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 AI for *de novo* design**

Introduction

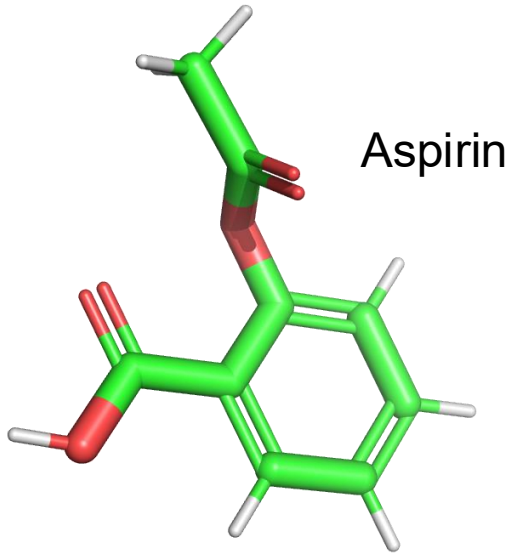
→ Proteins and Small Molecules
→ Engineering of Protein and Small Molecules

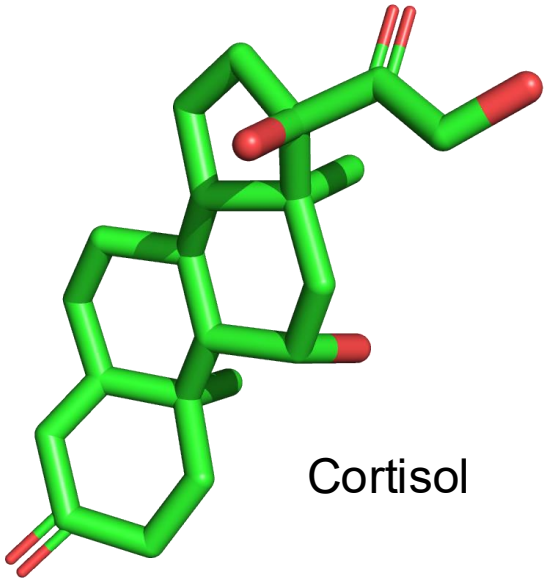
Introduction

Small molecules

- Organic compounds with a low molecular weight, usually less than 100 atoms
- Organisms produce small molecules naturally, as metabolites or signalling molecules (e.g. cortisol)
- Can easily diffuse across cell membranes due to their small size
- Often interact with proteins, influencing their function
- Most drugs are small molecules

- Their structure usually allows for **easy synthesis in the laboratory**


Aspirin

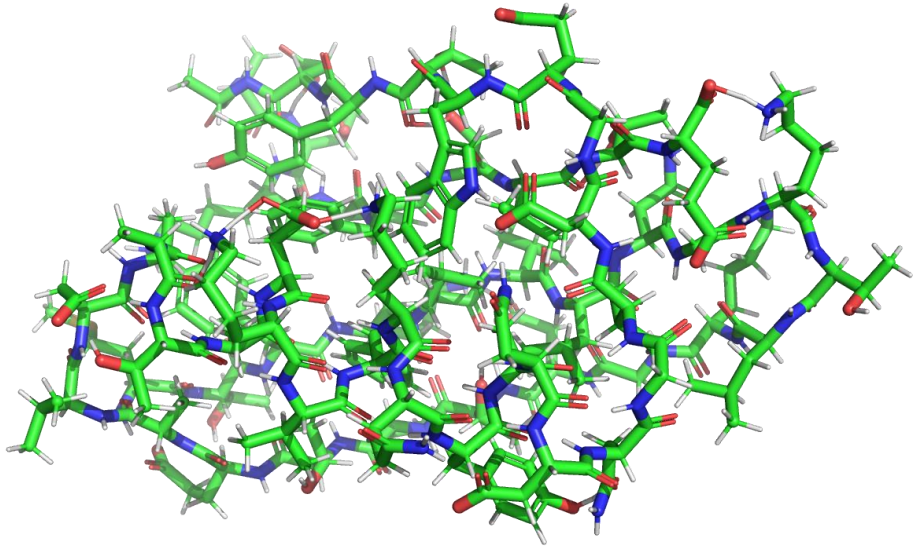

Cortisol

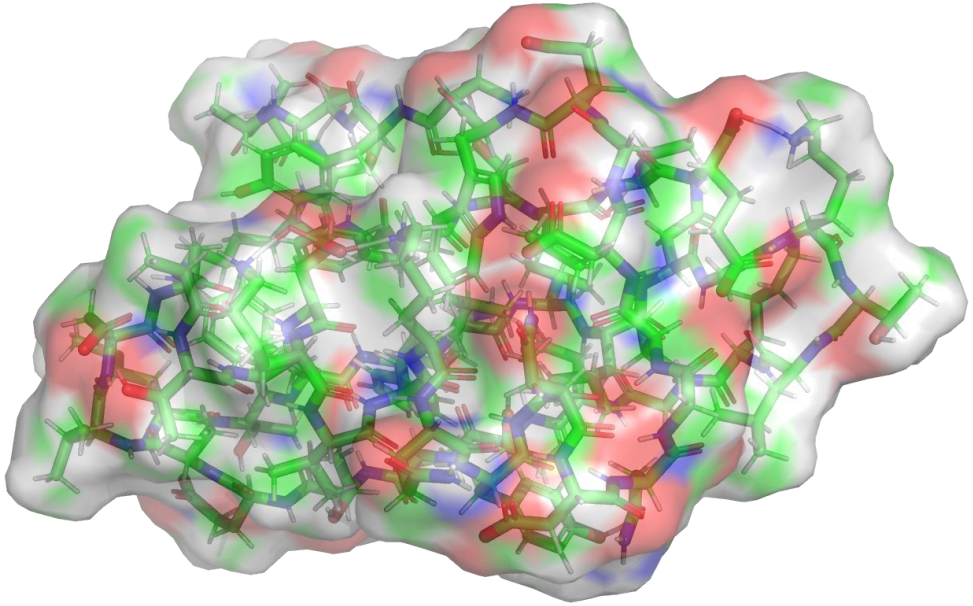
Introduction

Proteins

- Large, chain-like organic molecules made of a fixed set of building blocks
- These chains fold into a defined 3D structure known as the native conformation
- Proteins have many different functions
  - Enzymes catalyse chemical reactions
  - Structural components
  - Signalling molecules
  - Transporters

