

Combining AI and classical techniques to see proteins in motion

Václav Hanžl, UCT Prague, Czech Republic

Guided by **prof. Vojtěch Spiwok**, in collaboration with **wet labs**

With generous support of **prof. Daniel Svozil's** department

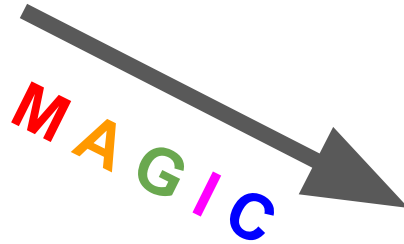


Central & Eastern European Cheminformatics Conference

For static structures, AlphaFold 2 was the breakthrough

GGMAGQNPNDQFEAYFKRADLDGDGRISGAEAVGFFQGSGLSKQVLAQIWSLSDRSHSGFLDRQNFYNSLRLVTVAQSKRDLTP EIVNAALNTPAAAKIPPPKINLSA

AA sequence



folded protein



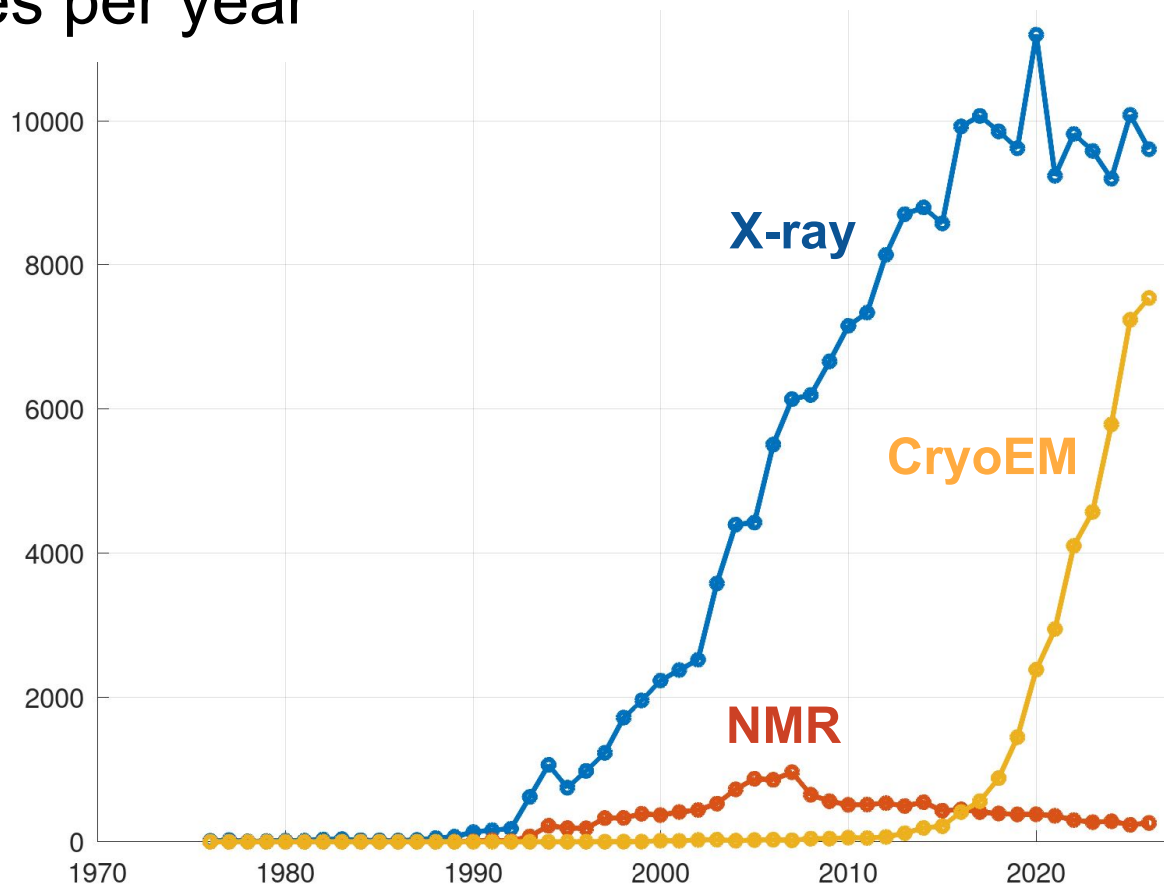
*Arabidopsis
thaliana*

A 3D molecular simulation of a protein (green ribbon) embedded in a lipid bilayer (brown mesh). The protein is surrounded by numerous sodium ions (Na+, purple spheres) and chloride ions (Cl-, green spheres). A single red sphere is also visible near the protein. The simulation box is bounded by a black plane at the bottom and a black plane at the top. The protein is a single chain with several alpha-helices and beta-strands. The lipid bilayer is composed of phospholipids with brown heads and yellow tails. The ions are distributed throughout the simulation box, with a higher concentration near the protein and the lipid headgroups.

New PDB structures per year

Data source:
www.rcsb.org/stats

(year 2026 estimated)



Journalism Quarterly

DEVOTED TO RESEARCH STUDIES IN THE FIELD OF MASS COMMUNICATIONS

FALL 1953

1. Deleting [redacted] equal number of words from each passage by some essentially random counting-out system. Such a system was [redacted] on a table of random numbers or else it simply counted out [redacted] *n*th word (every fifth one, for example) without [redacted] regard for the functions or meanings of specific [redacted].

Source:
Taylor, W.: "Cloze Procedure":
A New Tool For Measuring
Readability.
Journalism Quarterly, 1953

Journalism Quarterly

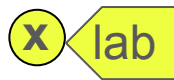
DEVOTED TO RESEARCH STUDIES IN THE FIELD OF MASS COMMUNICATIONS

FALL 1953

1. Deleting an equal number of words from each passage by some essentially random counting-out system. Such a system was based on a table of random numbers or else it simply counted out every n th word (every fifth one, for example) without any regard for the functions or meanings of specific words.

Guess the missing words - BERT uses this task for pre-training, **ESM** does the same with amino acids in proteins, HUBERT in speech, **before we even decide what we really want to do**

Labeled and unlabeled data points at inference time



Having just 3 data points (x, y) with labels (text),
which is easier:

- guess label for another 3 data points
- guess label for another 30 data points



Three labeled data points

GPS,
street name



Three unlabeled data points

just GPS



Three unlabeled data points

just GPS



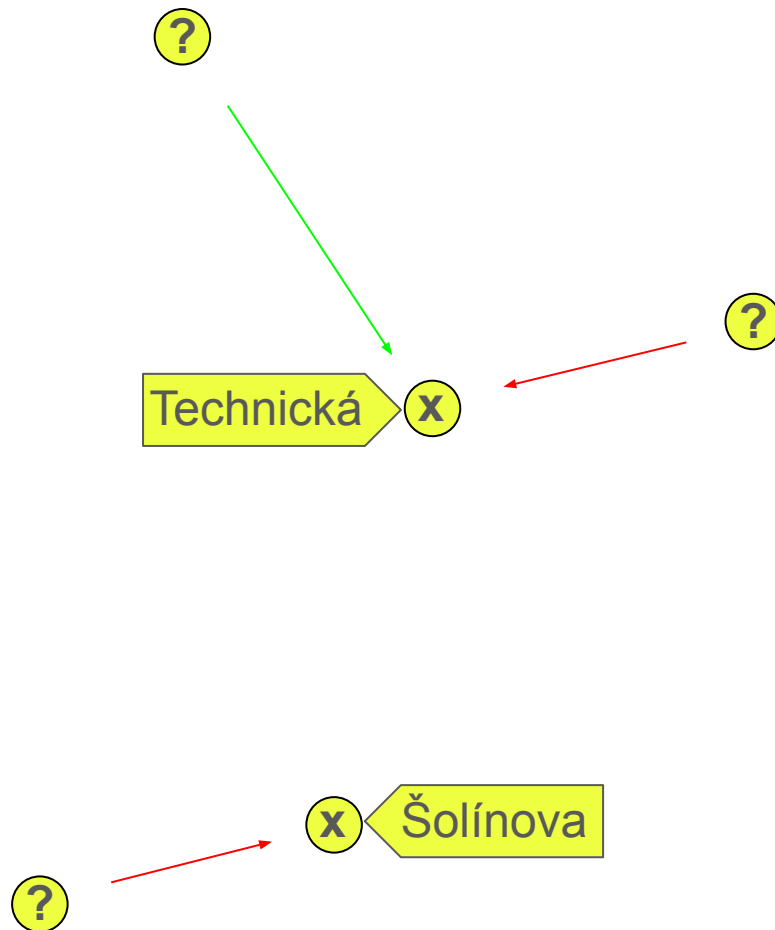
Three labeled data points

GPS,
street name

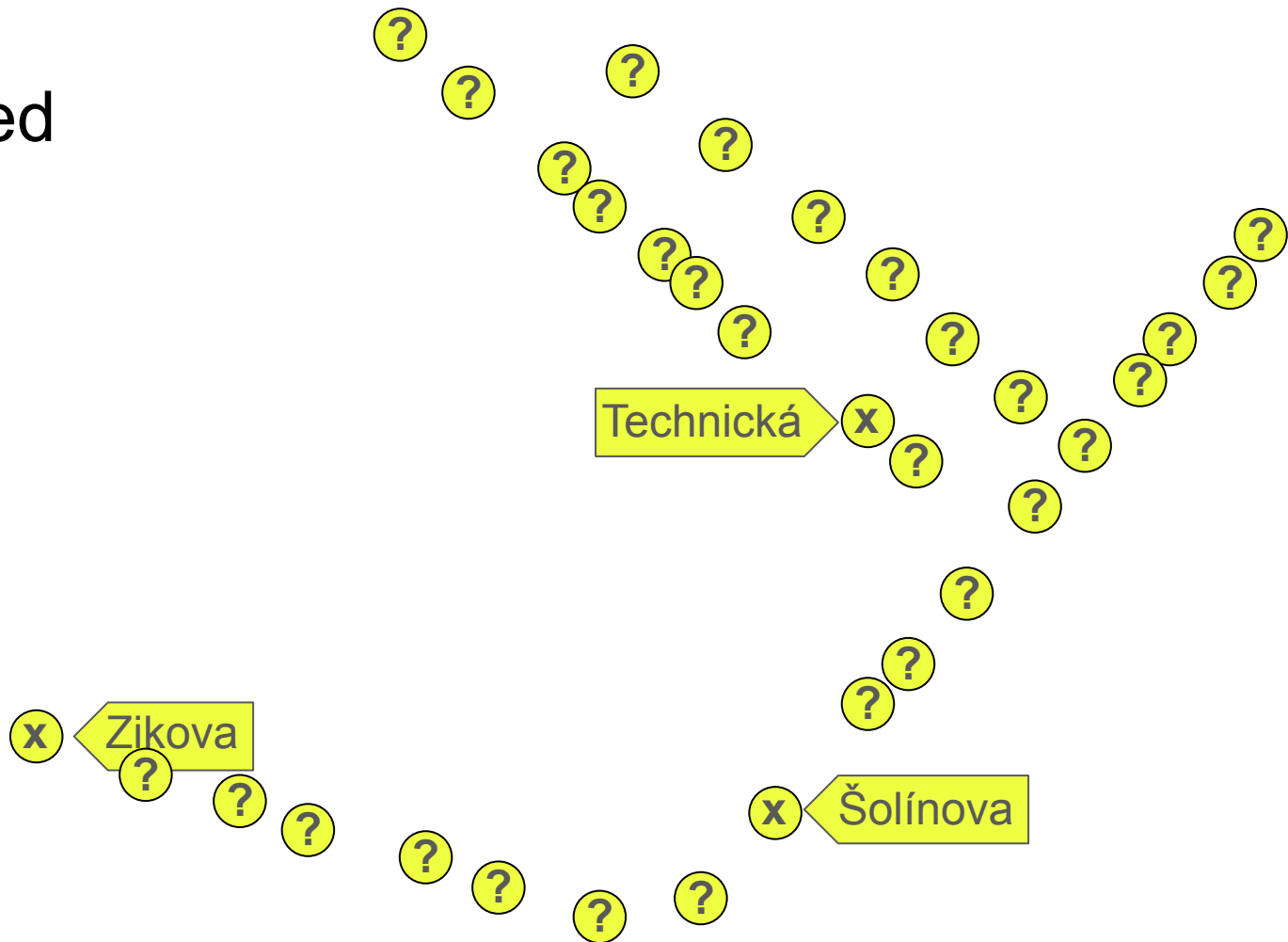


Three unlabeled data points

classified as nearest
labeled one
(mostly wrong in fact)



