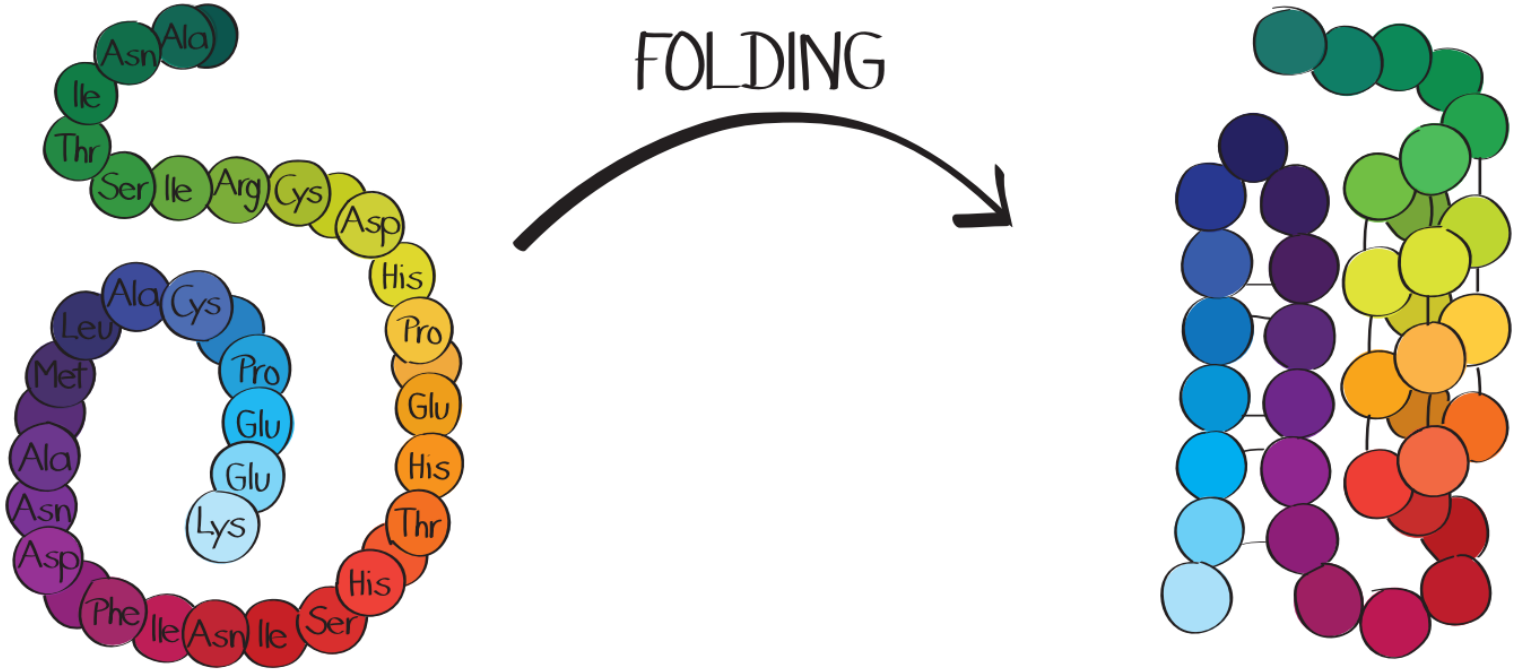
Machine Learning in Rosetta

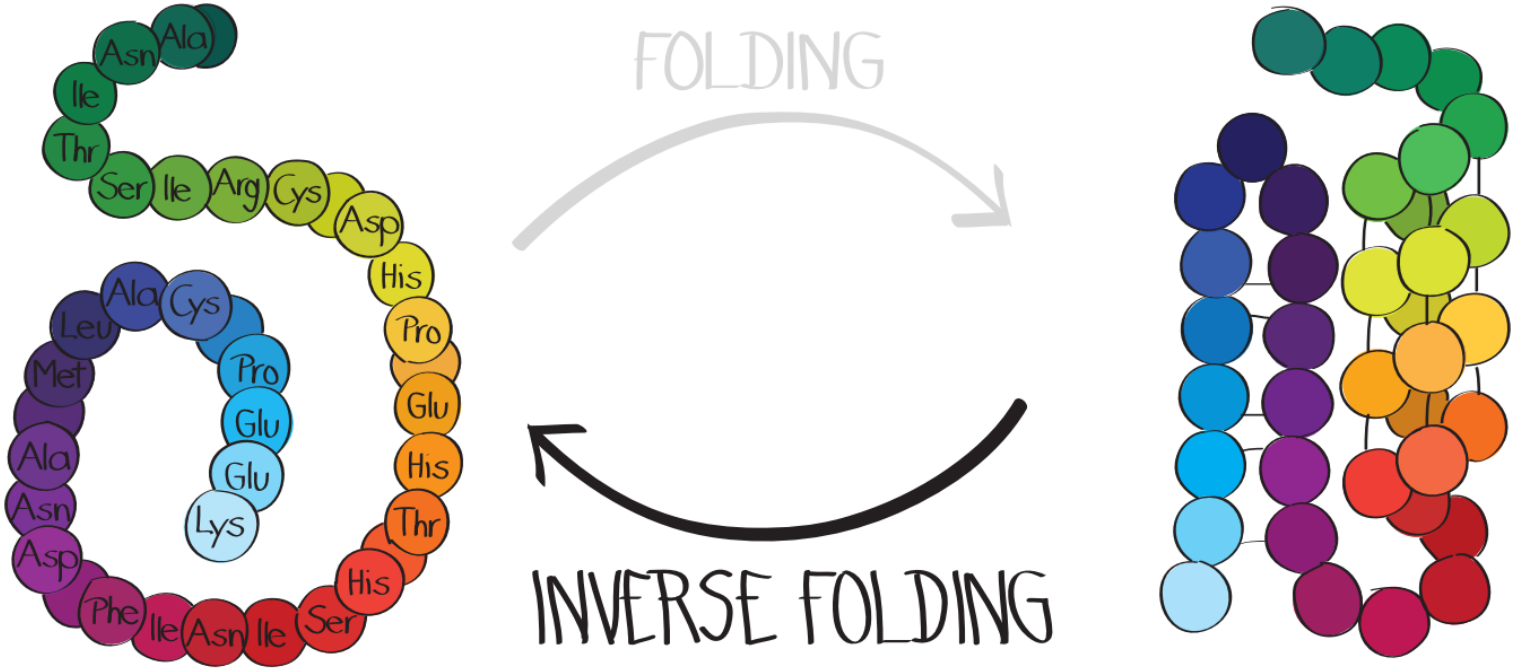
Presented by: Gustavo Araiza
Adapted from: Cristina Elisa Martina
Rosetta Workshop 2025
Meiler Lab

Revolution in Structural Biology:



(Art from Ruth Kellner)

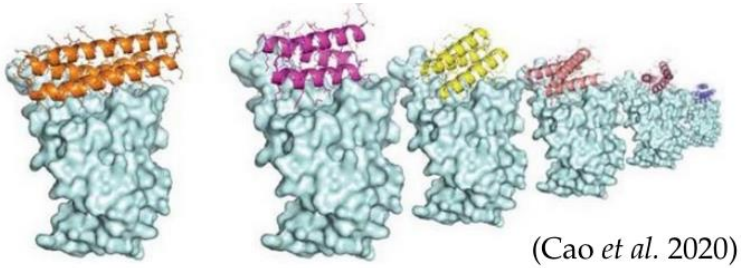
Protein design with ML:



(Art from Ruth Kellner)

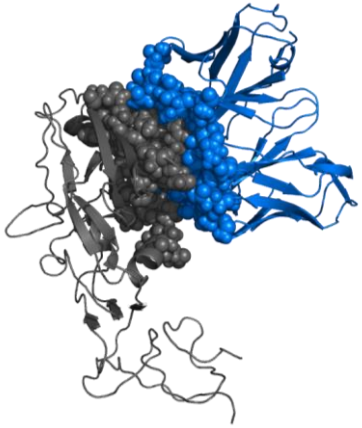
What is the best sequence to:

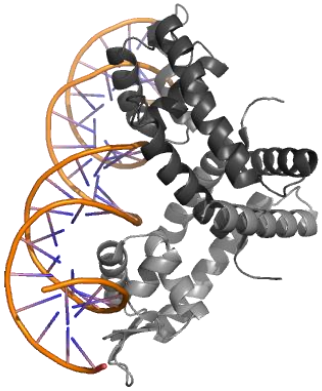
- **fold into this protein scaffold?**
  - new functions
  - new shapes (*de novo* design)



(Cao *et al.* 2020)

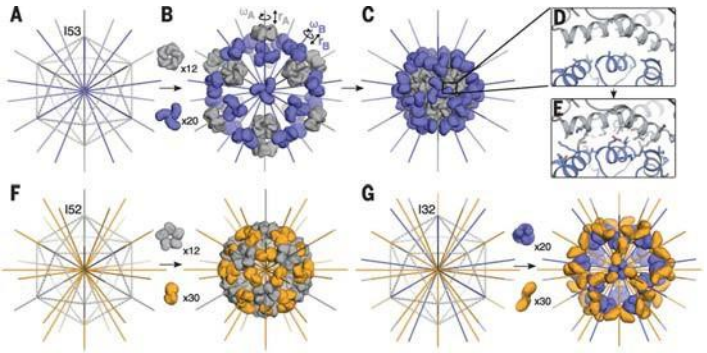
- **increase protein stability?**
  - half-life
  - thermostability
  - crystallizability
  - protein yields





- **increase binding to X?**
  - protein-protein
  - ligand-protein
  - supramolecular assemblies

- **increase enzymatic activity?**
  - activity
  - specificity



(Bale *et al.* 2016)

Computational tools for protein design:

Structure-based methods (*e.g.* Rosetta):

- Starting structure (experimental or model)
- Sampling component
- Scoring component

Machine Learning methods (*e.g.* ProteinMPNN):

- Large dataset for training
- Starting sequences, structures or both
- Very fast
- More accurate

General info on ML:

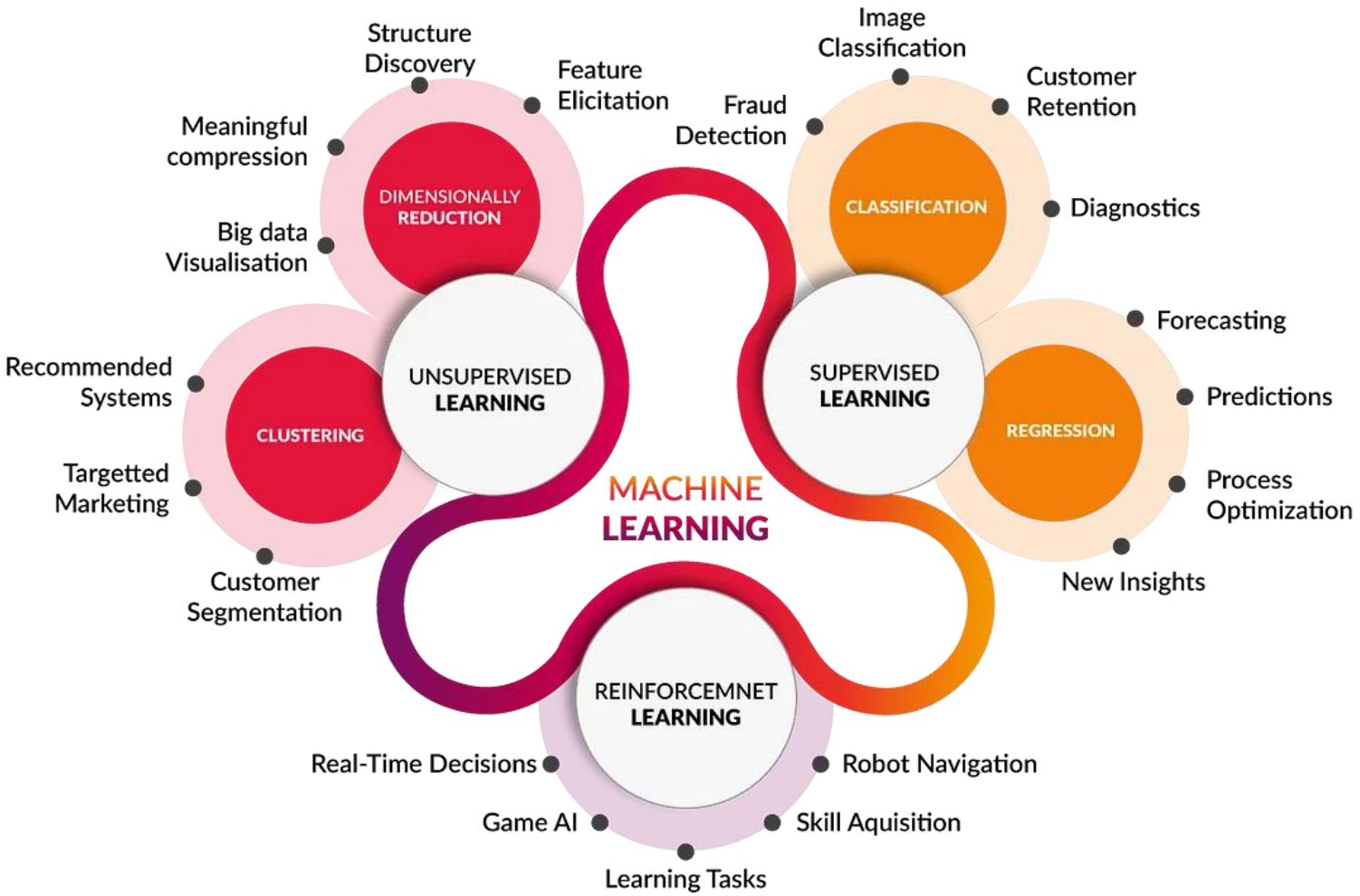
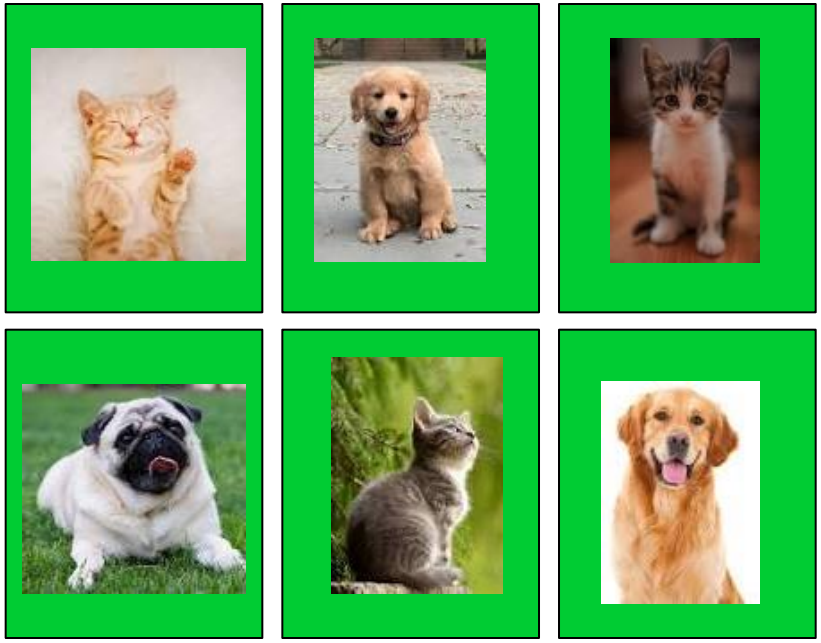


Image source: http://www.cognub.com/index.php/cognitive-platform/

How machines learn:

Training/validation data-set

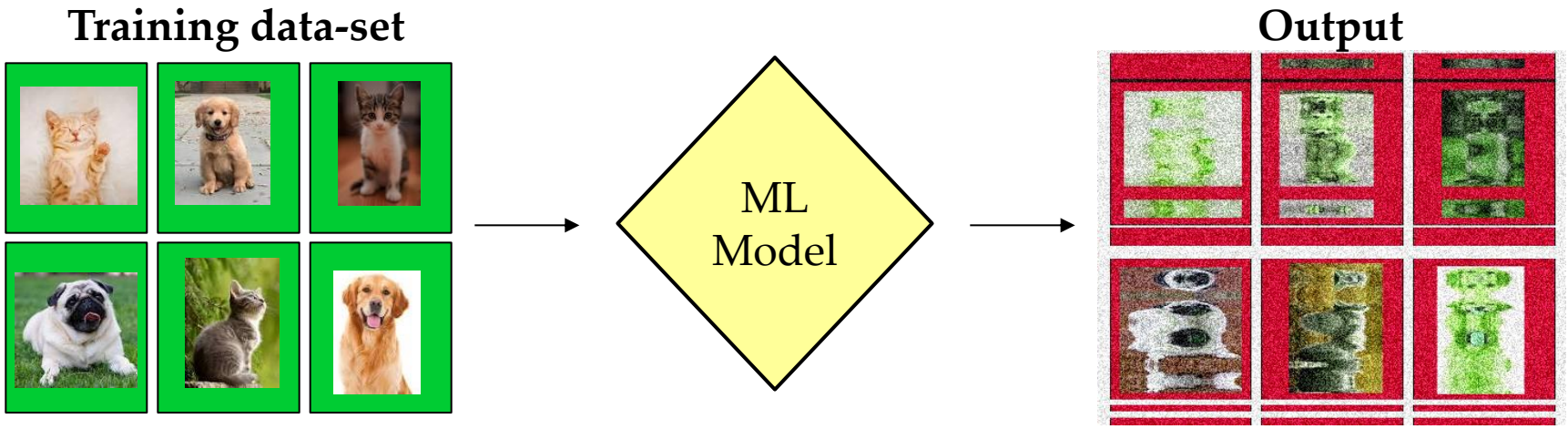


Testing data-set

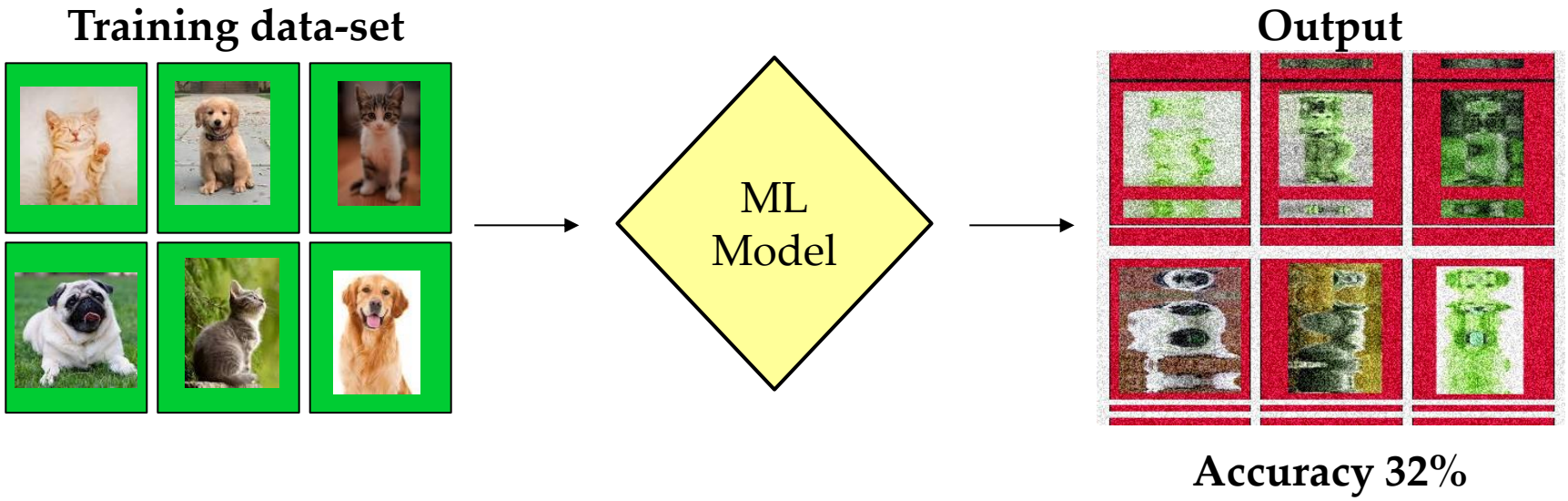
The data-set is divided into three groups:

- Training data-set (80%)
  - Trains the model (learning)
- Validation (10%)
  - Used to benchmark *during* learning
  - Enables 'fine-tuning'
- Testing data-set (10%)
  - used to evaluate the performances with unseen data *after* learning

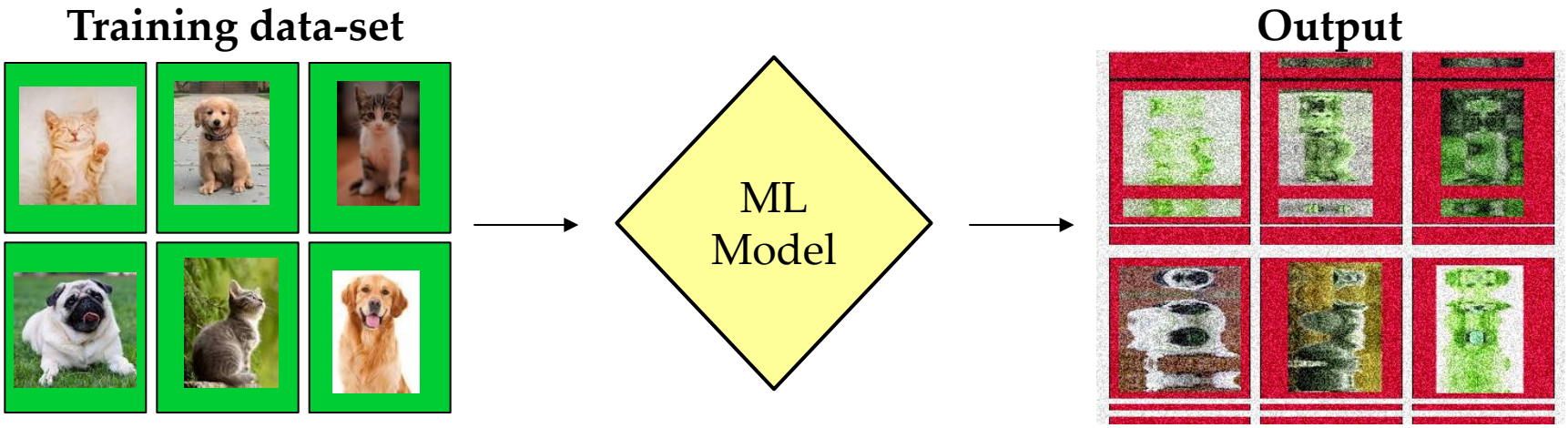
How machines learn:



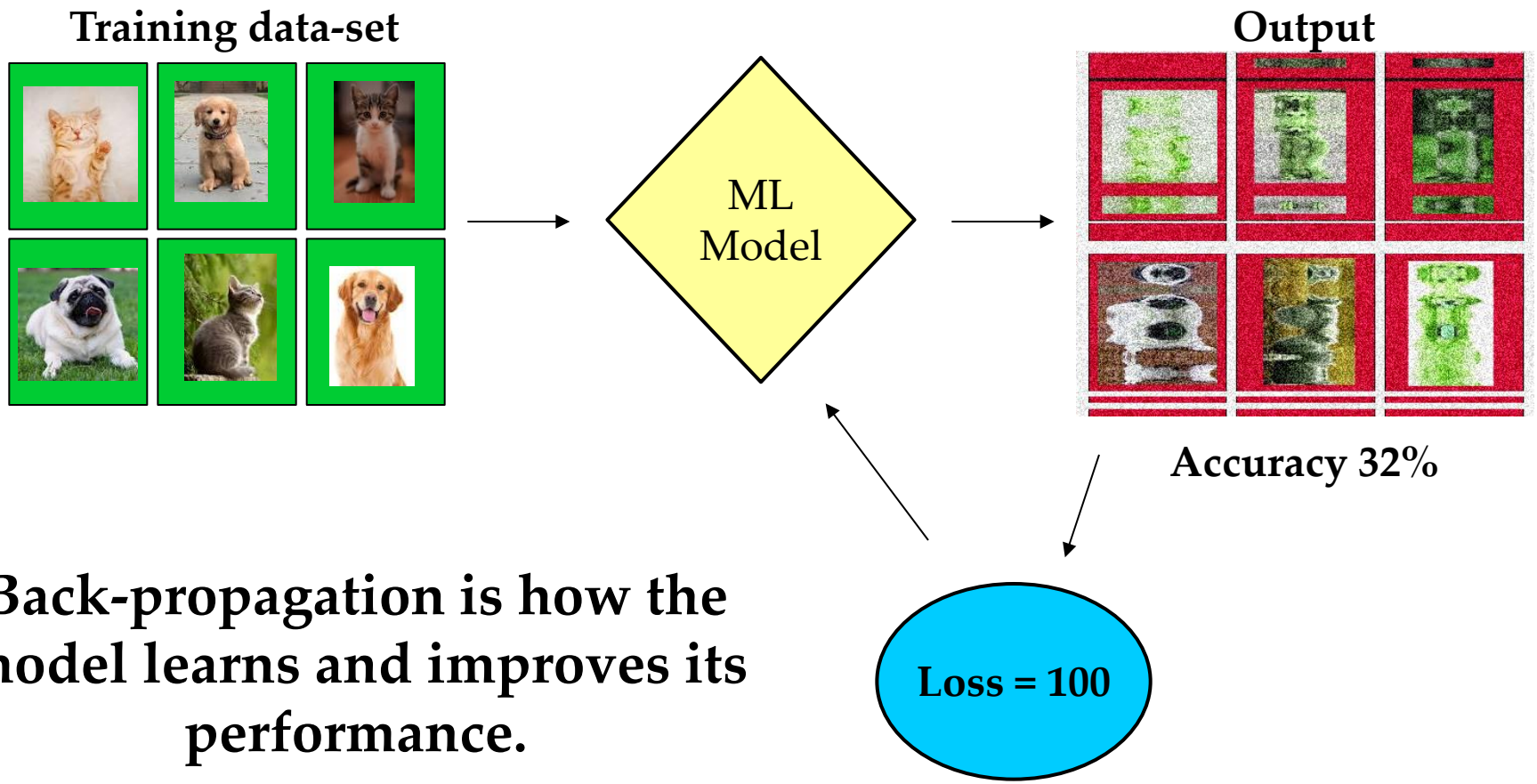
How machines learn:



How machines learn:

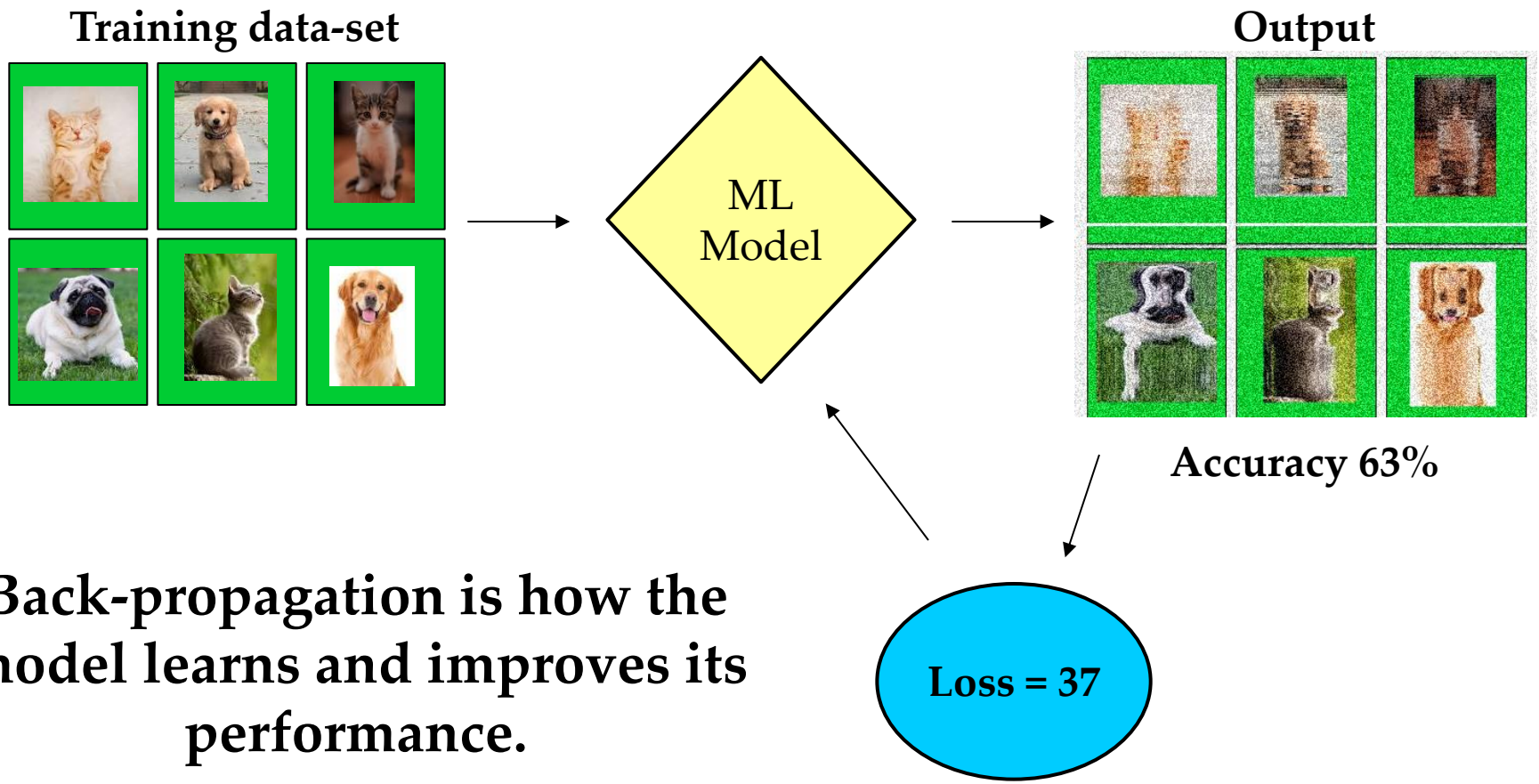


How machines learn:



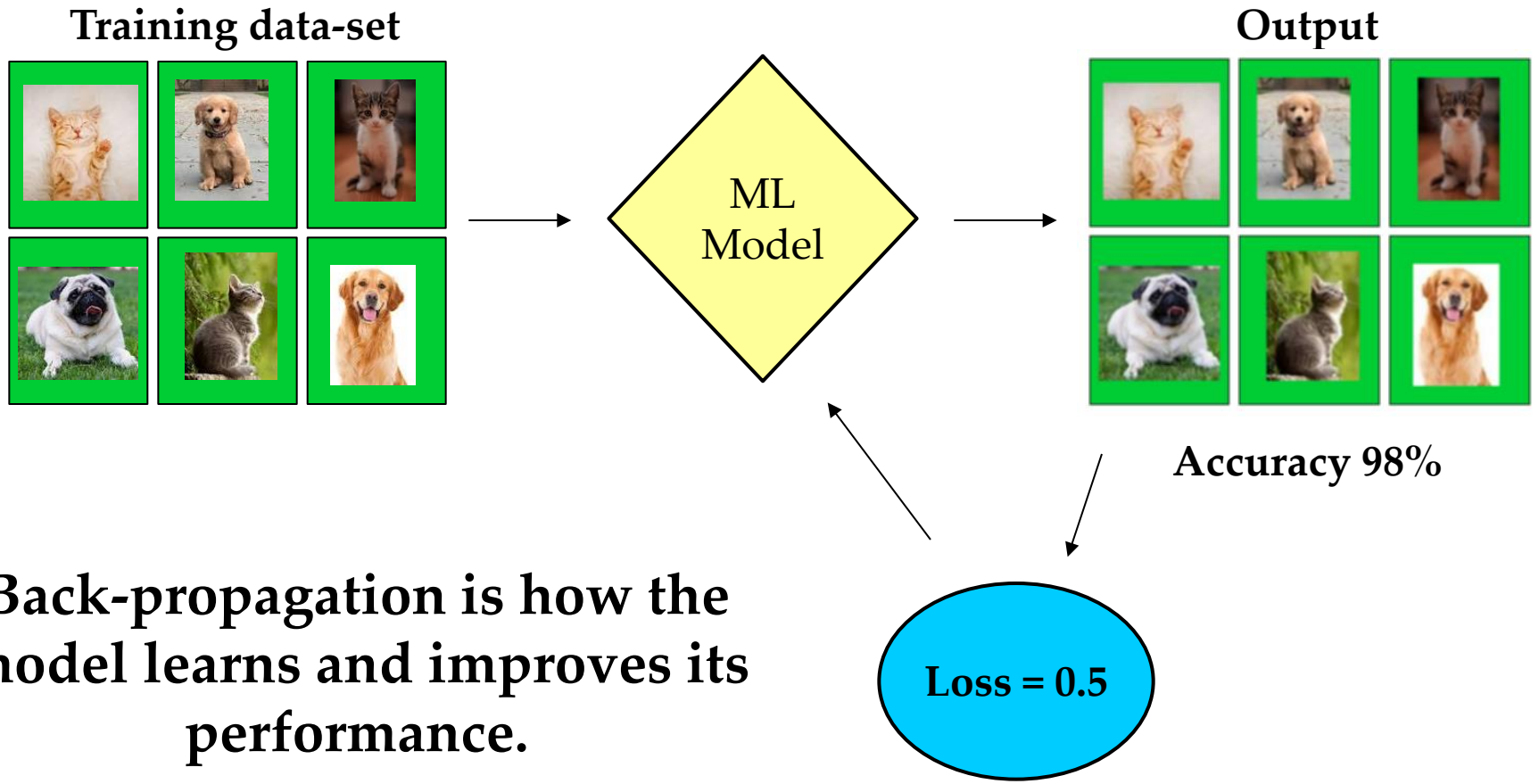
Back-propagation is how the model learns and improves its performance.

