Machine Learning in Rosetta

Presented by Kuniko Hunter
Adapted from Cristina Elisa Martina and Max Beining

Rosetta Workshop • November 2025

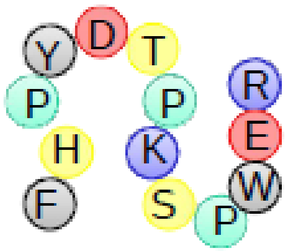
Meiler Lab • Vanderbilt University

VANDERBILT UNIVERSITY

meilerlab

Intro to ML in Protein Design

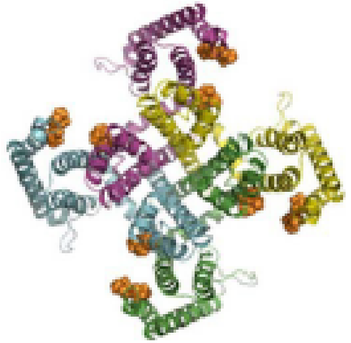
What 3D structure does a protein sequence adopt?



Protein Folding Problem

Sequence ⇄ Structure

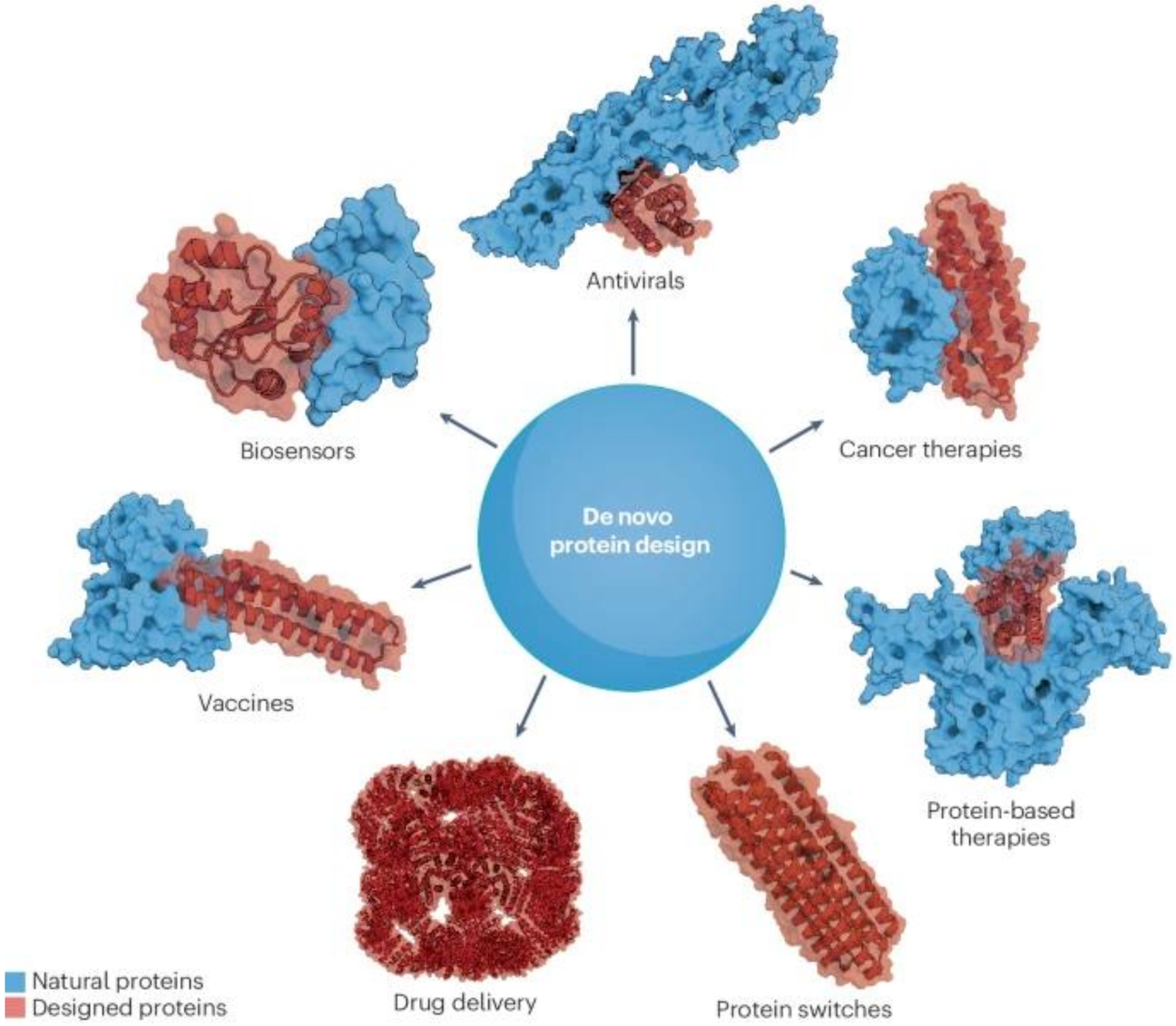
Inverse Folding Problem



Which protein sequence folds into the given 3D structure?

What is the best sequence to:

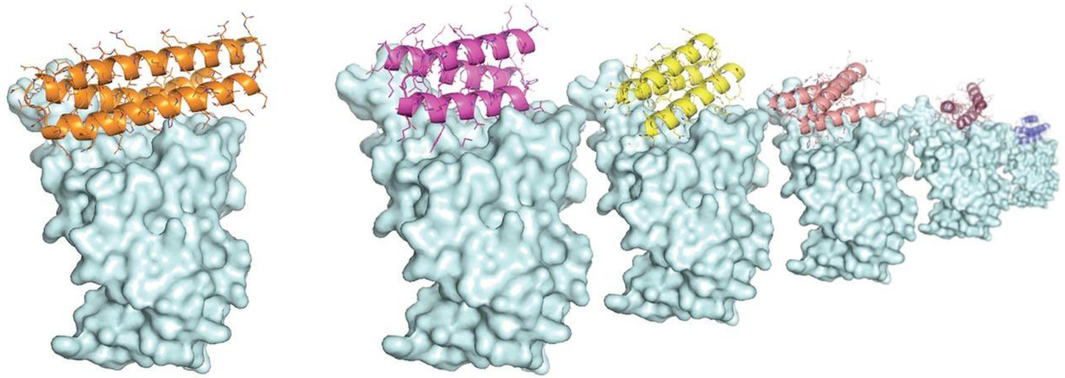
- Fold into this protein scaffold?
  - New functions
  - New shapes (de novo design)
- Increase protein stability?
  - Half-life
  - Thermostability
- Increase binding to X?
  - Protein-protein
  - Protein-ligand
  - Supramolecular assemblies
- Increase enzymatic activity?
  - Activity
  - Specificity

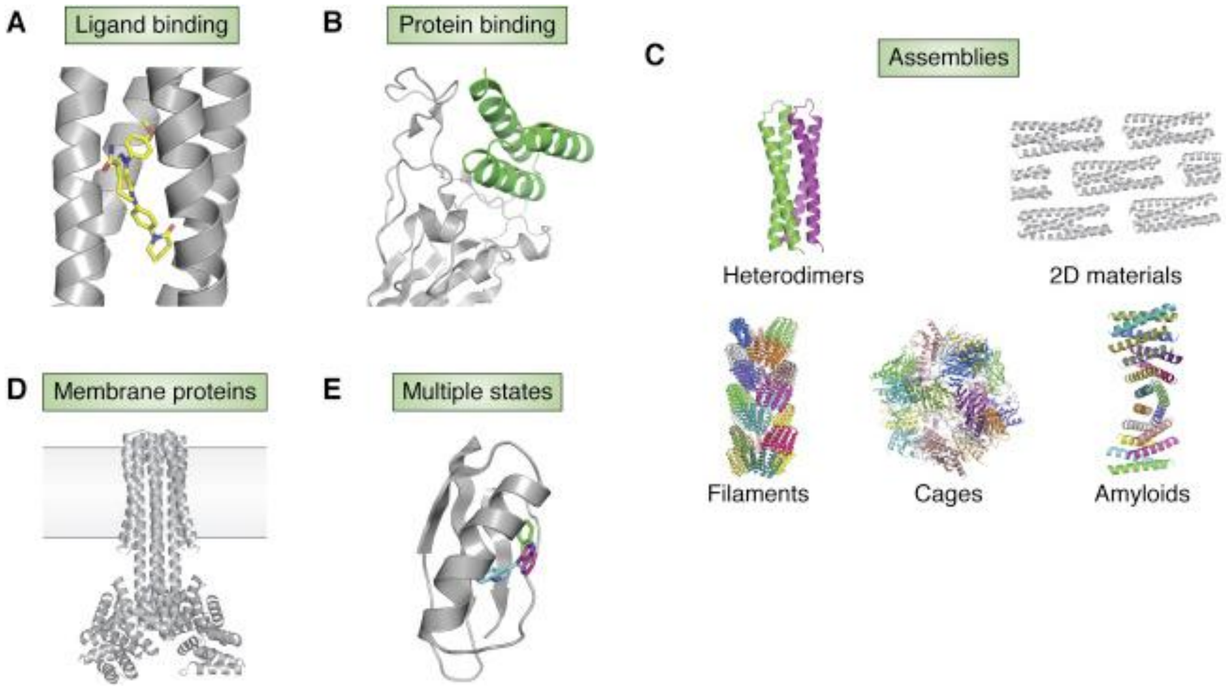


(Listov *et al.* 2024)

What is the best sequence to:

- Fold into this protein scaffold?
  - New functions
  - New shapes (de novo design)
- Increase protein stability?
  - Half-life
  - Thermostability
- Increase binding to X?
  - Protein-protein
  - Protein-ligand
  - Supramolecular assemblies
- Increase enzymatic activity?
  - Activity
  - Specificity



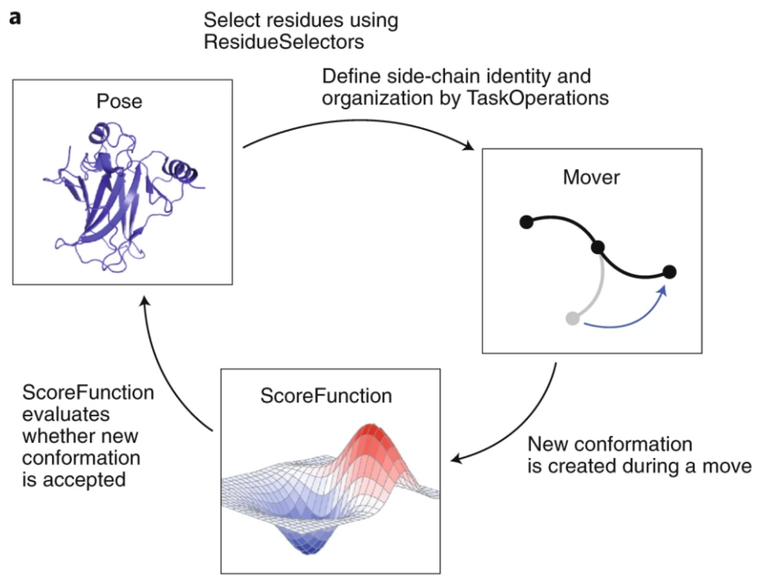


(Cao *et al.* 2020; Pan *et al.* 2021)

