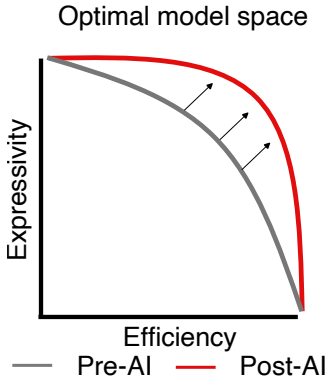


Figure: Learned substitution matrices from GMSM vs structure-only WL kernel.

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)

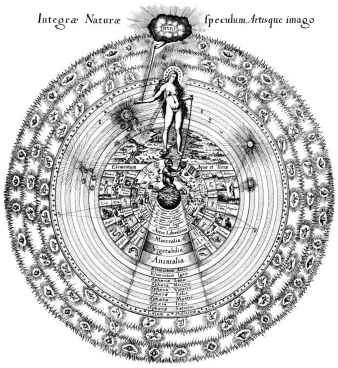


Acknowledgements

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 γ-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*.