How machines learn:



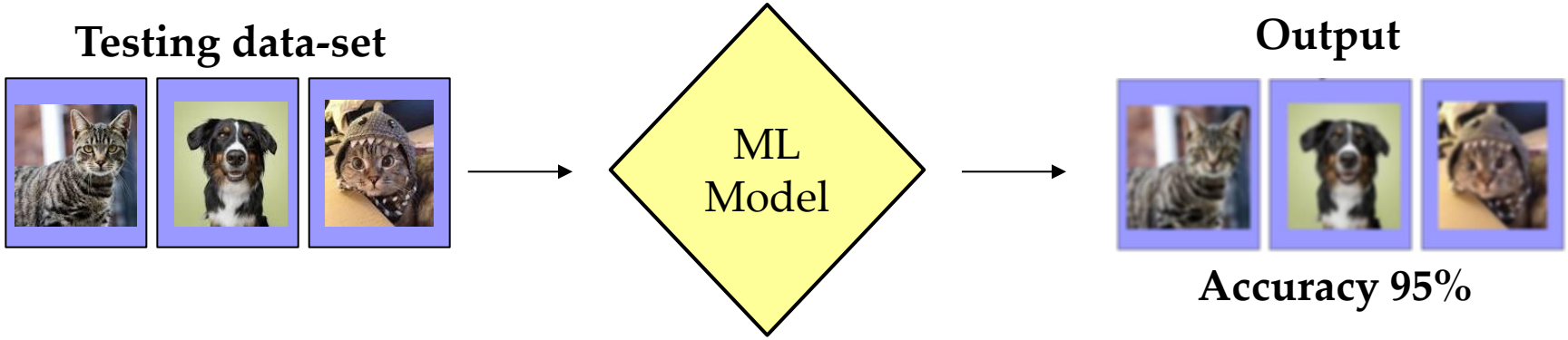
Back-propagation is how the model learns and improves its performance.

How machines learn:



Back-propagation is how the model learns and improves its performance.

How machines learn:



Today's ML methods:

ProteinMPNN – Structure only

- 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).

MIF-ST – Structure and sequence

- 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).

ESM – Sequence only

- 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).

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*:

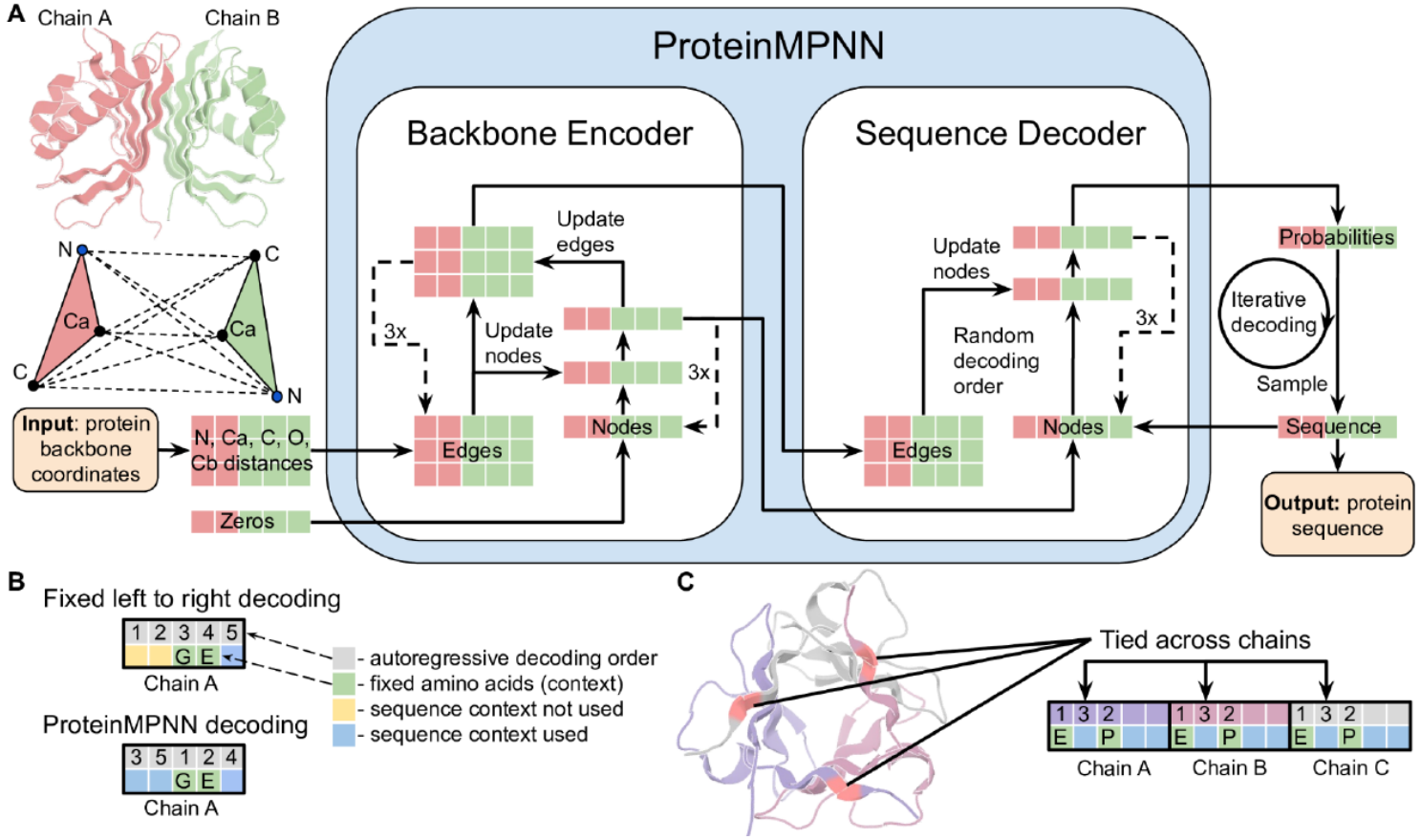
- 690 monomers

- 732 homomers

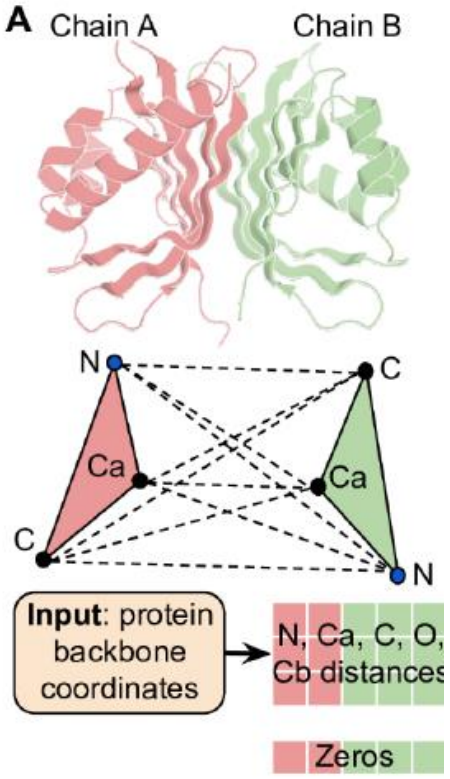
- 98 heteromers

Tested experimentally

ProteinMPNN:



ProteinMPNN, inputs:



RCSB-PDB database

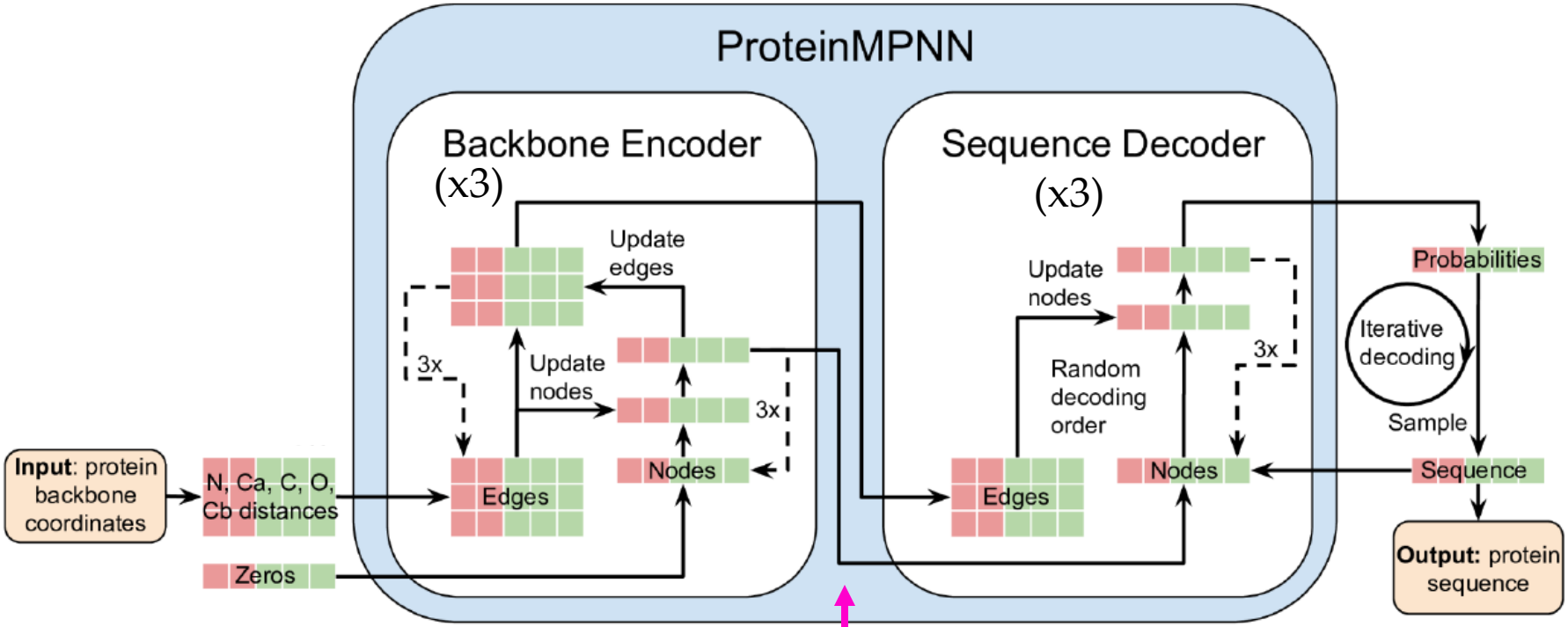
No evolutionary information!

