ETH zürich

Applications of AI in Chemistry & Biology

David Graber
david.graber@sam.math.ethz.ch
ETH Zürich

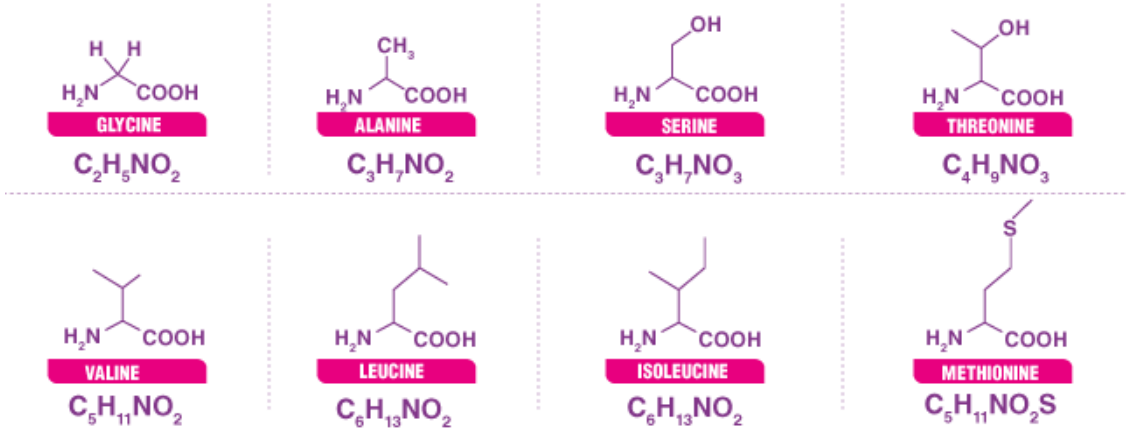
Lecture Outline

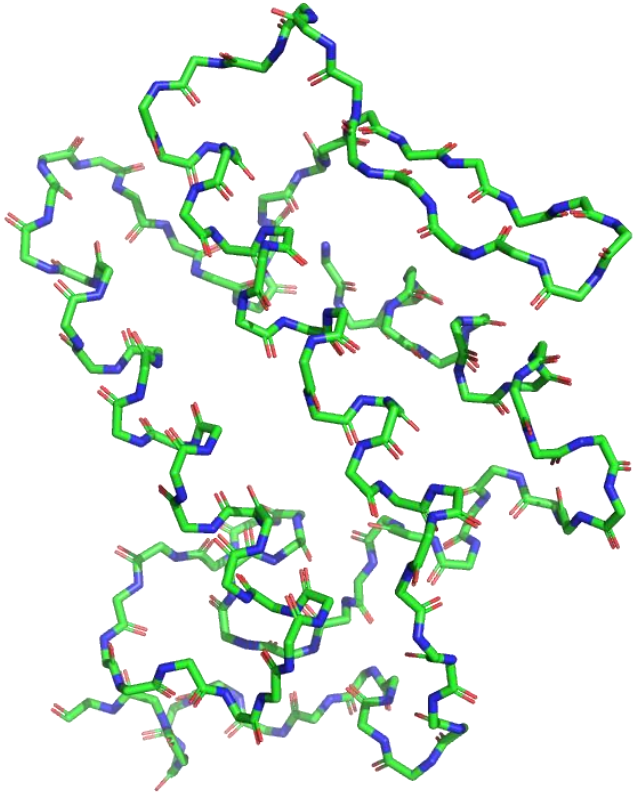
1. **Introduction - Proteins and Small Molecules**
2. **Sequence-based models**
   - Protein Language Models (PLMs)
3. **Protein Structure Prediction**
   - AlphaFold2
   - Evolutionary Scale Modelling (ESM)
   - Modelling of biomolecular assemblies
4. **Structure-based models**
   - 3D-Convolutional Neural Networks
   - Introduction to Graphs and Graph Neural Networks
   - Graph-based models
5. **Challenges of training on biological data**
6. **Generative models for *de novo* design**

Recap

Proteins are chains of amino acids

- Fold into a complex 3D shape based on amino acid interactions
- Backbone consists of the linked N-Cα-C triangles of all amino acids
- The difference between the 20 amino acids lies in their side chains





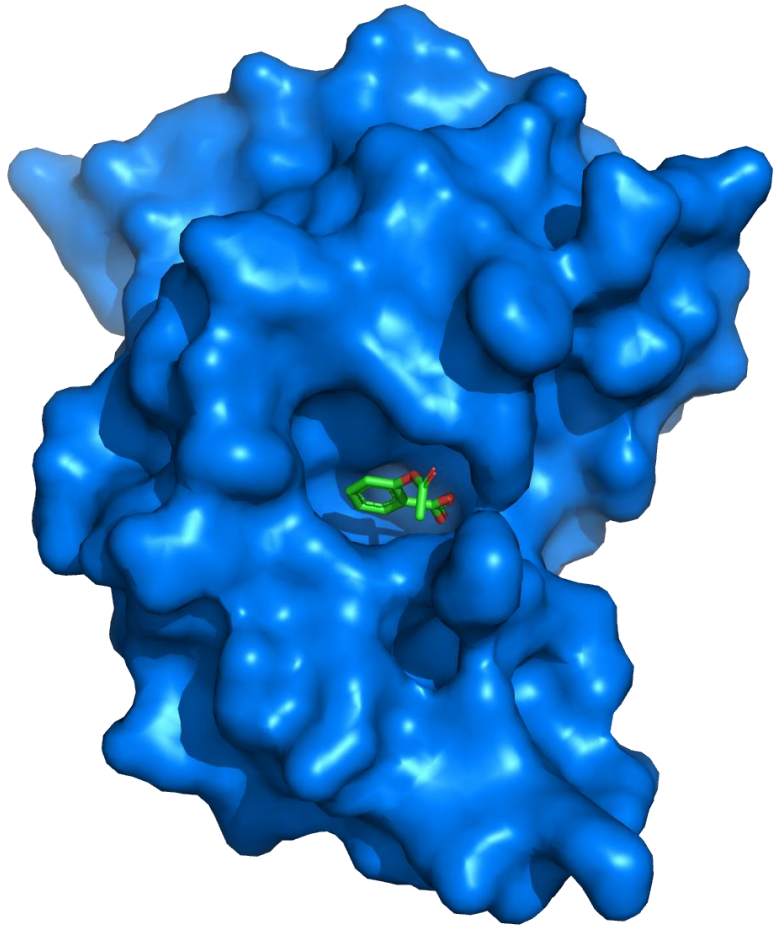
Recap

Small molecules

- Organic compounds with a low molecular weight, usually less than 100 atoms
- Their structure usually allows for **easy synthesis in the laboratory**

Most drugs are small molecules

- Can bind to proteins and affect a biological process
- The chemical structure and composition of the small molecule defines its function and binding preferences
- Small molecule drugs are engineered to interact with a specific target to modulate disease pathways



Aspirin (in green and red) inhibits the activity of the enzyme cyclooxygenase (COX) which leads to the formation of prostaglandins (PGs) that cause inflammation, swelling, pain and fever (PDB 1OXR)

Recap

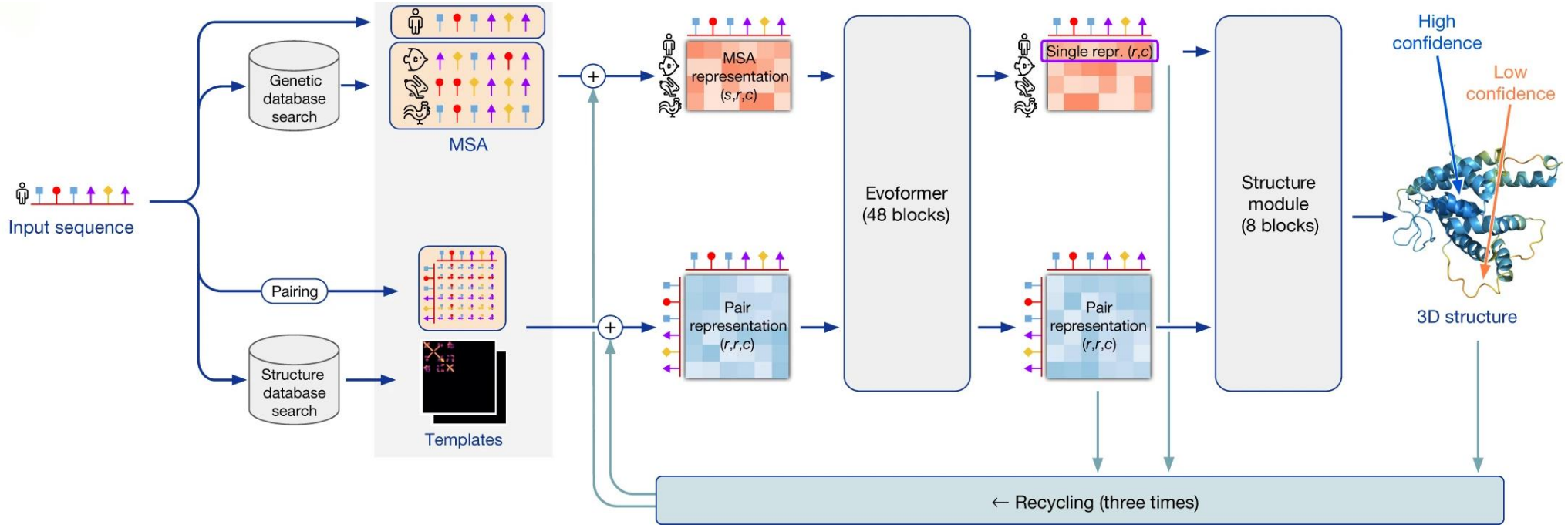


Figure: AlphaFold model architecture. Arrows show the information flow among the various components. Array shapes are shown in parentheses with s, number of sequences, r, number of residues and c, number of channels

AlphaFold2 predicts protein structures from amino acid sequences

- Genetic database search to generate a multiple sequence alignment (MSA)
- Structure database search to generate first structure prediction from similar proteins (templates)
- Generate final refined structure prediction with model trained on experimental structures

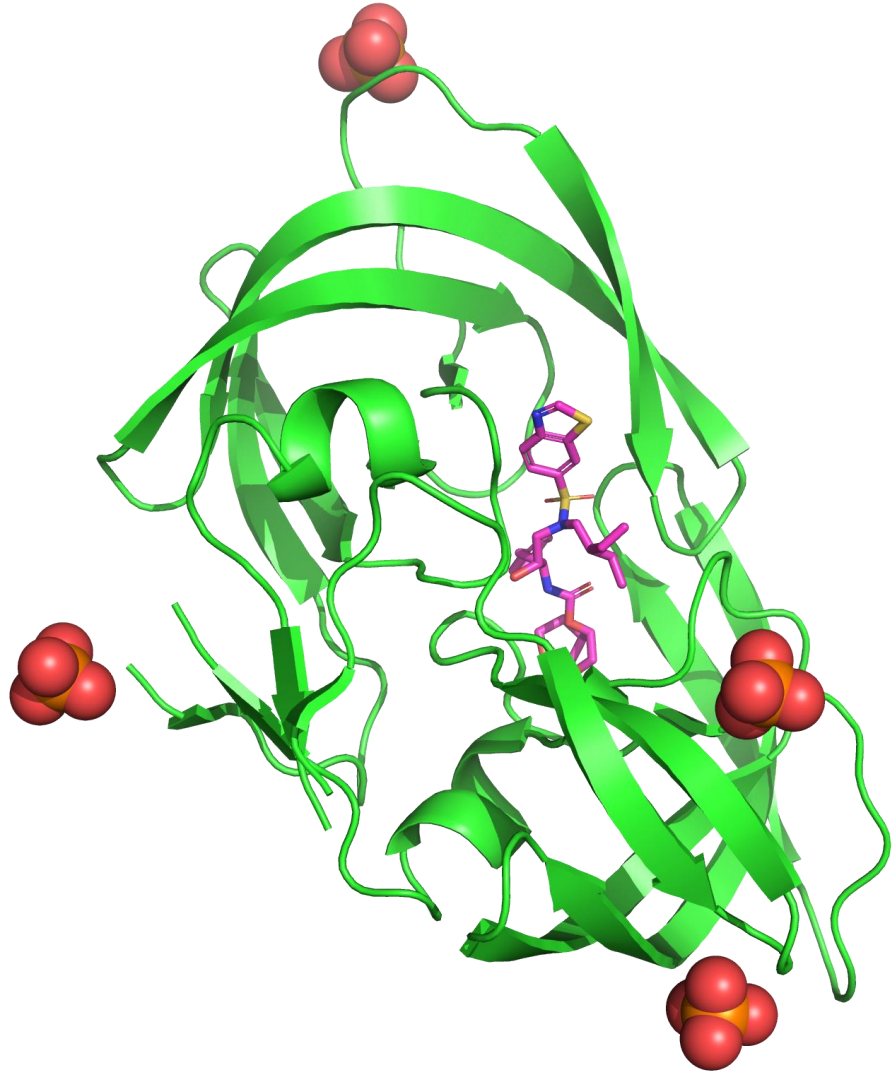
1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

Recap

Protein rarely act alone – they form complexes with other proteins, small molecules, DNA and RNA

Recent folding models include other biomolecules and form unified structure prediction tools predicting the structure of assemblies of proteins with

- Nucleic Acid (DNA)
- Metal Ions
- Small Molecules

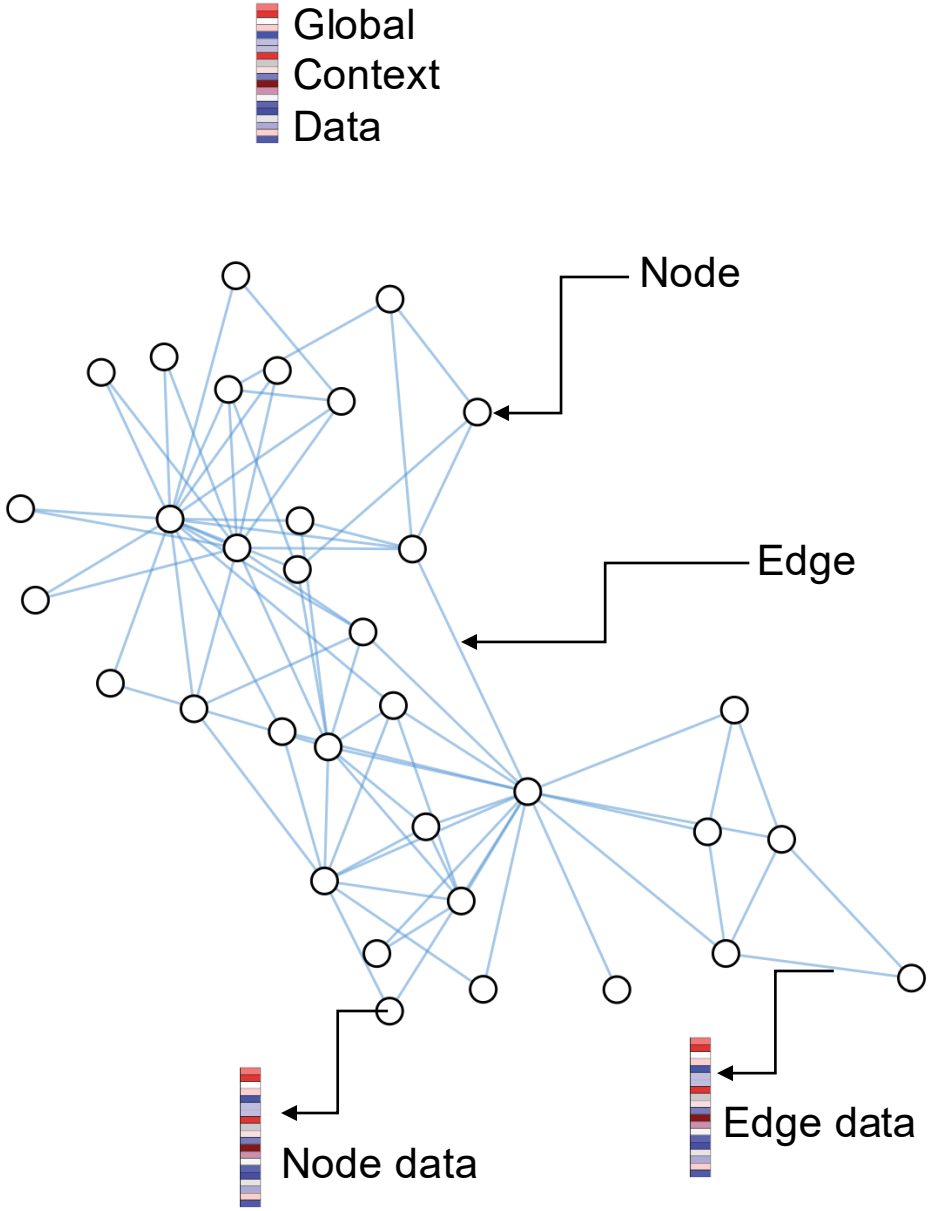


HIV-1 Protease bound to peptidomimetic inhibitor
Inhibitor binds to the active site of the HIV viral protease 1 without being cleaved, effectively blocking the active site and rendering the enzyme inactive.