Traditional Rosetta vs. ML-guided Rosetta

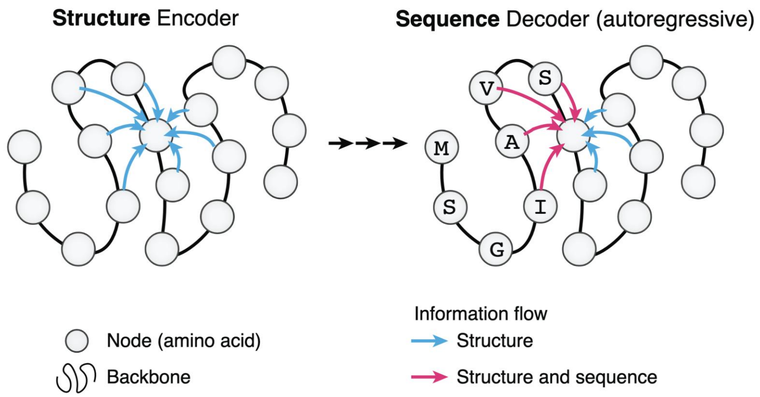
Traditional Rosetta

- Physics-based energy functions
- Rotamer sampling, backbone relaxation
- Evaluates sequences based on physical plausibility
- Can be slow for large sequence spaces



ML-guided Rosetta

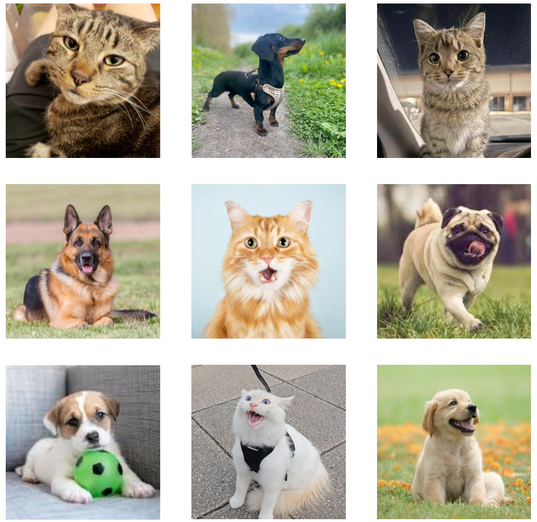
- Predicts residue probabilities at each position
- Uses sequence and/or structure data
- Captures evolutionary and structural patterns
- Guides or constrains design for faster, more accurate sampling



(Kohler Lehman *et al.* 2020; Ingraham *et al.* 2019)

How machines learn:

Example task: Classify images of dogs vs. cats



ML Model

“Dog” or “Cat”

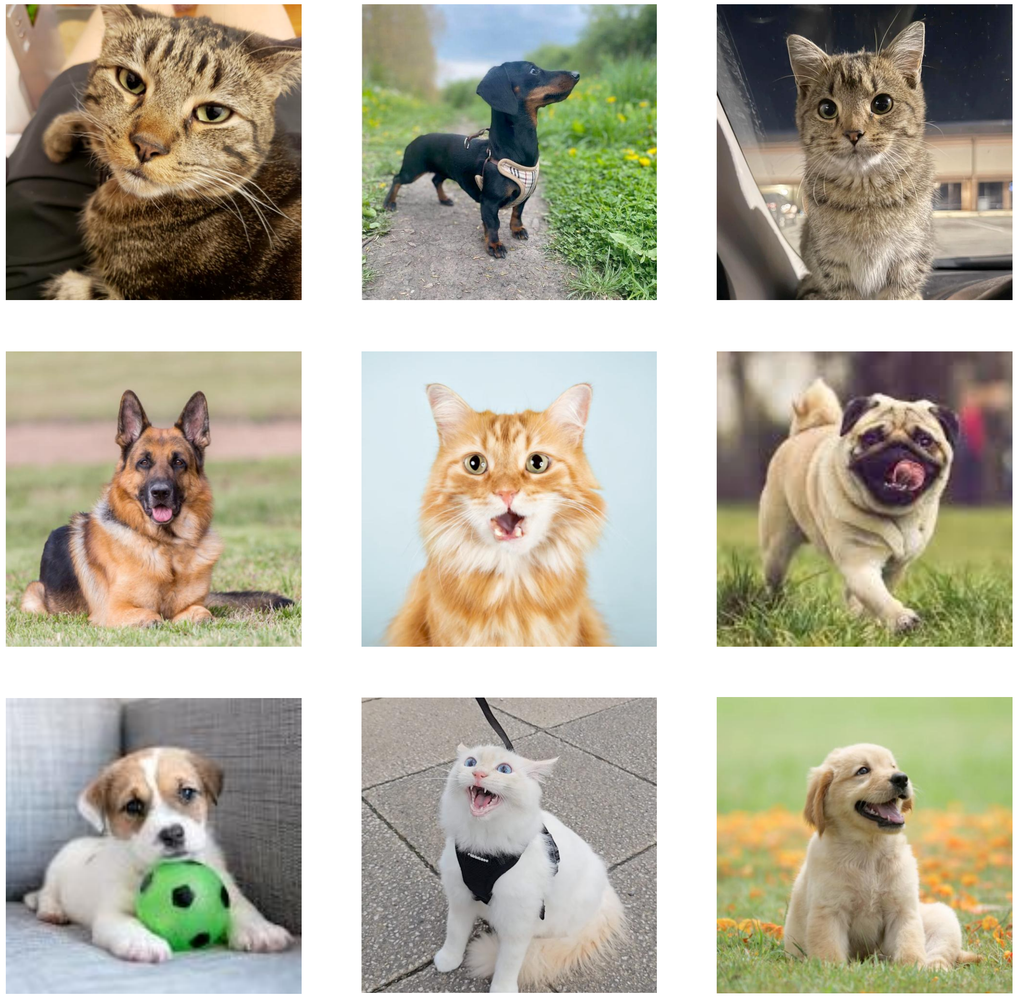
Input

Output

How machines learn:

- **Learning from examples:** ML models are trained on data (protein sequences, structures) to recognize patterns

Dataset



How machines learn:

Step 1: Dataset is divided into groups:

- **Train** (~75%): used to teach the model
- **Validation** (~15%): used to tune parameters *during* learning and avoid overfitting
- **Test** (~10%): used to evaluate performance on unseen data *after* learning

Train

Test

Proper split ensures the model generalizes, not just memorizes

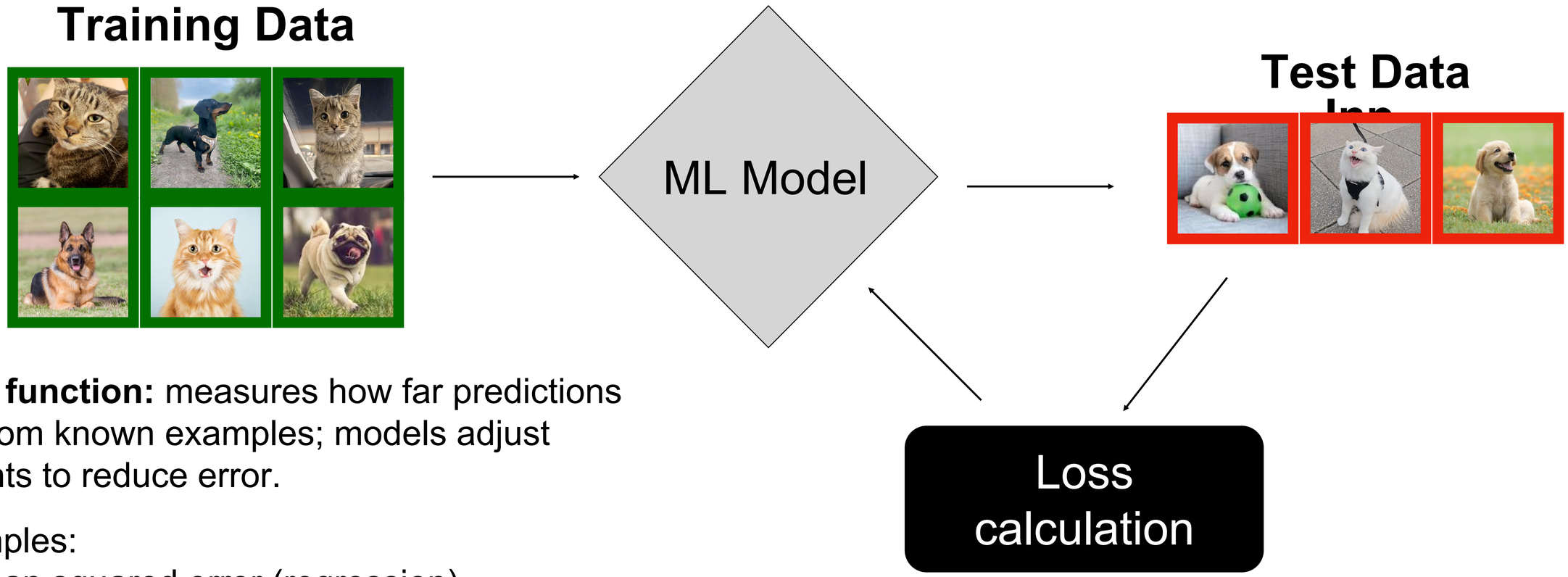
How machines learn: Loss (how the model knows its wrong)

Loss function: measures the difference between prediction and true label

- Example: predicting probabilities for dog/cat
  - Prediction1: True label = dog, model predicts 25% dog → high loss
  - Prediction2: True label = dog, model predicts 75% dog → lower loss
- Training = minimizing loss over all training examples

True label	DOG	CAT	DOG	DOG	DOG	CAT
						
Prediction1	CAT	CAT	CAT	DOG	CAT	CAT
Prediction2	CAT	CAT	DOG	DOG	DOG	CAT

How machines learn:

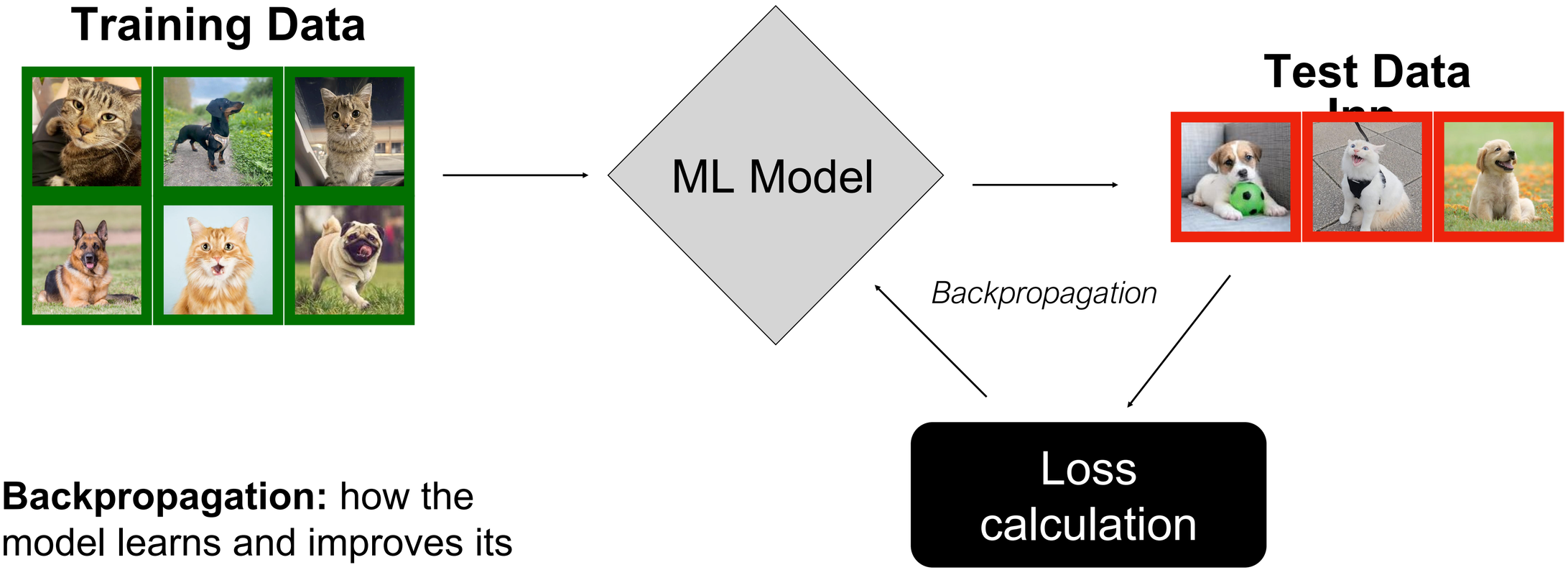


Loss function: measures how far predictions are from known examples; models adjust weights to reduce error.