Distances between N, Cα, C, O and virtual Cβ are encoded using graph theory:

- Nodes (atoms)
- Edges (distances)

(Dauparas et al., 2022)

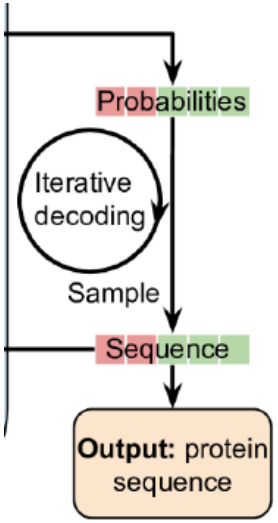
ProteinMPNN, the MPNN:



128 hidden dimensions
(here is where predictions happen)

(Dauparas et al., 2022)

ProteinMPNN, the outputs:

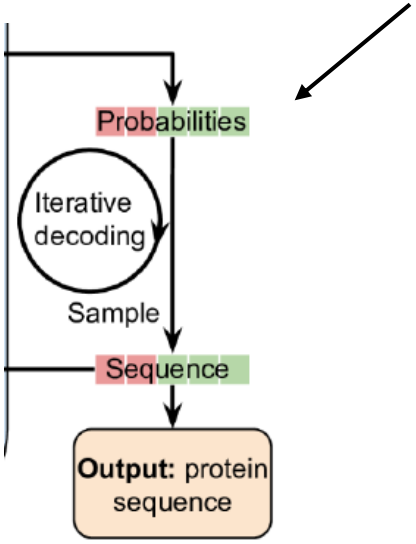


ProteinMPNN outputs re-designed sequences, not structures!

This means that you have predict a designed structure with an alternative method (AF, Rosetta)

(Dauparas et al., 2022)

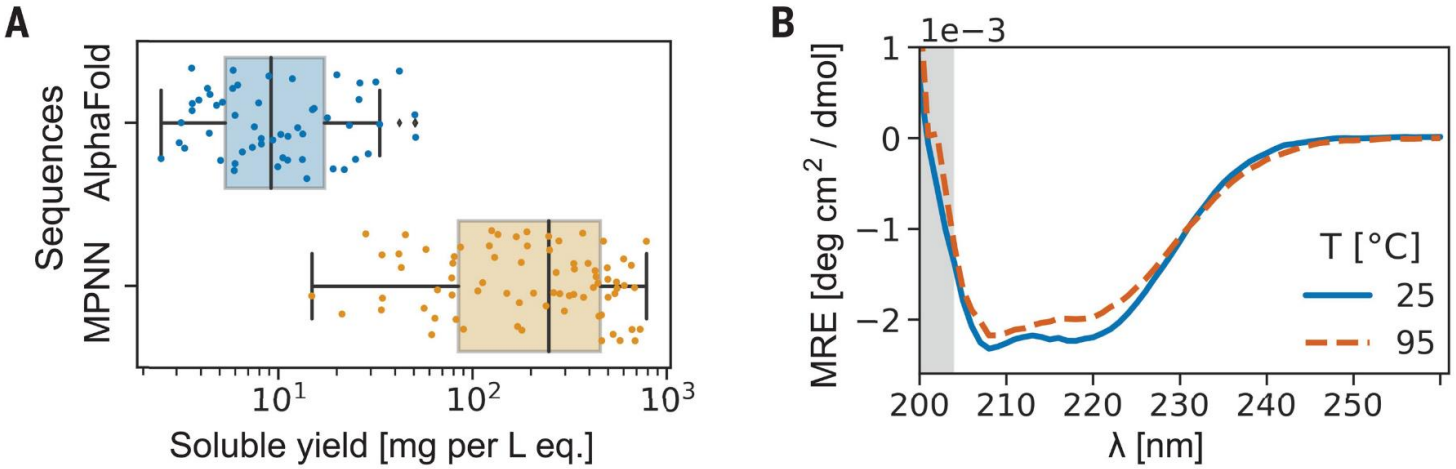
ProteinMPNN, the outputs:



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

(Dauparas et al., 2022)

ProteinMPNN, performance:

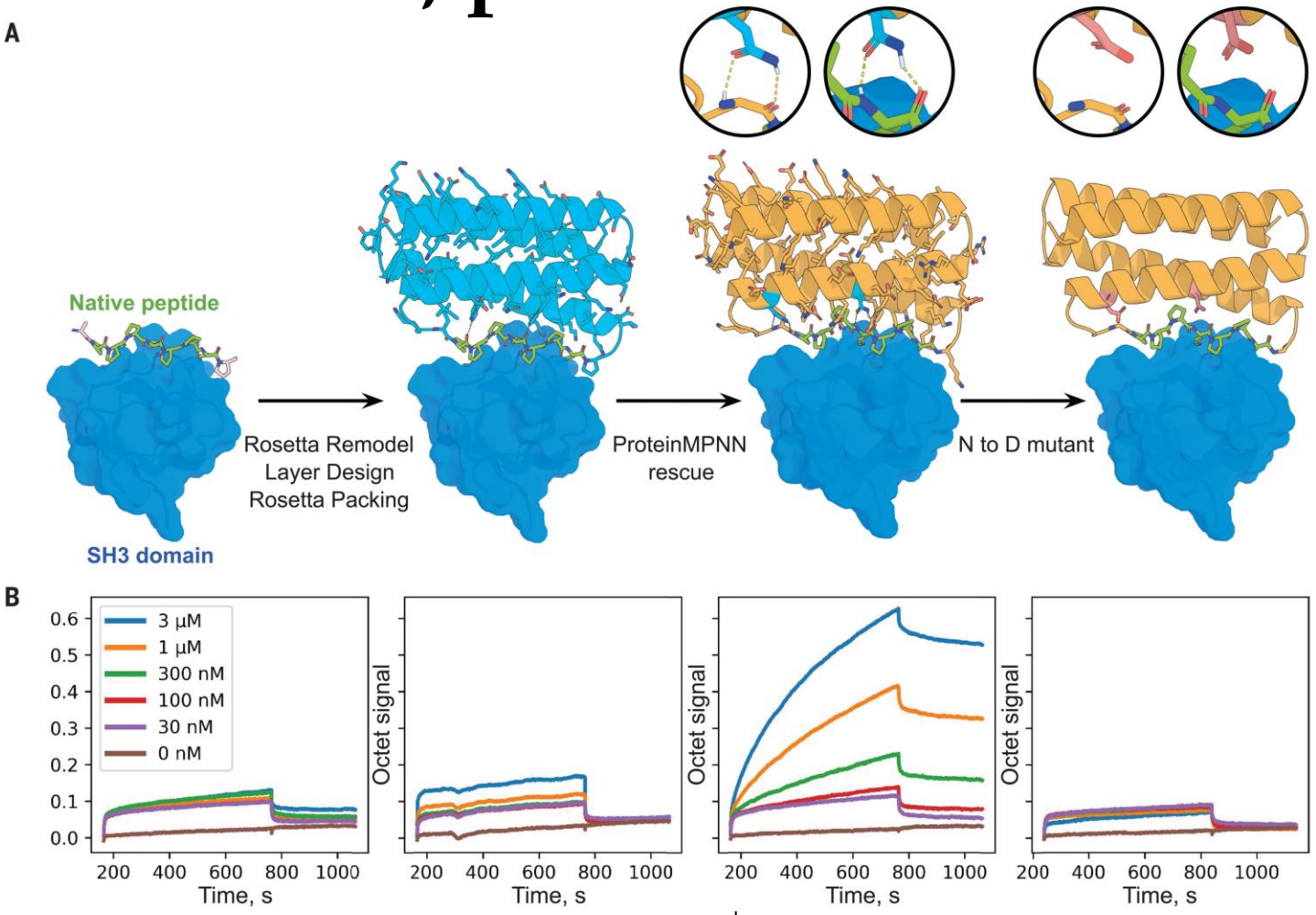


Recovered
solubility

Increased
thermostability

(Dauparas et al., 2022)

ProteinMPNN, performance:



Creates
new functions

(Dauparas et al., 2022)

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

Pre-trained on both protein structures and sequences:

- 19700 protein structures from RCSB-PDB
- 42 M sequences from UniRef50
- sequences are partially masked
- model must predict masked residues

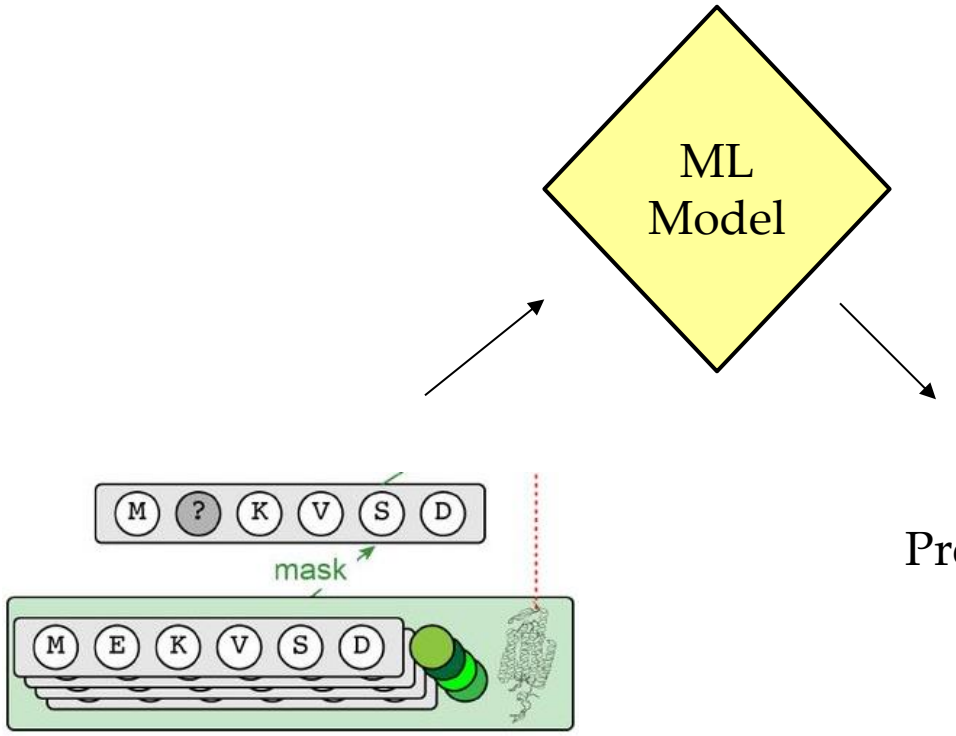
Training for downstream task

- trained on single mutants and predicts multiple mutants
- predict experimental measurements

Tested *in silico* on small and large data-sets:

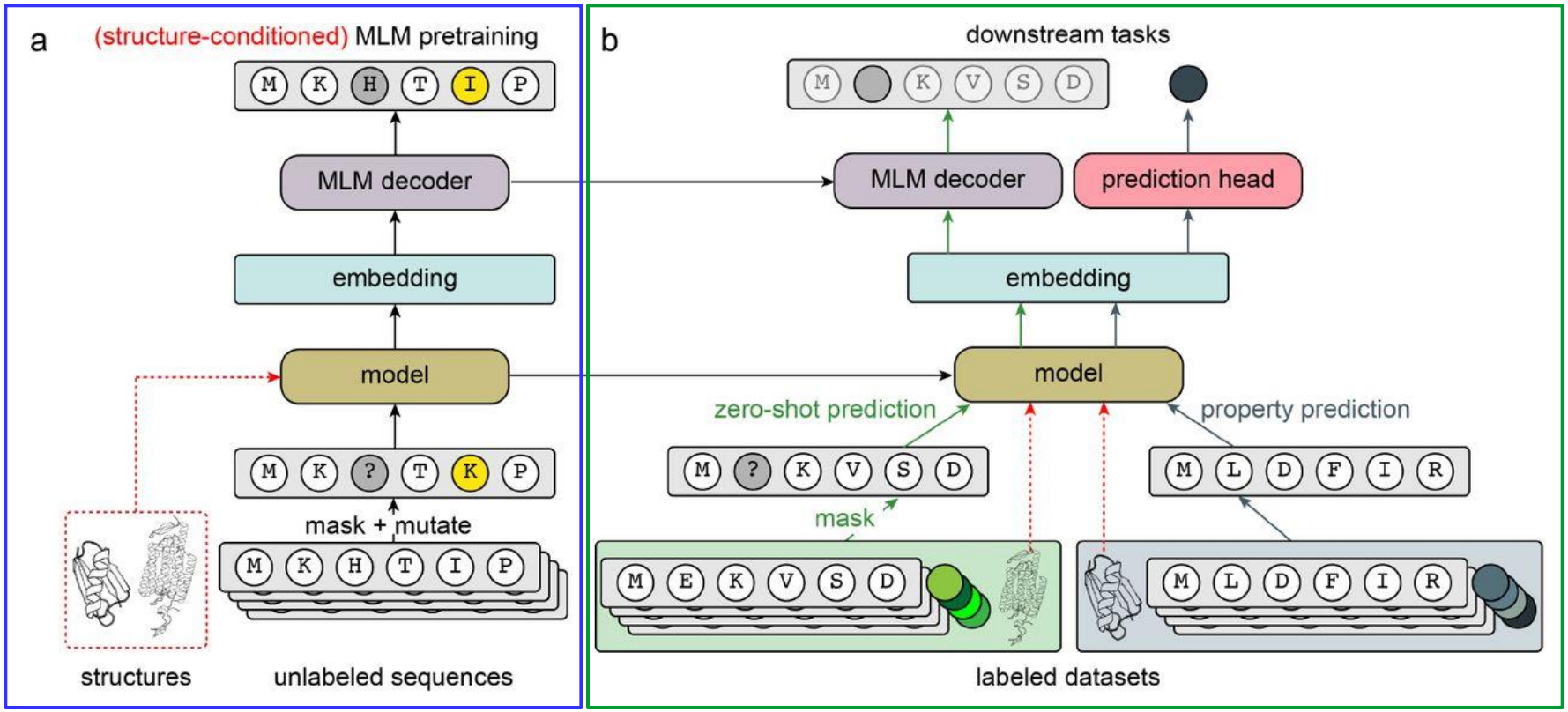
- Deep mutational scans
- Enzymatic activity
- Stability
- Binding

Masking protein sequences in ML:



Prediction: M**E**KVSD

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 better)

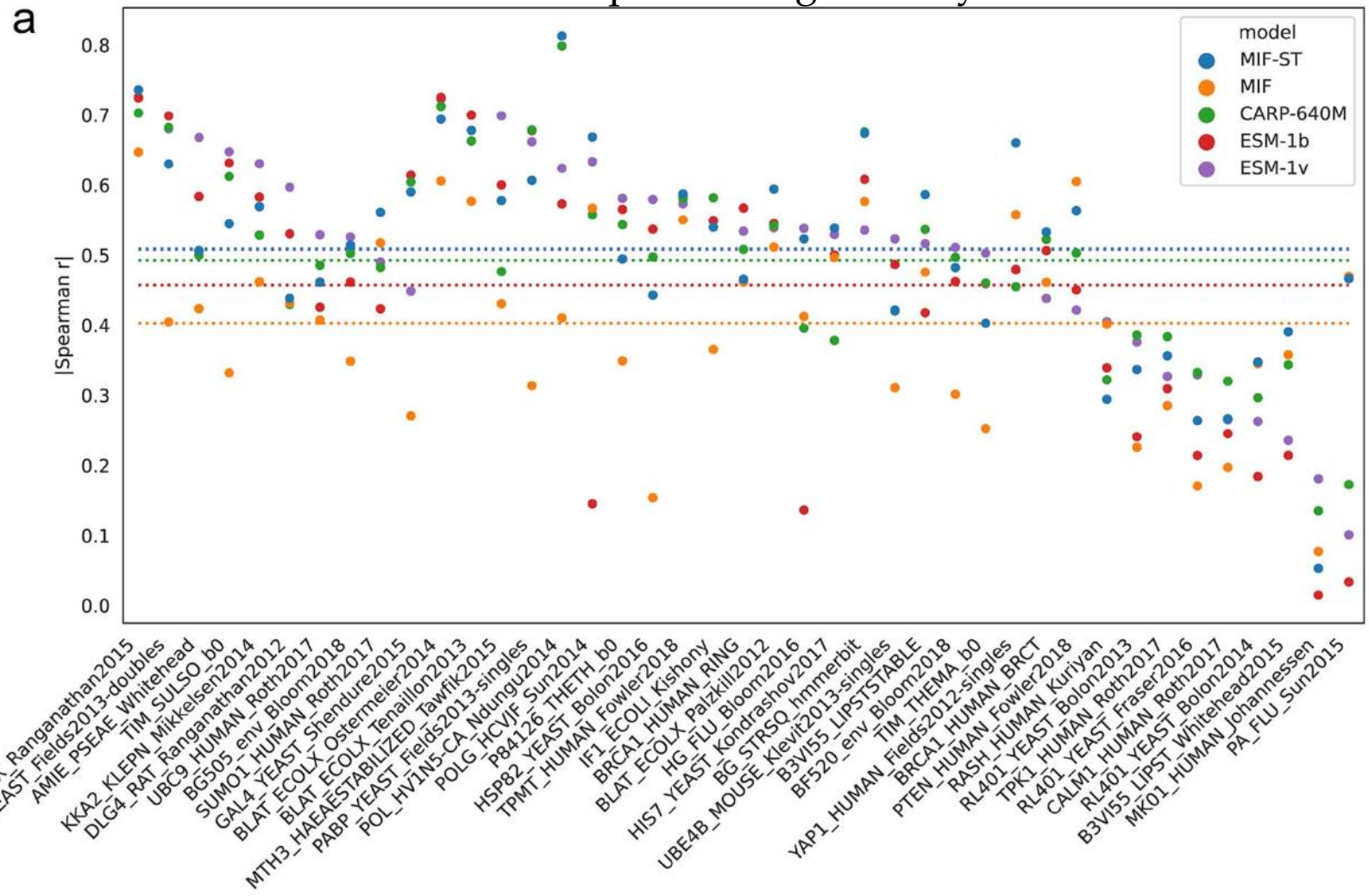
Sequence Recovery:
How well the model recovers native sequences. (higher is better)

(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:

- 250 M sequences from UniParc
- Also uses masking techniques

Evaluated on sequences from UniRef:

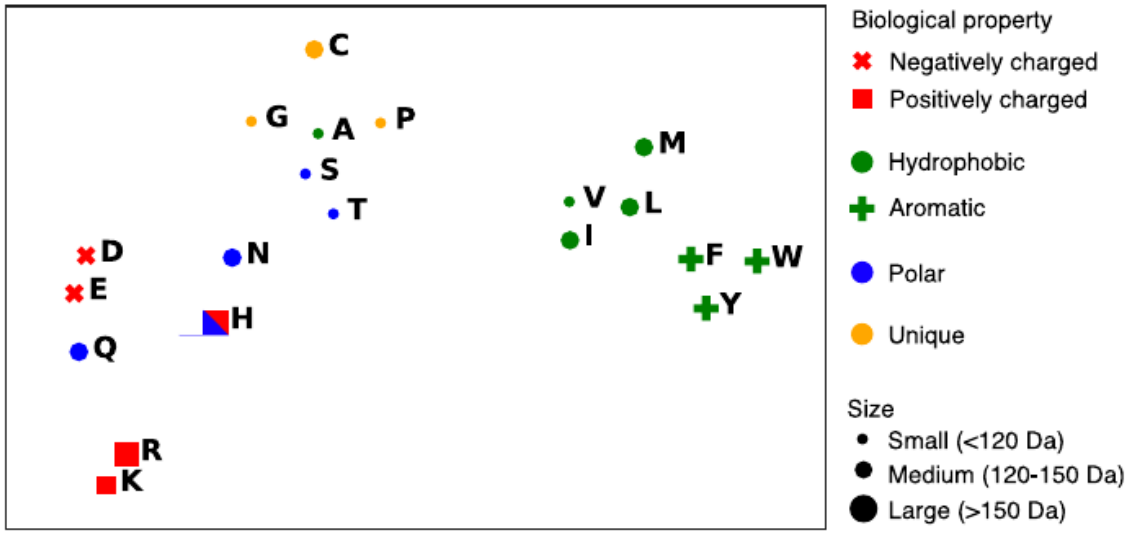
- Low-diversity data-set with UniRef100
- High-diversity sparse data-set with UniRef50 representative
- High-diversity dense data-set with UniRef50 clusters

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 - BioRvix) (Verkuil et al., 2022)

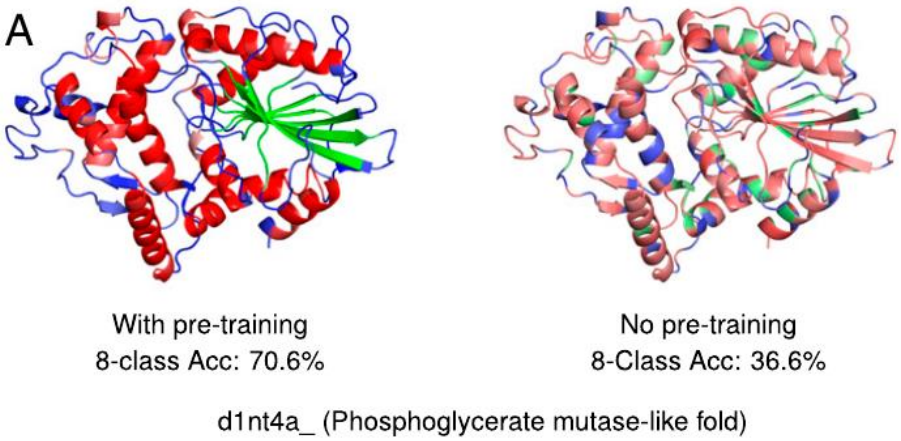
ESM, performance:



Cluster amino acids
by properties

(Rives et al., 2021)

ESM, performance:

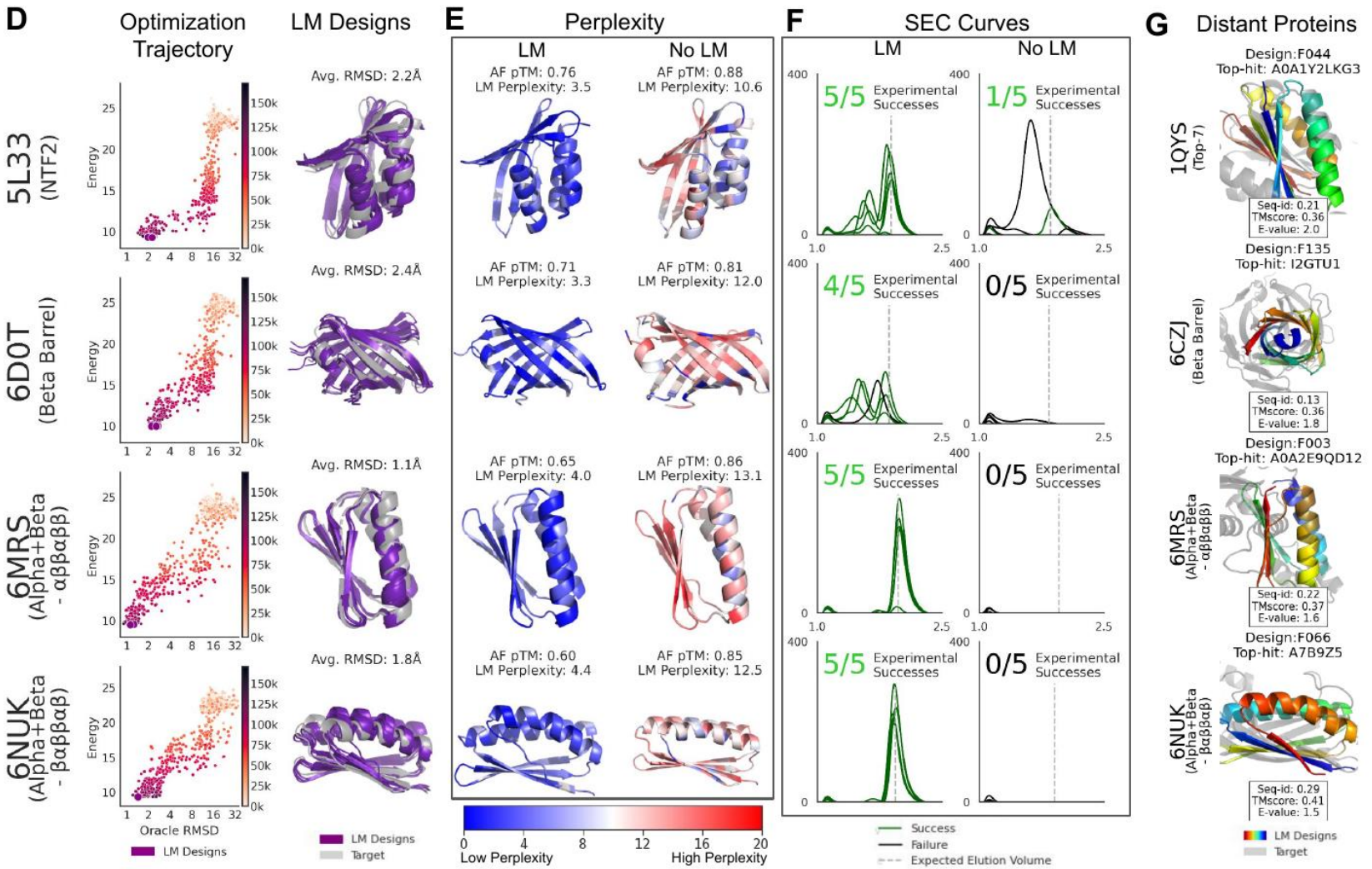


→ Predict secondary structures

Helices
Strands
Loops

(Rives et al., 2021)

ESM, performance:



(Verkuil et al., 2022)

ML in Rosetta:

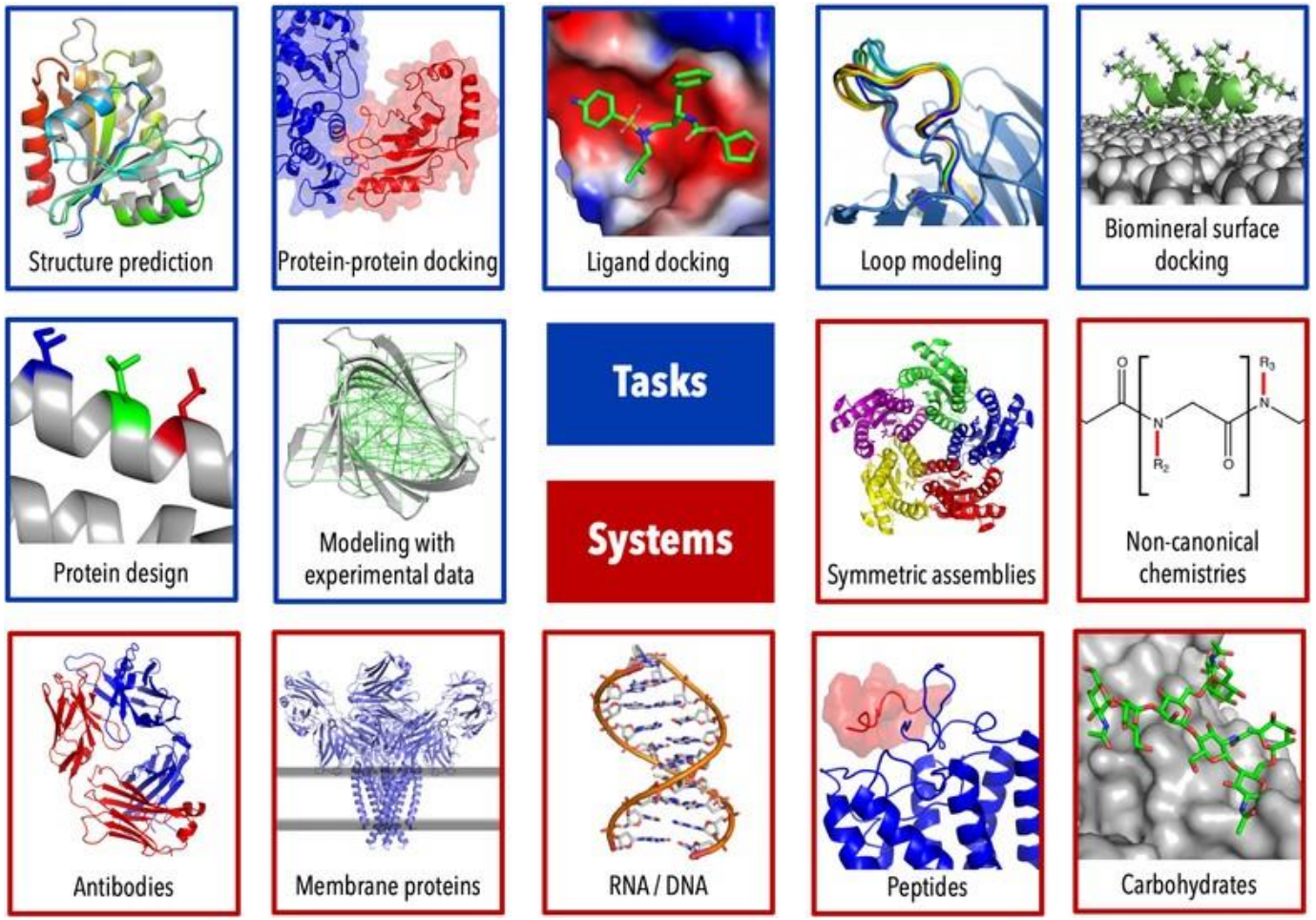
The hero here:

Moritz Ertelt, Ph. D.
Meiler lab alumn from
Leipzig University

Contact:
moritz.ertelt@uni-leipzig.de

ML in Rosetta:

Why integrating protein ML methods in Rosetta?



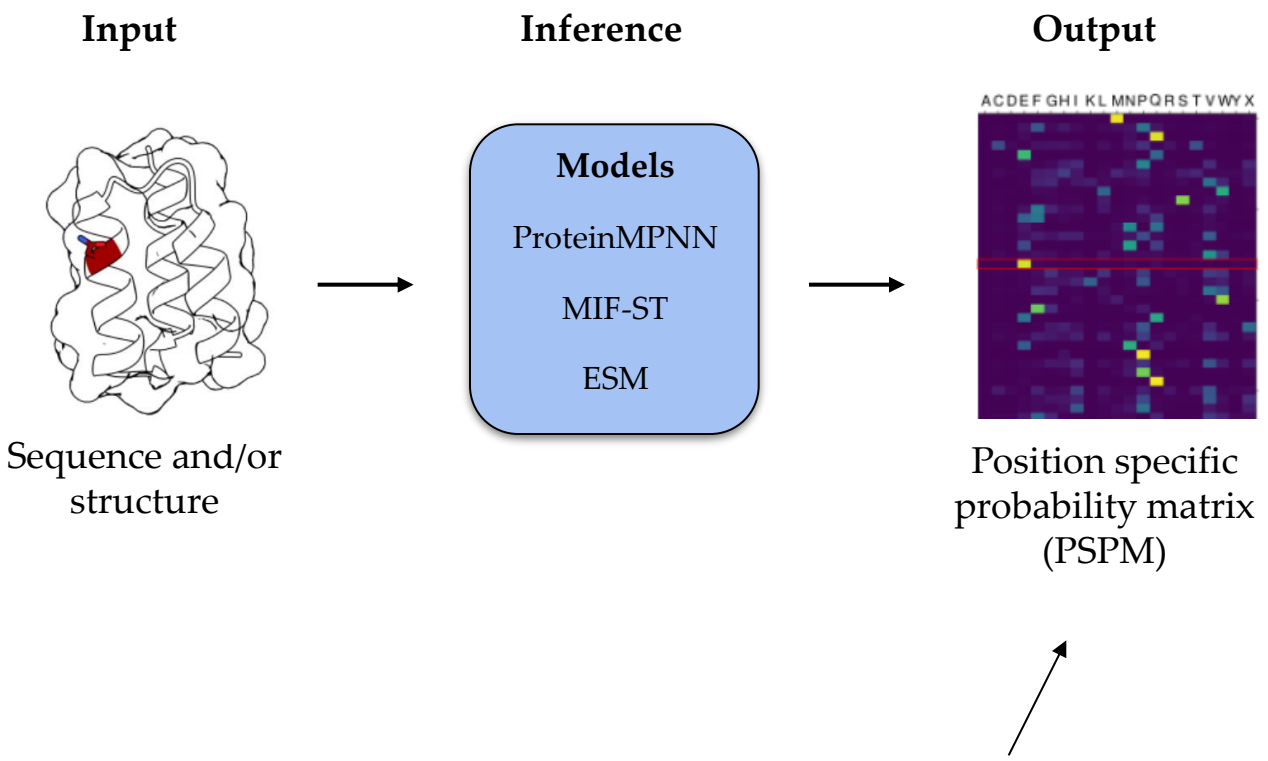
(Koehler-Lehman *et al.* 2020)

ML in Rosetta:

Why integrate ML design methods in Rosetta?