Recap

Graph Convolutional Neural Networks

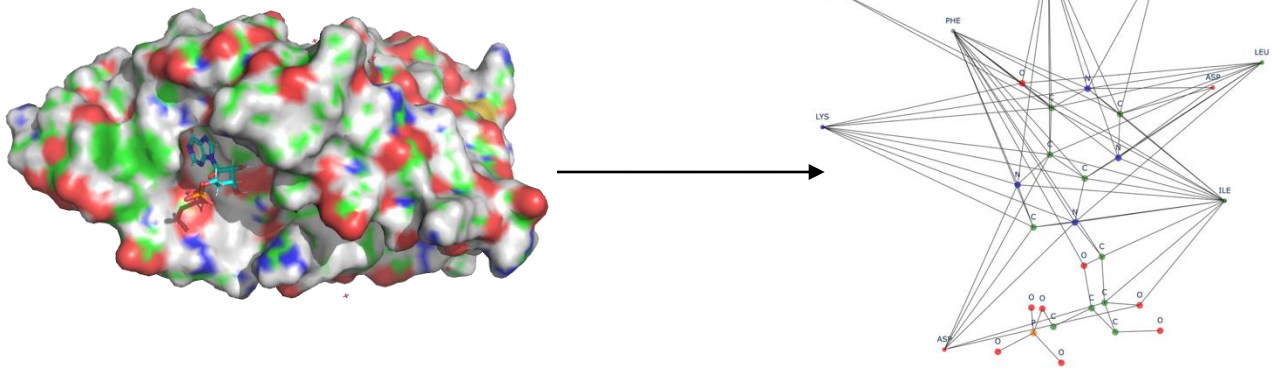
- Graphs are ideally suited to model unordered data such as social networks and molecules
- Graph Neural Networks transform node and edge embeddings of graphs through message-passing
- Resulting embeddings include information about the connectivity and the neighborhood of nodes and edges



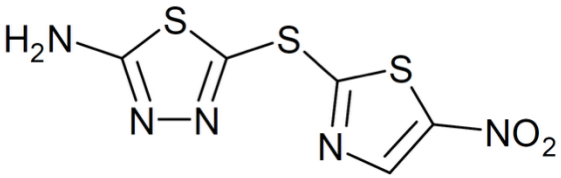
Recap

Graph Neural Networks are useful for modeling small molecules and proteins and are successful in predicting:

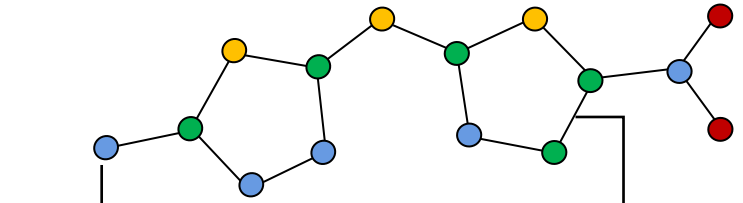
- Properties of small molecules
- Protein function
- Protein-Ligand Binding Affinity



PDB 5KA



Modelling molecules as graphs



Node data
- atomic number
- number of bonds
- formal charge
- chirality
- number of hydrogens

…

Edge data
- bond type
- conjugation
- ring membership
- stereochemistry

Stokes, J. M. et al. A Deep Learning Approach to Antibiotic Discovery. Cell 180, 688-702.e13 (2020).

Challenges of Training on Biological Data

→ Data Challenges in Machine Learning for Biology
→ Examples

Data Quality and Scarcity

Common sources of biological data include many databases based on community contributions of biology and chemistry researchers.

- **UniProt:** Database of protein sequences and functional annotation.
- **Protein Data Bank (PDB):** Database of protein and other macromolecular structures.
- **PubChem:** Chemicals database

Data Quality and Noise:
Biological datasets often contain errors, missing values, or inconsistencies due to experimental variability, measurement errors, or incomplete annotations.

Data Scarcity and Imbalance:
High-quality labelled biological data is often scarce and imbalanced, with overrepresentation of certain classes (e.g., well-studied proteins) and underrepresentation of others.

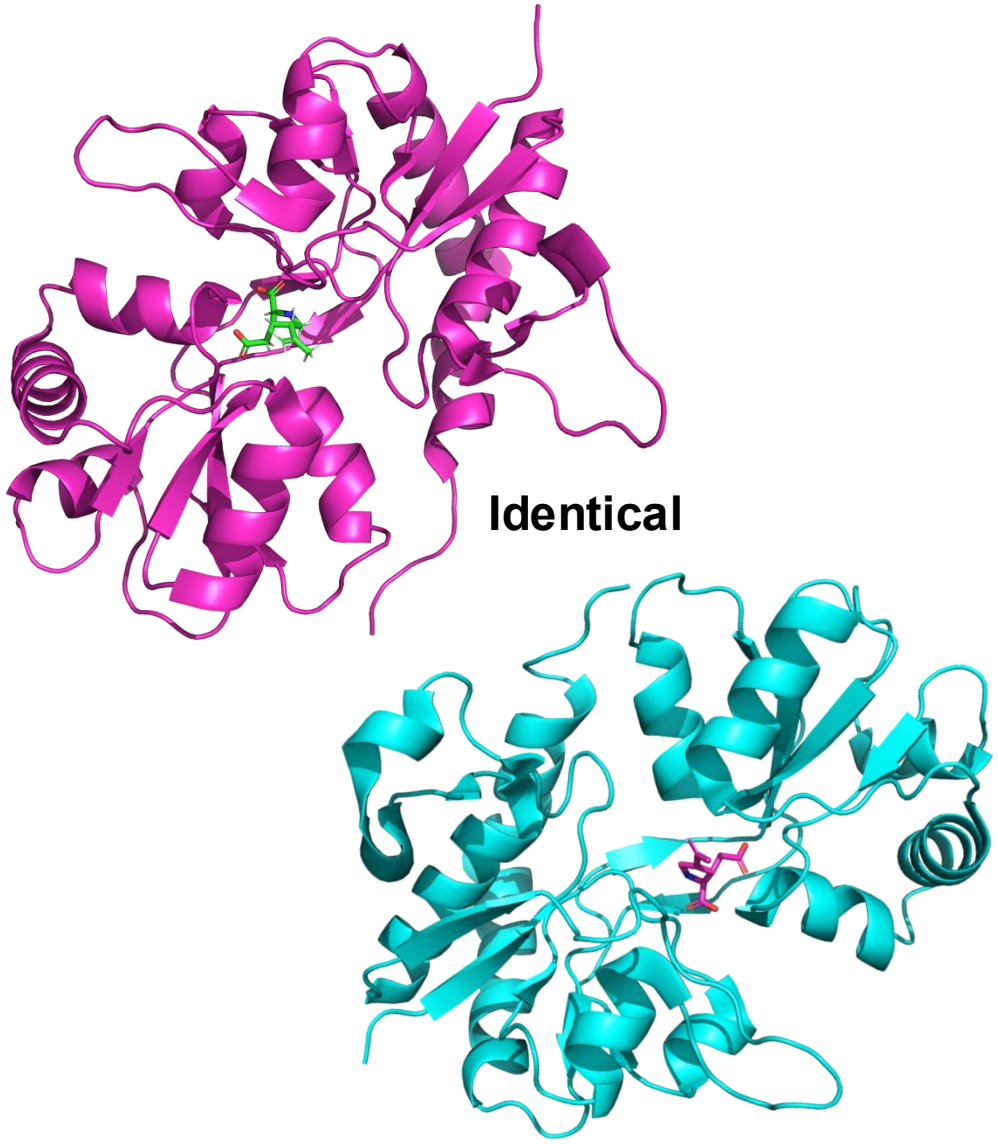
Bias and Redundancy

Datasets may be biased towards or well-studied proteins and contain significant redundancies

Redundancies are often not recognized due to

- Complex data type
- Wrong proxies for similarity, such sequence identity as a measure of protein-ligand complex similarity
- Complexes with low sequence identity can have identical interaction patterns

➔ **Models are trained on highly redundant (but not very diverse) datasets, often with undetected overlap between training and testing data**



Example 1: Structure-based binding affinity prediction

Task: Protein Ligand Binding Affinity Prediction (Scoring Functions)
Predict interaction strength of protein-ligand complexes from their 3D structure

Train

PDBbind database (n = 19’000)
- All affinity-labelled 3D protein-ligand complexes in the Protein Data Bank
- Many redundancies arising from multiple experiments on the same protein

Test

CASF2016 benchmark (n=285, subset of PDBbind)
- initially designed to benchmark classical physics-based scoring algorithms
- sampled from sequence similarity clusters in PDBbind

More than 20 peer-reviewed and published models trained in this train-test setup (GNNs, 3D-CNNs)

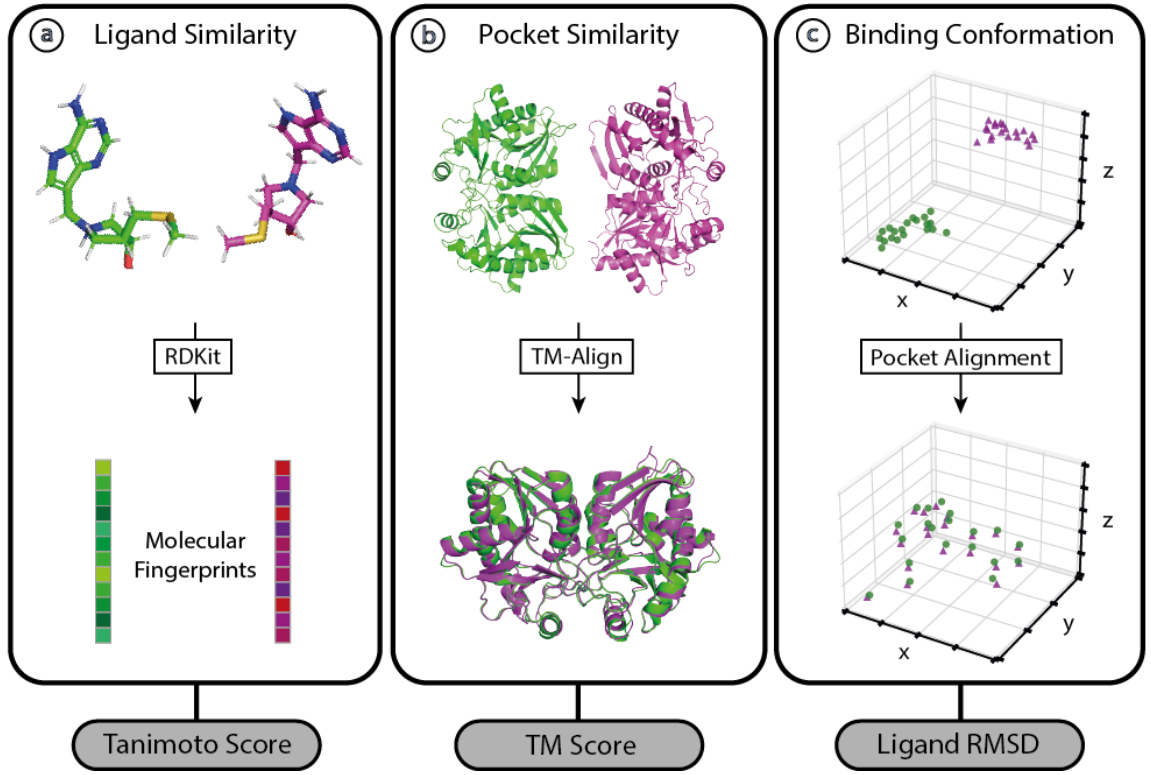
→ Data Leakage Concerns

Example 1: Structure-based binding affinity prediction

Computing three pairwise metrics for the entire PDBbind database (including CASF benchmark)

- **Ligand Similarity** - Tanimoto similarity between count-based molecular fingerprints
- **Protein Similarity** – TM Scores using TM-align (finding optimal alignment of 3D structures)
- **Binding Conformation** – RMSD of ligand atoms after pocket alignment (by TM-align)

Three pairwise similarity matrices, each based on 189mio computed comparisons.

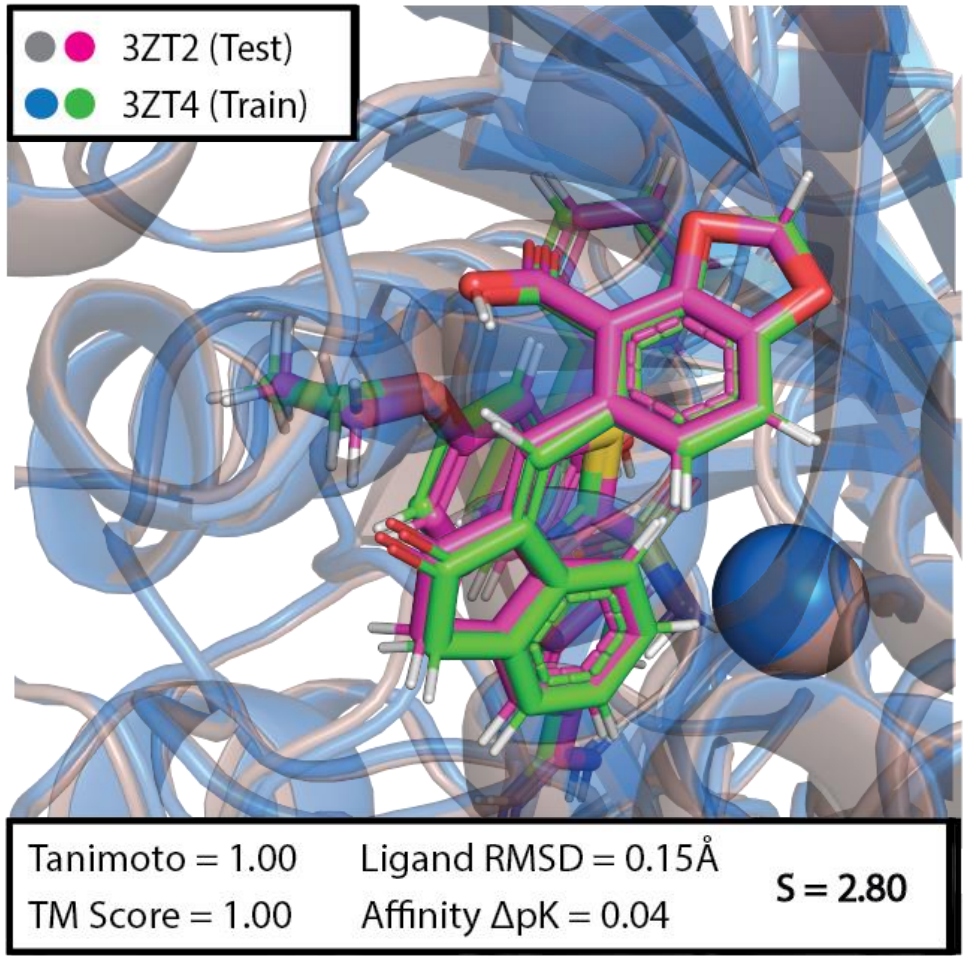


Graber, D. *et al.* Resolving data bias improves generalization in binding affinity prediction. *Nat. Mach. Intell.* **7**, 1713–1725 (2025).