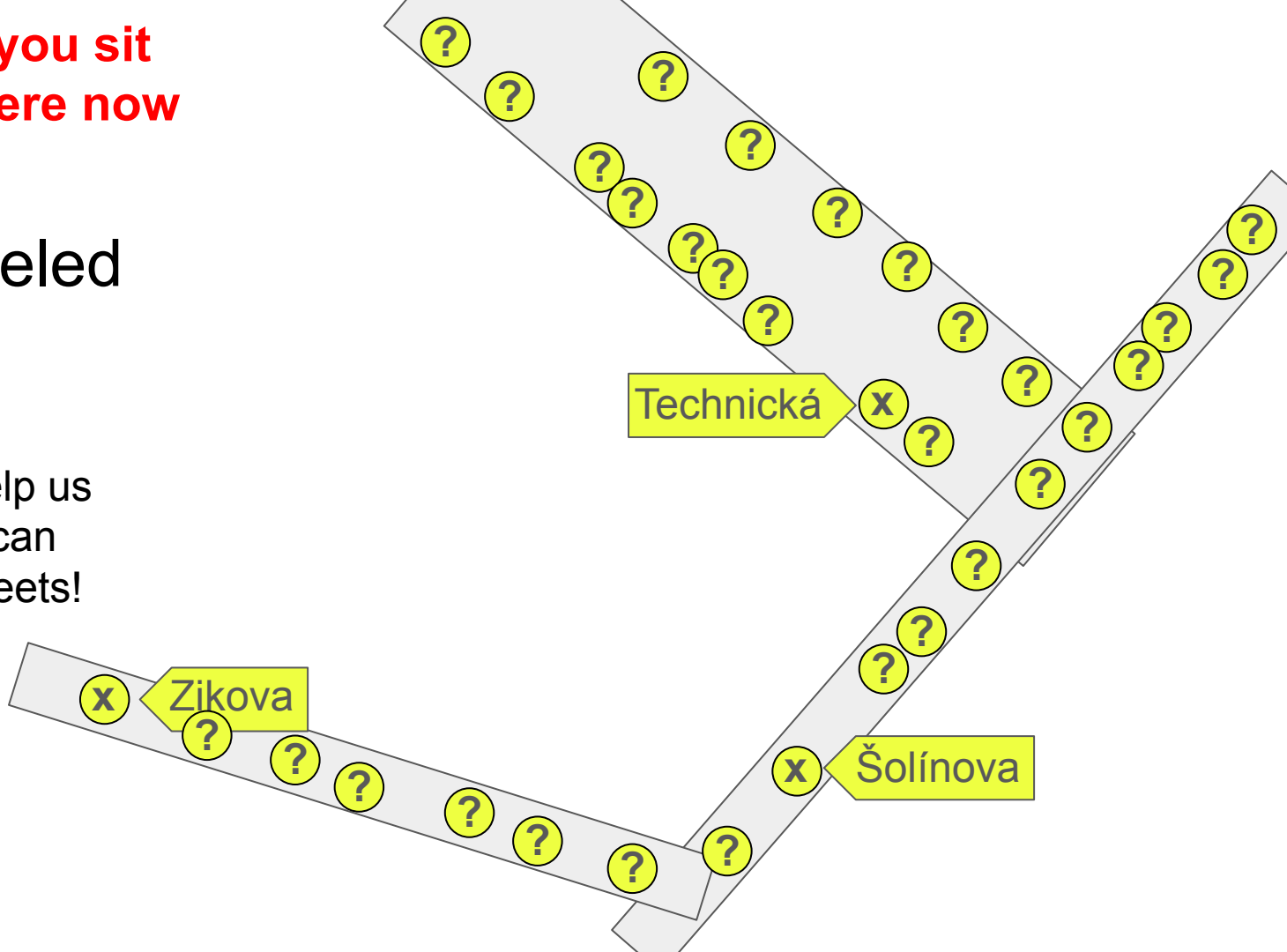
Thirty unlabeled
data points



you sit
here now

Thirty unlabeled data points

More questions help us
with answers, we can
infer shapes of streets!



you sit
here now

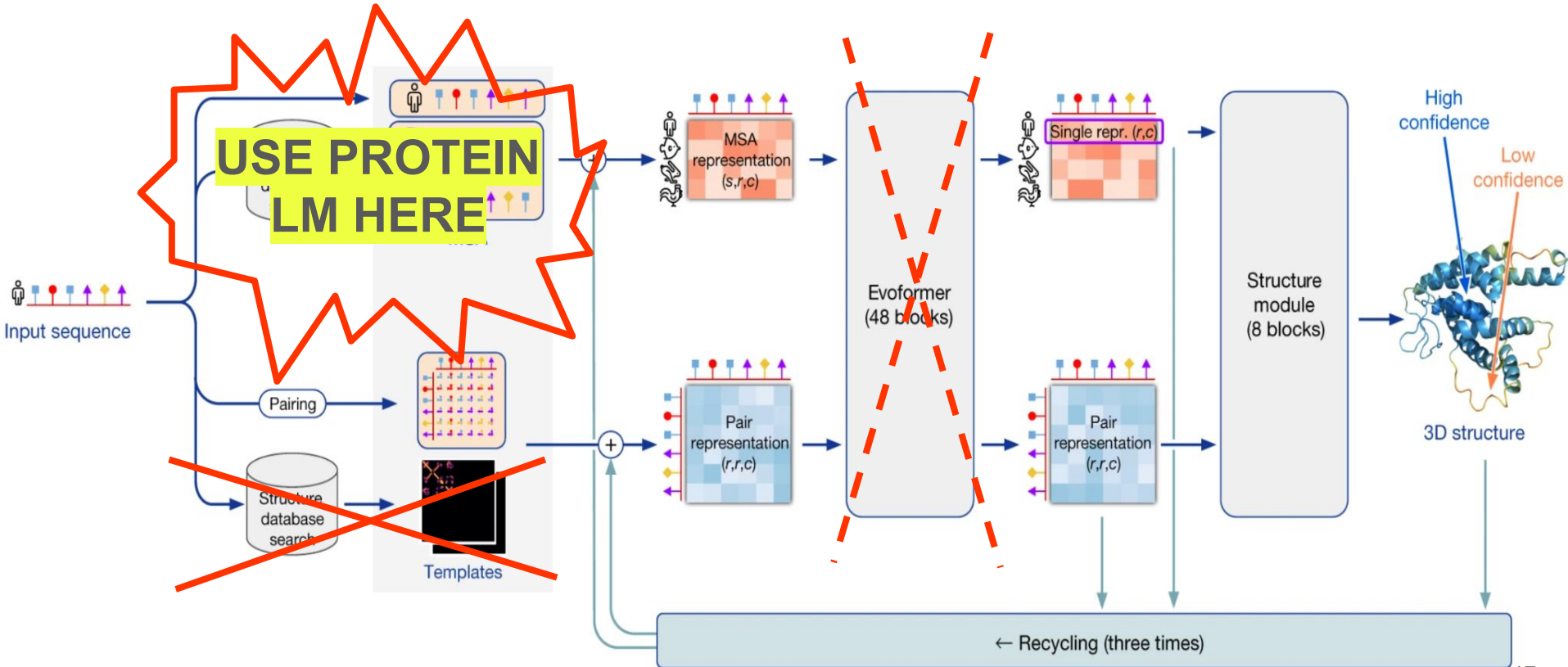
Full
disclosure
of data



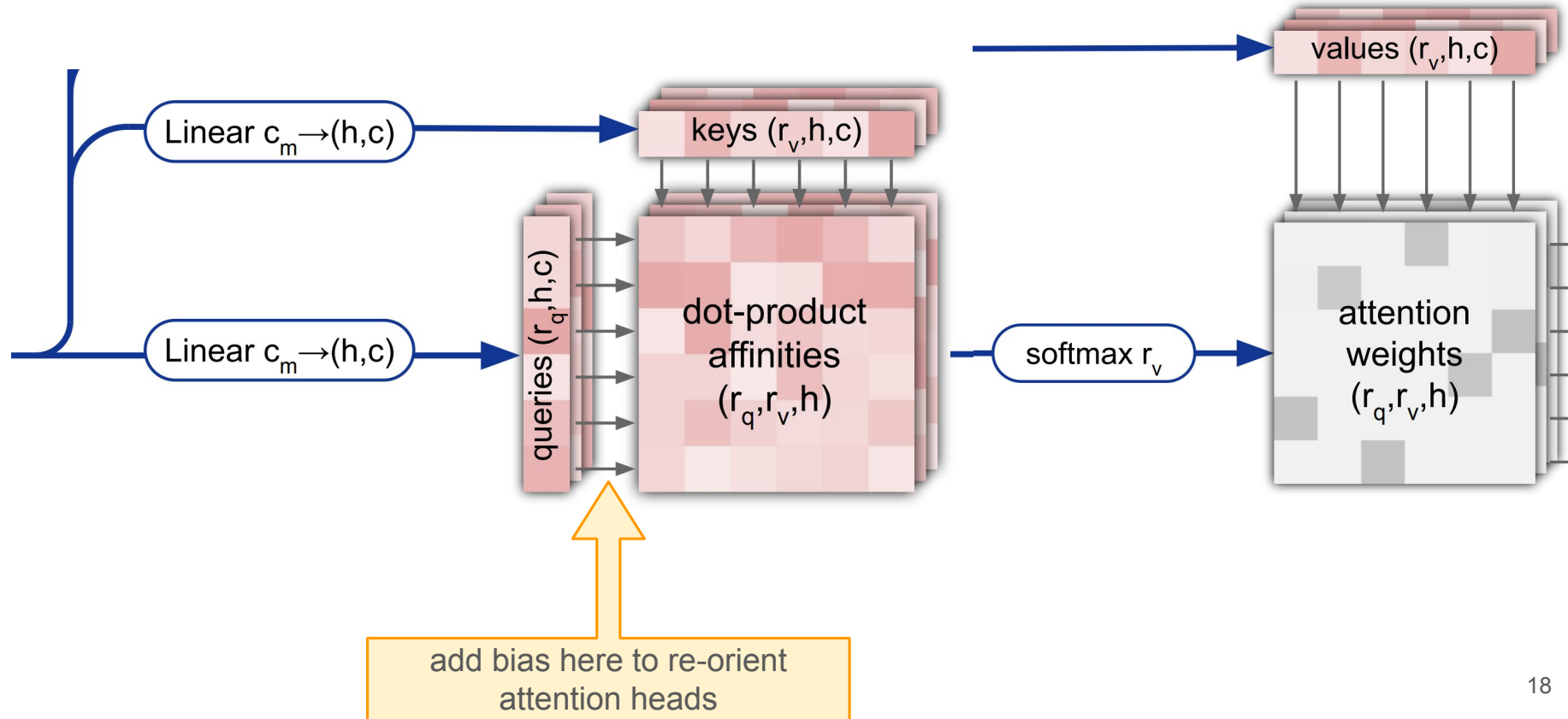
Even **unlabeled** data points tell us something, and given the price and scarcity of labels, unlabeled ones **can be even more important**