Examples:

- Mean squared error (regression)
- Cross-entropy (classification)
- Perplexity (language models)

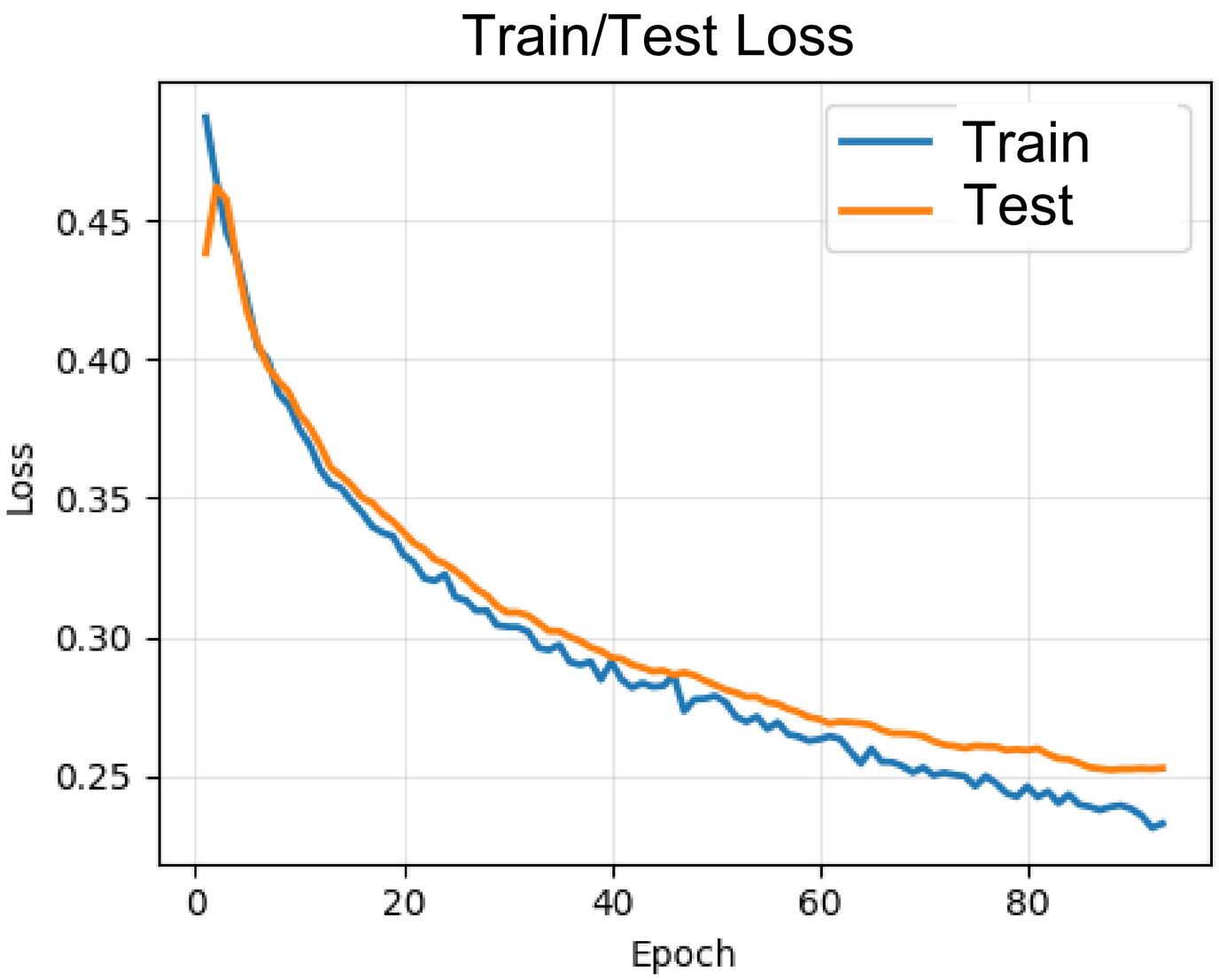
How machines learn:



- **Backpropagation:** how the model learns and improves its performance.

How machines learn:

- **Iterative learning:** gradually improves accuracy over many sequences/structures



Today's ML methods:

<u>ProteinMPNN</u>: structure-aware, predicts amino acids for a given structure

- Dauparas, J. et al. Robust deep learning based protein sequence design using ProteinMPNN. 2022.06.03.494563 Preprint at https://doi.org/10.1101/2022.06.03.494563 (2022).

<u>ESM</u> (Evolutionary Scale Modeling): sequence-only, captures evolutionary conservation

- Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 118, e2016239118 (2021).
- Rao, R. M. et al. MSA Transformer. in Proceedings of the 38th International Conference on Machine Learning 8844–8856 (PMLR, 2021).
- Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).

<u>MIF-ST</u>: sequence & monomeric structure, hybrid approach

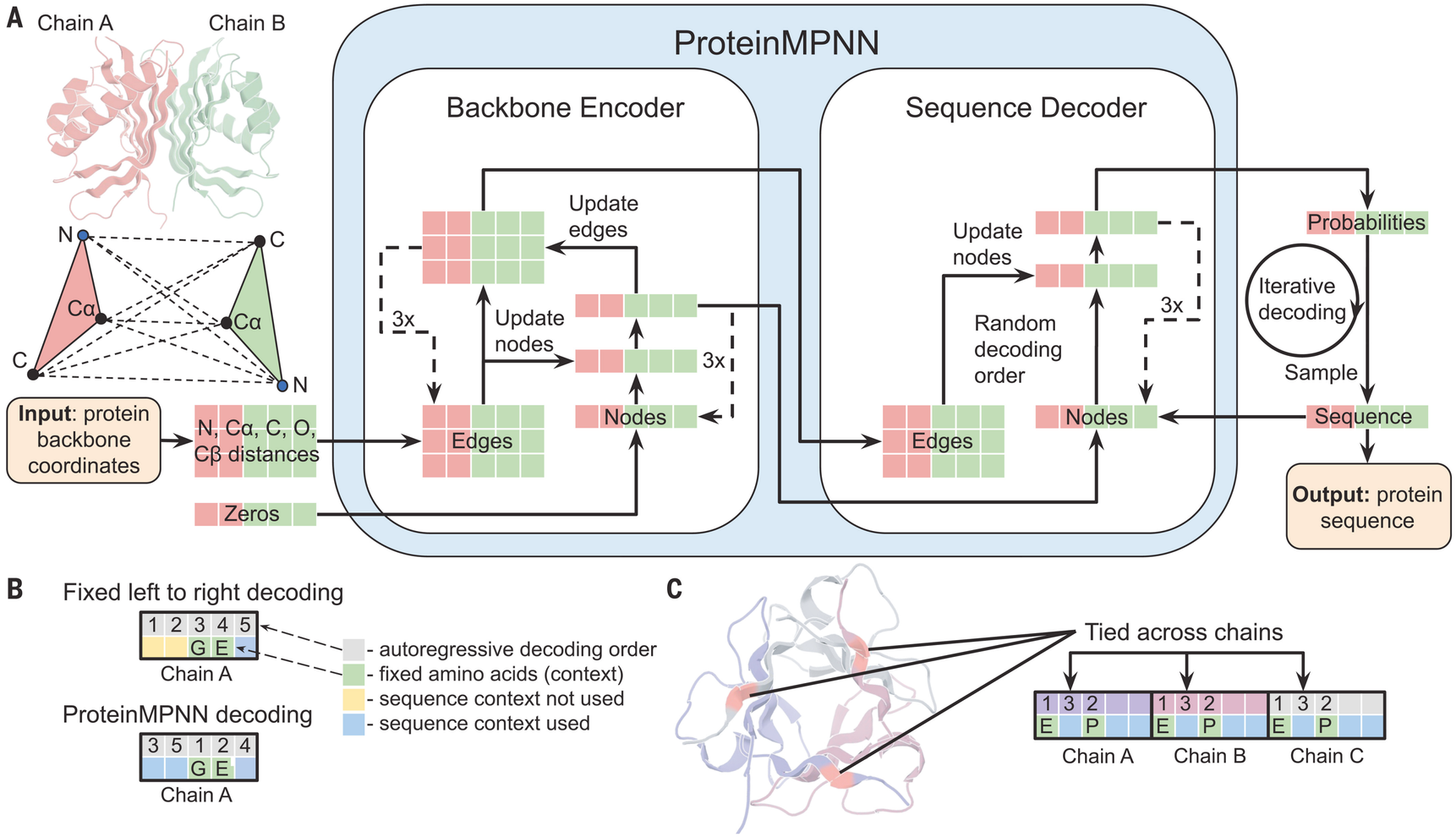
- Yang, K. K., Zanichelli, N. & Yeh, H. Masked inverse folding with sequence transfer for protein representation learning. Protein Engineering, Design and Selection 36, gzad015 (2022).

ProteinMPNN (Message Passing Neural Network)

- Trained on protein structures from RCSB-PDB:
  - 19,700 single-chain protein structures
  - Further trained on clustered high-res multichain structures
- Predict probabilities of each natural aa for each position
- Use probabilities to design sequences
- Tested in silico:
  - 690 monomers
  - 732 homomers
  - 98 heteromers
- Tested experimentally

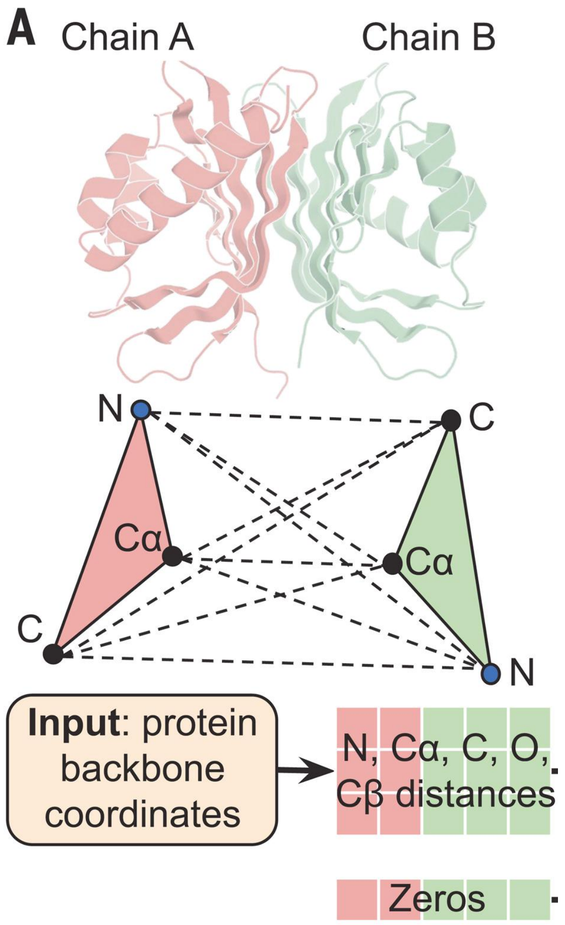
(Dauparas *et al.* 2022)

ProteinMPNN (Message Passing Neural Network)



(Dauparas *et al.* 2022)