The building blocks of cells and play key roles in almost every function necessary for life





Proteins

Each protein is a linear chain of amino acids

There are twenty standard amino acids, which are the fundamental building blocks of all proteins

Backbone
Identical for all amino acids (the links of the chain)

- Amino group ($-H_2N$)
- Carboxyl group (-COOH)

Side chains
Varies between amino acids and determines their chemical properties

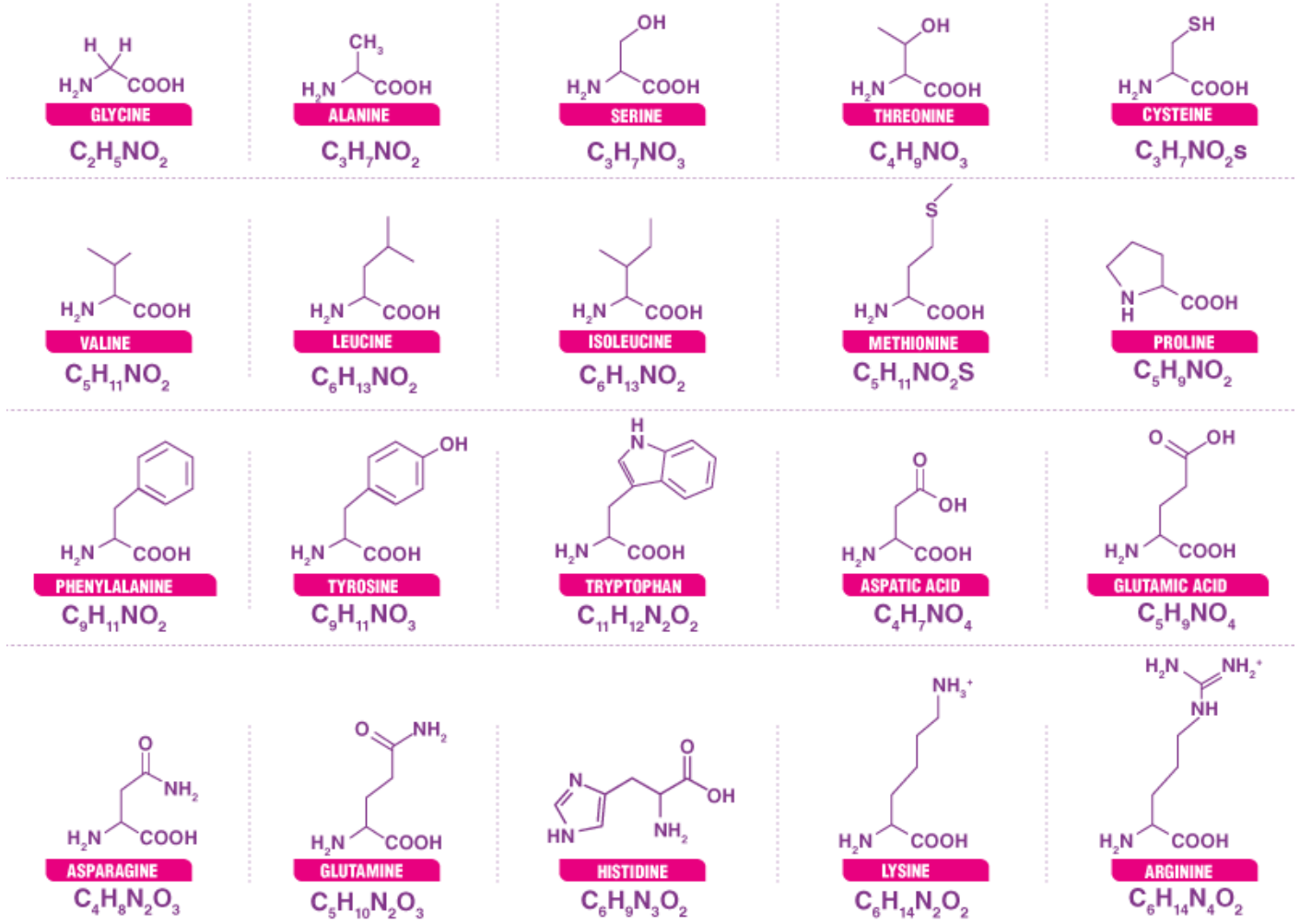


Image Credits: https://byjus.com/biology/amino-acids/

From DNA to Proteins

The genetic code of an organism dictates the precise sequence of amino acids in proteins