Example 1: Structure-based binding affinity prediction

Using the pairwise similarity matrices to

- find the **train-test** pairs that were most similar in terms of all three metrics
- Compare their binding affinity label



Train-test data leakage between PDBbind and CASF

Almost 600 training complexes are nearly identical to a test complex in terms of

- Protein structure
- Ligand structure
- Ligand binding conformation
- Affinity label

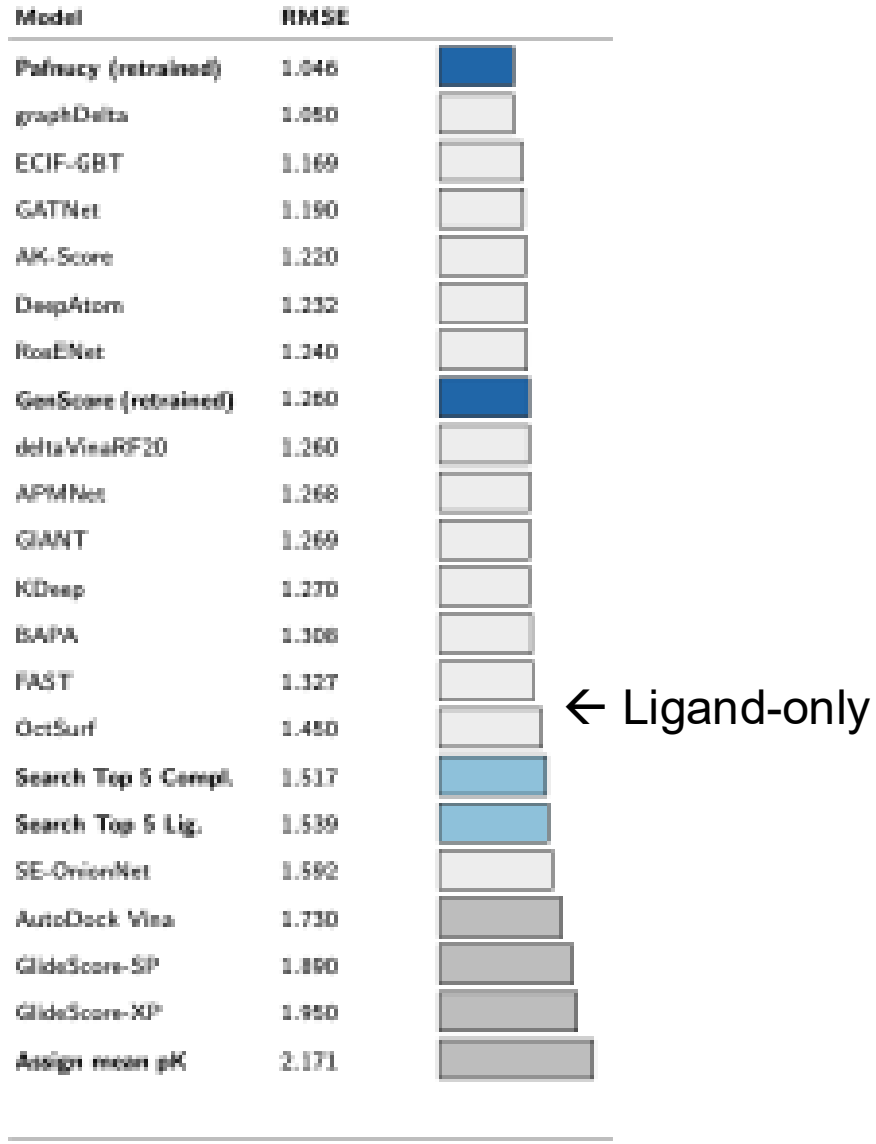
Almost half of the test complexes are not presenting new challenge to models trained on PDBbind

Graber, D. *et al.* Resolving data bias improves generalization in binding affinity prediction. *Nat. Mach. Intell.* **7**, 1713–1725 (2025).

Example 1: Structure-based binding affinity prediction

Almost all ML-based scoring functions were trained on PDBbind in recent years, using the CASF2016 benchmark to measure generalization and compare to other models

- All models are outperformed by the oldest model when it is retrained on the newest data (3D-CNN, Pafnucy, 2018)
- Some models are outperformed by models trained on PDBbind with all protein information removed from the input data (ligand-only)
- One model is outperformed by look-up algorithms, which…
  - Predict test set affinities by averaging the labels of the 5 most similar training **complexes**
  - Predict test set affinities by averaging the labels of the 5 training complexes with the most similar **ligand**



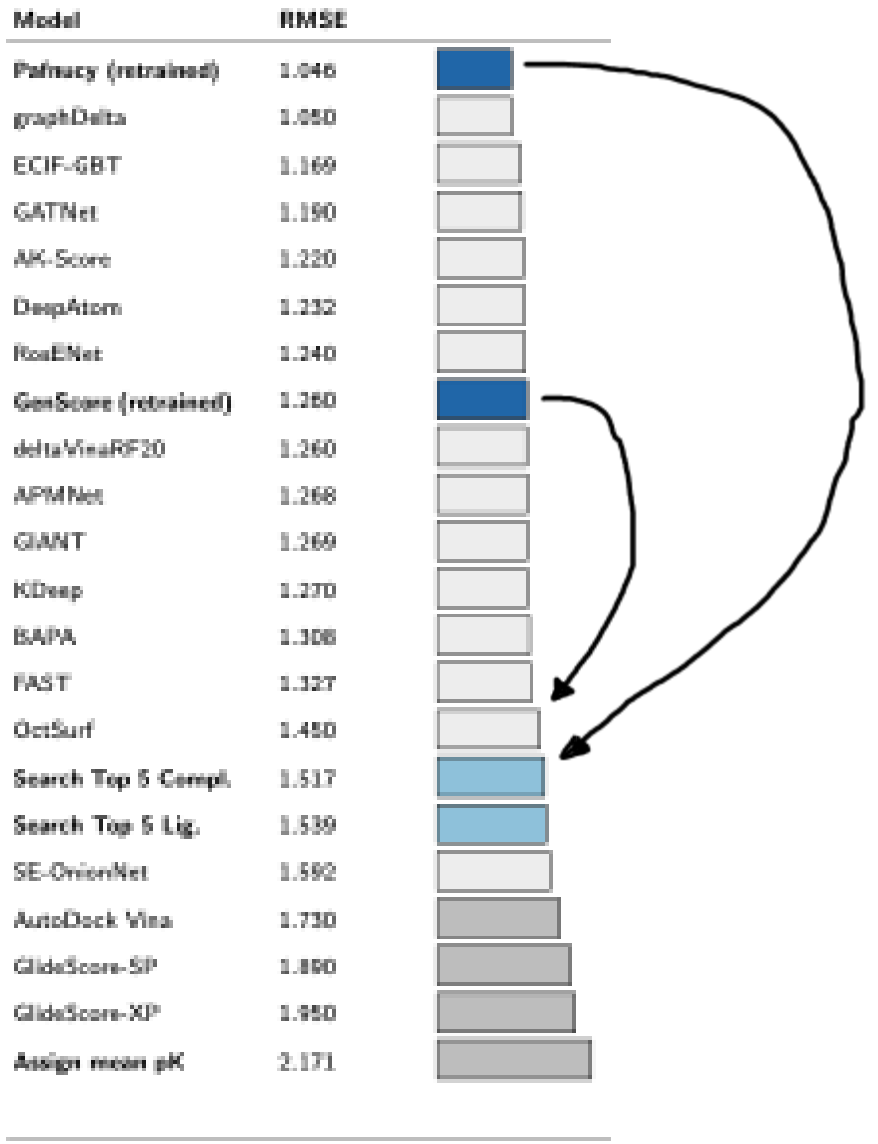
Graber, D. *et al.* Resolving data bias improves generalization in binding affinity prediction. *Nat. Mach. Intell.* **7**, 1713–1725 (2025).

Example 1: Structure-based binding affinity prediction

Filter training dataset (PDBbind) to remove train-test data leakage into CASF benchmark

Filter the training dataset until all **test** complexes contain unique ligands and are structurally distinct from all training complexes.

- **Removes around 600/19'000 <u>training</u> complexes**
- Retraining state-of-the-art models on this 4% smaller training dataset **causes considerable drops in CASF2016 performance.**



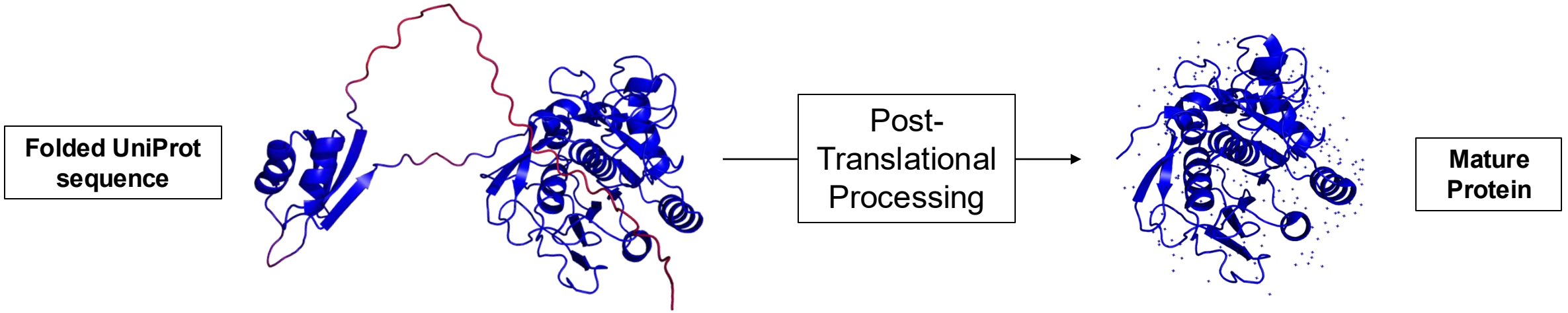
Graber, D. *et al.* Resolving data bias improves generalization in binding affinity prediction. *Nat. Mach. Intell.* **7**, 1713–1725 (2025).

Example 2: Protein Thermostability Prediction

Task: End-to-end protein melting point prediction from protein sequences

Training Data: Meltome Atlas – Experimental melting points for 48’000 proteins across 13 species. Corresponding protein sequences for model training are fetched from **UniProt** database.

Problem 2 – Signalling Peptides:
Uniprot contains complete amino acid sequences from sequencing experiments, including regions that are cleaved off during post-translational processing.



→ Protein Stability is a property of the mature protein (signalling peptides cleaved off)

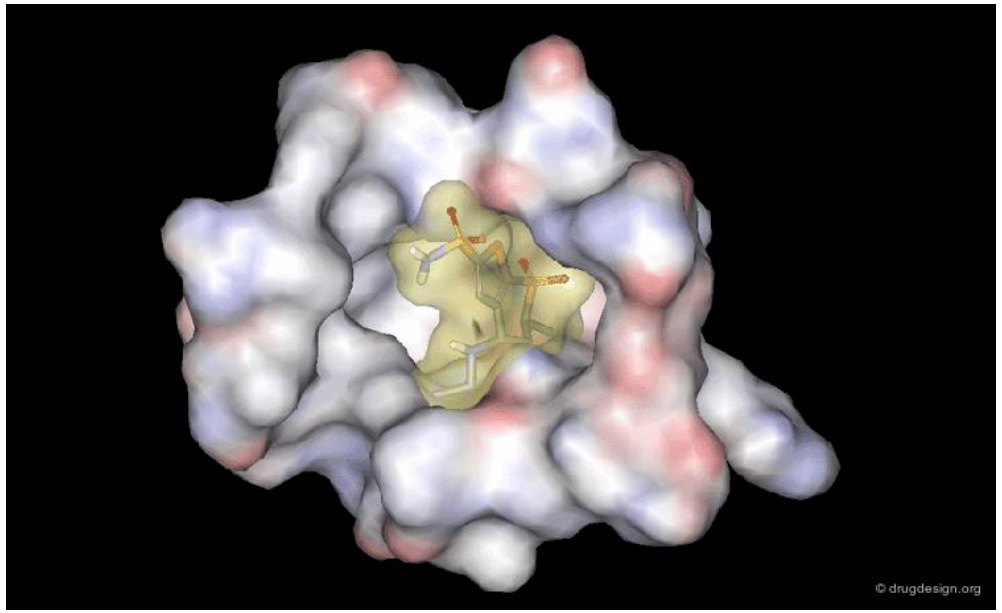
Jarzab, A., Kurzawa, N., Hopf, T. et al. Meltome atlas—thermal proteome stability across the tree of life. Nat Methods 17, 495–503 (2020).