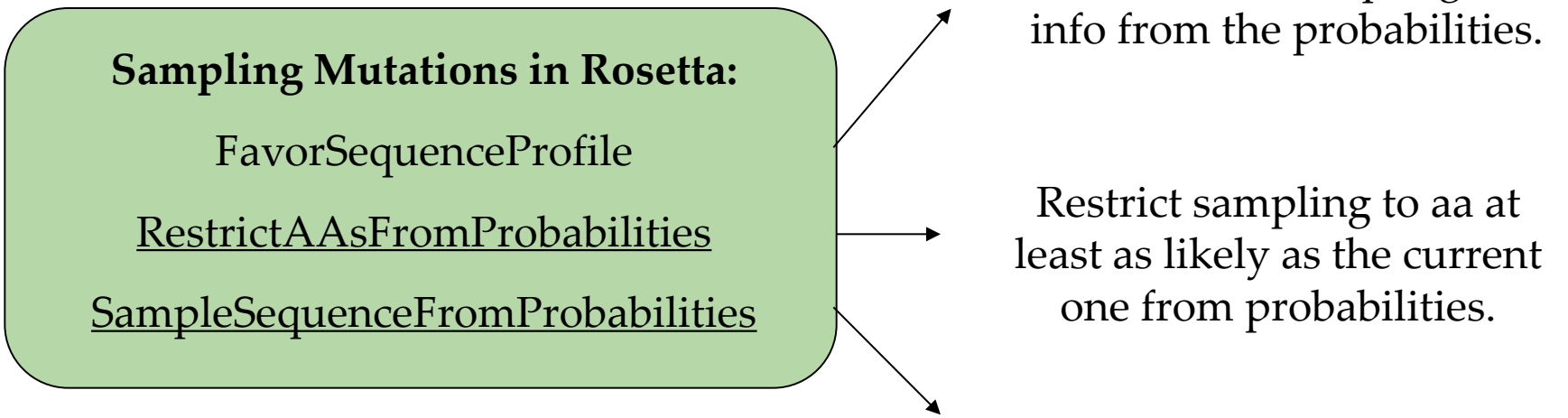
+ Feature calculation is fast in C++
+ No knowledge of Python needed for RosettaScripts
+ Makes it easy to combine ML with Rosetta elements
+ No need to reinvent the wheel for sampling, scoring, etc.
+ Provides an established testing framework

ML in Rosetta Design:



Referred in the tutorial as "Probabilities"

ML in Rosetta Design, design tools:



ML in Rosetta Design, design tools:

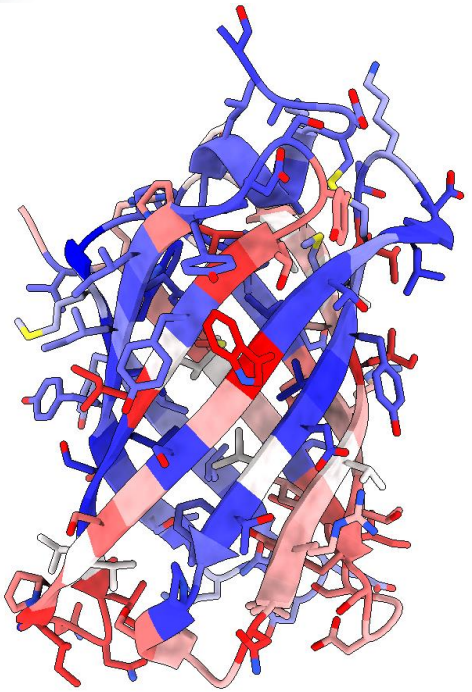
```
<TASKOPERATIONS>
    <ReadResfile name="rrf" filename="./resfile.resfile"/>
</TASKOPERATIONS>
<SIMPLE_METRICS>
    <PerResidueEsmProbabilitiesMetric name="esm" residue_selector="res"
    model="esm2_t33_650M_UR50D"/>
</SIMPLE_METRICS>
<MOVERS>
    <SampleSequenceFromProbabilities name="sample" metric="esm" 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"`)

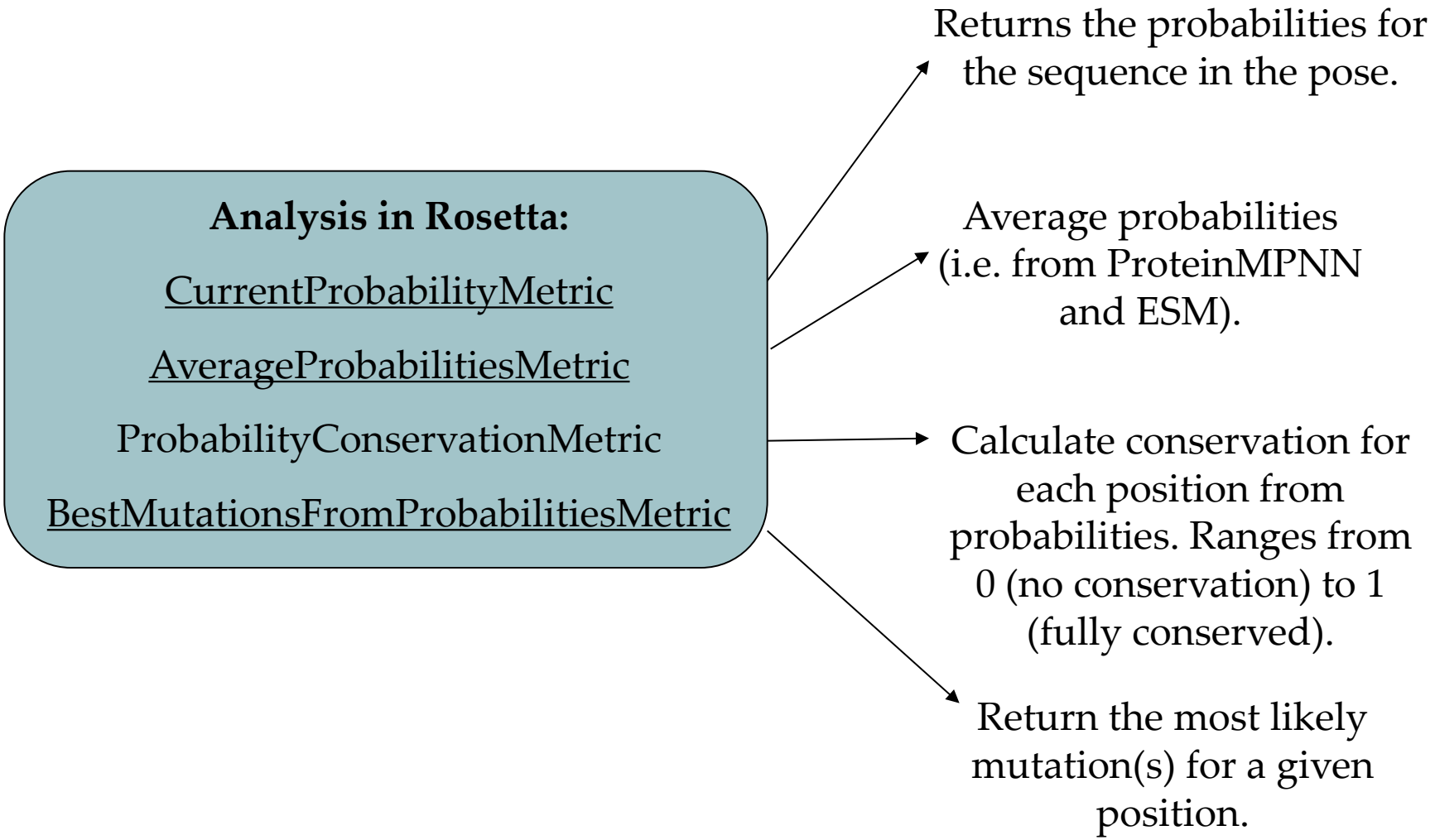
ML in Rosetta Design, analysis tools:

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

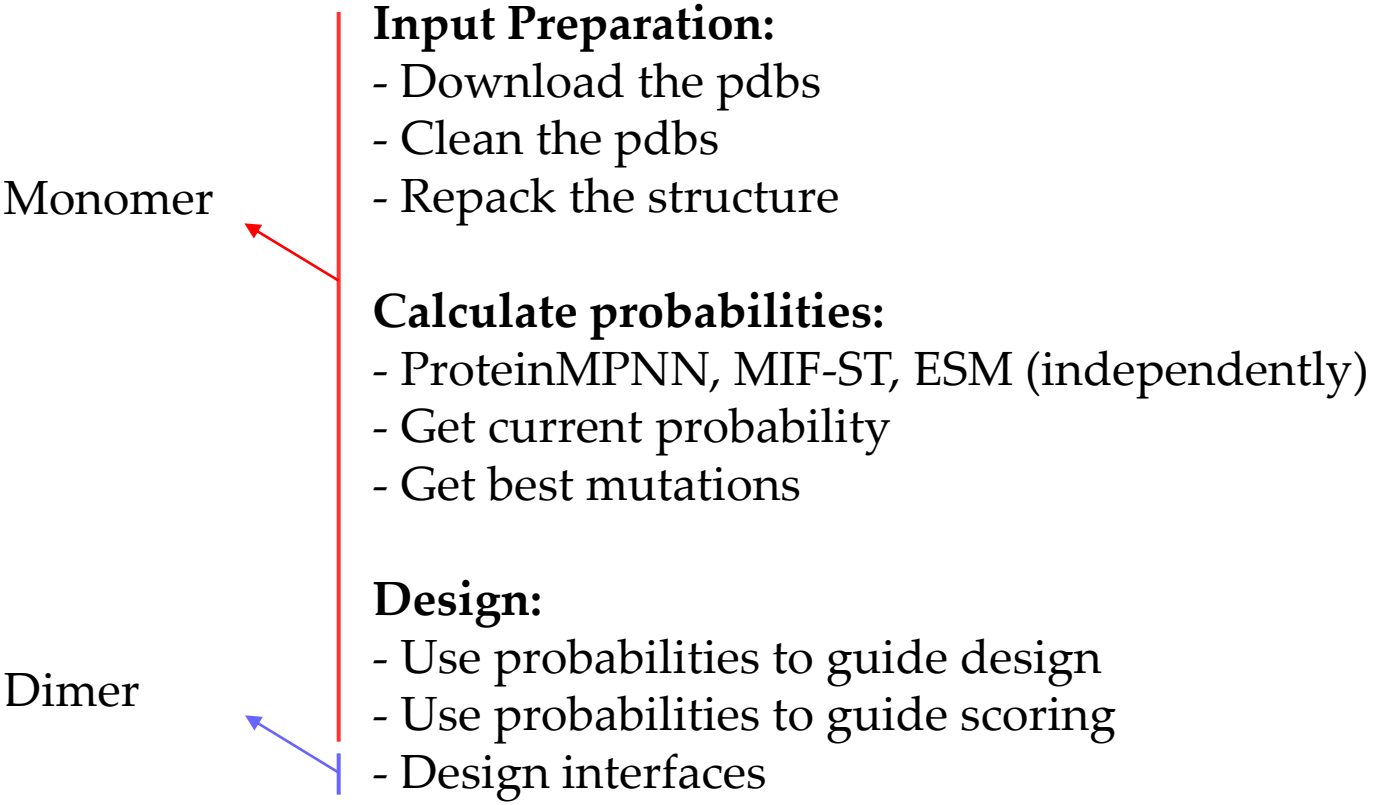
The probabilities for the sequence are saved in the b-factor column of the pdb and can be easily visualized with pymol/chimera.



ML in Rosetta Design, analysis tools:



The tutorial:



Bibliography - ML in Rosetta:

- 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.**