- Each triplet of DNA nucleotides (codons) encodes for a specific amino acid

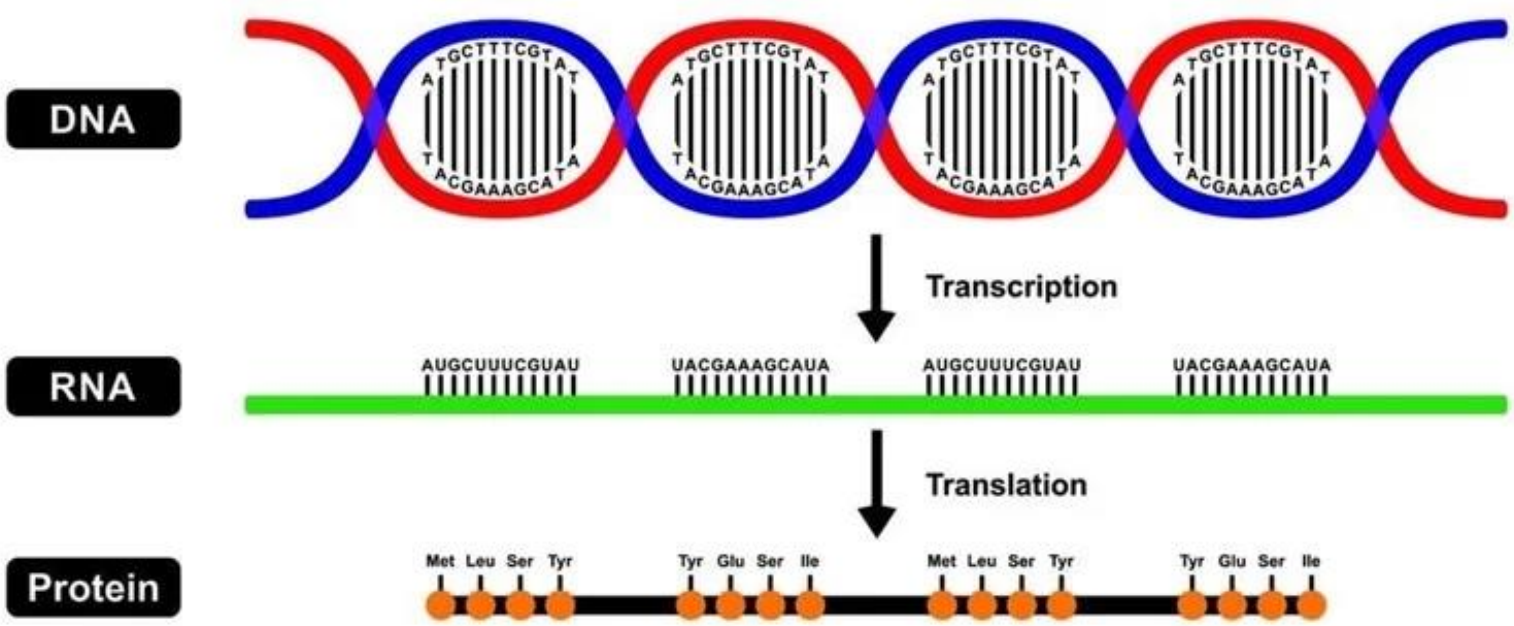
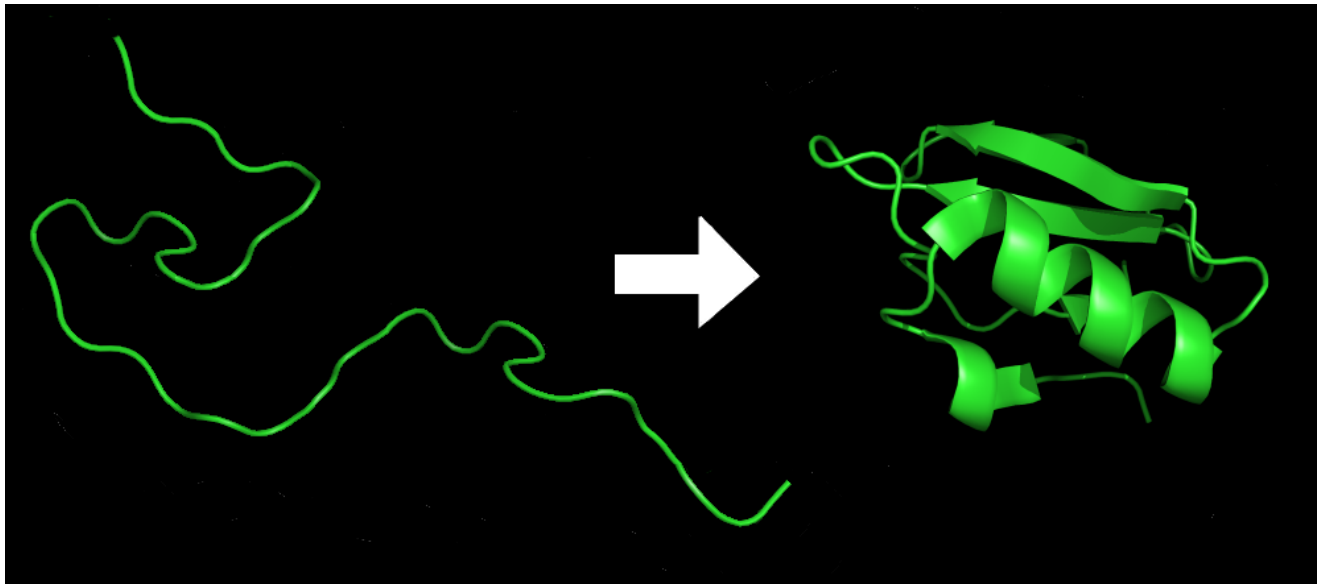


Image Credits: https://www.azolifesciences.com/article/From-DNA-to-Protein3b-The-Central-Dogma-of-Molecular-Biology.aspx

Protein Folding

After synthesis, the linear chain of amino acids folds into a more ordered three-dimensional structure, which defined the protein's biological function



- The folded structure is defined by the sequence of amino acids
- Folding is a spontaneous process of energy minimization that is guided by interactions between the amino acid such as hydrophobic interactions, hydrogen bonds and van der Waals forces