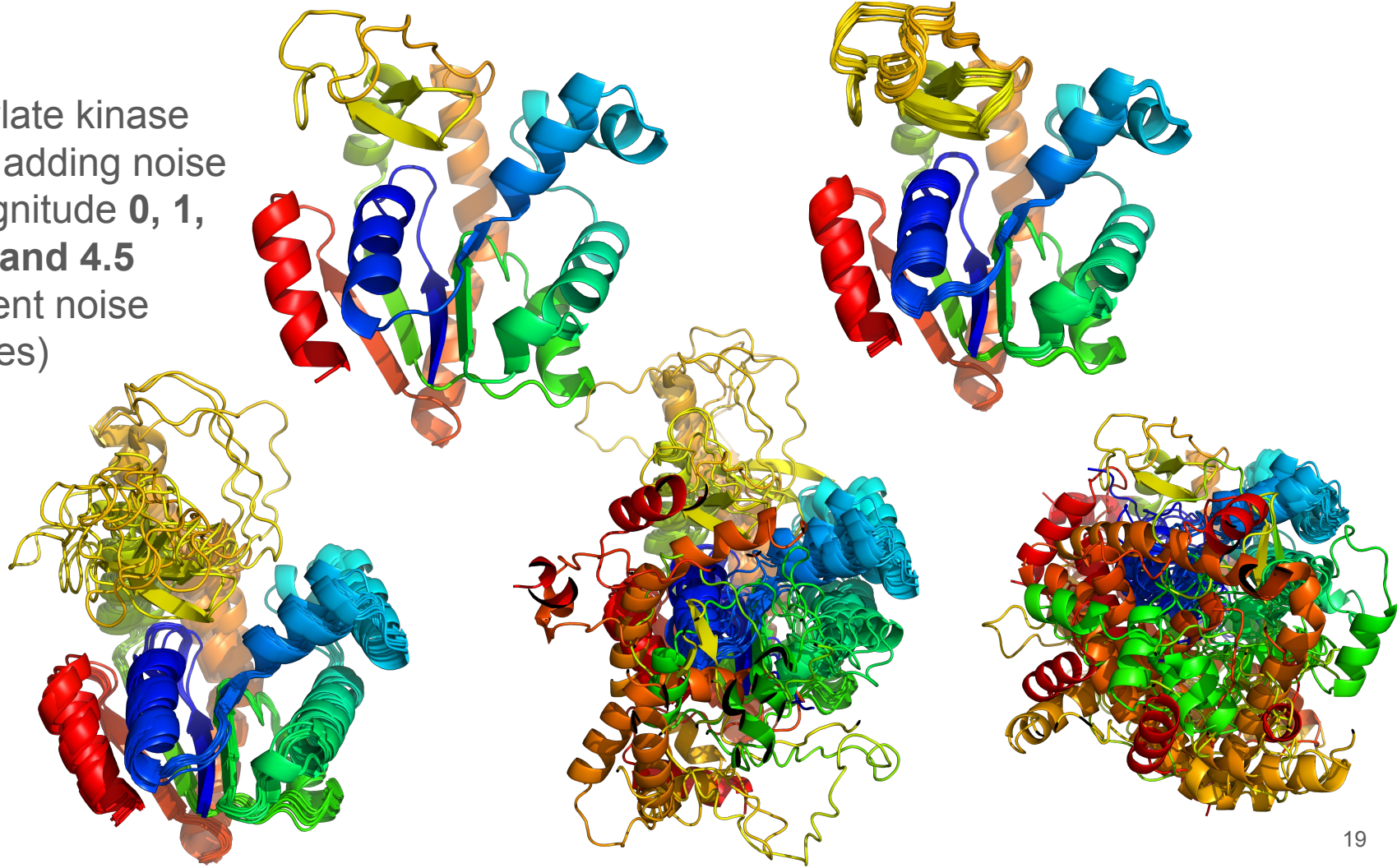
ESMFold replaced MSA in OpenFold with ESM PLM and simplified Evoformer and left out Templates



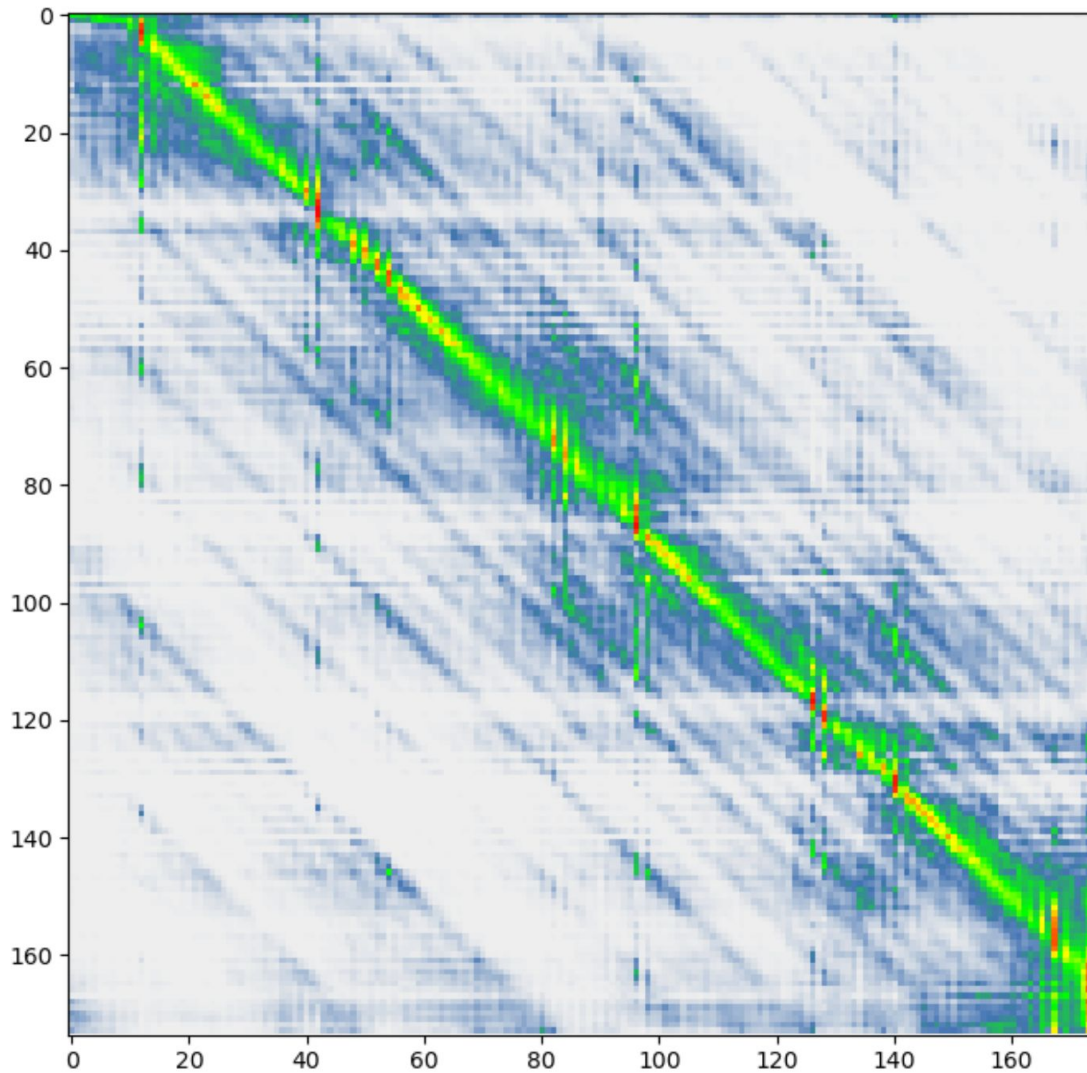
Adding noise to ESM LM leads to structure diversity



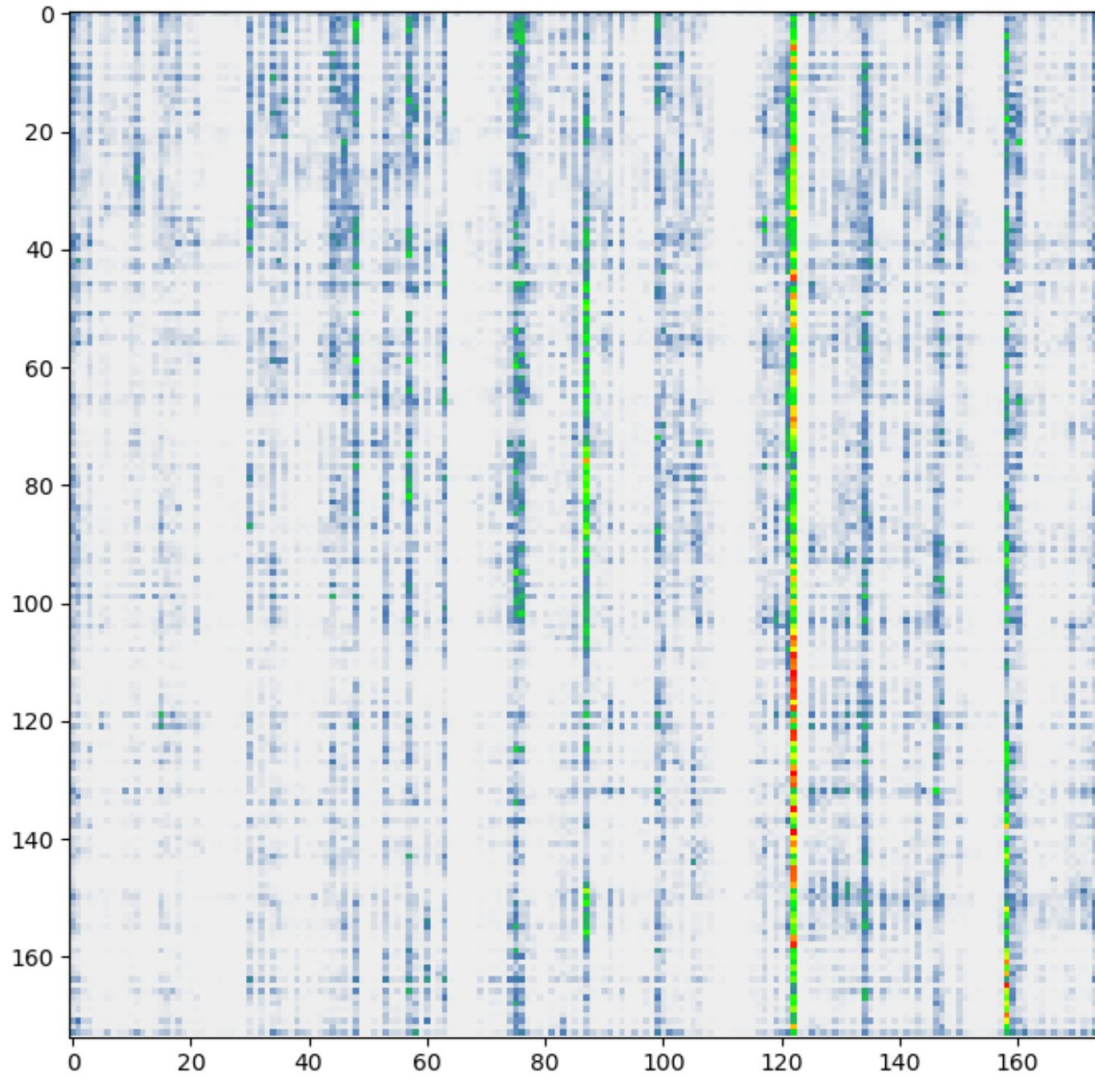
Adenylate kinase
4JZK, adding noise
of magnitude **0**, **1**,
2, **2.5** and **4.5**
(different noise
samples)



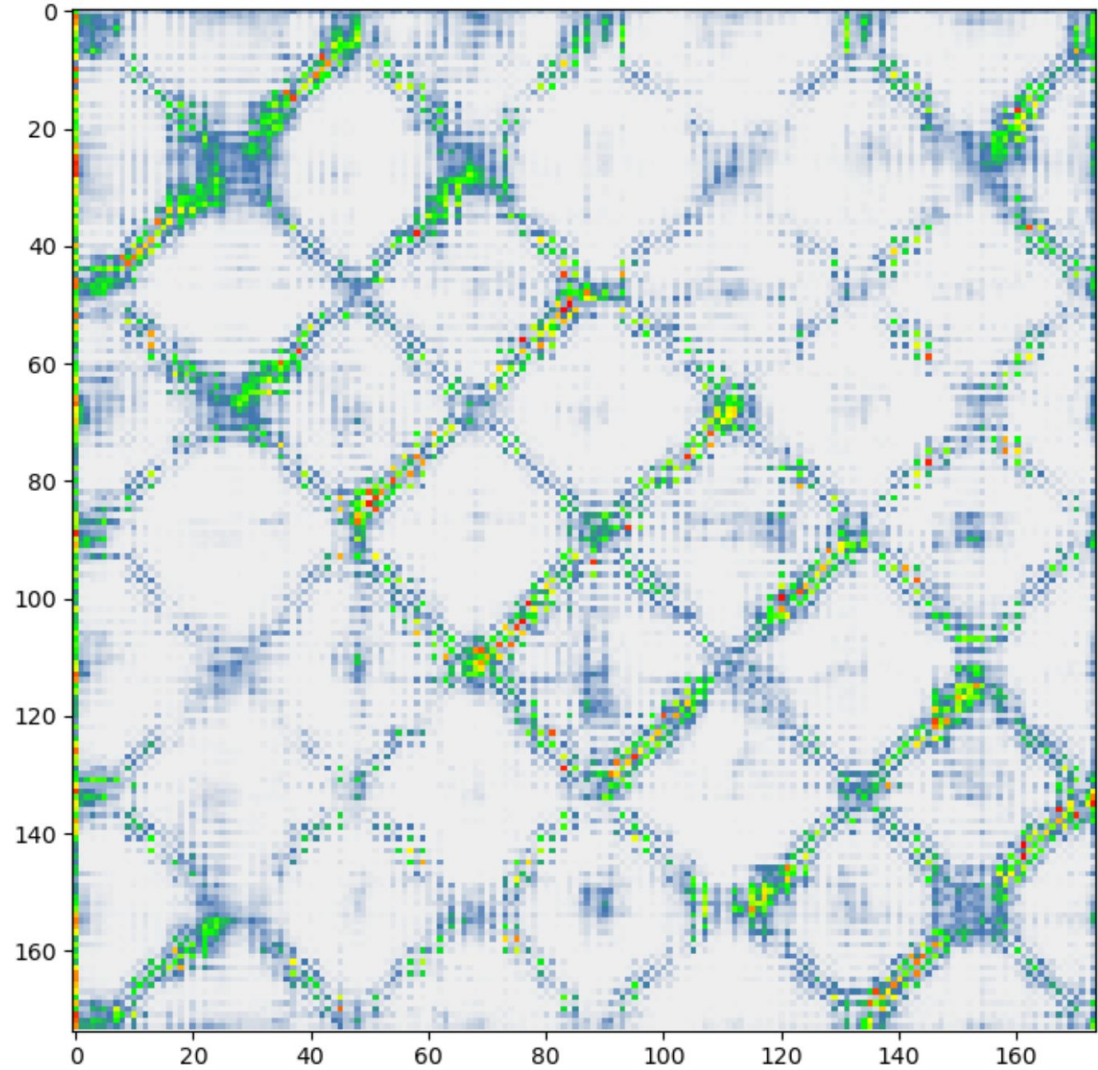
Attention to
unfocused
relative
positions