Example 3: Molecular Docking / Co-folding

Task: Predict the binding conformation of a small molecule to a protein

Classical docking algorithms are simulation-based

- Protein is static
- Scoring functions predict the binding free energy of the sampled binding poses
- Find the location of binding and the preferred orientation = Docking Pose



Drugdesign.org

Example 3: Molecular Docking / Co-folding

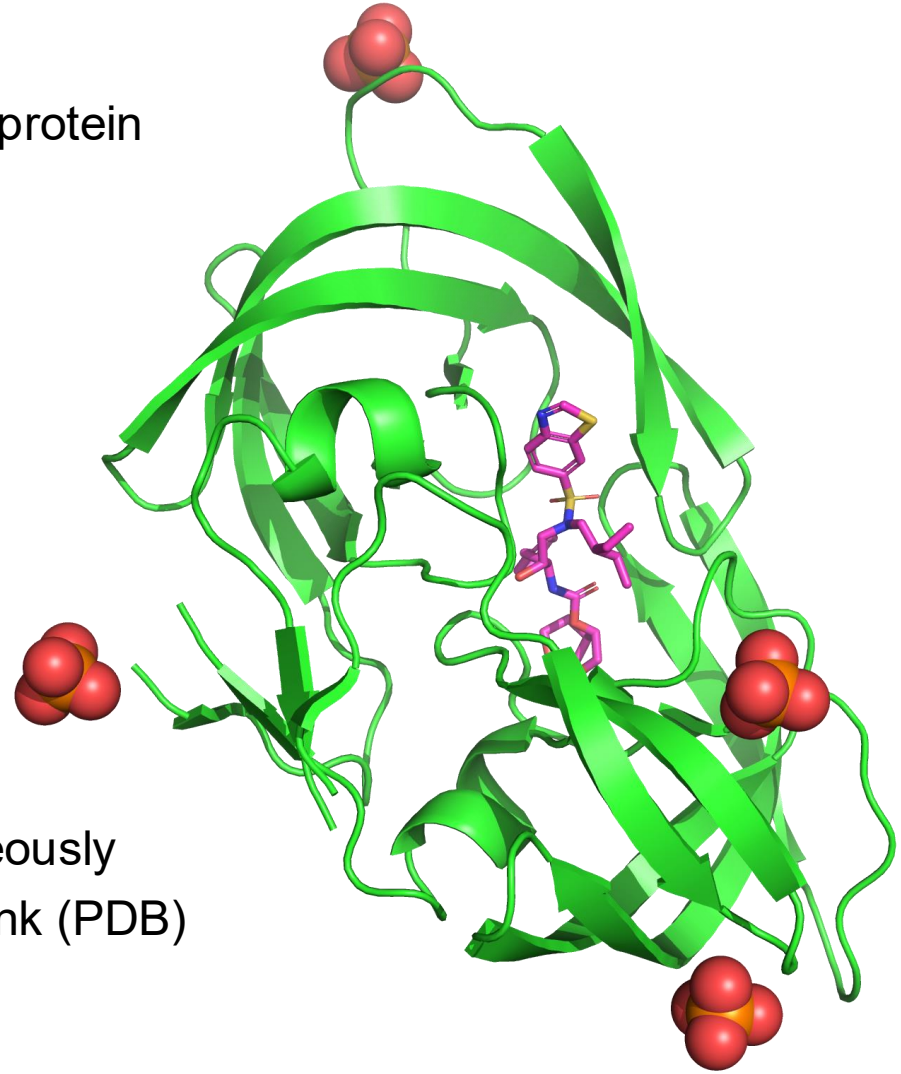
Task: Predict the binding conformation of a small molecule to a protein

Classical docking algorithms are simulation-based

- Protein is static
- Scoring functions predict the binding free energy of the sampled binding poses
- Find the location of binding and the preferred orientation = Docking Pose

Co-folding models are mostly diffusion-based

- Predict protein structure and ligand binding pose simultaneously
- Trained on all molecular complexes in the Protein Data Bank (PDB)
- Small-molecule ligands, DNA, metal ions



HIV-1 Protease bound to peptidomimetic inhibitor
Inhibitor binds to the active site of the HIV viral protease 1 without being cleaved, effectively blocking the active site and rendering the enzyme inactive.

Example 3: Molecular Docking / Co-folding

Co-folding models:

- AlphaFold3, RosettaFold All-atom, Chai-1, Boltz
- Protein folding performance similar to AlphaFold2

Ligand Positioning:

- Co-folding models were shown to outperform classical physics-based docking algorithm (redocking)
- Models interpolate robustly, but struggle to extrapolate to inputs not present during training
- Artefacts such as steric clashes occur frequently
- Models are not sensitive to adversarial challenges

Primarily rely on data-driven pattern recognition, not on an understanding of physics

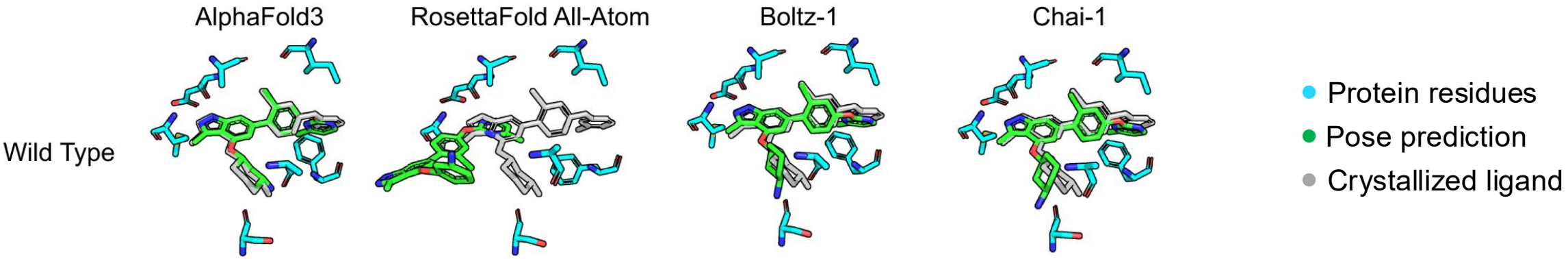
Masters, M. R., Mahmoud, A. H. & Lill, M. A. Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions. *Nat. Commun.* **16**, 8854 (2025).

Example 3: Molecular Docking / Co-folding

Docking MEK1 inhibitor (binding to ATP-binding site)

Challenge: Wild-type docking

- Most co-folding models are able to predict the native binding pose correctly



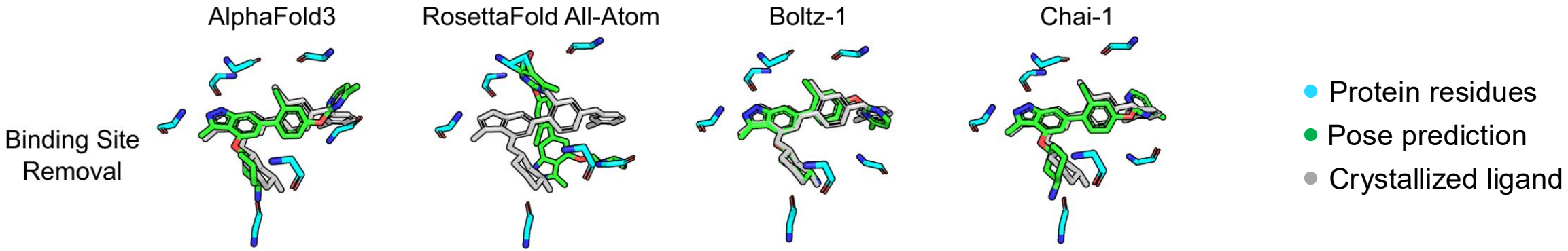
Masters, M. R., Mahmoud, A. H. & Lill, M. A. Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions. *Nat. Commun.* **16**, 8854 (2025).

Example 3: Molecular Docking / Co-folding

Docking MEK1 inhibitor (binding to ATP-binding site)

Challenge: Binding Site Removal

- All binding site residues are mutated to glycine, removing all possibility for ligand-protein interactions
- Most co-folding still predict the native binding pose



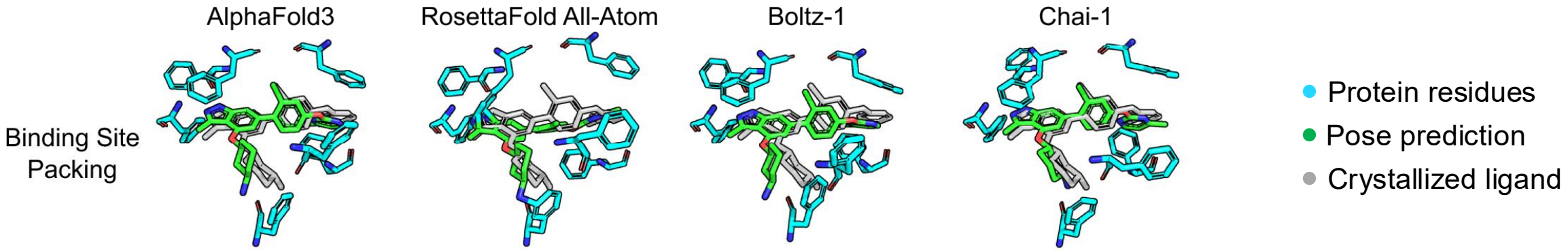
Masters, M. R., Mahmoud, A. H. & Lill, M. A. Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions. *Nat. Commun.* **16**, 8854 (2025).

Example 3: Molecular Docking / Co-folding

Docking MEK1 inhibitor (binding to ATP-binding site)

Challenge: Binding Site Packing

- All binding site residues are mutated to phenylalanine, removing all native interactions with side-chains and further occupying the pocket with bulky, hydrophobic groups.
- Most co-folding still predict the native binding pose
- Cases of unphysical overlapping atoms and large steric clashes



Masters, M. R., Mahmoud, A. H. & Lill, M. A. Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions. *Nat. Commun.* **16**, 8854 (2025).

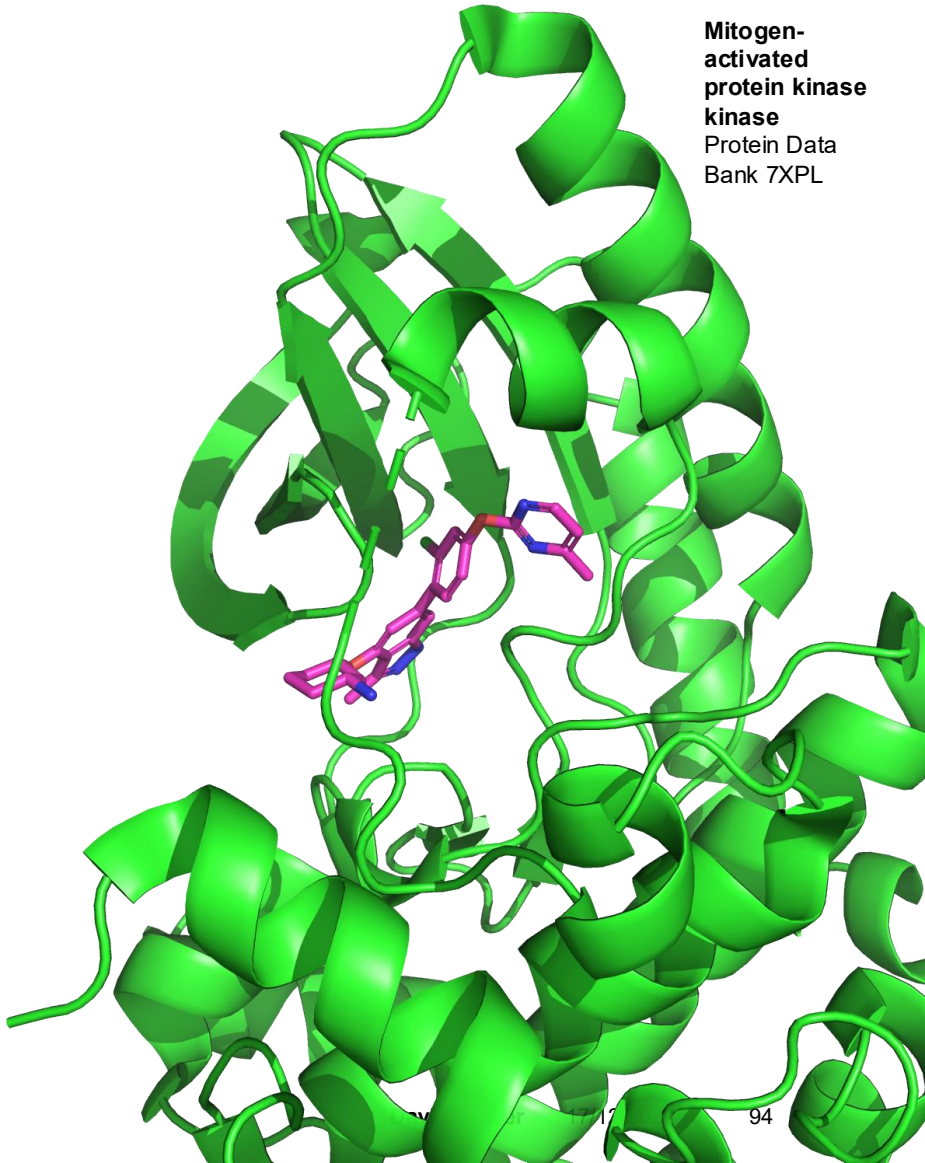
Example 3: Molecular Docking / Co-folding

Docking MEK1 inhibitor (binding to ATP-binding site)

- Results are based on only 2 protein-ligand systems
- Tendency for unphysical predictions, showing only partial capacity to adapt to disruptions
- Co-folding models are not predicting poses based on physics of interactions, but rather learning patterns in global protein structures and sequences.

Traditional Benchmarks primarily measure how well these models capture structural patterns

Better physics integration is needed



Mitogen-activated protein kinase kinase
Protein Data Bank 7XPL

Masters, M. R., Mahmoud, A. H. & Lill, M. A. Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions. *Nat. Commun.* **16**, 8854 (2025).

Challenges of biological datasets - Conclusions

- **Lack of understanding of biological context,** e.g. failing to incorporate domain-specific preprocessing, which can lead to models learning irrelevant or misleading features.

- **Dataset redundancy**
  - Random or temporal splits are often insufficient due to intrinsic redundancies, which can lead to test sets that are not independent of the training data.

- **Lack of physical/chemical constraints**
  - Models tend to reproduce patterns observed during training, even when the results are unphysical
  - Validating model outputs against established scientific principles
  - We need to integrate robust chemical and physical priors into such tools

We need better benchmarks that quantify generalization and test adherence to fundamental physical principles

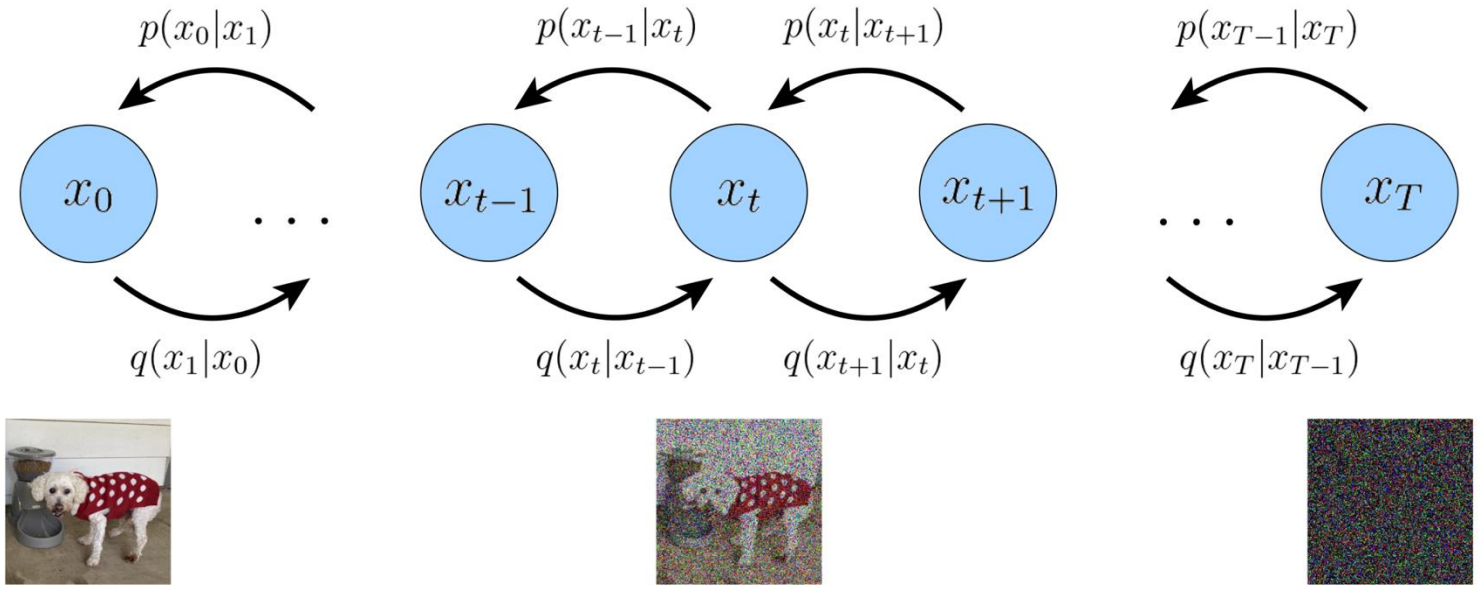
Generative AI for *de novo* design

→ *De novo* design of small molecules
→ *De novo* design of proteins

Diffusion Models

Probabilistic Diffusion Models achieve state of the art performance in image generation

- Add increasing amounts of gaussian noise over a certain number of timesteps
- Train a neural network to undo the steps of noise
- Sample from your known gaussian distribution and feed this pure noise to your models
- Realistic image from the distribution you trained on



Luo, C. Understanding Diffusion Models: A Unified Perspective. *arXiv* (2022) doi:10.48550/arxiv.2208.11970.