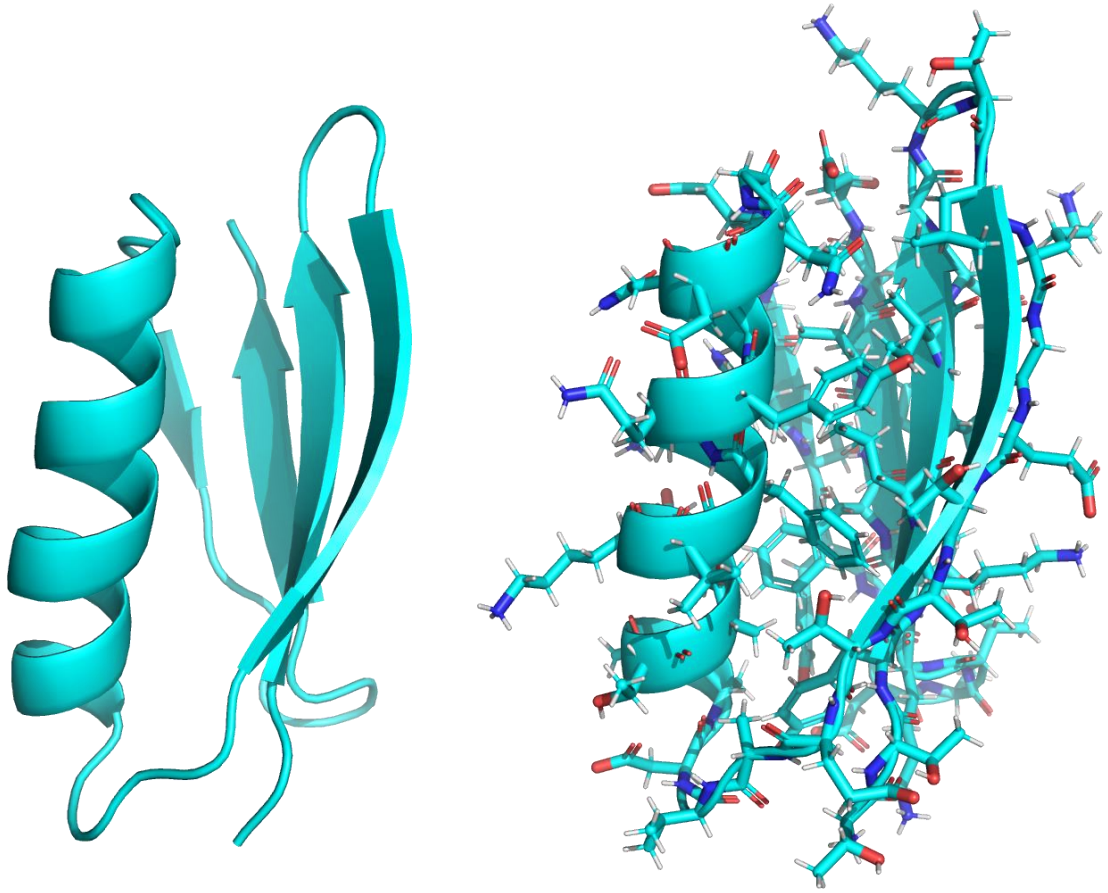
Image Credits: https://en.wikipedia.org/wiki/Protein_folding

Protein Structure

Primary Structure
The sequence of amino acids in the chain

Secondary Structure
Regularly repeating local structures stabilized by hydrogen bonds. The most common examples are the α-helix and the β-sheet

Tertiary Structure
The spatial relationship of the secondary structures to one another, which defines the overall shape of the protein molecule



Protein Structure: Protein 3D-structures are often depicted in a simplified cartoon format, which shows the location of the backbone of the amino acid chain and the presence of characteristic secondary structures like alpha-helices and beta-sheets. The side chains of the amino acids are not visible.

Introduction

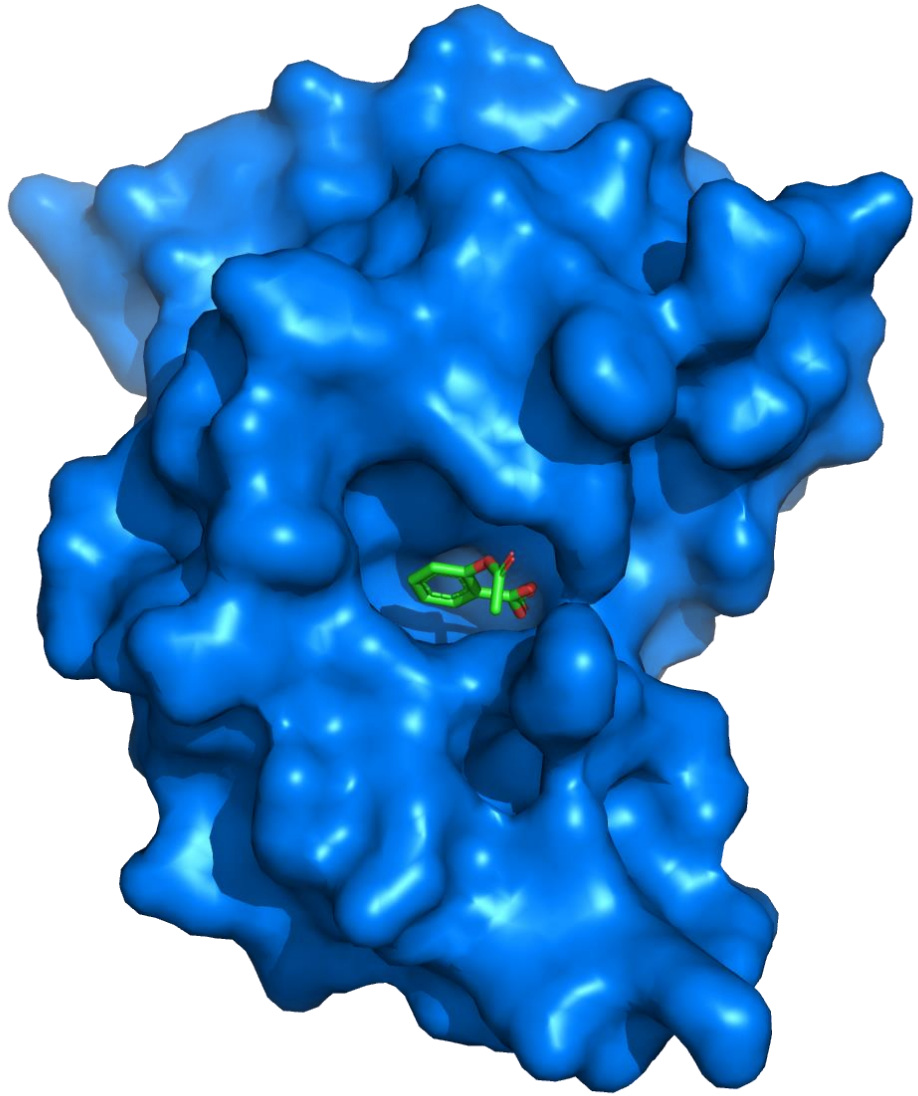
→ Proteins and Small Molecules
→ Engineering Proteins and Small Molecules