Attention to
fixed positions



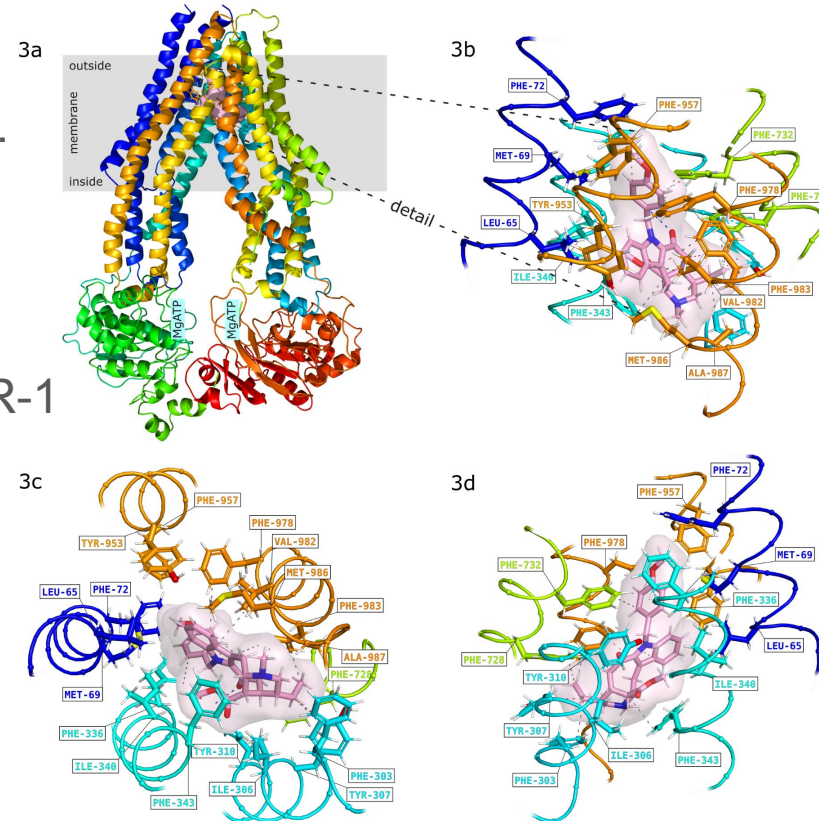
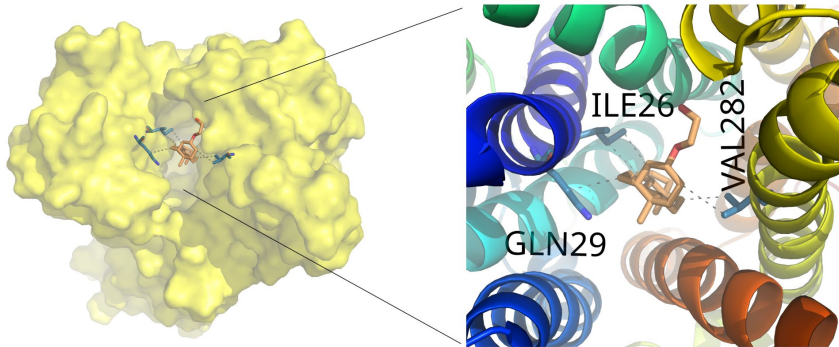
Attention to beta sheet contacts



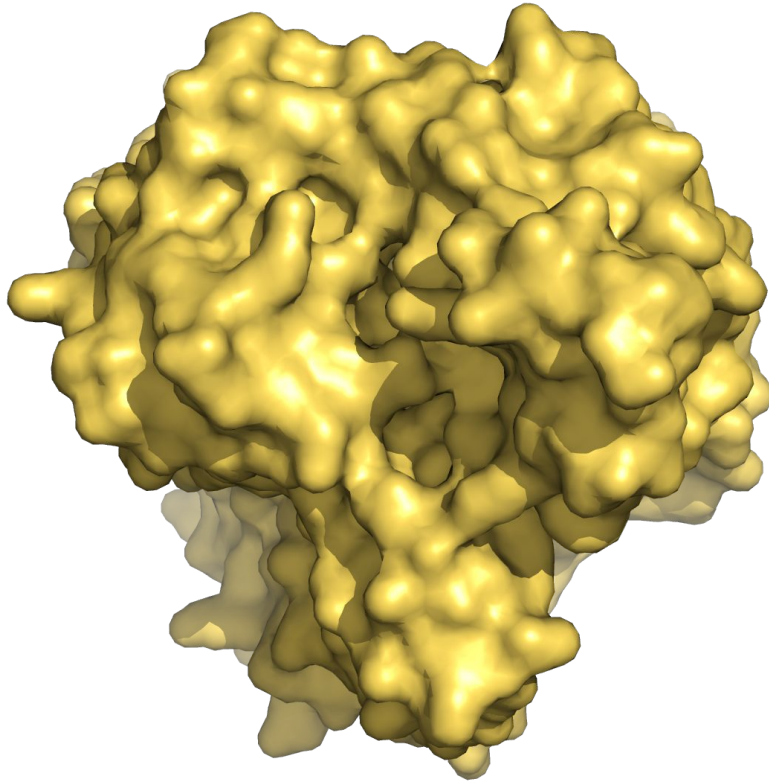
Make it useful

Trying in on whatever moves in a test tube here...

- LmrS and NorA pumps in *S. aureus*
- P-glycoprotein pump in cancer cells
- *S. a.* phosphoethanolamine transferase MCR-1

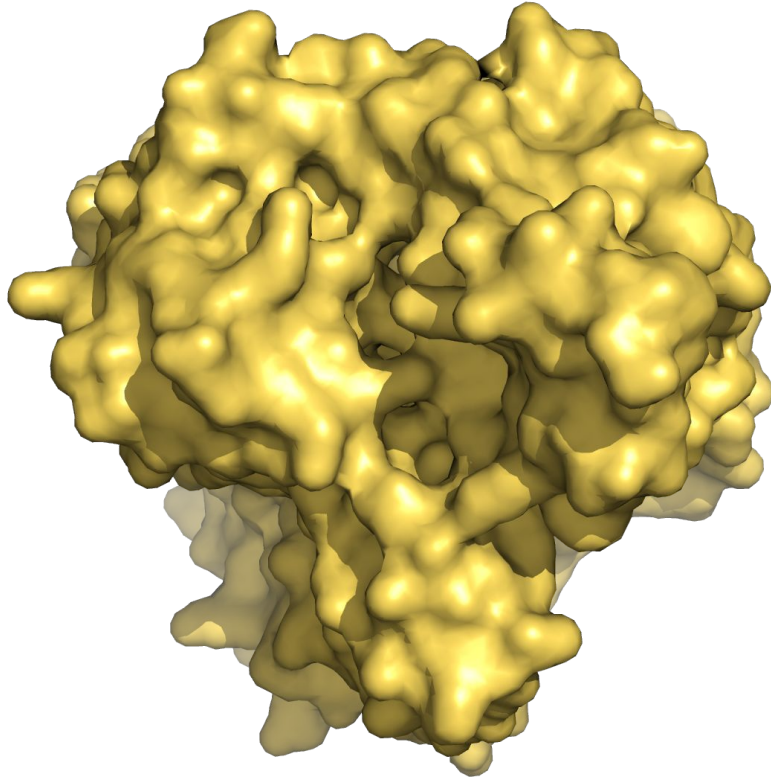


LmrS efflux pump, *Staphylococcus aureus* (top view)



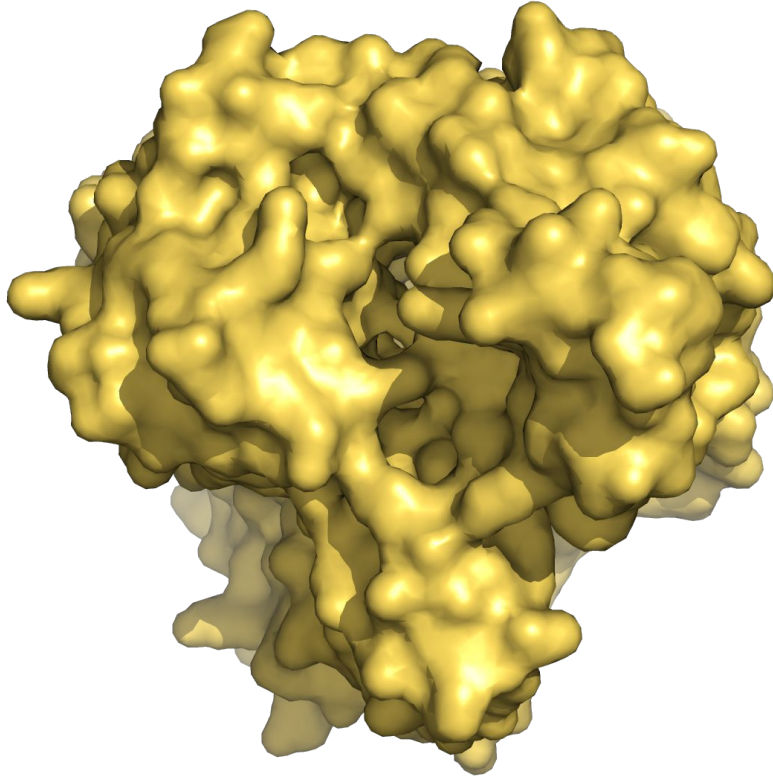
“conscious” static
structure prediction,
slightly outward open

LmrS efflux pump, *Staphylococcus aureus*



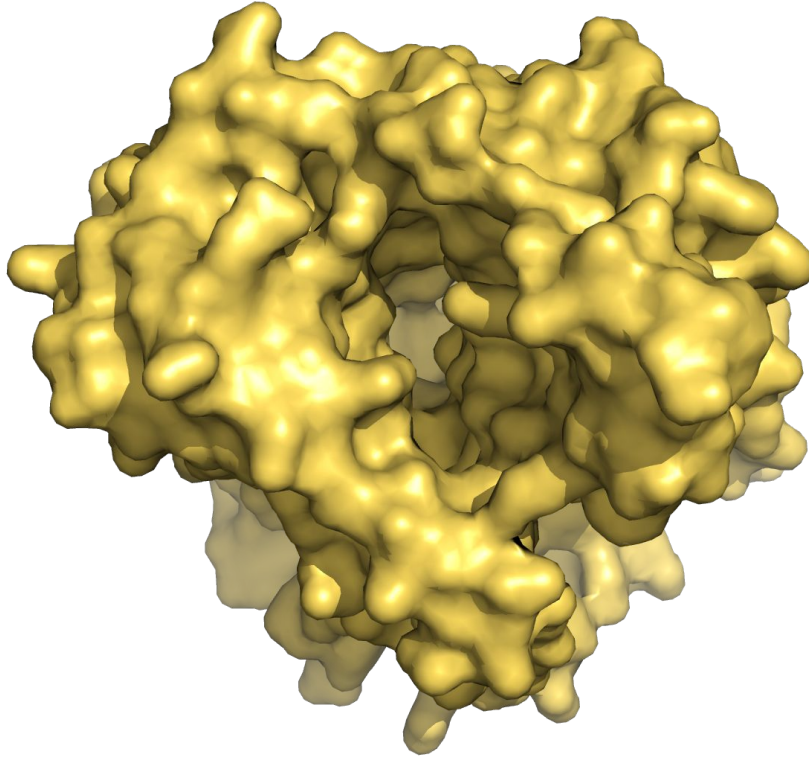
“subconscious”
structure prediction,
more outward open