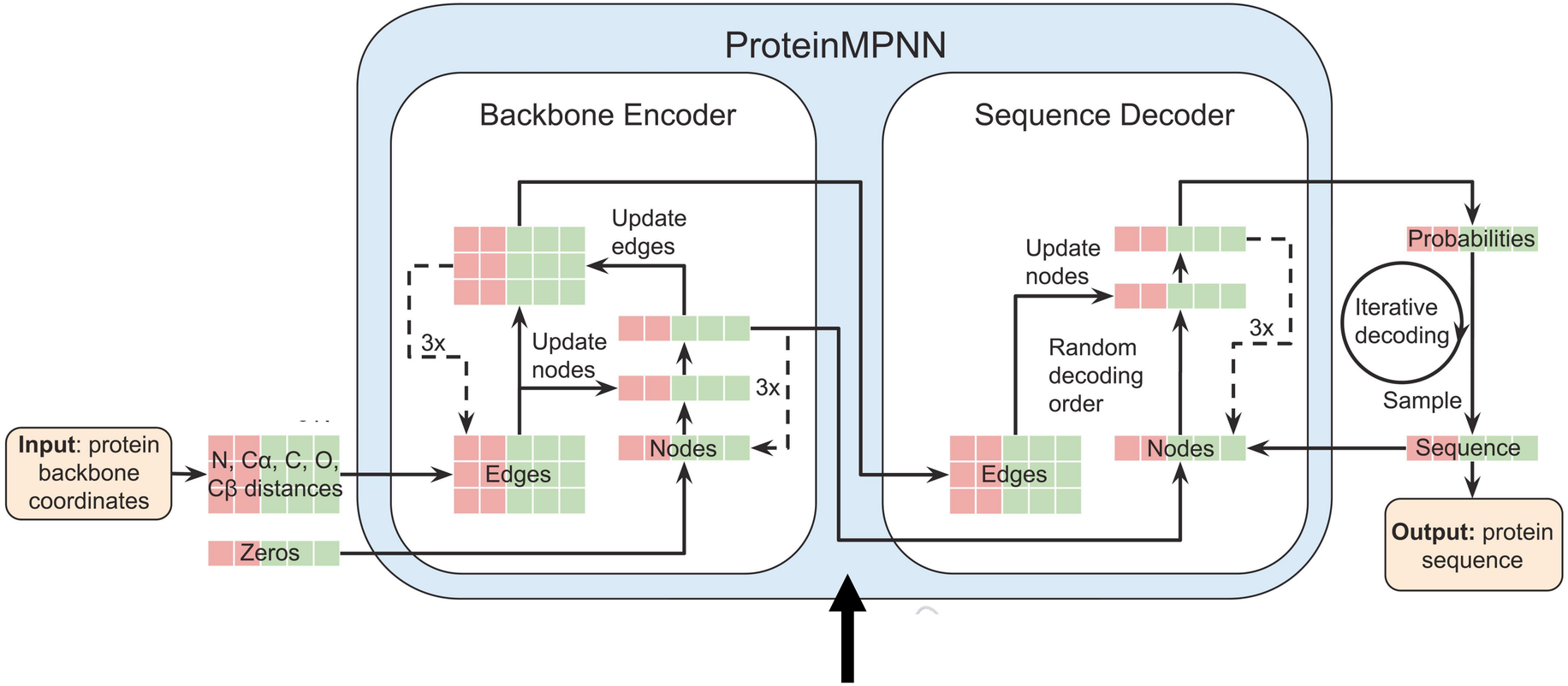
ProteinMPNN Inputs:



- RCSB-PDB database
- No evolutionary information
- Distances between N, Ca, C, O and virtual CB are encoded using graph theory:
  - Nodes (atoms)
  - Edges (distances)

(Dauparas *et al.* 2022)

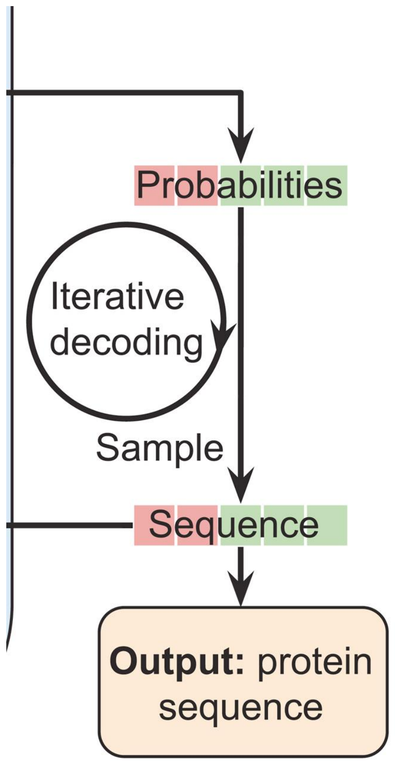
ProteinMPNN: the Message Passing Neural Network



128 hidden dimensions
(Where predictions occur)

(Dauparas *et al.* 2022)

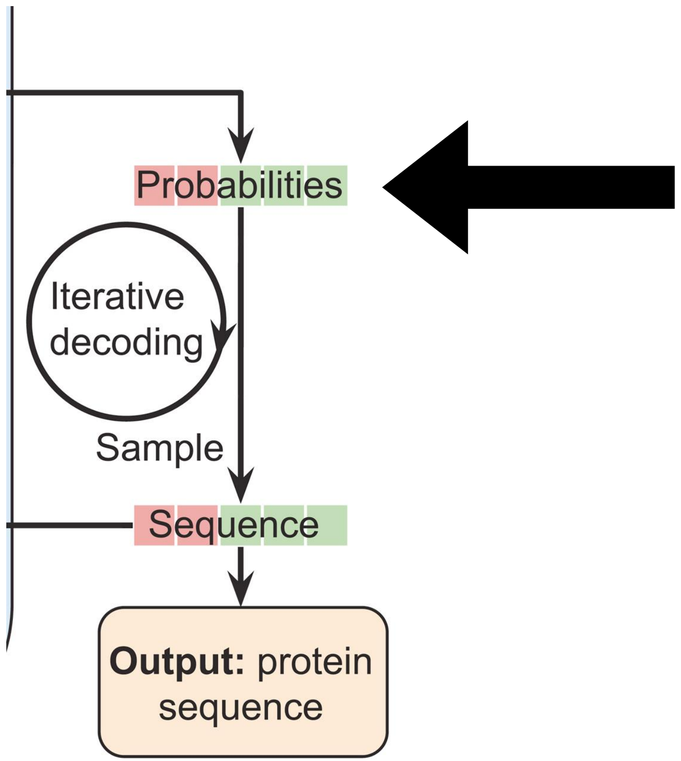
ProteinMPNN Outputs:



- ProteinMPNN outputs re-designed sequences, not structures!
- This means that you must predict a designed structure with an alternative method (AlphaFold, Rosetta, Chai, Boltz)

(Dauparas *et al.* 2022)

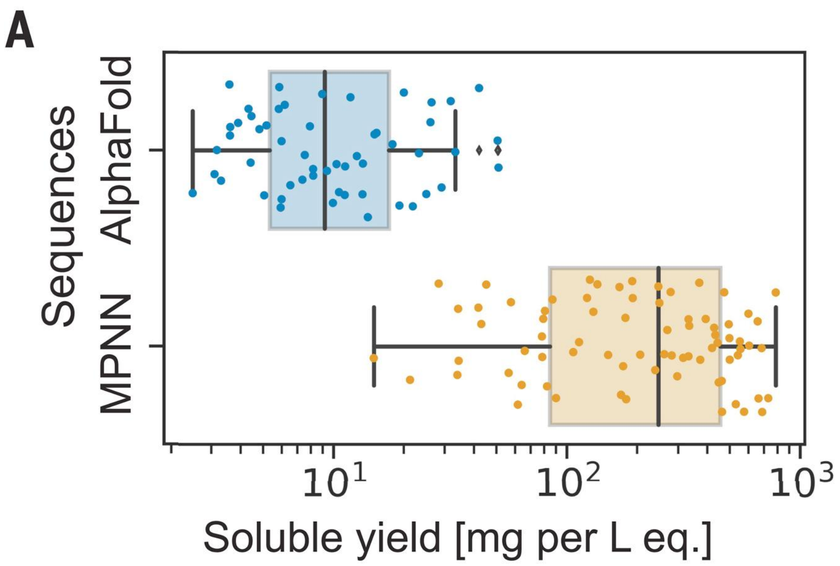
ProteinMPNN Outputs:

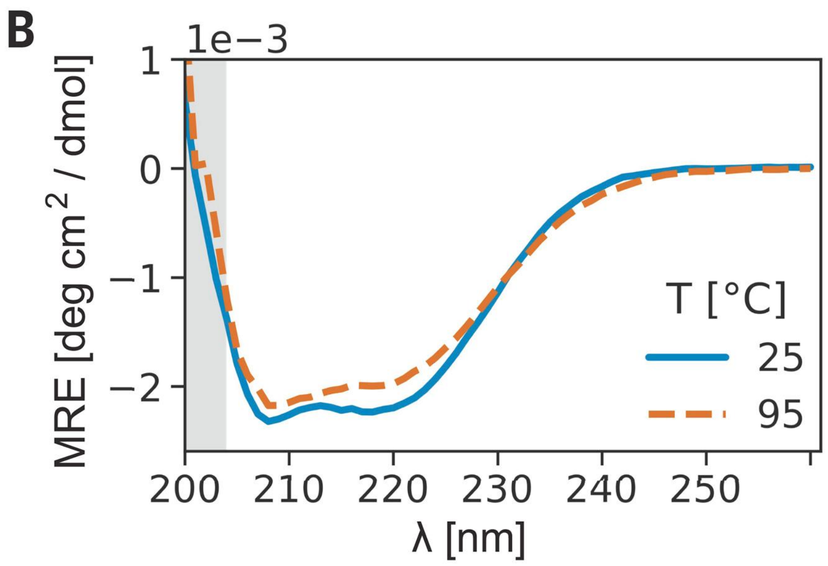


ProteinMPNN in Rosetta takes the probabilities as outputs, and uses it for designing the structure directly.

(Dauparas *et al.* 2022)

ProteinMPNN Performance:



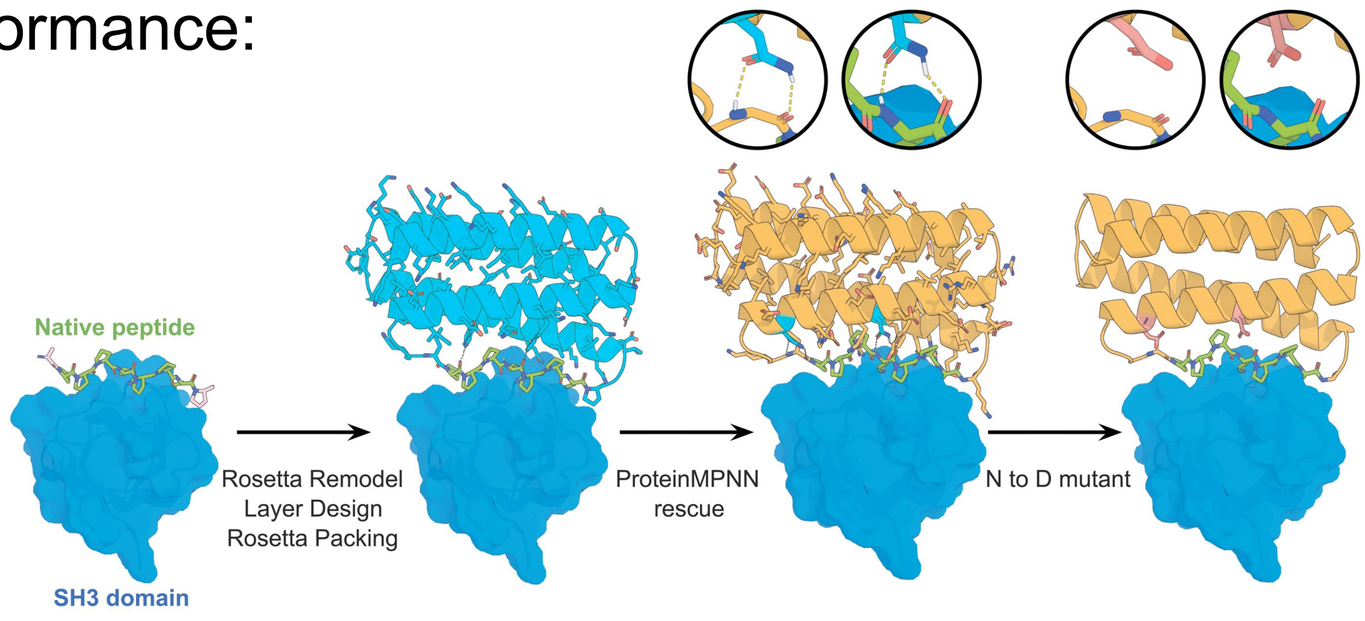


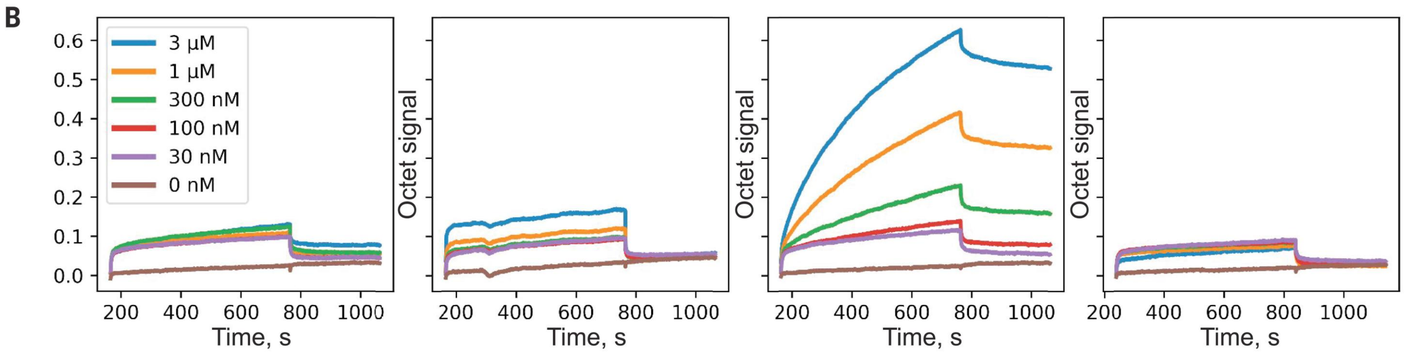
Recovered
solubility

Increased
thermostability

(Dauparas *et al.* 2022)

ProteinMPNN Performance:





Creates **new functions**

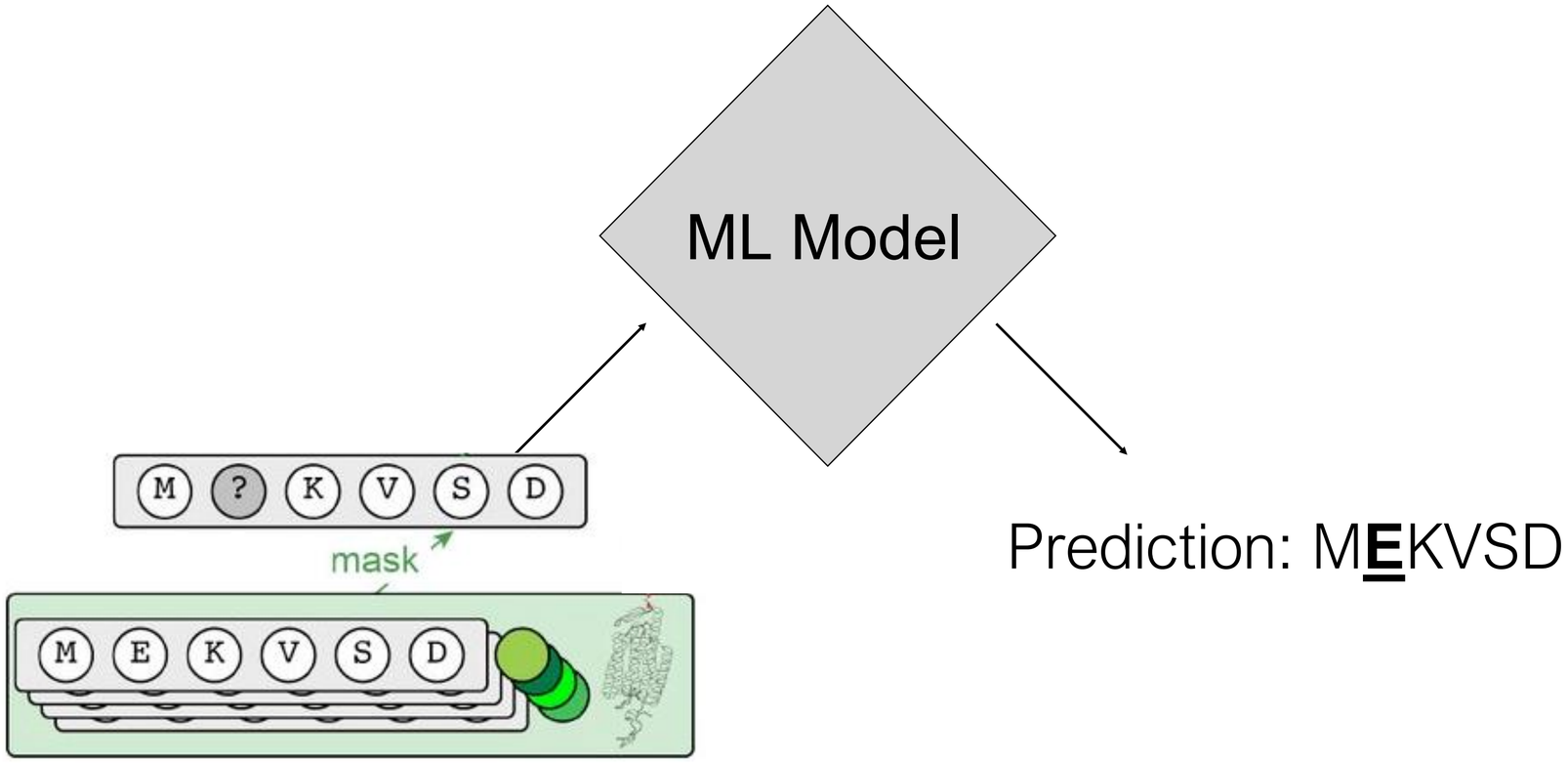
(Dauparas *et al.* 2022)

MIF-ST (Masked Inverse Folding with Sequence Transfer):

- Trained on **both** protein structures and sequences:
  - 19,700 single-chain protein structures from RCSB-PDB
  - 42M sequences from UniRef50
  - Sequences are partially masked
  - Model must predict masked residues
- Training for downstream task
  - Trained on single mutants and predicts multiple mutants
  - Predict experimental measurements
- Tested *in silico* on small and large datasets:
  - Deep mutational scans
  - Enzymatic activity
  - Stability
  - Binding

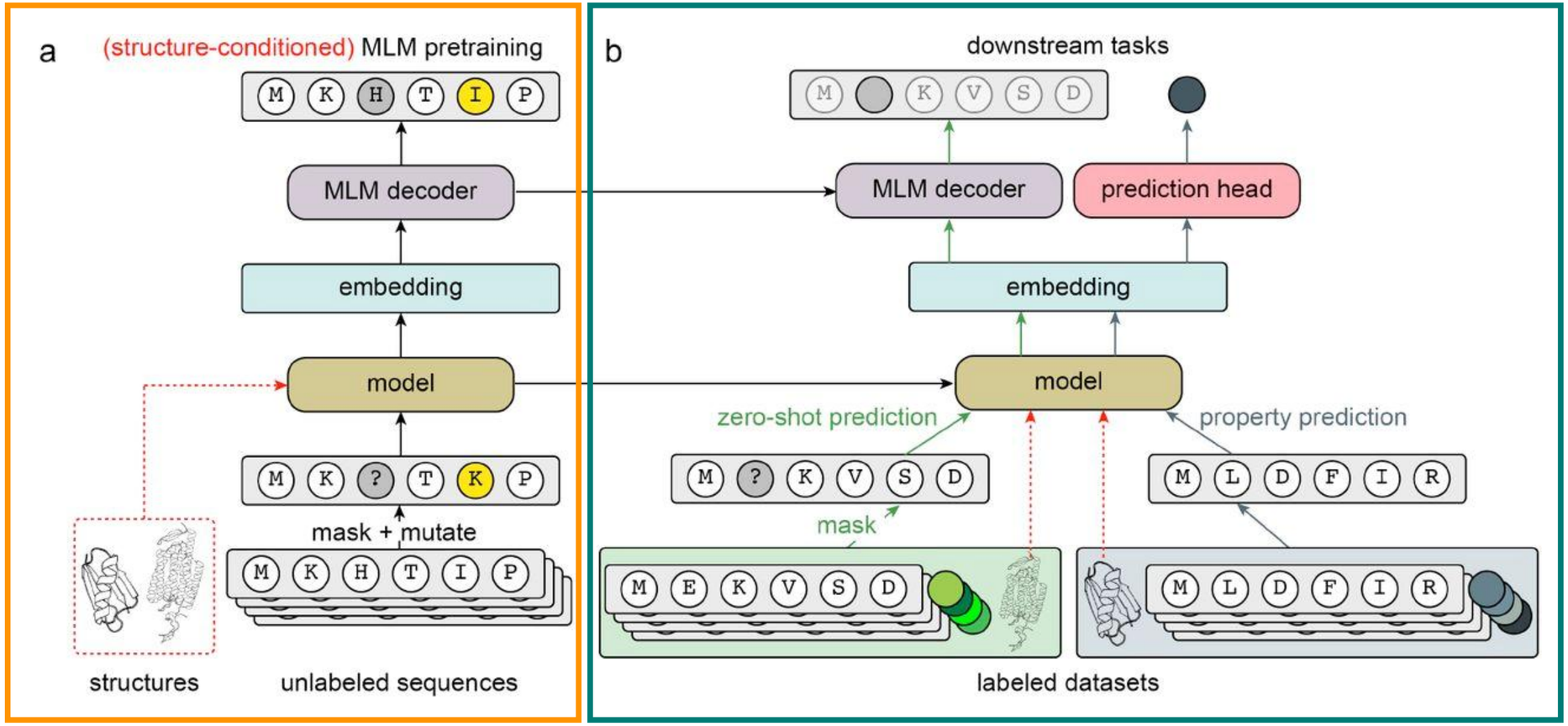
(Yang *et al.* 2023)

Masking Protein Sequences in ML:



(Yang *et al.* 2023)

MIF-ST:



Pre-training
(Structures, sequences, masking)

Training
(Sequences, masking)

(Yang *et al.* 2023)

MIF-ST Performance:

Regime	Model	Parameters	Perplexity	Recovery
sequence only	CARP-640M	640M	7.06	40.5%
sequence & structure	MIF-4	3.4M	4.95	49.9%
	MIF-8	6.8M	5.00	46.7%
	GVPMIF	3.5M	4.68	51.2%
+sequence transfer	MIF-ST	3.4M	**4.08**	**55.6%**
-UniRef50 pretraining	MIF-ST	3.4M	5.70	45.4%