Engineering of Small Molecules

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
  - Inhibit enzymes
  - Block receptors

→ **Engineering of Small Molecules:**
Altering the size and composition of the organic compound to get desired properties (e.g. different carbon backbone and change functional groups)



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)

Engineering of Proteins

Altering the amino acid sequence of proteins to achieve specific functions or properties

Biocatalysis
Example: Change amino acid sequence of an enzyme to accept a new substrate

Therapeutics
Example: Change amino acid sequence of an antibody to bind to SARS-CoV-2 spike protein

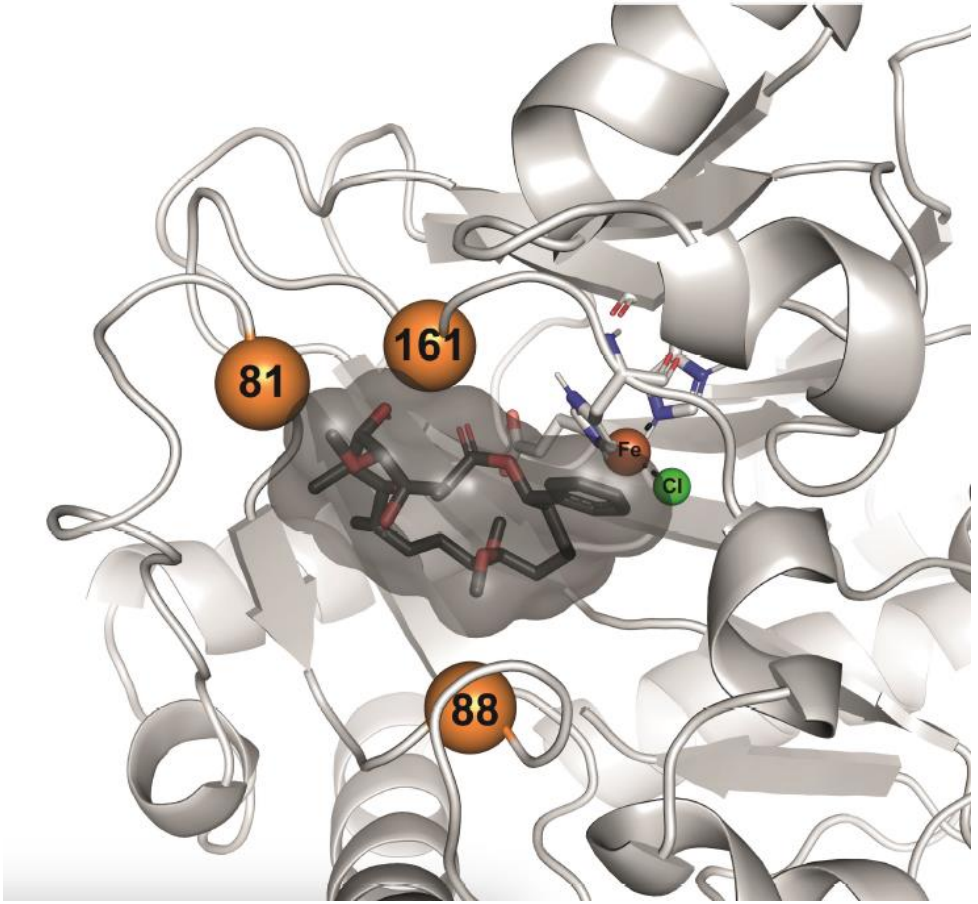


Image Credits: Büchler, J. *et al.* Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens. *Nat Commun* **13**, 371 (2022).

Strategies for Engineering of Proteins

(Semi-)Rational Design:

- Use detailed knowledge of the structure of a protein to make desired changes
- Generate combinatorial libraries of a protein with varying amino acids at a specific location