LmrS efflux pump, *Staphylococcus aureus*



“subconscious”
structure prediction,
yet more outward open

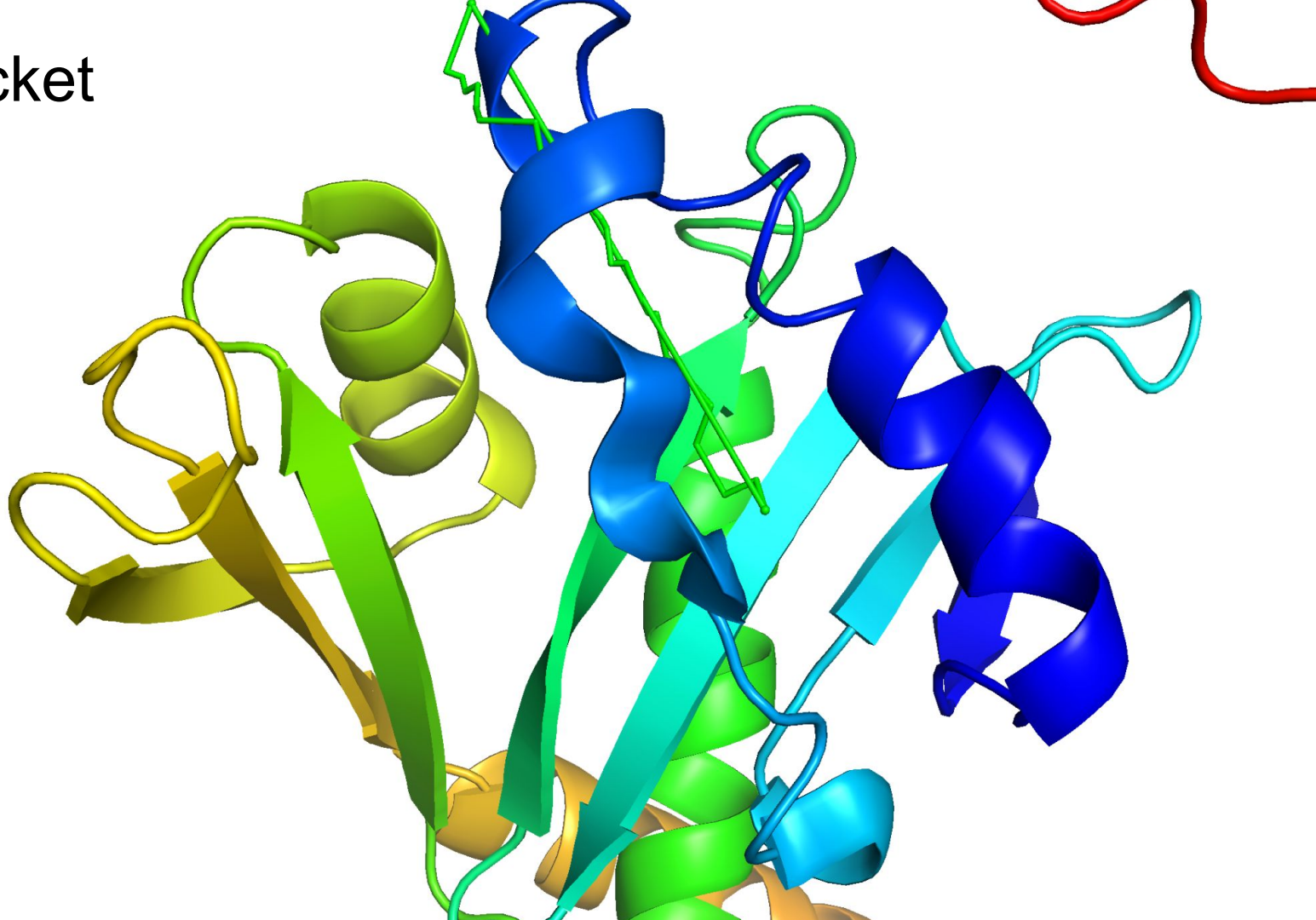
LmrS efflux pump, *Staphylococcus aureus*



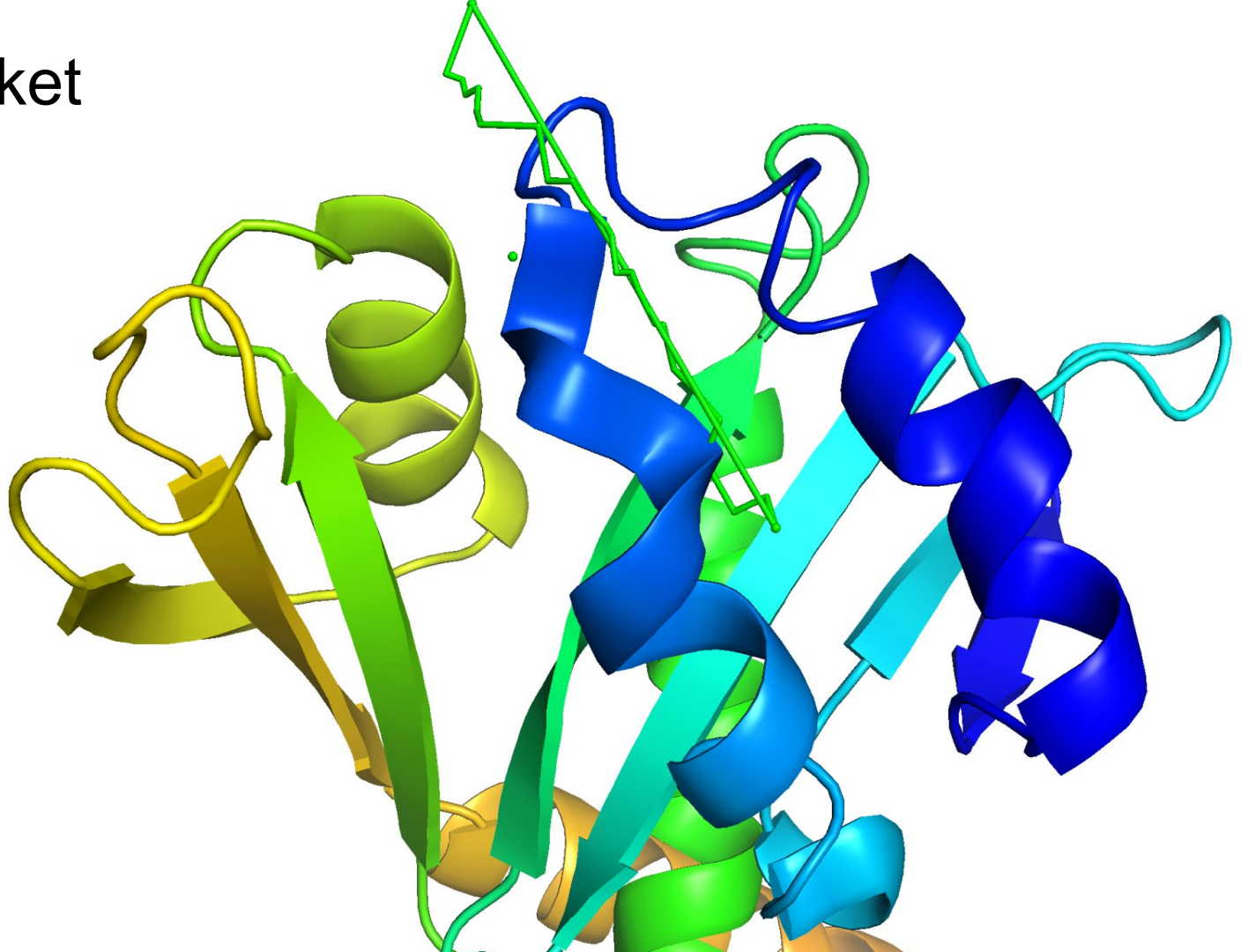
“subconscious”
structure prediction,
wide outward open

(end of the video)