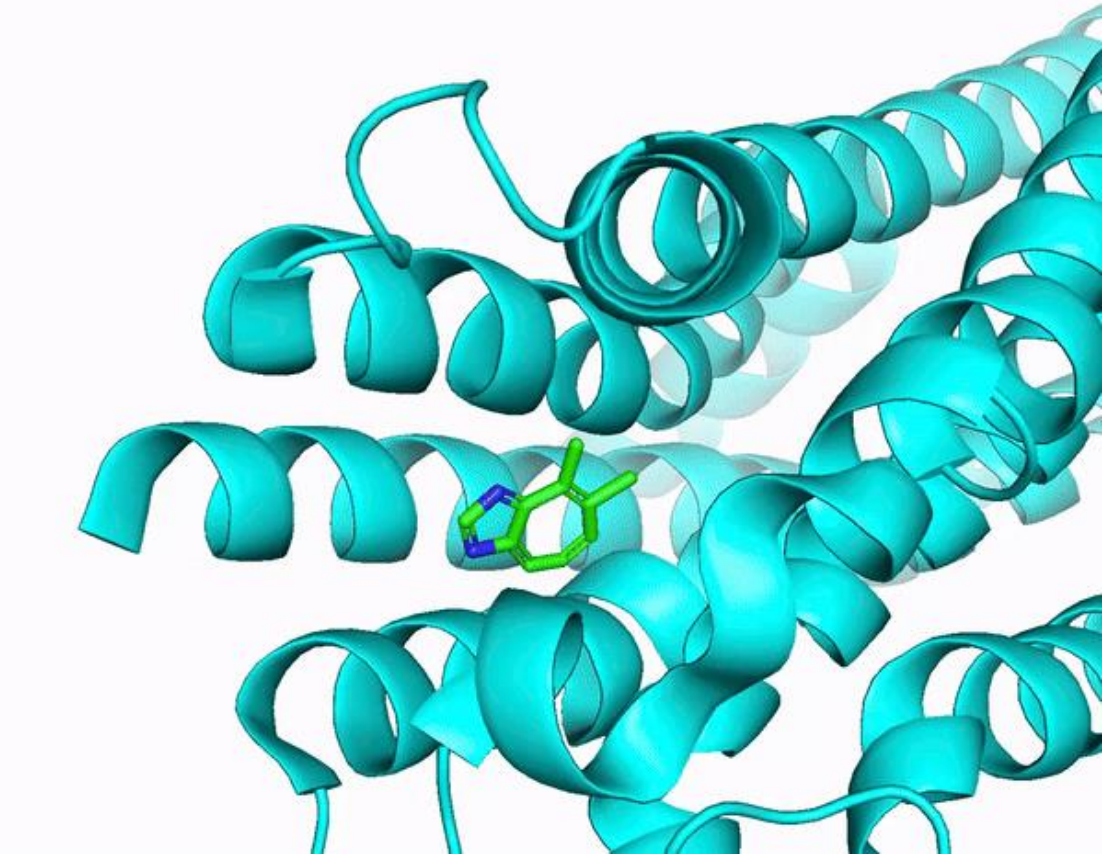
DiffSBDD for design of small molecule binders

Denoising diffusion probabilistic model (DDPM) that generates novel ligands conditioned on protein pockets

- It aims to generate small molecule ligands that bind to a specific 3D protein structure
- Considers the protein pocket as a fixed context
- Ligands are novel, diverse, and match the shape of the protein's binding pocket

➔ **Traditional drug screening is limited to the chemical space of previously studied molecules.**

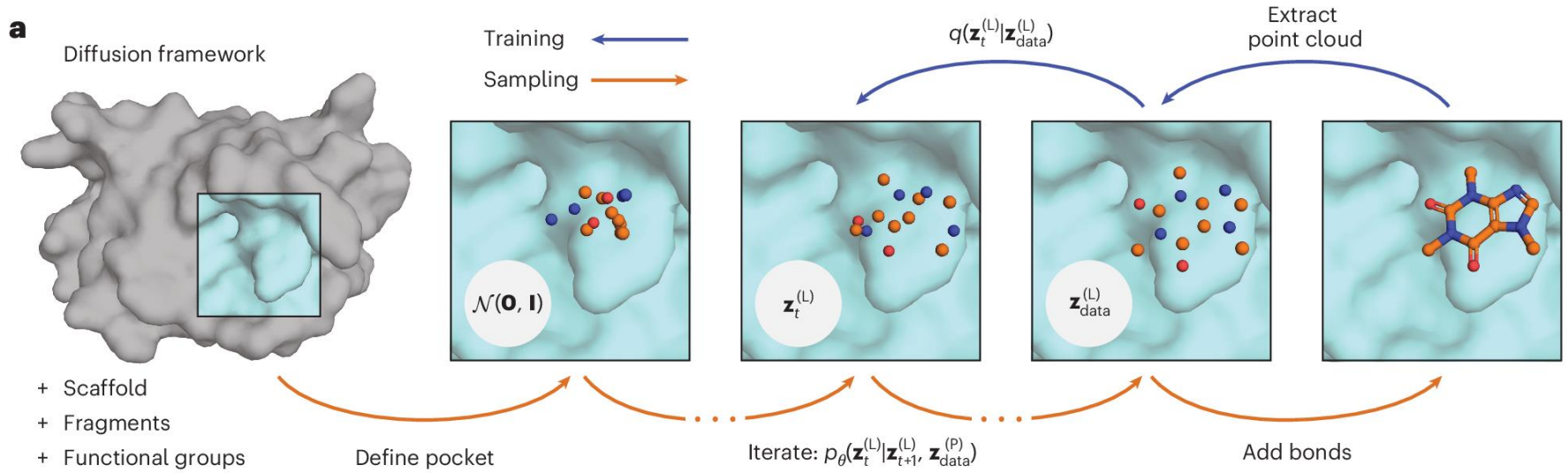
➔ **Diffusion can generate entirely new molecules**



Schneuing, A., Harris, C., Du, Y. *et al.* Structure-based drug design with equivariant diffusion models. *Nat Comput Sci* (2024). https://doi.org/10.1038/s43588-024-00737-x,

DiffSBDD for design of small molecule binders

Denoising diffusion probabilistic model (DDPM) that generates novel ligands conditioned on protein pockets



- Generated training data by noising protein-ligand structures from the CrossDocked dataset
- Model is trained to recover ground-truth structures from noised inputs
- Generate new structures through iterative denoising of random noise

Schneuing, A., Harris, C., Du, Y. *et al.* Structure-based drug design with equivariant diffusion models. *Nat Comput Sci* (2024). https://doi.org/10.1038/s43588-024-00737-x,

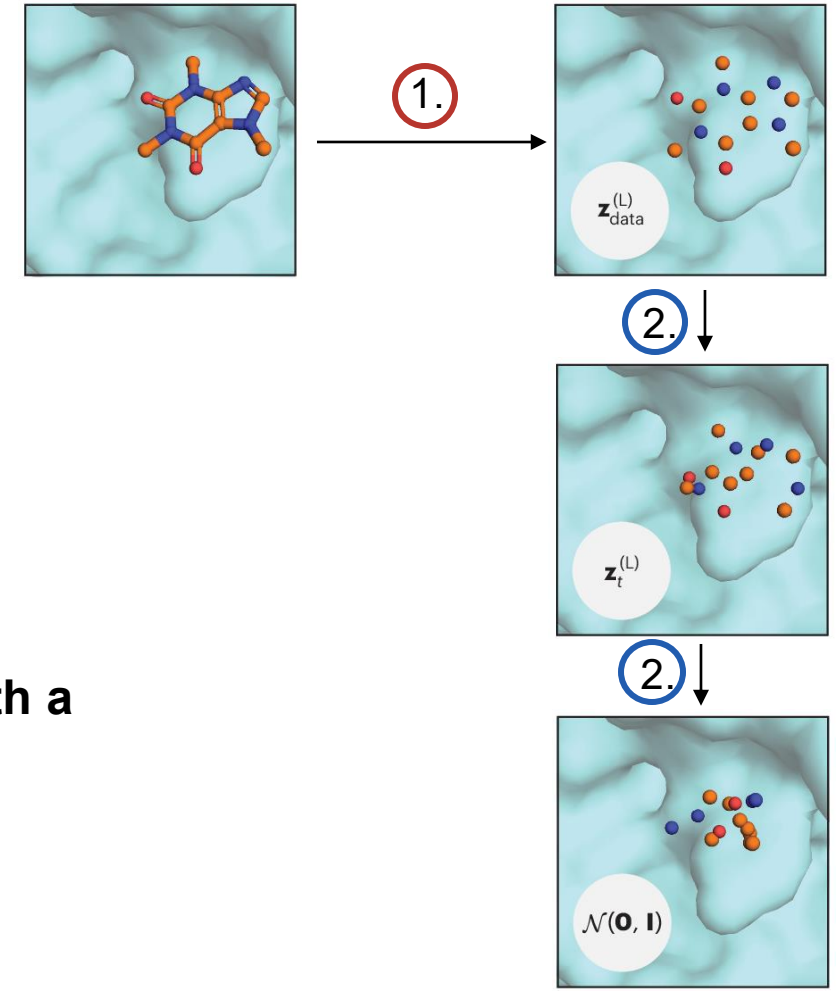
DiffSBDD for design of small molecule binders

How to add gaussian noise to small molecules bound to a protein?

1. Represent ligand as an atomic point cloud $z_{data}^{(L)} = [x, h]$
   - 3D coordinates x
   - Categorical features h (atom type)
   - Molecular bonds are neglected

2. Add noise to $z_{data}^{(L)} = [x, h]$ for t noising steps to obtain $z_t^{(L)} = [x, h]$

   → Noising of coordinates **and** atom types



Noising process transforms atomic point cloud form aligned with a molecule structure into randomly distributed

→ Train a neural network to reverse this process

Schneuing, A., Harris, C., Du, Y. *et al.* Structure-based drug design with equivariant diffusion models. *Nat Comput Sci* (2024). https://doi.org/10.1038/s43588-024-00737-x,

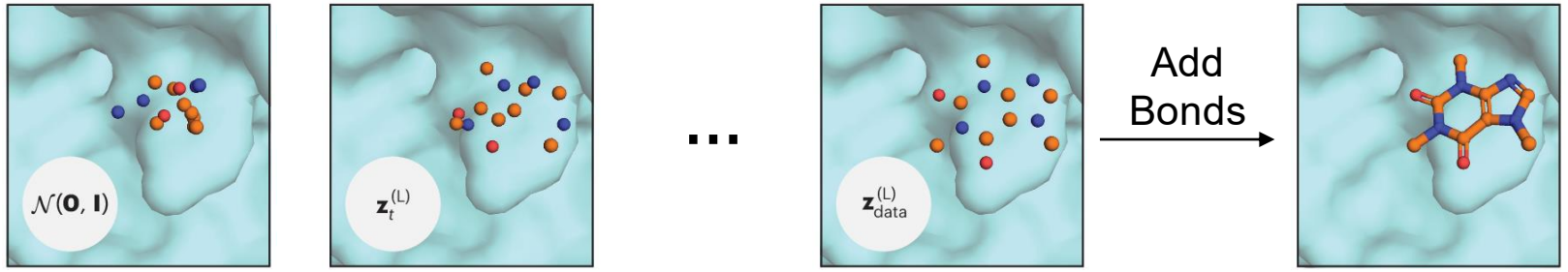
DiffSBDD for design of small molecule binders

Training of DiffSBDD to recover original ligand structure from noised atomic point clouds

1. Noised atomic cloud $z_t^{(L)}[x, h]$ as a fully connected graph
2. Merged with a protein graph $z_{data}^{(P)}$ (not noised)
3. GNN computes updates of the ligand's features h and coordinates x needed to reverse a noising step. Features of the protein graph $z_{data}^{(P)}$ remain unchanged during the reverse diffusion process

$$z_t^{(L)} = [x_t, h_t] \rightarrow \boldsymbol{GNN} \rightarrow z_{t-1}^{(L)} = [x_{t-1}, h_{t-1}]$$

$z_t = [x, h]$ $z_{t-1} = [x, h]$ $\hat{z}_{data}[x, h]$



Prediction of ground truth distribution $\hat{z}_{data}[x, h]$ of atom point cloud after reversing t noising steps

Schneuing, A., Harris, C., Du, Y. *et al.* Structure-based drug design with equivariant diffusion models. *Nat Comput Sci* (2024). https://doi.org/10.1038/s43588-024-00737-x,

Beyond today's scope

Generative Modelling for Ligand Design (conditioned on target pocket)

- **TargetDiff:** Similar to DiffSBDD, but results in higher quality ligands
- **DrugFlow:** Flow-matching implementation of pocket-conditioned ligand design, using a virtual node type to allow the model to dynamically add or remove atoms and avoid steric clashes
- **FLOWR.root:** Flow-matching implementation of pocket-conditioned ligand design including binding affinity prediction and confidence estimation

Generative AI for *de novo* design

→ *De novo* design of small molecules
→ *De novo* design of proteins

De novo design of proteins

Protein engineering

- Explore the protein space by modifying the sequence of a natural proteins to optimize its properties
- Resulting proteins share most of their sequence with the protein that served as starting point
- Resulting proteins are structurally similar to a natural proteins