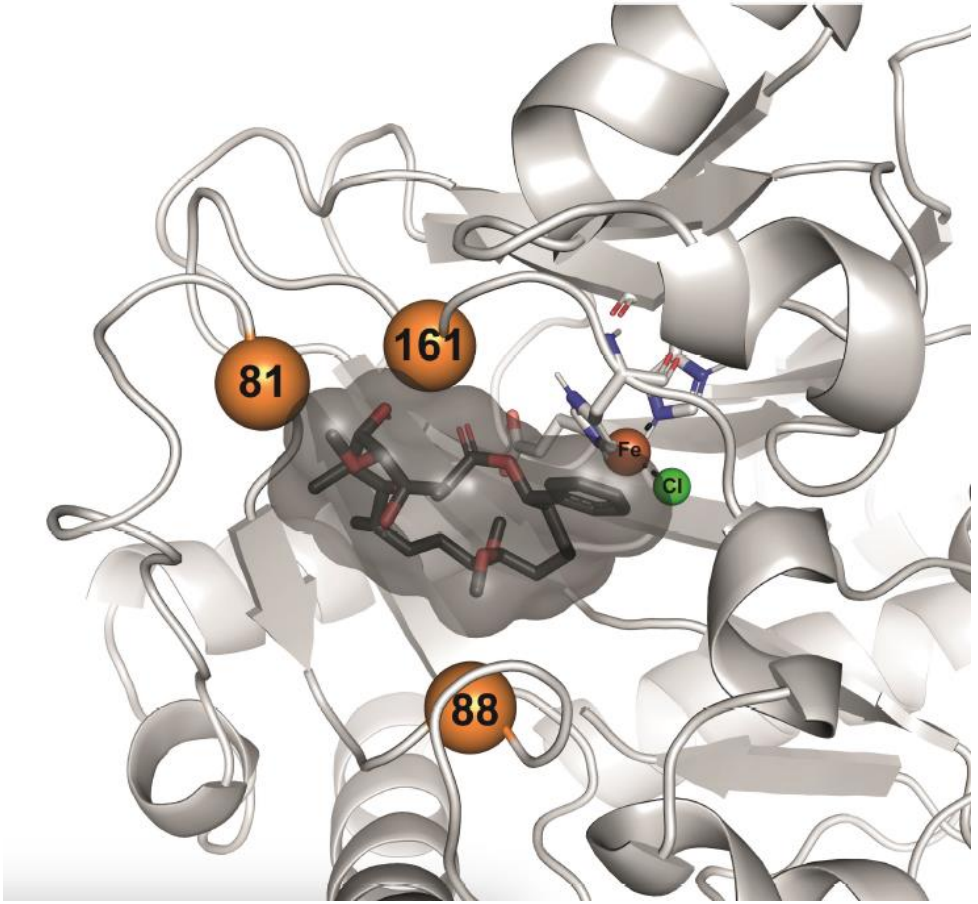


Image Credits: Büchler, J. *et al.* Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens. *Nat Commun* **13**, 371 (2022).

Strategies for Engineering of Proteins

(Semi-)Rational Design:

- Use detailed knowledge of the structure of a protein to make desired changes
- Generate combinatorial libraries of a protein with varying amino acids at a specific location

Directed Evolution:

- Cycles of introducing random mutations and selecting most performant variant from the pool of mutants

Very time consuming and expensive and are only suited for optimizing existing proteins

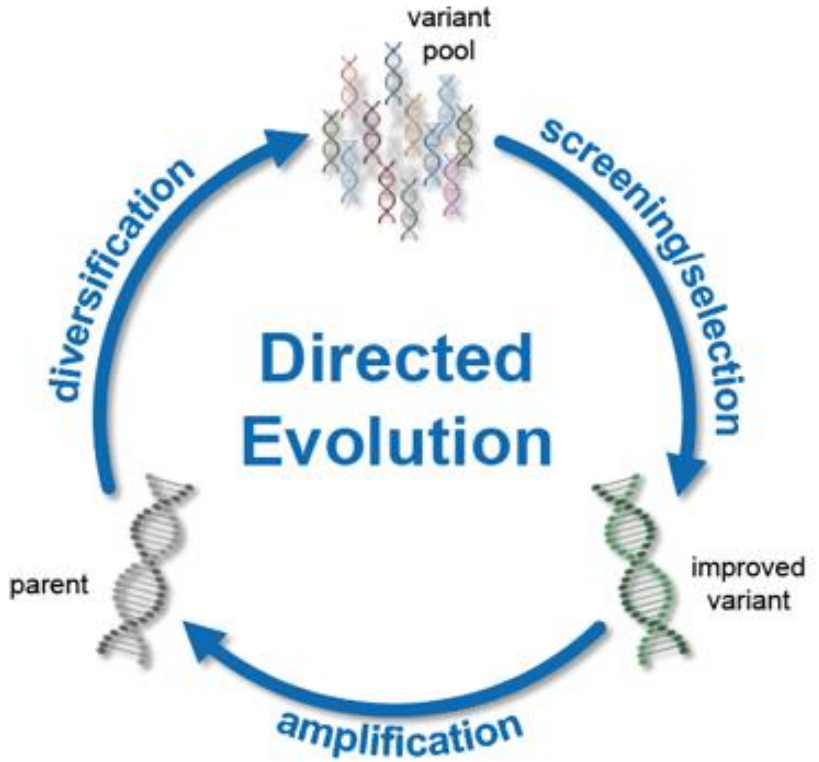


Image Credits: Bioprocess Laboratory, ETH Zürich (https://bsse.ethz.ch/bpl/research/directed-evolution.html)

Combinatorial Search Space

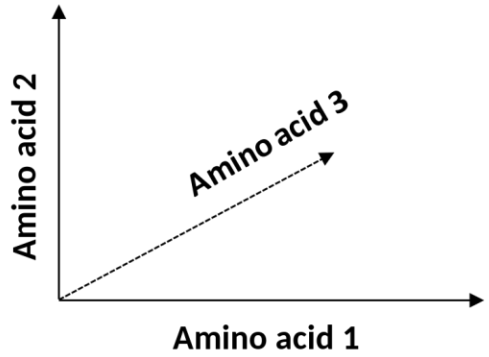
The search space for optimizing amino acid sequences in protein engineering is extremely large

Considering that there are 20 naturally occurring amino acids, each of which can occupy any position within the protein sequence, a protein of just 100 amino acids has 20^{100} possible sequences