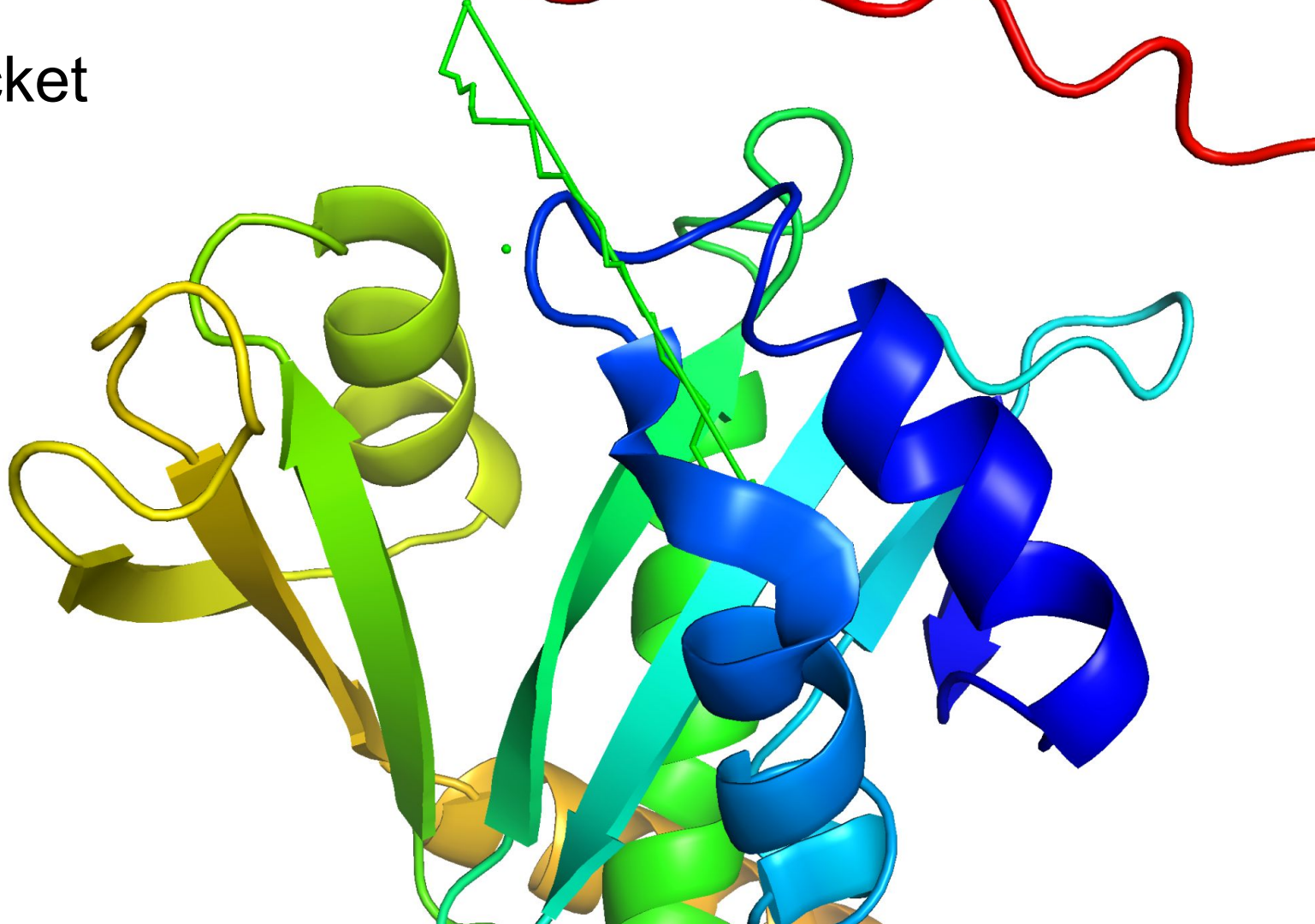
NatA pocket
opening



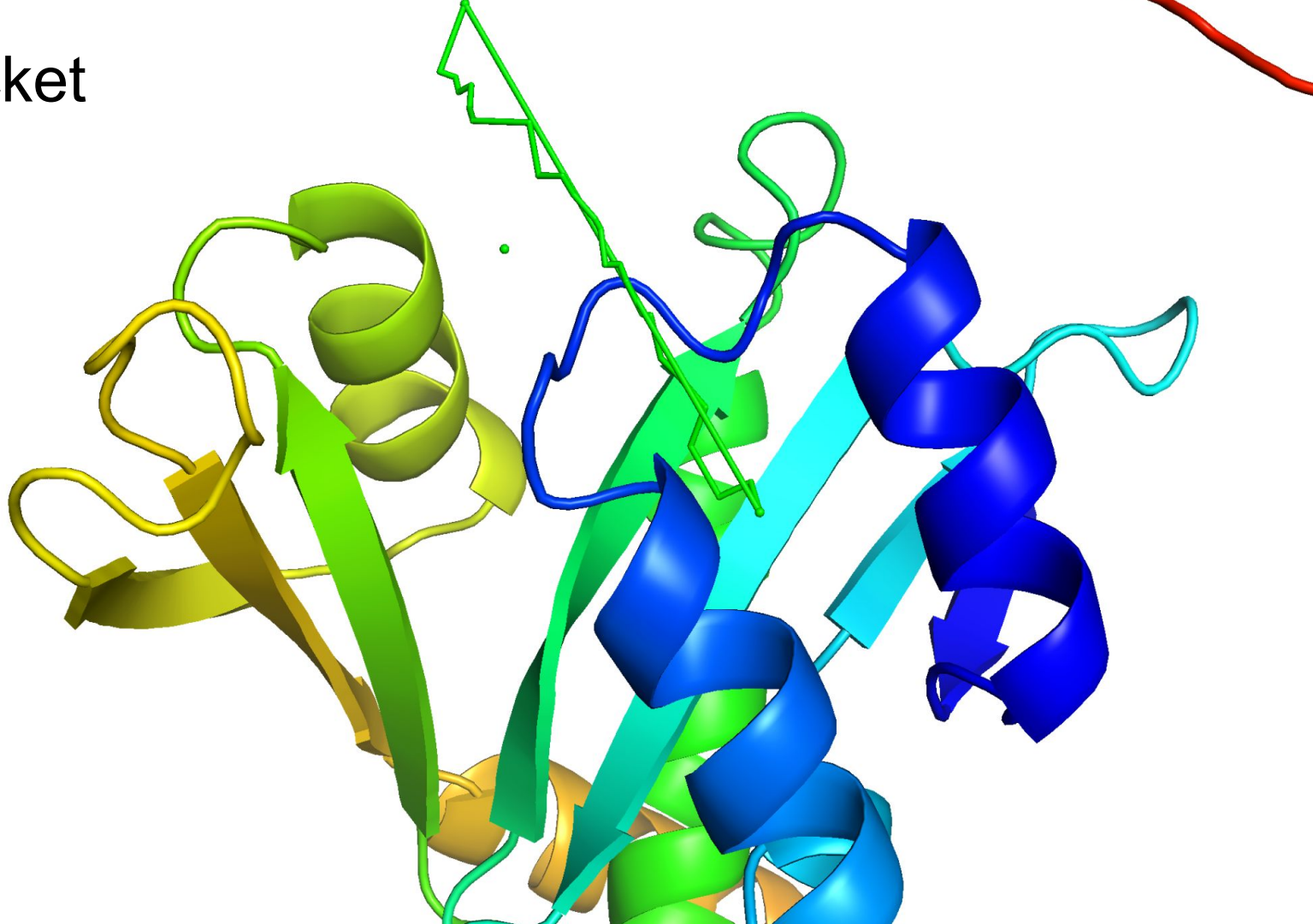
NatA pocket
opening



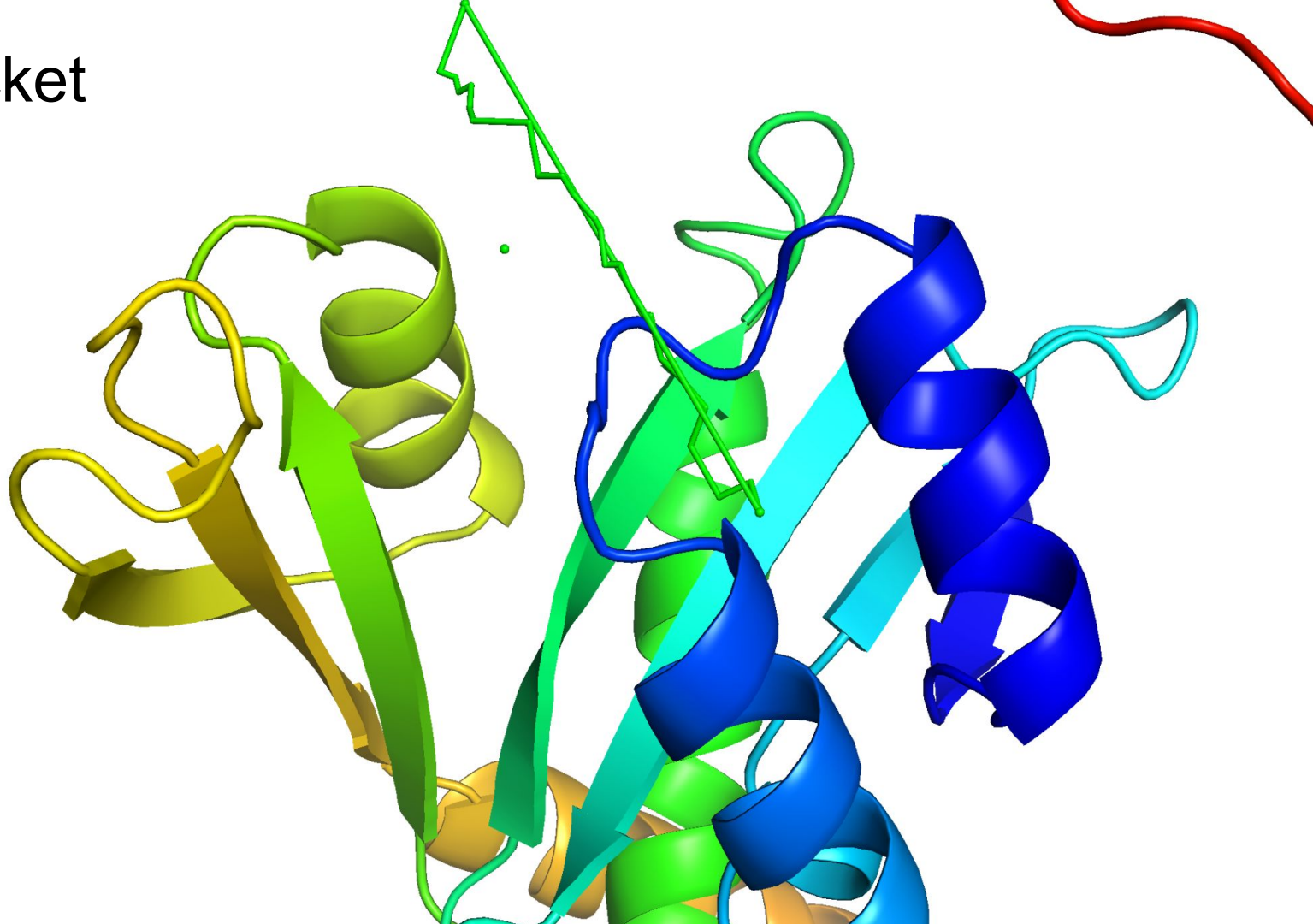
NatA pocket
opening



NatA pocket
opening



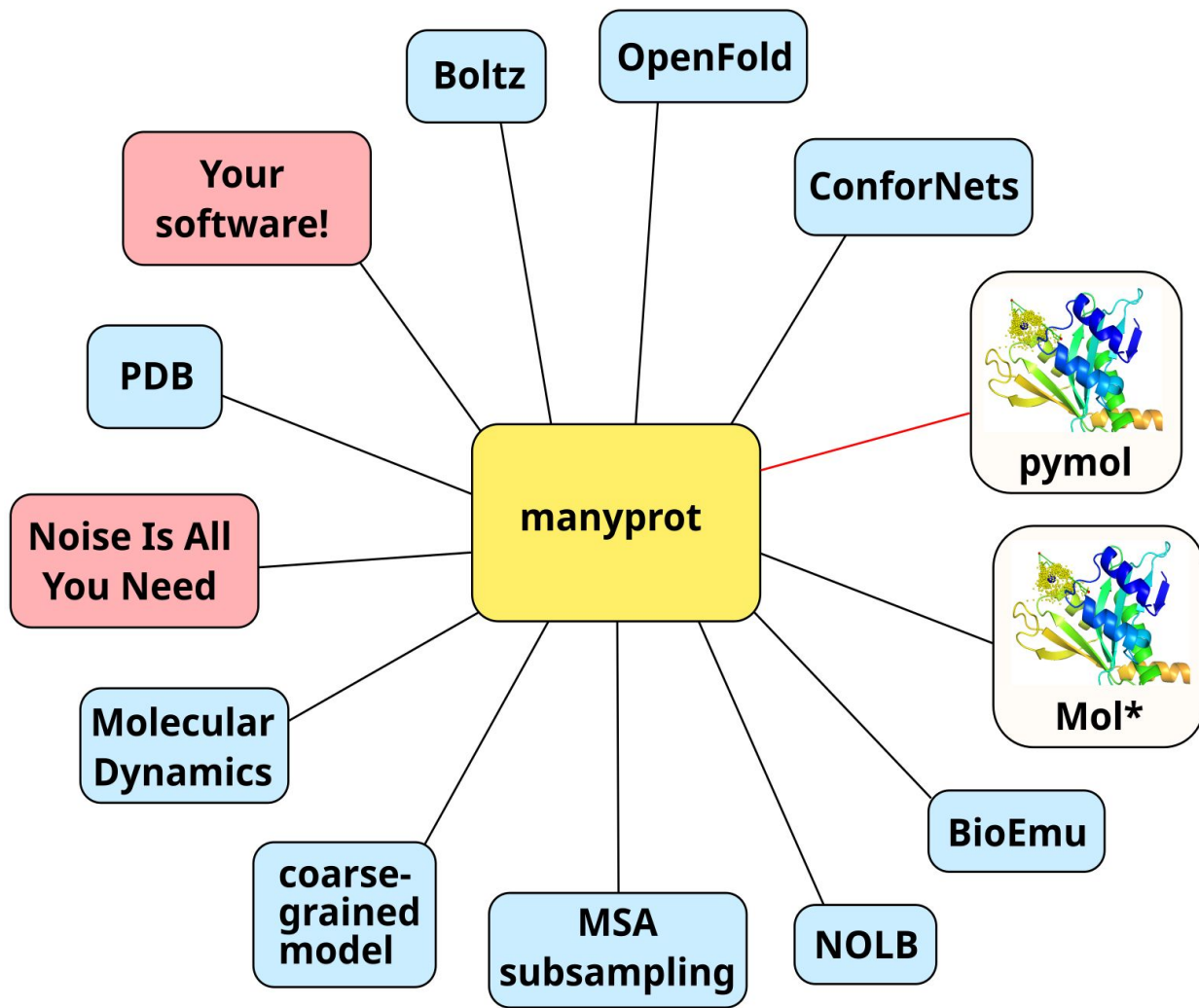
NatA pocket
opening



(end of the video)

manyprot project

- have thousands conformations in files & pytorch
- send only some to pymol or Mol*



Jupytermol - control pymol from Jupyter cells

`pymol()`

`%pymol`

`%%pymol`

and some 3D graphics

The screenshot displays a JupyterLab environment. On the left, a Jupyter notebook titled 'Allosteric analysis' is open. It contains the following code cells:

```
[2]:
import os
%load_ext jupytermol
from jupytermol import pymol, pymol_3d_scatter, pymol_3d_line, pymol_reinitialize

pymol(), %pymol and %%pymol now send to pymol -R

[3]:
pymol_reinitialize()
%pymol bg_color white

Sending 3 lines to pymol

[1]:
%%pymol
#everything in this cell
#will be sent to pymol
```

On the right, a PyMOL window is open, showing a 3D molecular model of a protein structure. The protein is represented by a cyan ribbon, and a cluster of yellow spheres is visible within the protein's binding pocket. The PyMOL interface includes a menu bar (File, Edit, Build, Movie, Display, Setting, Scene, Mouse, Wizard, Plugin, Help), a toolbar with buttons like Reset, Zoom, Orient, Draw/Ray, Unpick, Deselect, Rock, Get View, and a command line at the bottom.