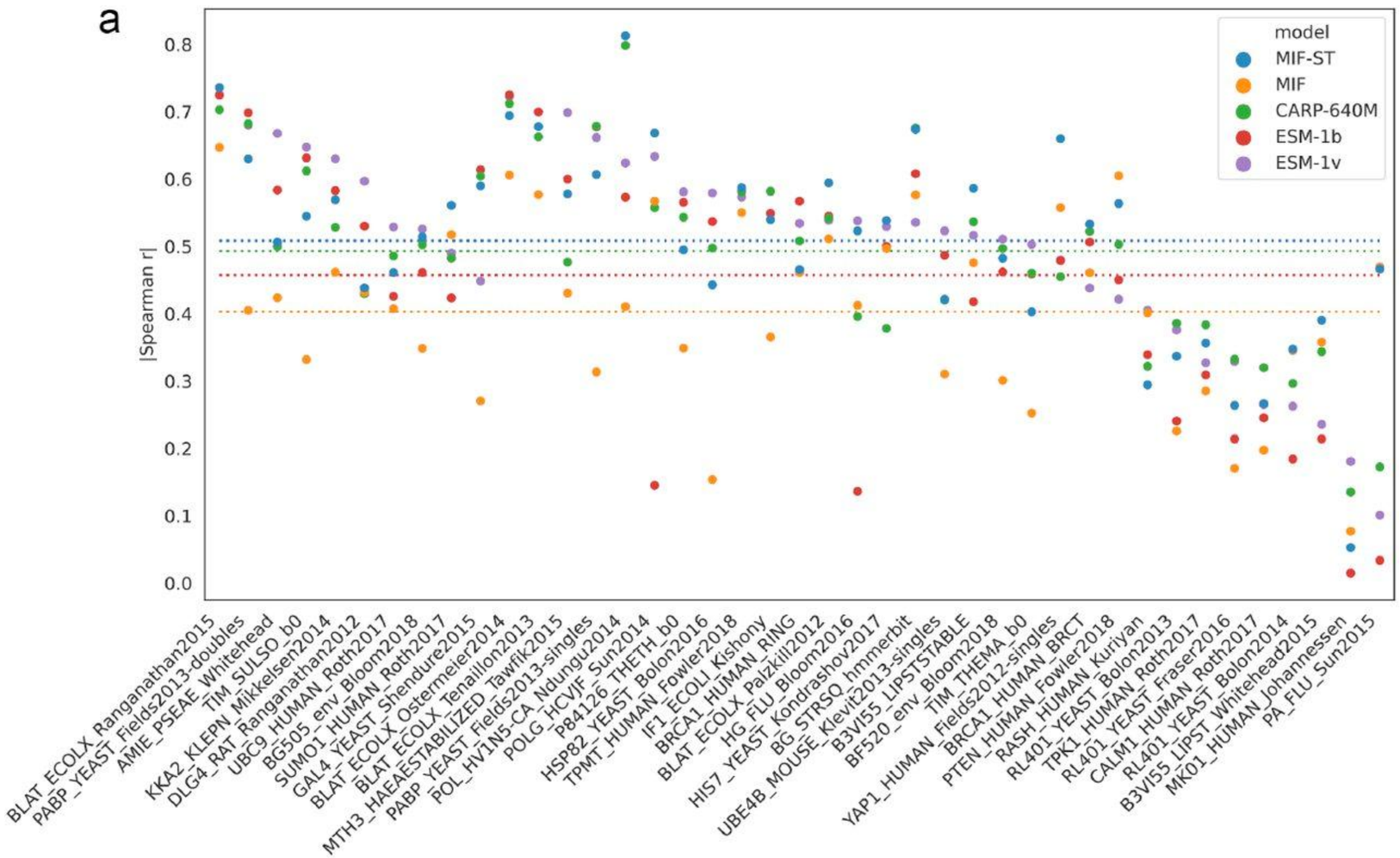
Perplexity
Model's uncertainty in prediction
(lower is batter)

Sequence Recovery
How well the model recovers native sequences
(higher is better)

(Yang *et al.* 2023)

MIF-ST Performance:

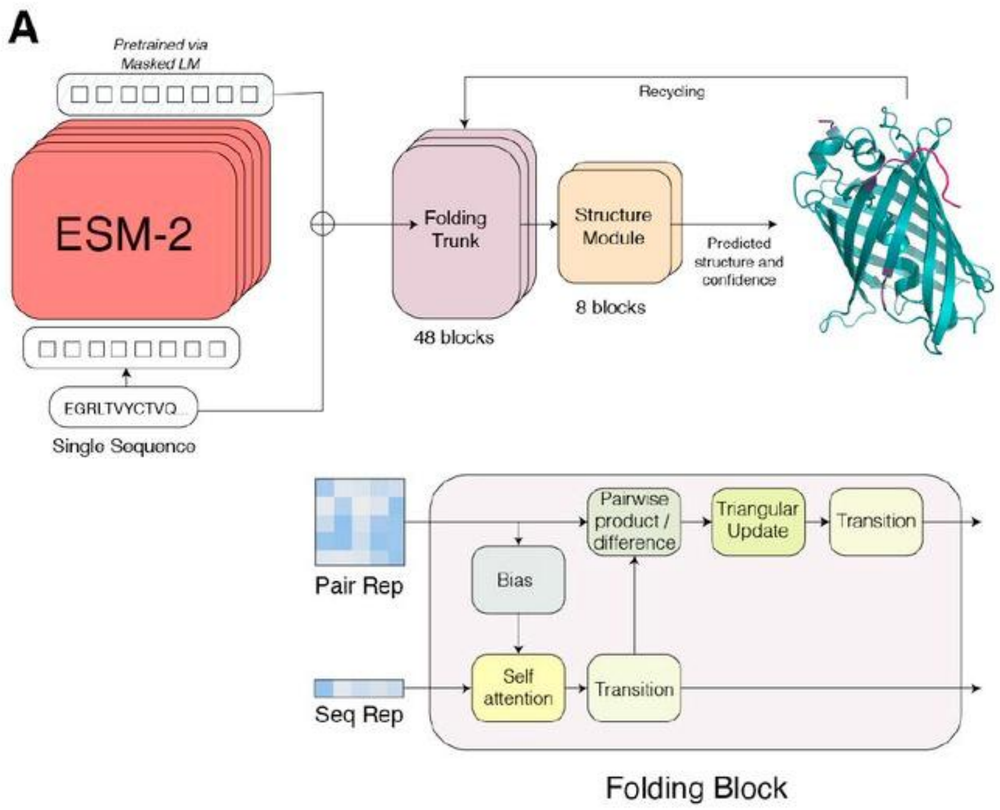
Predictions on DMS datasets:
MIF-ST is outperforming in many cases



(Yang *et al.* 2023)

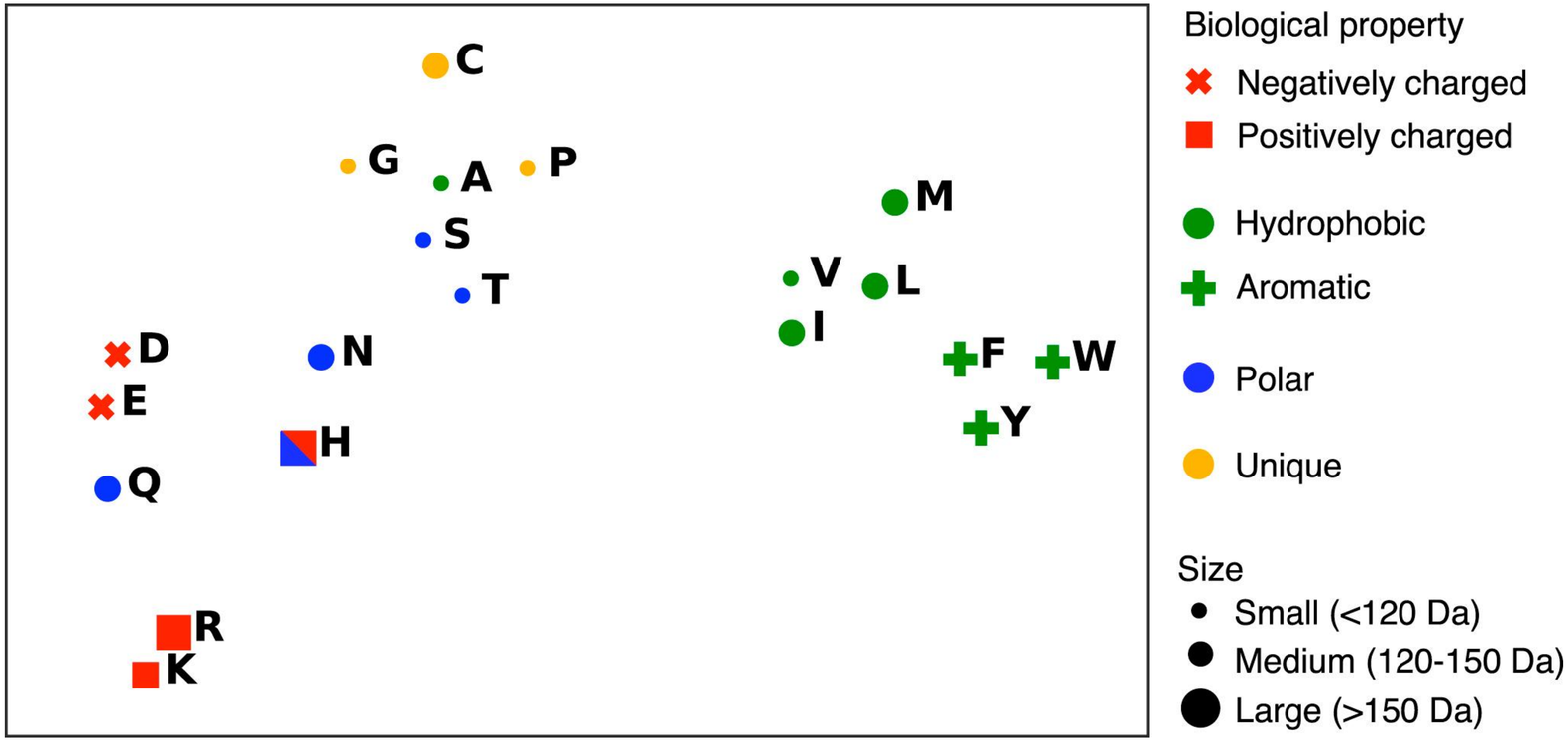
ESM (Evolutionary Scale Modeling):

- Trained on protein sequences:
  - 250M sequences from UniParc
  - Also uses masking techniques
- Evaluated on sequences from UniRef:
  - Low-diversity dataset with UniRef100
  - High-diversity sparse dataset with UniRef50 representative
  - High-diversity dense dataset with UniRef50 clusters
- Tested *in silico* to predict:
  - Physio-chemical residue properties
  - Biological variation
  - Protein homology
  - Secondary and tertiary structure (Lin et al., 2023)
  - Effects of mutations
- Experimental validation: *De novo* design (Verhuil et al., 2022)



(Rives *et al.* 2021; Lin *et al.* 2023)

ESM Performance:

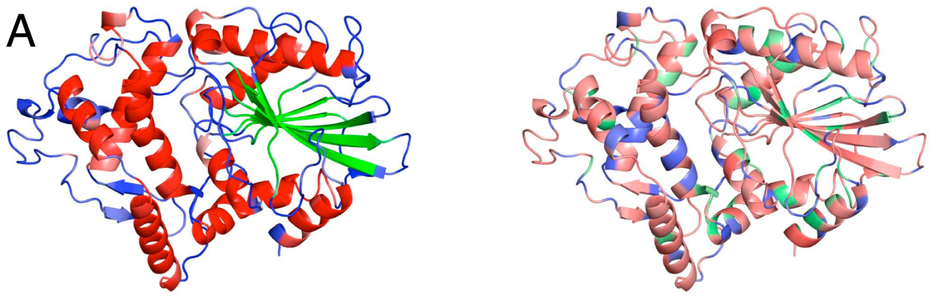


Cluster amino acids by properties

(Rives *et al.* 2021)

ESM Performance:

A

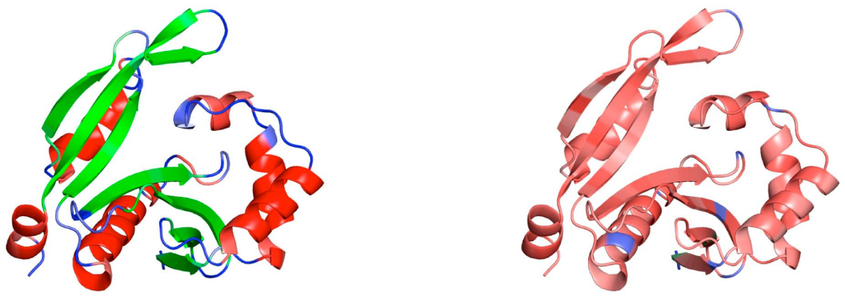


With pre-training
8-class Acc: 70.6%

No pre-training
8-Class Acc: 36.6%

d1nt4a_ (Phosphoglycerate mutase-like fold)

B



With pre-training
8-class Acc: 82.4%

No pre-training
8-class Acc: 32.4%

d3wr7a_ (Acyl-CoA N-acyltransferases fold)

→ Predict secondary structures

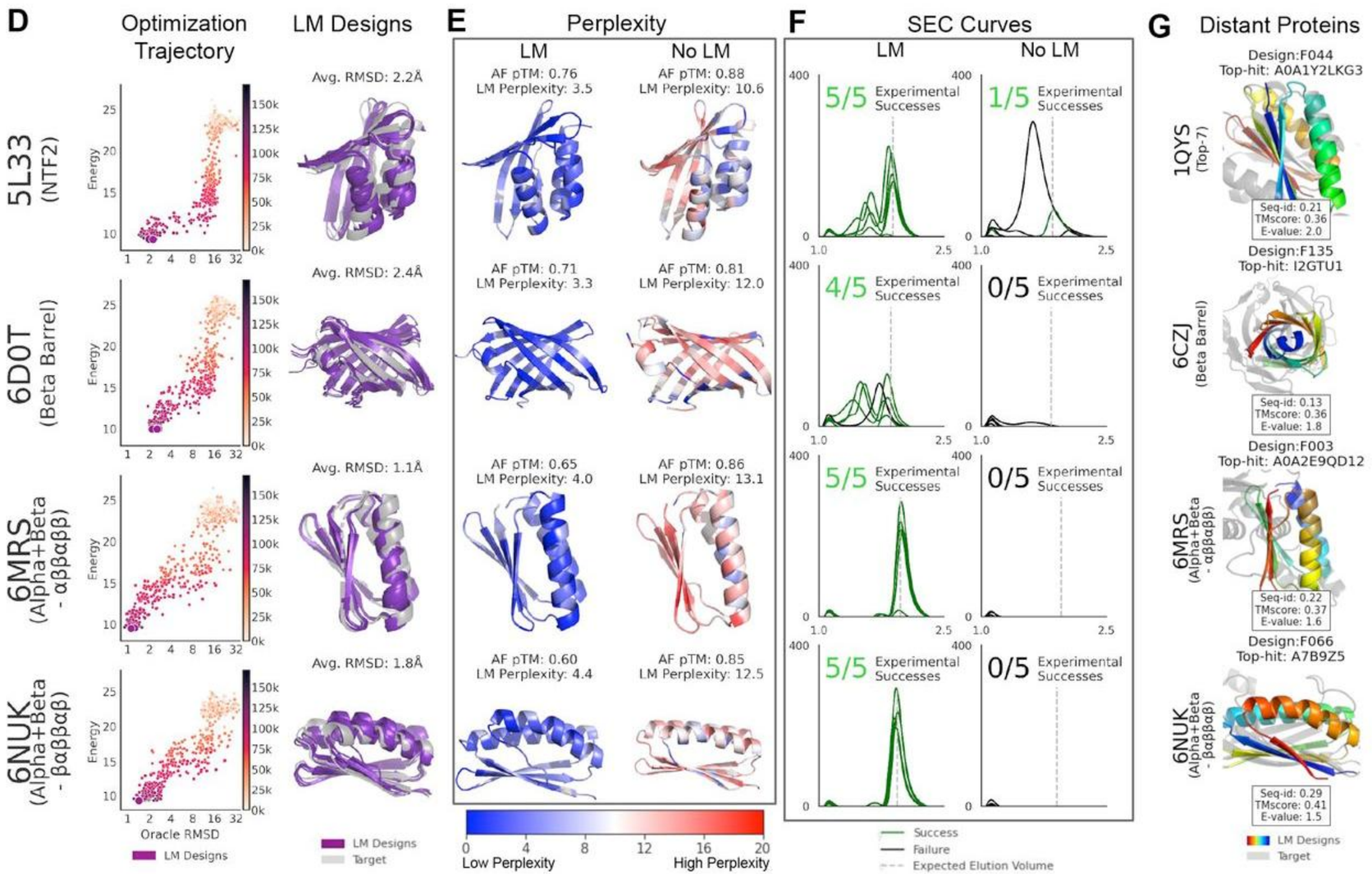
Helices

Strands

Loops

(Rives *et al.* 2021)

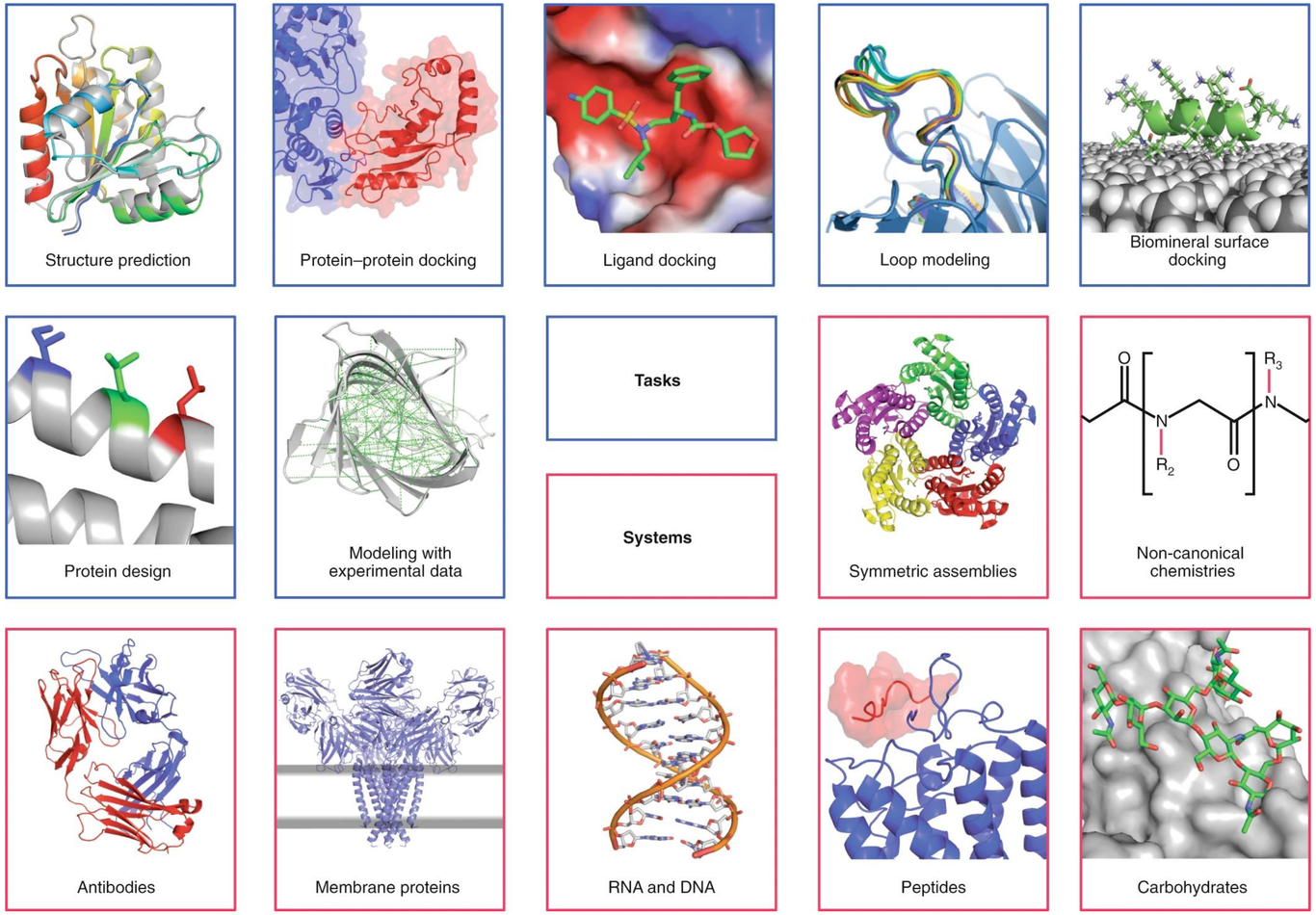
ESM Performance:



(Verkuil *et al.* 2022)

ML in Rosetta Examples

Why integrate ML design methods in Rosetta?



(Koehler-Lehman *et al.* 2020)

Why integrate ML design methods in Rosetta?

+ Feature calculation is fast in C++
+ No knowledge of Python needed for RosettaScripts
+ Makes it easy to combine ML with Rosetta elements
+ No need to reinvent the wheel for sampling, scoring, etc.
+ Provides an established testing framework

ML in Rosetta Design, design tools:

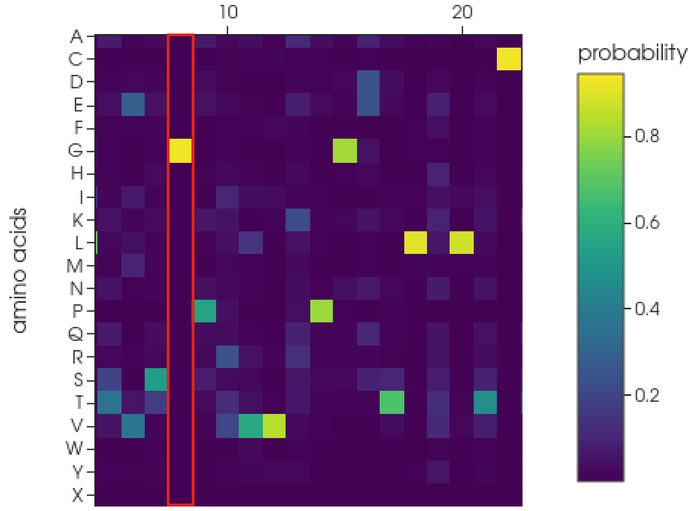
Input

Sequence
and/or structure

Inference

Models
ProteinMPNN
MIF-ST
ESM

Output



Position specific
probability matrix (PSPM)

(Referred to in the tutorial
as “Probabilities”)

ML in Rosetta Design, design tools:

Sampling Mutations in Rosetta

FavorSequenceProfile

RestrictAAsFromProbabilities

SampleSequenceFromProbabilities

Constrain the sampling with info from probabilities

Restrict sampling to aa at least as likely as the current one from probabilities

Sample aa from probabilities

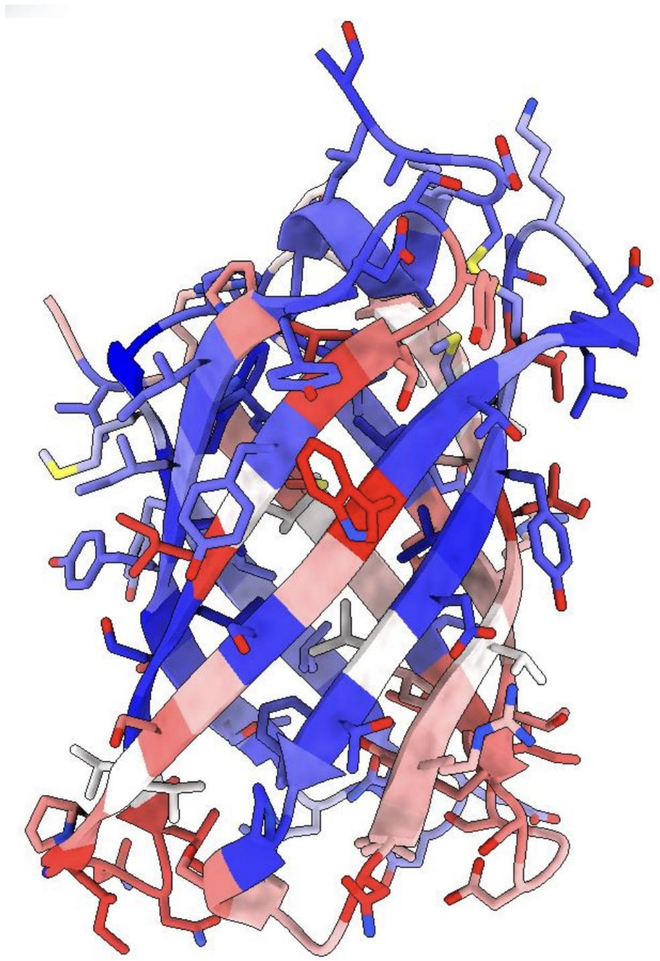
ML in Rosetta Design, design tools:

```
<TASKOPERATIONS>
    <ReadResfile name="rrf" filename="./resfile.resfile"/>
</TASKOPERATIONS>
<SIMPLE_METRICS>
    <PerResidueEsmProbabilitiesMetric name="esm" residue_selector="res"
    model="esm2_t33_650M_UR50D"/>
</SIMPLE_METRICS>
<MOVERS>
    <SampleSequenceFromProbabilities name="sample" metric="esm" pos_temp="0.1"
    aa_temp="0.1" prob_cutoff="0.1" delta_prob_cutoff="0.0" max_mutations="10"
    task_operations="rrf" use_cached_data="true"/>
</MOVERS>
```

- Sample 10 positions: **(max_mutations=”10”)**
- Sample aa with p>0.1: **(prob_cutoff=”0.1”)**
- At least as likely as the current aa: **(delta_prob_cutoff=”0.0”)**

ML in Rosetta Design, analysis tools:

```
<SIMPLE_METRICS>
    <ProteinMPNNProbabilitiesMetric name="prediction"/>
    <CurrentProbabilityMetric name="current" metric="prediction"/>
</SIMPLE_METRICS>
```



The probabilities for the sequence are saved in the b-factor column of the PDB and can be easily visualized with Pymol/Chimera

ML in Rosetta Design, analysis tools:

Analysis in Rosetta

CurrentProbabilityMetric

AverageProbabilitiesMetric

ProbabilityConservationMetric

BestMutationsFromProbabilitiesMetric

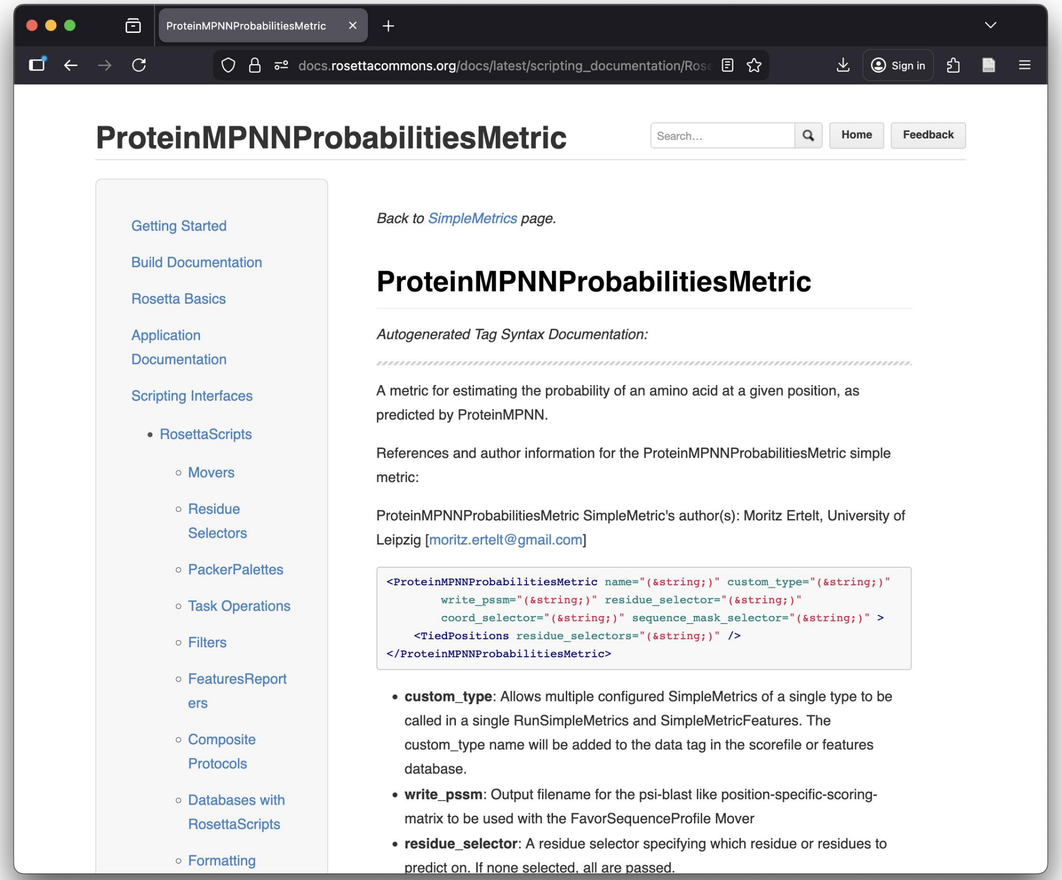
Returns the probabilities for the sequence in the pose

Average probabilities.(i.e. from ProteinMPNN and ESM)

Calculate conservation for each position from probabilities. Ranges from 0 (no conservation) to 1 (fully conserved)

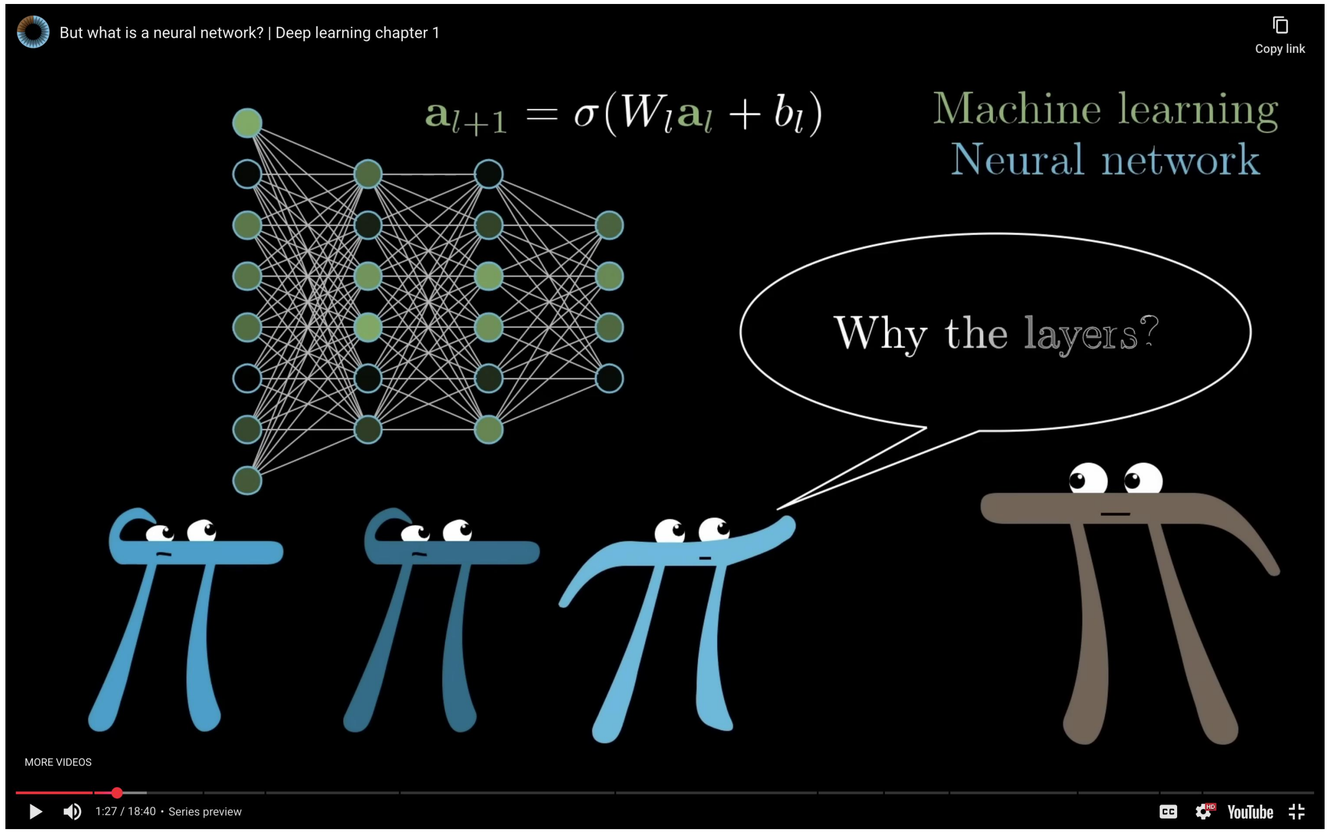
Return the most likely mutation(s) for a given position

ML in Rosetta Design, design and analysis tools:



See RosettaCommons documentation for more info on each metric.

For more on ML (particularly neural networks and deep learning):



3blue1brown has a series on Youtube and their site:
https://www.3blue1brown.com/topics/neural-networks

The tutorial: ~/rosetta_workshop/tutorials/ml_in_rosetta

Monomer

Input preparation:

- Download the PDBs
- Clean the PDBs
- Repack the structure

Calculate probabilities:

- ProteinMPNN, MIF-ST, ESM
- Get current probability
- Get best mutations

Design

- Use probabilities to guide design
- Use probabilities to guide scoring

Dimer

- Design interfaces

The tutorial: ~/rosetta_workshop/tutorials/ml_in_rosetta

Monomer

Input preparation:

- Download the PDBs
- Clean the PDBs
- Repack the structure

Calculate probabilities:

- ProteinMPNN, MIF-ST, ESM
- Get current probability
- Get best mutations

Design

- Use probabilities to guide design
- Use probabilities to guide scoring

Dimer

- Design interfaces

NOTE

For longer steps (e.g. MIF-ST prediction), copy the provided output files from /output_files and proceed with the next step

Bibliography:

- Yang, K. K., Zanichelli, N. & Yeh, H. **Masked inverse folding with sequence transfer for protein representation learning. Protein Engineering, Design and Selection 36,** gzad015 (2023).
- Lin, Z. et al. **Evolutionary-scale prediction of atomic-level protein structure with a language model.** Science 379, 1123–1130 (2023).
- Hie, B. L. et al. **Efficient evolution of human antibodies from general protein language models.** Nat Biotechnol 1–9 (2023)
- Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. **DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking.** (2023).
- Verkuil, R. et al. **Language models generalize beyond natural proteins.** 2022.12.21.521521 (2022).
- Dauparas, J. et al. **Robust deep learning based protein sequence design using ProteinMPNN.** 2022.06.03.494563 (2022).
- Rives, A. et al. **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences.** Proceedings of the National Academy of Sciences 118, e2016239118 (2021).
- Rao, R. M. et al. **MSA Transformer.** in Proceedings of the 38th International Conference on Machine Learning 8844–8856 (PMLR, 2021).
- Jumper, J. et al. **Highly accurate protein structure prediction with AlphaFold.** Nature 1–11 (2021) doi:10.1038/s41586-021-03819-2.
- Sculley, D. et al. **Machine Learning: The High-Interest Credit Card of Technical Debt.**