All dipeptides in 2D space = 400

Amino acid 2 \ Amino acid 1	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	AA	RA	NA	DA	CA	EA	QA	GA	HA	IA	LA	KA	MA	FA	PA	SA	TA	WA	YA	VA
Arg	AR	RR	NR	DR	CR	ER	QR	GR	HR	IR	LR	KR	MR	FR	PR	SR	TR	WR	YR	VR
Asn	AN	RN	NN	DN	CN	EN	QN	GN	HN	IN	LN	KN	MN	FN	PN	SN	TN	WN	YN	VN
Asp	AD	RD	ND	DD	CD	ED	QD	GD	HD	ID	LD	KD	MD	FD	PD	SD	TD	WD	YD	VD
Cys	AC	RC	NC	DC	CC	EC	QC	GC	HC	IC	LC	KC	MC	FC	PC	SC	TC	WC	YC	VC
Glu	AE	RE	NE	DE	CE	EE	QE	GE	HE	IE	LE	KE	ME	FE	PE	SE	TE	WE	YE	VE
Gln	AQ	RQ	NQ	DQ	CQ	EQ	QQ	GQ	HQ	IQ	LQ	KQ	MQ	FQ	PQ	SQ	TQ	WQ	YQ	VQ
Gly	AG	RG	NG	DG	CG	EG	QG	GG	HG	IG	LG	KG	MG	FG	PG	SG	TG	WG	YG	VG
His	AH	RH	NH	DH	CH	EH	QH	GH	HH	IH	LH	KH	MH	FH	PH	SH	TH	WH	YH	VH
Ile	AI	RI	NI	DI	CI	EI	QI	GI	HI	II	LI	KI	MI	FI	PI	SI	TI	WI	YI	VI
Leu	AL	RL	NL	DL	CL	EL	QL	GL	HL	IL	LL	KL	ML	FL	PL	SL	TL	WL	YL	VL
Lys	AK	RK	NK	DK	CK	EK	QK	GK	HK	IK	LK	KK	MK	FK	PK	SK	TK	WK	YK	VK
Met	AM	RM	NM	DM	CM	EM	QM	GM	HM	IM	LM	KM	MM	FM	PM	SM	TM	WM	YM	VM
Phe	AF	RF	NF	DF	CF	EF	QF	GF	HF	IF	LF	KF	MF	FF	PF	SF	TF	WF	YF	VF
Pro	AP	RP	NP	DP	CP	EP	QP	GP	HP	IP	LP	KP	MP	FP	PP	SP	TP	WP	YP	VP
Ser	AS	RS	NS	DS	CS	ES	QS	GS	HS	IS	LS	KS	MS	FS	PS	SS	TS	WS	YS	VS
Thr	AT	RT	NT	DT	CT	ET	QT	GT	HT	IT	LT	KT	MT	FT	PT	ST	TT	WT	YT	VT
Trp	AW	RW	NW	DW	CW	EW	QW	GW	HW	IW	LW	KW	MW	FW	PW	SW	TW	WW	YW	VW
Tyr	AY	RY	NY	DY	CY	EY	QY	GY	HY	IY	LY	KY	MY	FY	PY	SY	TY	WY	YY	VY
Val	AV	RV	NV	DV	CV	EV	QV	GV	HV	IV	LV	KV	MV	FV	PV	SV	TV	WV	YV	VV

All tripeptides in 3D space = 8000



All 10 a.a proteins in 10D space ≈ 10^{13}

All 50 a.a proteins in 50D space ≈ 10^{65}

All 100 a.a proteins in 100D space ≈ 10^{130}

All 300 a.a proteins in 300D space ≈ 10^{390}

Image Credits: https://en.wikipedia.org/wiki/Sequence_space_%28evolution%29

Sequence-based models

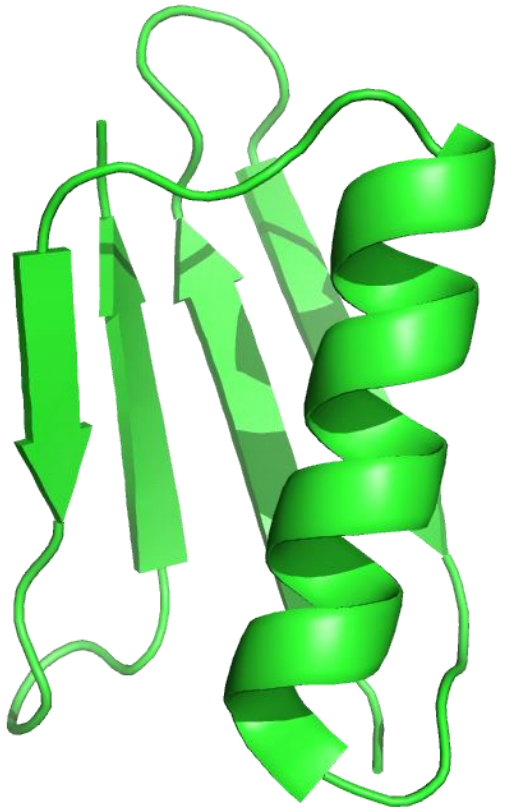
→ Sequence representations of proteins and small molecules
→ Examples of Machine Learning Models

Proteins: sequence representation

Amino Acid Sequences: A linear sequence of letters representing the order of amino acids in the protein

For machine learning, amino acid sequences are encoded into numerical representations by

- One-Hot-Encoding
- Tokenization - Each amino acid is treated as a separate token, like words in text processing.