Human-computer interface still feels like a bottleneck...

... until you meet prof. Andrea Brancale and see his work on haptic devices

J Mol Model (2009) 15:193–196
DOI 10.1007/s00894-008-0387-8

ORIGINAL PAPER

Accessible haptic technology for drug design applications

Nicola Zonta · Ian J. Grimstead · Nick J. Avis ·
Andrea Brancale

Received: 12 September 2008 / Accepted: 29 September 2008 / Published online: 2 December 2008
© Springer-Verlag 2008

Abstract Structure-based drug design is a creative process that displays several features that make it closer to human reasoning than to machine automation. However, very often the user intervention is limited to the preparation of the input and analysis of the output of a computer simulation. In some cases, allowing human intervention directly in the process could improve the quality of the results by applying

tested in a virtual screening simulation processed by a desktop machine in a few days. Continuous improvement in user interface design has simplified user interaction to a few mouse clicks; the development of many modeling software packages is often oriented toward limiting the human intervention, considered to be the bottleneck of the process. The user carries out preparation of the input and analysis of

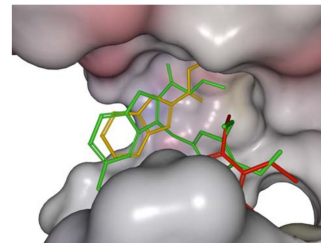


Fig. 2 Overlapping of an arylthioindole inhibitor (green) with the indole-2-methylcarboxylate (yellow)

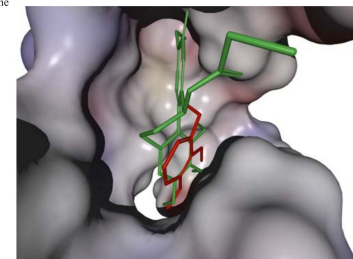
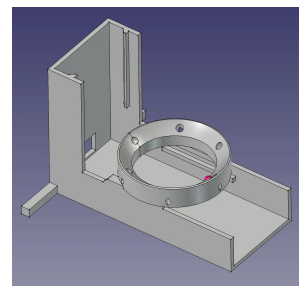
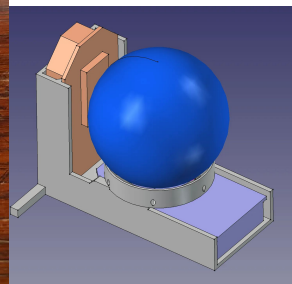
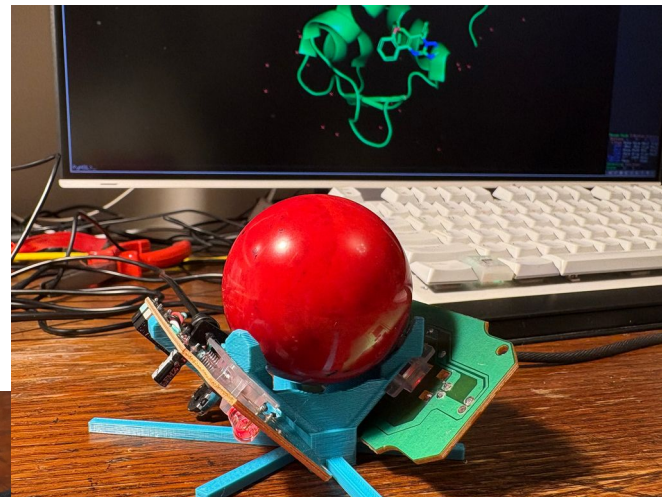
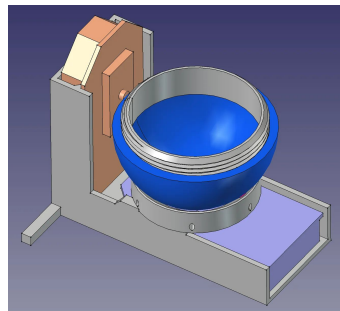


Fig. 1 Overlapping of DAMA-Colchicine (green) and the trimethoxyphenyl ring (red) placed with ZODIAC

3D printed 3D trackball to control pymol and Avogadro



References and Thanks

AlphaFold2: Jumper et al., Nature (2021)

OpenFold: Ahdritz et al., bioRxiv (2022) and Nature Methods (2024)

ESMFold: Lin et al., Science (2023), ESMFold 2 with ESM-C LM: Candido et al., [bioRxiv](#) (2026)

MSA subsampling: Sampling alternative conformational states of transporters and receptors with AlphaFold2, del Alamo et al., eLife (2022)

BioEmu (2024): <https://www.biorxiv.org/content/10.1101/2024.12.05.626885v1.full.pdf>

AlphaFLOW and ESMFLOW (2024):

<https://cbirt.net/alphafLOW-and-esmflow-a-new-approach-to-modeling-protein-ensembles-with-combined-power-of-alphafold-and-flow-matching/>

ConforNets (2026): <https://arxiv.org/abs/2604.18559>

Zonta, Grimstead & Brancale: Accessible haptic technology for drug design applications (2008)

<https://link.springer.com/article/10.1007/s00894-008-0387-8>

3D printed 3D trackball for pymol and Avogadro (me, 2026): <https://www.printables.com/model/1787278-trackball3d>

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. Computational resources were provided by the ELIXIR-CZ project (ID:90255), part of the international ELIXIR infrastructure.

Thank you for your attention!

Any questions?

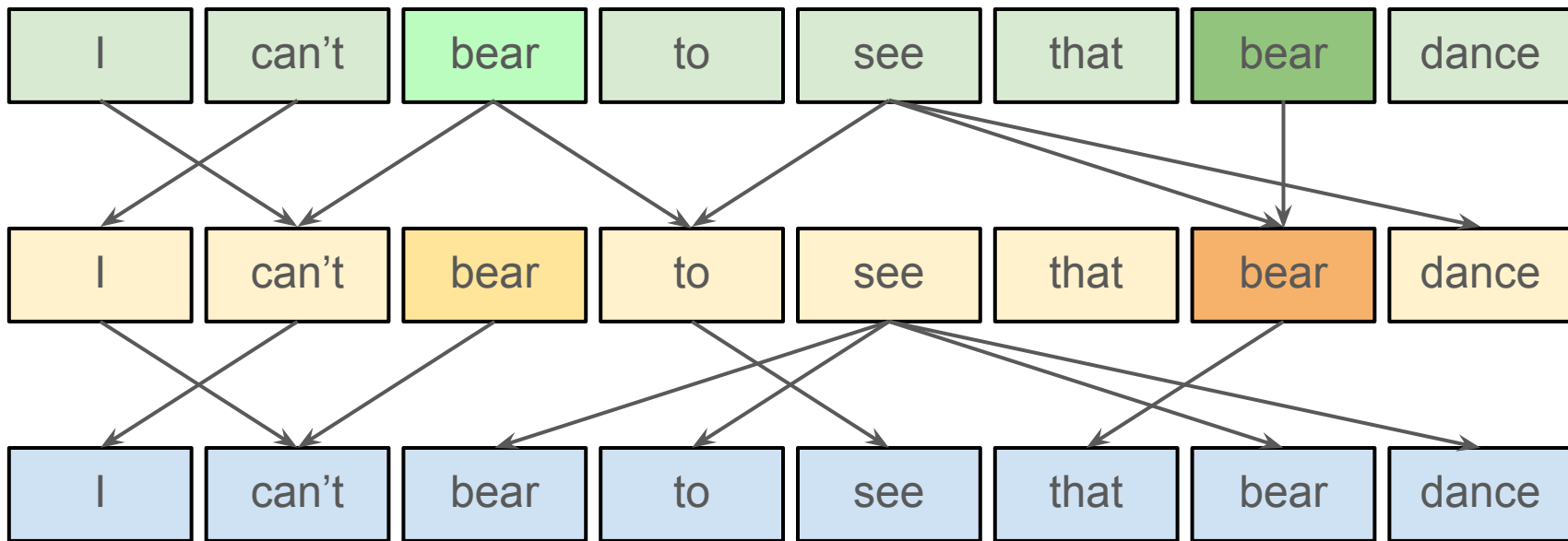
[slides](#)
are freely
available
here:

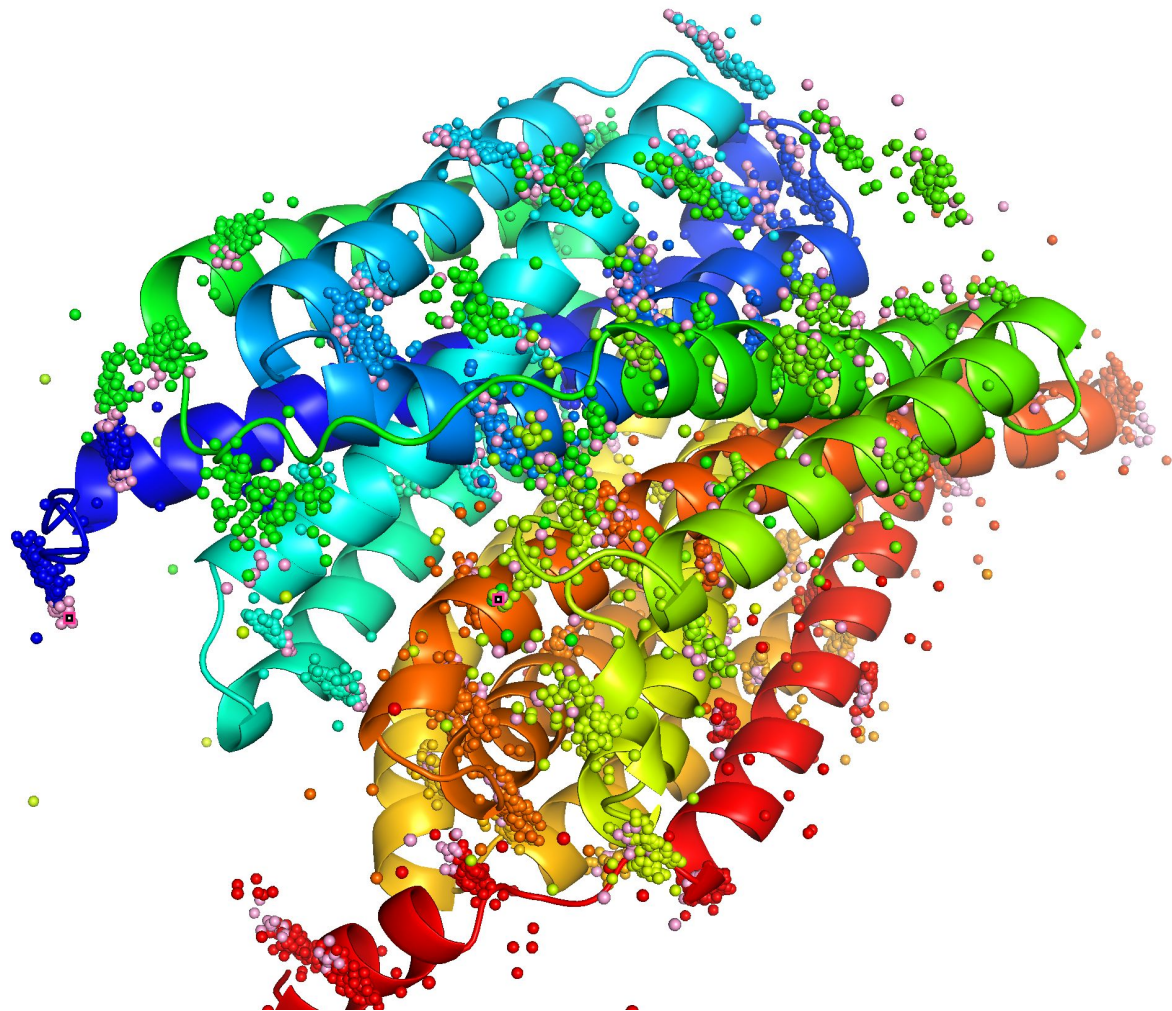


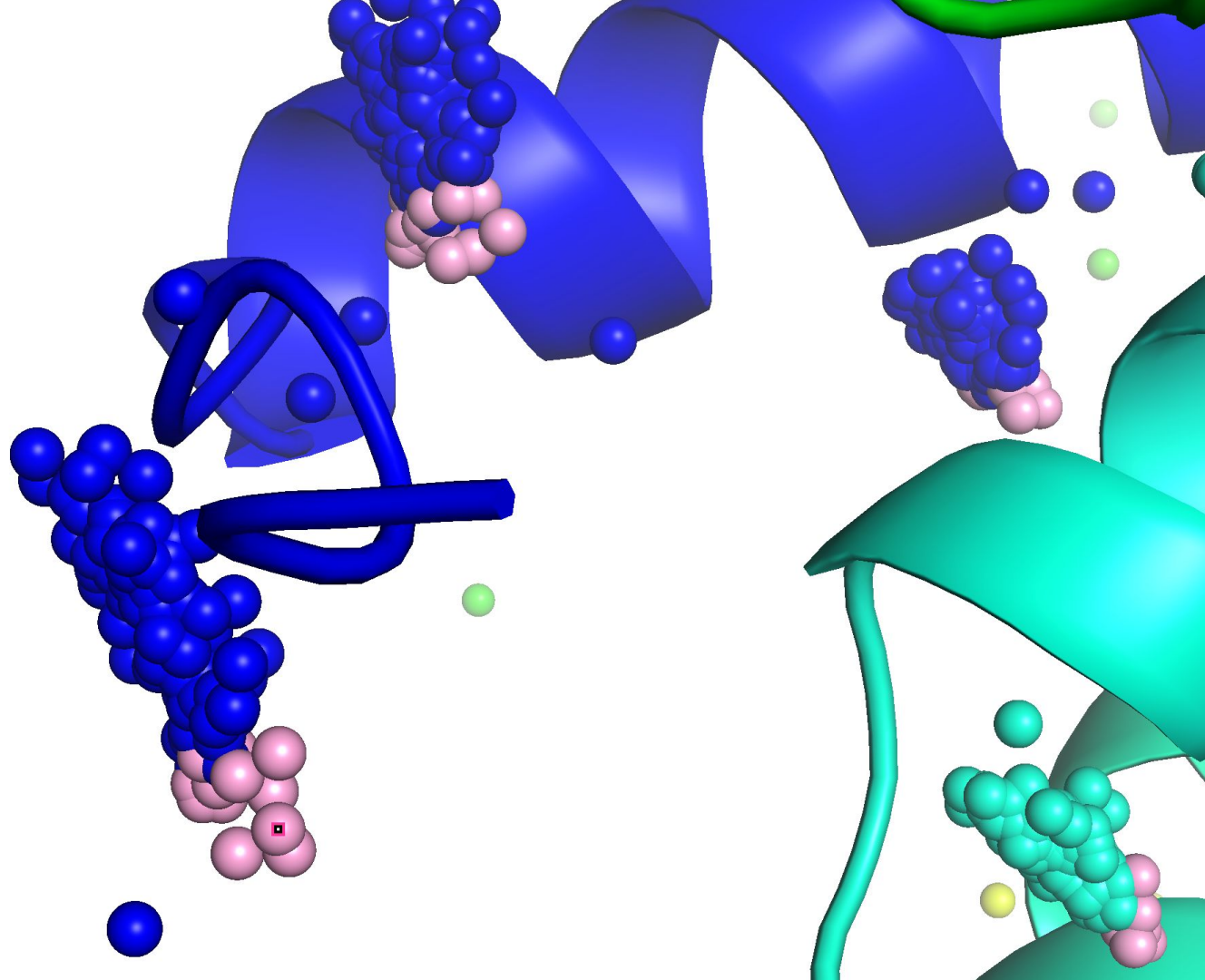
Some bonus slides for eventual
discussions follow...

How the Protein Language Model (PLM) works

It uses model structure from natural language processing, just applied to amino acid sequences in proteins - and it works great, too. Attention mechanism:



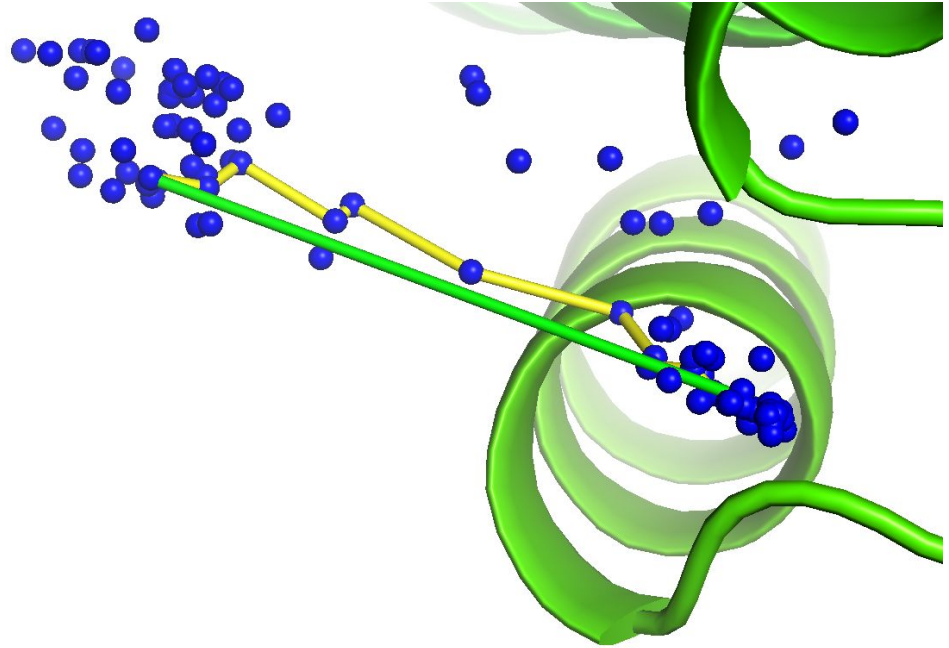




Interpolate between extreme positions

- Make path smooth in local detail
- Make changes small in global
- This way we can isolate a sensible mode (for human eye)

Interpolation: Gradual placement of best mid-points between existing points on path



BERT (Bidirectional encoder representations from transformers) uses the 1953 method on Internet texts, getting **huge amount of training data without manual labeling**.

Once computers learns to fill in missing words, it virtually knows the language.

1. Deleting an equal number of words from each passage by some essentially random counting-out system.

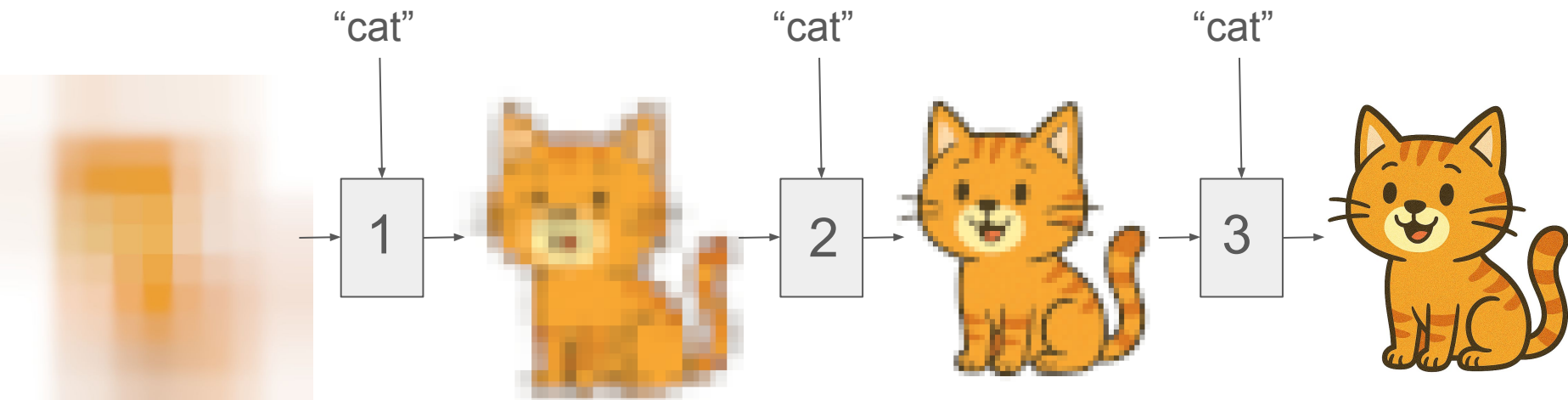
Such a system was based on a table of random numbers or else it simply counted out every n th word (every fifth one, for example) without any regard for the functions or meanings of specific words.

Does the **green** sentence really follow after the **orange** one in the original text?

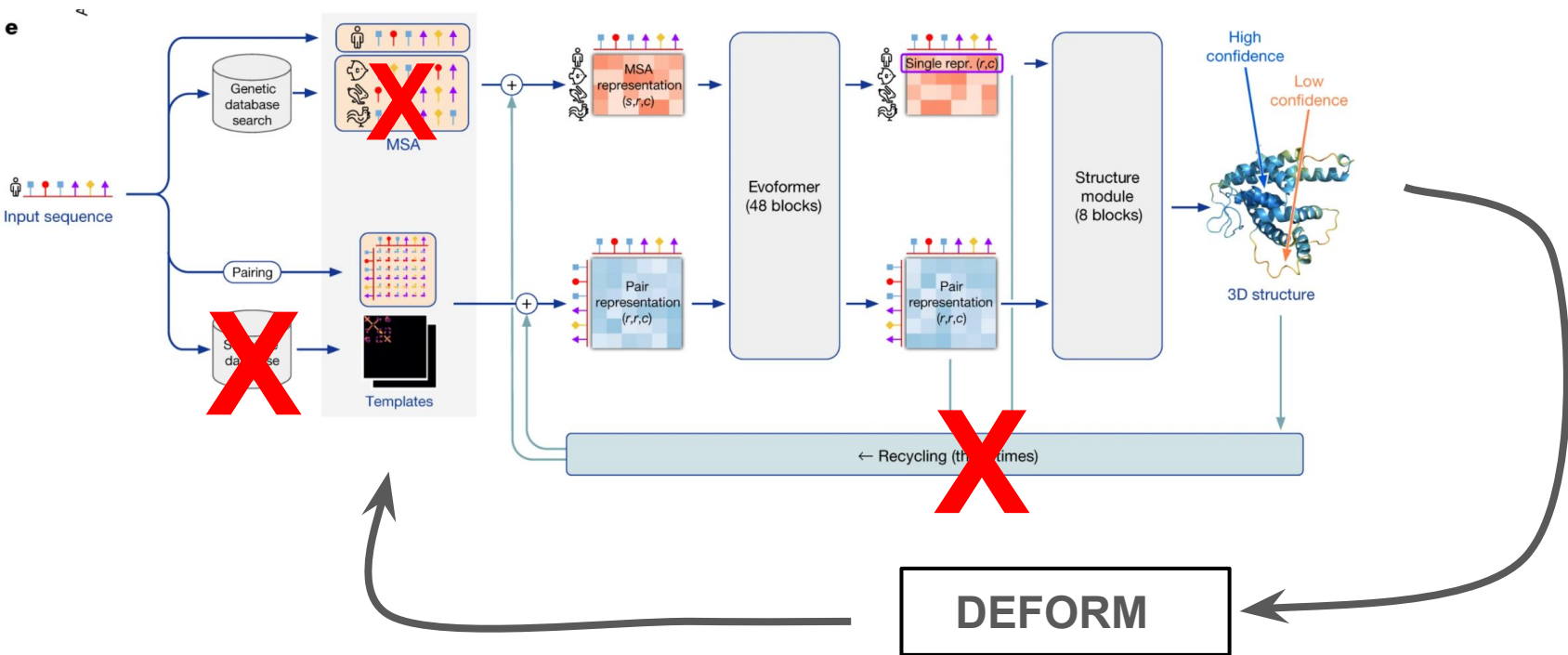
BERT uses this additional task in pre-training.

Diffusion needs just a label “cat”, can hallucinate many

trained to remove noise, knowing what is on the picture (cat)



Templates used to manipulate shape using OpenFold 2



Templates created by altering the output of the model itself

2007, 2012: word as vector

Tomáš Mikolov in his Ing. and PhD. diploma works stubbornly insists on using vectors of continuous numbers to represent words, instead of using graph nodes as everybody else.

His work is later released as **word2vec**, revolutionizing language models. Vectors of continuous numbers represent **mostly everything now**.



king - man + woman = queen