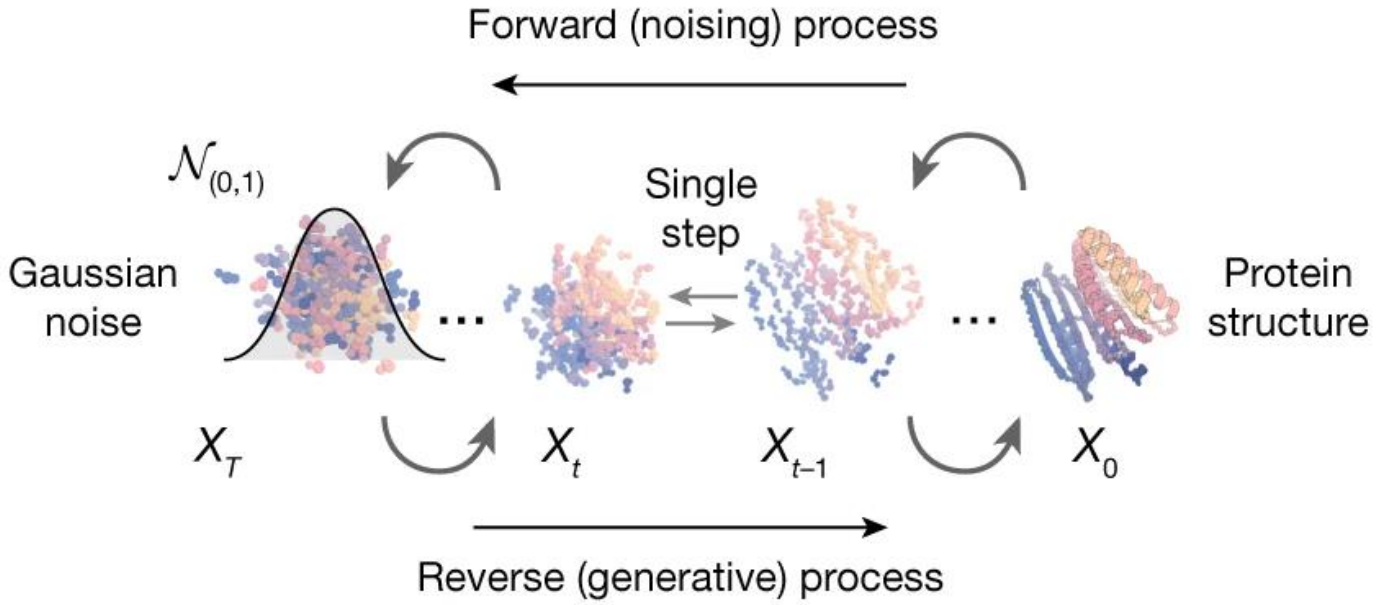
***De novo* protein design**

- Designing a novel sequence from scratch
- Does not rely on natural protein sequences, but uses computational methods to design sequences that are potentially very different from all existing sequences

“Evolution has only explored a tiny subset of all proteins that could exist”

RFdiffusion for protein design

Diffusion models applied to proteins: Generate realistic protein backbones *de novo*



- Generated training data by noising structures from the Protein Data Bank (PDB)
- Model is trained to recover ground-truth structures from noised inputs
- Generate new structures X_0 through iterative denoising of random noise X_T

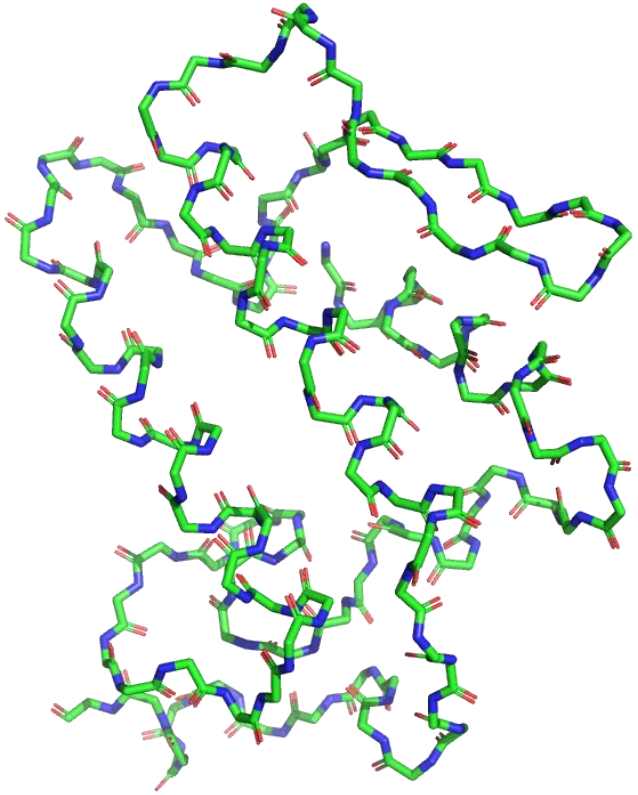
Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

RFdiffusion

How to add gaussian noise to protein backbones

1. Frame-based representation of protein backbone with L residues
   - The atoms N, Cα, and C form a nearly planar, rigid triangle
   - Each triangle can be described by a translation and a rotation
   - **Protein Backbone Structure = Set of L translations and rotations**

2. Add 3D gaussian noise to translations (perturb Cα coordinates)
3. Use Brownian motion on the manifold of rotation matrices



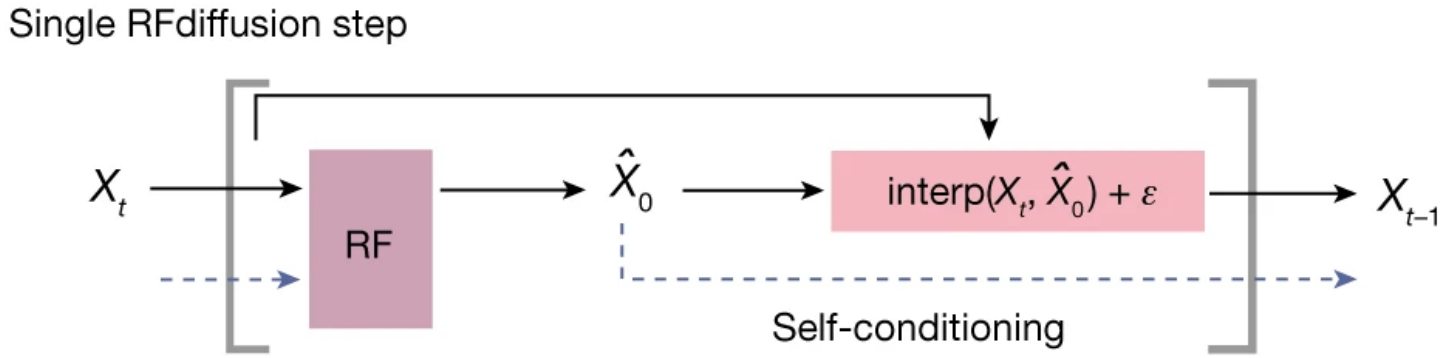
Noising process transforms set of triangles form neatly aligned along a protein backbone into randomly distributed and oriented

→ Train neural network to reverse this process

Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

RFdiffusion training

Training of RFdiffusion to recover original protein structure from noised protein structures



1. RFdiffusion takes the current coordinates X_t (initially random noise) and makes a prediction of the ground truth coordinates $\hat{X}_0$
2. The next coordinate input to the model (X_{t-1}) is generated by a noisy interpolation towards $\hat{X}_0$
3. Ground-truth prediction $\hat{X}_0$ is not discarded but handed to the next timestep

Self-conditioning: The model can use the positions of the coordinates at the current timestep, and the predicted ground truth coordinates of the previous timestep

→ **Markedly improved performance of RFdiffusion**

RoseTTAFold → RFdiffusion

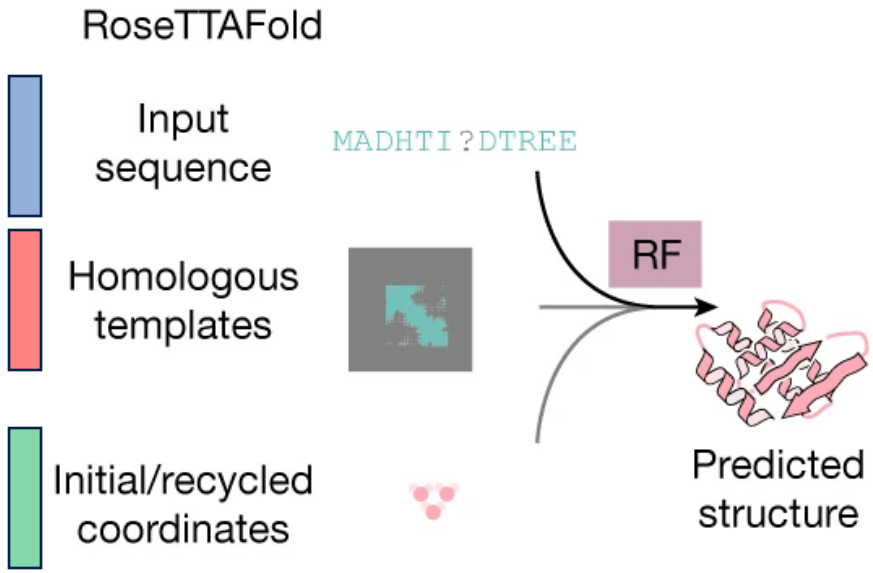
RoseTTAFold (similar to AlphaFold) accurately predicts protein structures

→ Improved diffusion models for protein design could be developed by taking advantage of the deep understanding of protein structure in structure prediction models

→ RFdiffusion was developed through fine-tuning of the RoseTTAFold structure prediction network

RoseTTAFold modelling pipeline

- Given input sequence
- Search for homologous proteins in a databases
- Creates a first structure prediction based on the homologous protein's structure
  - Create pairwise residue distance matrix
  - Create frame-based representation of protein backbone
- Output predicted structure



Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

RoseTTAFold → RFdiffusion

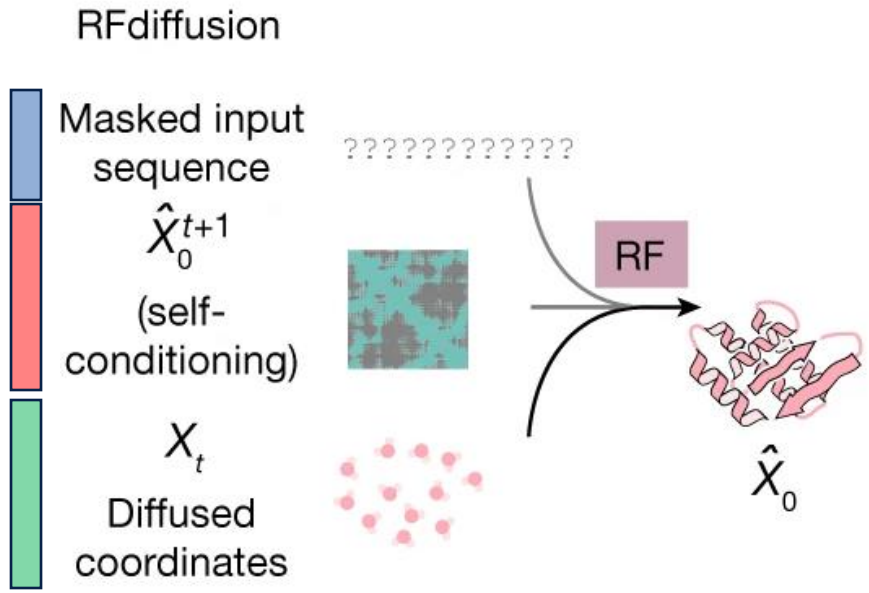
RoseTTAFold (similar to AlphaFold) accurately predicts protein structures

→ Improved diffusion models for protein design could be developed by taking advantage of the deep understanding of protein structure in structure prediction models

→ RFdiffusion was developed through fine-tuning of the RoseTTAFold structure prediction network

RFdiffusion modelling pipeline inputs:

- The sequence input in masked
- Current coordinate prediction at this timestep
- Prediction of the ground-truth coordinates $X_0^{(t+1)}$ of the previous timestep (self-conditioning)
- Output predicted structure



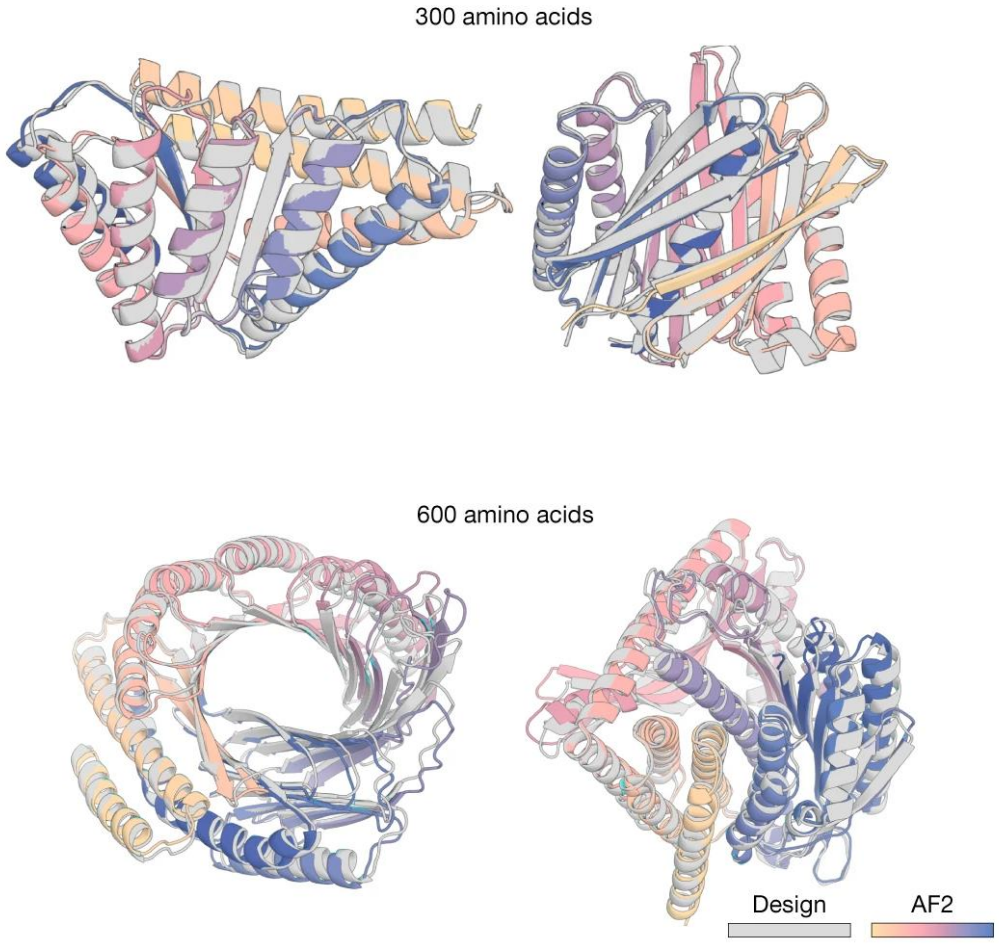
Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

RFdiffusion - Summary

→ Starting from random noise, RFdiffusion can generate protein backbones that are highly diverse

→ Many unconstrained designs show little overall similarity to structures seen during training, indicating that the model can generalize beyond its training data

→ Experimental characterization showed that most generated proteins are stable and soluble



RFdiffusion can generate new monomeric proteins of different lengths with no conditioning information

Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

RFdiffusion - Summary

Conditioning of RFdiffusion for specific designs

→ RFdiffusion can be customized to specific design challenges by addition of external potentials and fine-tuning

→ Guiding at each step of the iterative denosing process towards specific design objectives

Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

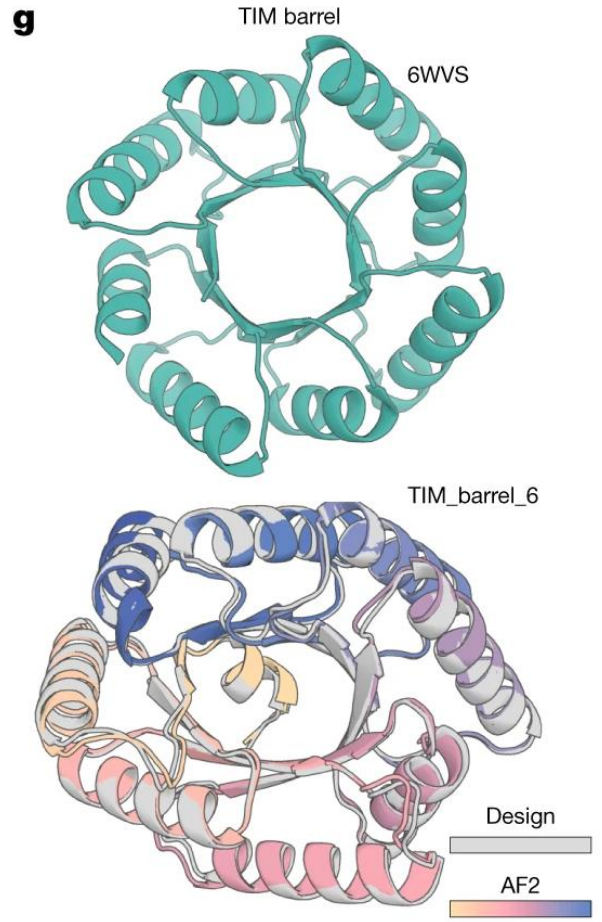
RFdiffusion - Summary

Conditioning of RFdiffusion for specific designs

→ RFdiffusion can be customized to specific design challenges by addition of external potentials and fine-tuning

→ Guiding at each step of the iterative denosing process towards specific design objectives

→ RFdiffusion conditioned with structure of TIM barrel protein 6WVS produced diverse designs with the desired topologies



RFdiffusion can condition on fold information. An example TIM barrel is shown (bottom left), conditioned on the secondary structure and block adjacency of a previously designed TIM barrel, PDB 6WVS. Designs have very similar circular dichroism spectra to PDB 6WVS

Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

RFdiffusion - Summary

Conditioning of RFdiffusion for specific designs

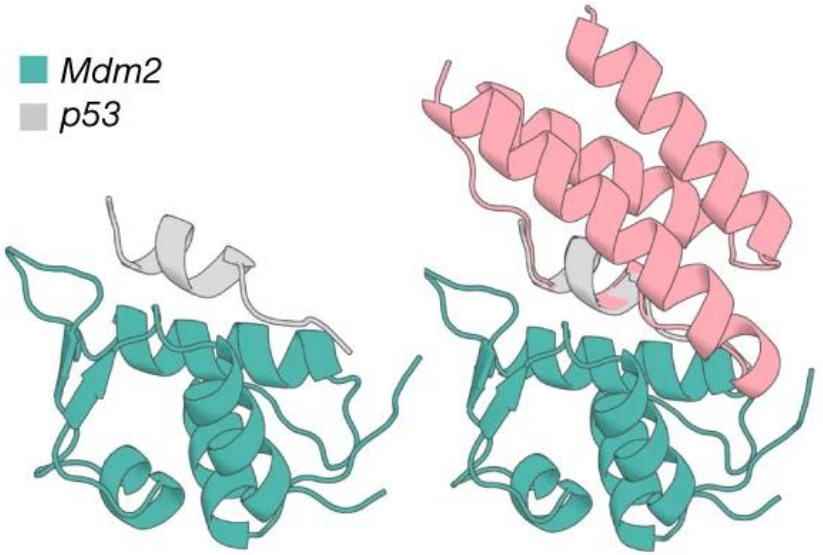
→ RFdiffusion can be customized to specific design challenges by addition of external potentials and fine-tuning

→ Guiding at each step of the iterative denosing process towards specific design objectives

→ RFdiffusion conditioned with structure of TIM barrel protein 6WVS produced diverse designs with the desired topologies

→ RFdiffusion can build scaffolds to hold a desired motif

- In tumours, MDM2 is often overexpressed and binds to p53, preventing p53-mediated cell death
- Designing a competitor protein to bind to MDM2 is attractive
- Take the MDM2-binding helix of the p53 protein into new designs
- RFdiffusion conditioned to include this helix generates designs that strongly bind to MDM2
- Potential cancer drug candidate



RFdiffusion has been used to design a MDM2-binding protein that can compete with p53 and thus prevent p53-MDM2 interaction, which is important for cell-death evasion in many tumours. To incorporate the MDM2-binding helix of p53 in a new design, RFdiffusion was conditioned to include this helix.

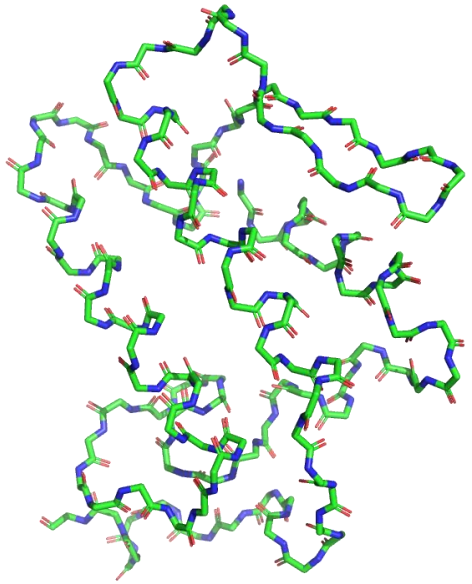
Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

From designed backbones to proteins

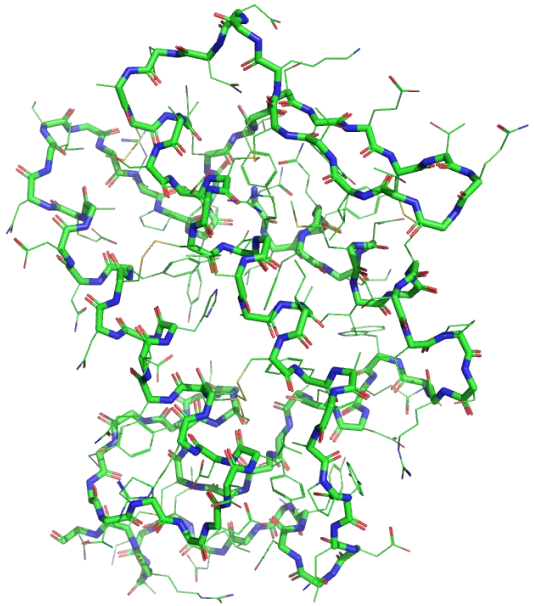
Designed backbones need amino acids

- Protein backbone designs (e.g., using RFdiffusion) do not include specific amino acid sequences
- To complete the protein design, we need to identify an amino acid sequence that will correctly fold into the designed backbone

Backbone Design



Sequence Design

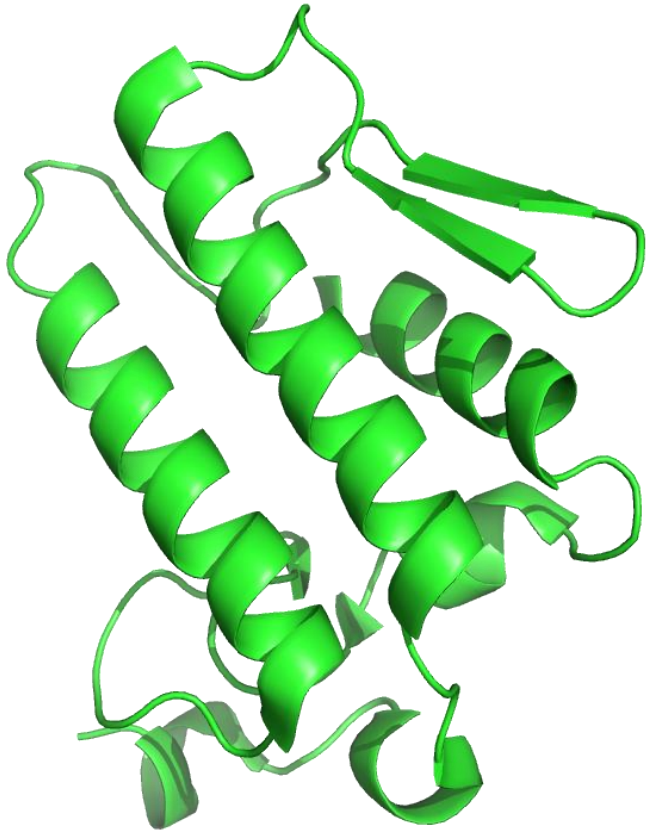


Folding with AlphaFold2
Check if the generated sequence folds into the correct backbone design

ProteinMPNN for sequence design

Generative model for sequence design that integrates graph-based representations of protein structure

- Sequence-from-structure objective: Design a new sequence for a given fixed backbone
- Inverse Folding Problem
- Find a sequence that will fold into the given structure

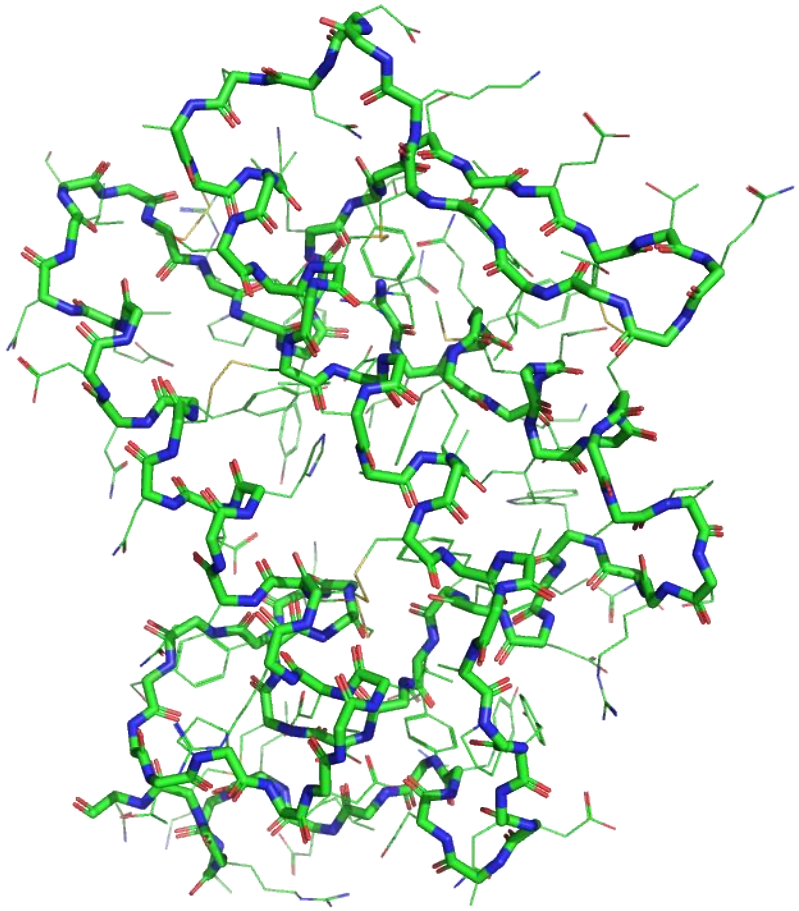


Dauparas, J. et al. Robust deep learning based protein sequence design using ProteinMPNN. bioRxiv 2022.06.03.494563 (2022) doi:10.1101/2022.06.03.494563.

ProteinMPNN for sequence design

Generative model for sequence design that integrates graph-based representations of protein structure

- Sequence-from-structure objective: Design a new sequence for a given fixed backbone
- Inverse Folding Problem
- Find a sequence that will fold into the given structure

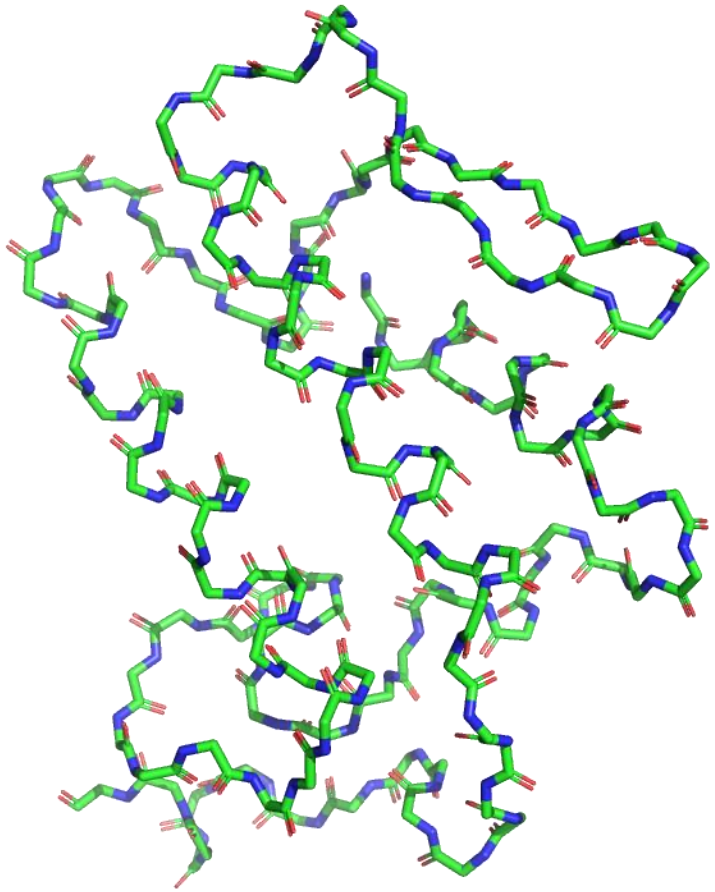


Dauparas, J. et al. Robust deep learning based protein sequence design using ProteinMPNN. bioRxiv 2022.06.03.494563 (2022) doi:10.1101/2022.06.03.494563.

ProteinMPNN for sequence design

Generative model for sequence design that integrates graph-based representations of protein structure

- Sequence-from-structure objective: Design a new sequence for a given fixed backbone
- Inverse Folding Problem
- Find a sequence that will fold into the given structure



Fixed protein structure is modelled as a graph:

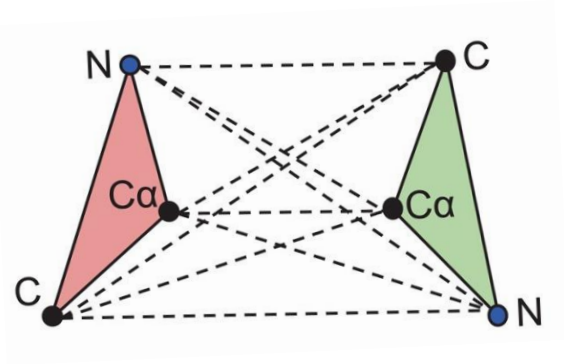
- **Nodes** = amino acids in the fixed structure (specifically the Cα-atom of the unknown amino acid)
- **Edges:** Each node is connected to its 32 nearest neighbor nodes by an edge (each amino acid is connected to its 32 neighboring amino acids)

Dauparas, J. et al. Robust deep learning based protein sequence design using ProteinMPNN. bioRxiv 2022.06.03.494563 (2022) doi:10.1101/2022.06.03.494563.

ProteinMPNN for sequence design

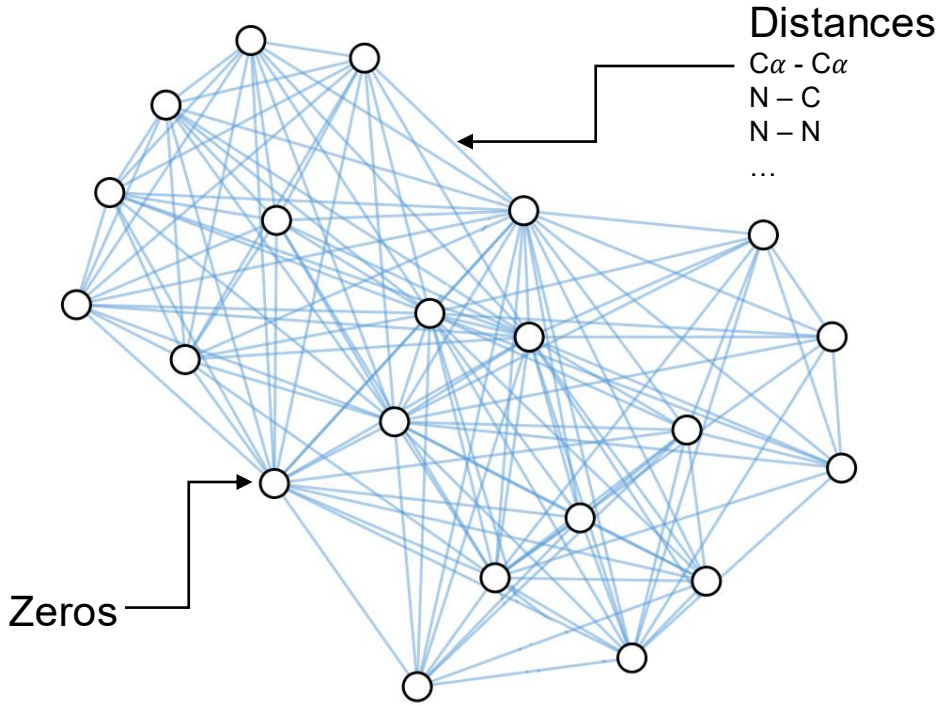
Fixed Protein Structure is modelled as a graph:

- **Node features** = Zeros
- **Edge features** = All distances between the backbone atoms of both amino acids (to encode the spatial orientation of the amino acids)



Encoder: Message-Passing Neural Network

- Initial node features (zeros) and initial edge features (distances) are transformed with graph convolutions
- Information about the spatial environment of a node is encoded in its node features

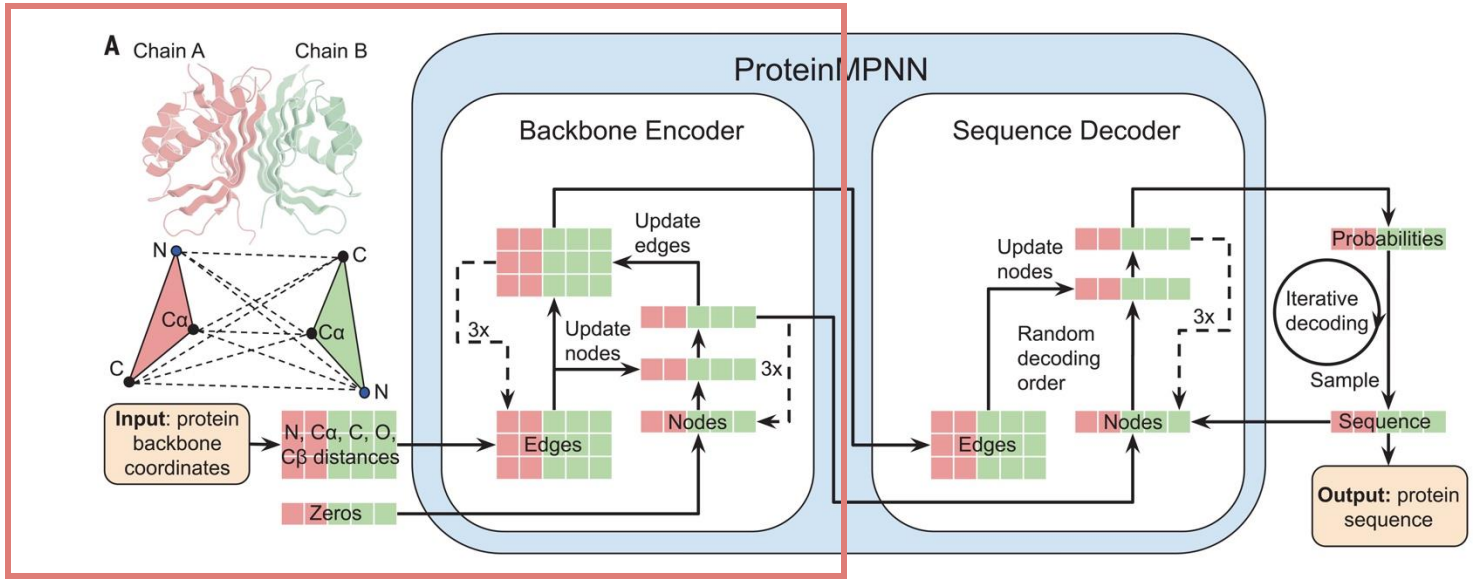


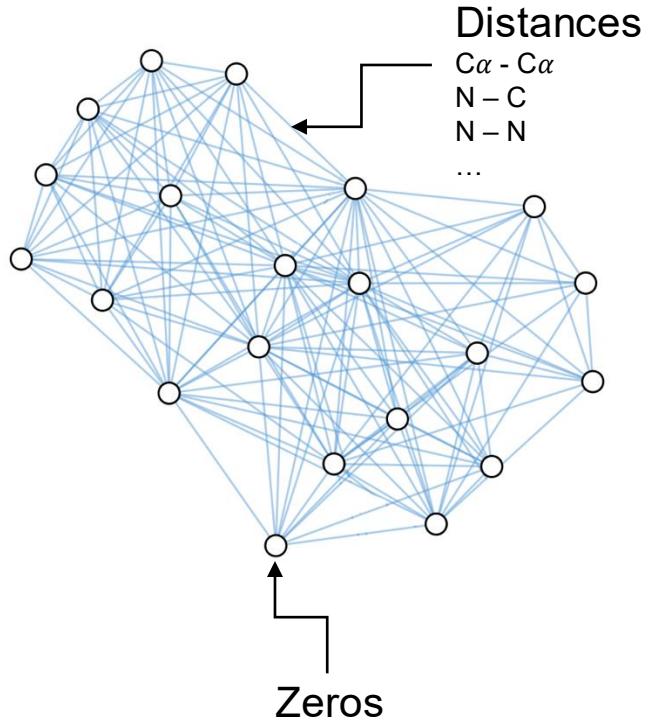
Dauparas, J. et al. Robust deep learning based protein sequence design using ProteinMPNN. bioRxiv 2022.06.03.494563 (2022) doi:10.1101/2022.06.03.494563.

ProteinMPNN for sequence design

Encoder: Message-Passing Neural Network

- Updates the nodes
- Updates the edges
- Repeat 3x





→ Graph Convolution generates new edge and node embeddings

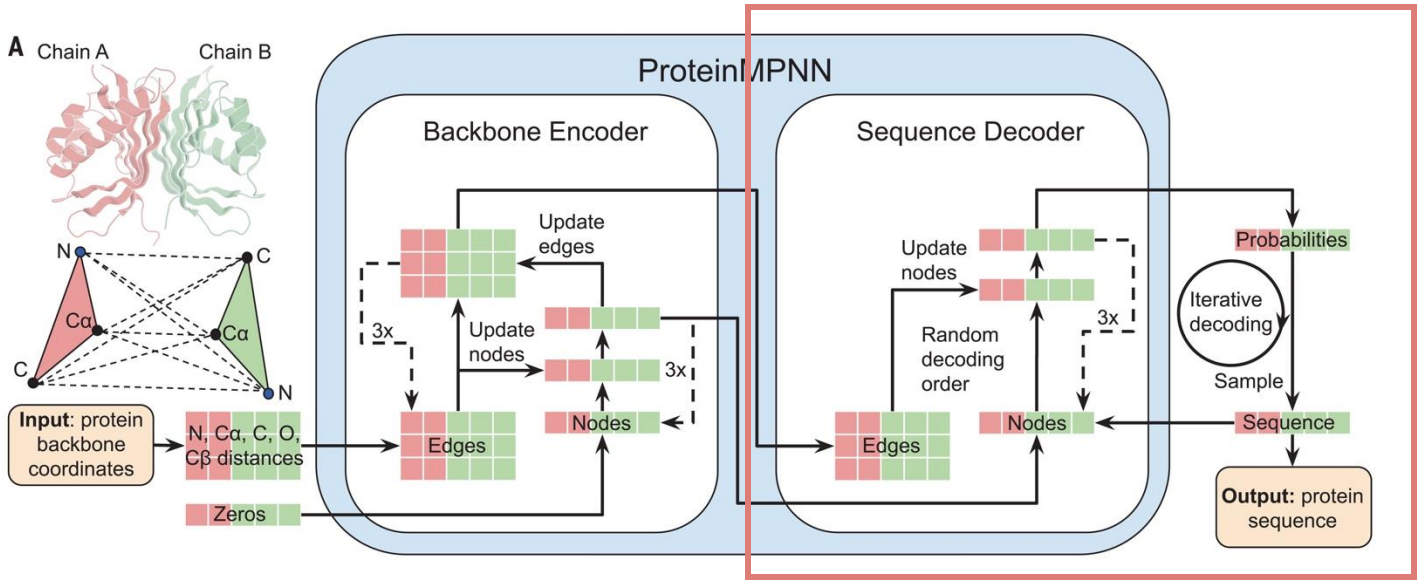
→ Encoding information of the spatial neighborhood of an amino acid

Dauparas, J. et al. Robust deep learning based protein sequence design using ProteinMPNN. bioRxiv 2022.06.03.494563 (2022) doi:10.1101/2022.06.03.494563.

ProteinMPNN for sequence design

Decoder: Transformer Model

- Autoregressive decomposition
- One amino acid at a time – Run decoder as many times as there are amino acids



For each amino acid:

- Input: Updated **edge and node embeddings** and **previously predicted sequence**
- Returns a categorical probability distribution from which the next amino acid is sampled

Dauparas, J. et al. Robust deep learning based protein sequence design using ProteinMPNN. bioRxiv 2022.06.03.494563 (2022) doi:10.1101/2022.06.03.494563.

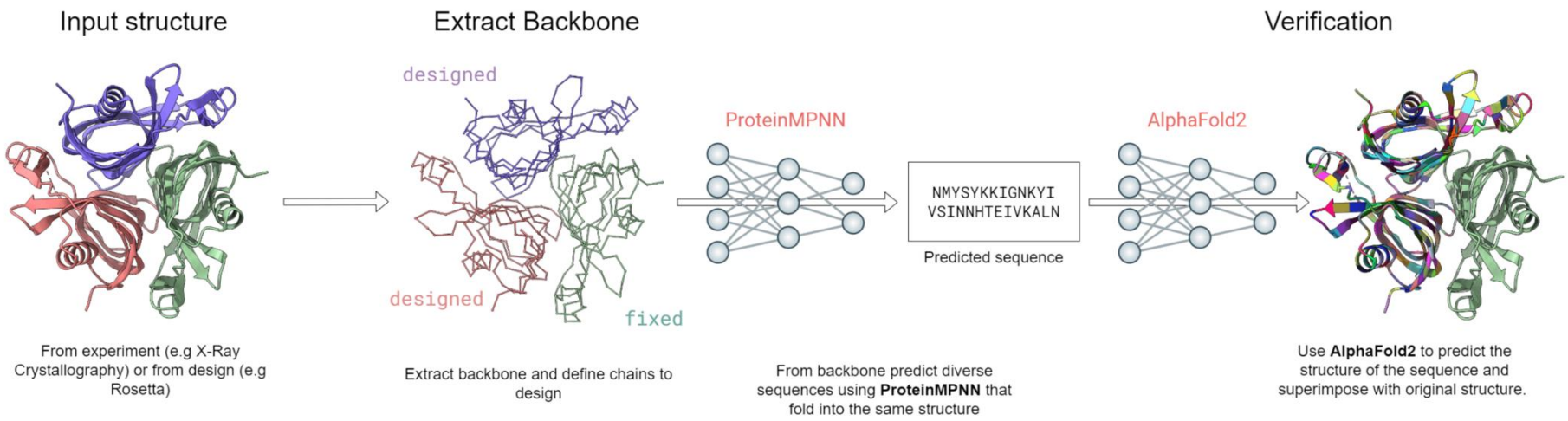
ProteinMPNN for sequence design

Training Data: Proteins with known sequence and experimentally solved structure

- Amino acid side chains are removed and ProteinMPNN is required to rediscovered them
- Loss = Categorical cross entropy per amino acid $CCE = -\sum_{i=1}^{20} p_i \log q_i$

Training Performance:

- Sequence recovery of ≈ 50% (Percentage of correct amino acids recovered from original sequence)
- Structure Validation – Do the predicted sequences fold into the right structure?



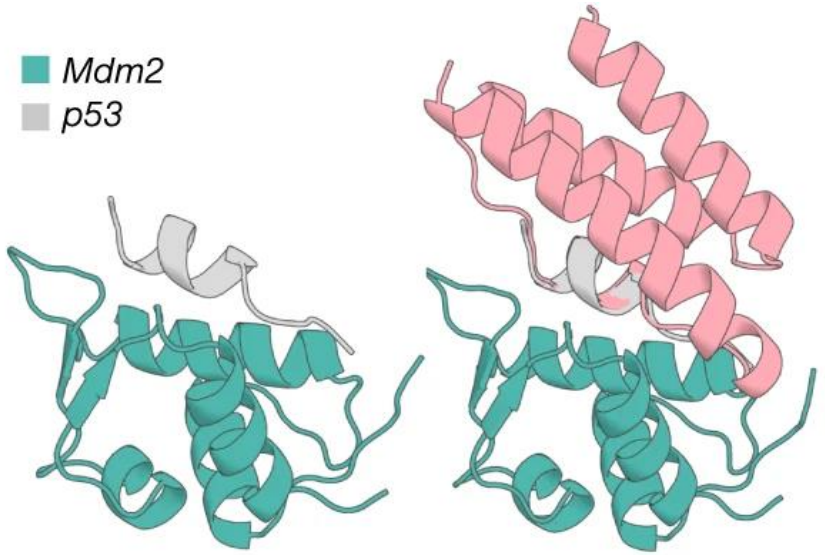
ProteinMPNN - Summary

ProteinMPNN solves sequence design problems, finding sequences that fold into specific backbone structures

- In silico and experimental validation show high accuracy
- Graph-based modelling of 3D backbone structures and MPNN-derived node features are crucial for model performance

Desired motifs can be integrated by keeping a part of the sequence fixed during sequence generation

- For design of MDM2-binding protein with the p53 helix, one could keep the residues of the helix fixed as they are in p53
- ProteinMPNN then designs the sequence around the fixed motif



Sequence design for backbones generated with RFdiffusion: Given backbone coordinates of a MDM2-binding protein generated by RFdiffusion, ProteinMPNN can find sequences that fold into the desired structures. Knowing that the original sequence of the p53-helix binds to MDM2, we can fix these residues during sequence generation with ProteinMPNN, and the model will design a sequence around the fixed amino acids.

Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

Summary Generative AI for *de novo* design

DiffSBDD is a diffusion model to generate novel small-molecule binders for specific binding pockets.

The combination of RFdiffusion and ProteinMPNN is one potential approach for de novo protein structures design

→ **RFdiffusion** is a generative model for protein backbones that generates highly diverse protein backbones from random noise

→ **ProteinMPNN** is a generative model for sequence design, that finds sequences that fold into specific structures

Both models can be conditioned to include desired motifs/topologies, allowing to design proteins for specific applications.

Thank you for your attention