MTYKLILNGKTLKGETTTEAVDAATAEKVF
KQYANDNGVDGEWTYDAATKTFTVTE

Small molecules: sequence representation

SMILES (Simplified Molecular Input Line Entry System)

- A linear notation representing the molecular structure using ASCII characters
- Atoms are represented by their chemical symbols (C, N, O, P, S, F, Cl, Br, I)
- Single bonds are omitted; double bonds are denoted by “=“
- Branches are described using parentheses
- Rings are noted by using numbers to indicate the connection points

For machine learning, SMILES codes are tokenized and mapped to a fixed vocabulary.

Image Credits: https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

Natural Language Processing in Chemistry and Biology

Analogy between the building blocks of language and those of proteins & small molecules

Proteins: Like natural language, protein sequences have their own kind of grammar and contain long-range dependencies

- Amino acids ≈ words
- Sequences of amino acids ≈ sentences
- Proteins ≈ text paragraph

Small molecules: SMILES codes are also linear sequences that consist of a fixed vocabulary of symbols with patterns and rules

→ **Both protein sequences and SMILES codes are compatible with language model architectures designed to handle sequences**

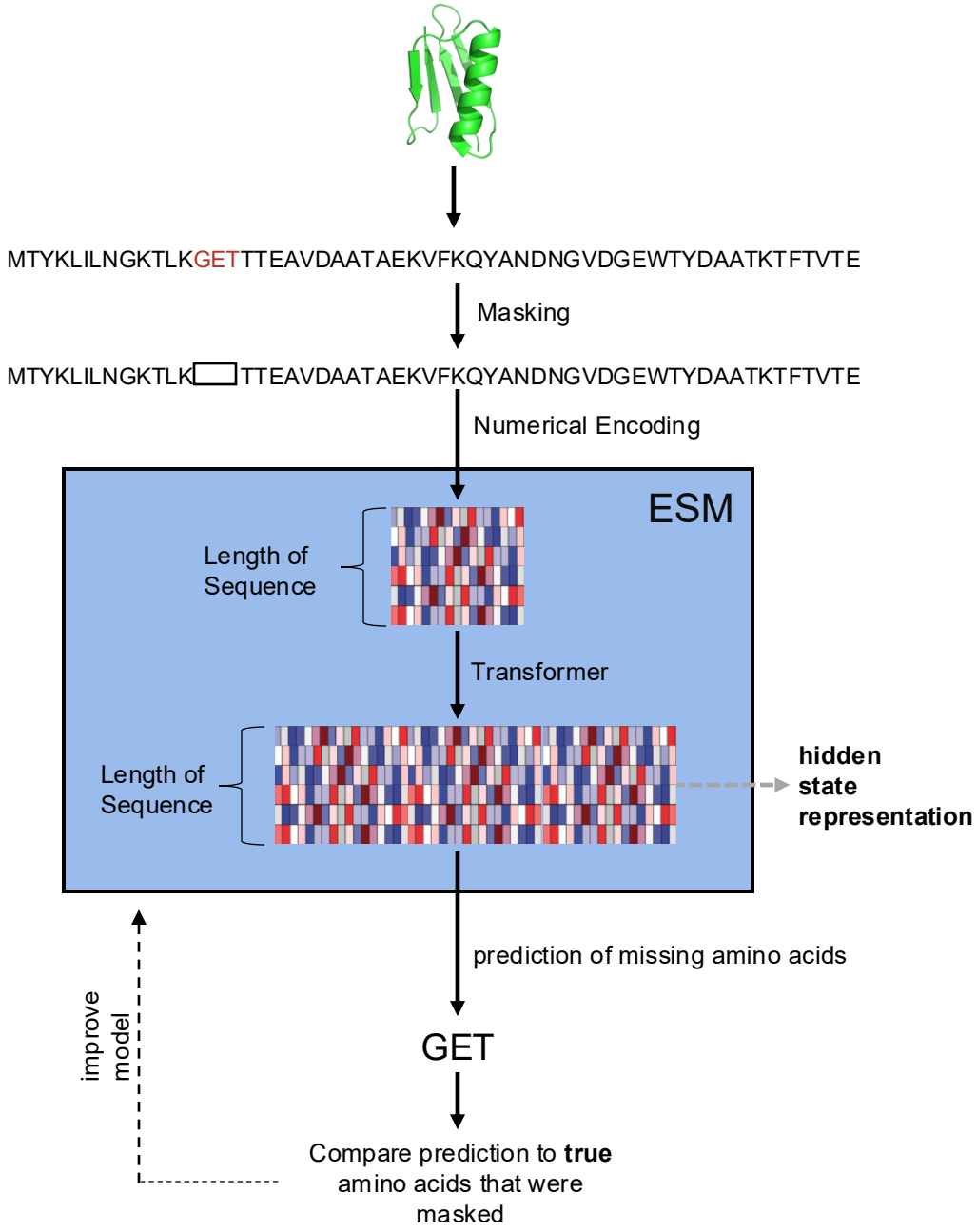
→ **Methods of Natural Language Processing (NLP) are applied to proteins & small molecules**

Evolutionary Scale Modelling (ESM)

Self-supervised transformer language model

- Self-supervised learning with unlabelled datasets, far larger amounts of data available
- Trained with the masked language modelling objective on 250 million protein sequences
- Each input sequence is corrupted by replacing a fraction of the amino acids with a mask
- ESM transformer generates prediction for the masked amino acids from the sequence context

→ **Knowledge of intrinsic biological properties of the proteins emerges without supervision, encoded in the hidden representations of the language model**



Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**, e2016239118 (2021).

Evolutionary Scale Modelling (ESM)

Self-supervised transformer language model

- Self-supervised learning with unlabelled datasets, far larger amounts of data available
- Trained with the masked language modelling objective on 250 million protein sequences
- Each input sequence is corrupted by replacing a fraction of the amino acids with a mask
- ESM transformer generates prediction for the masked amino acids from the sequence context

→ **Knowledge of intrinsic biological properties of the proteins emerges without supervision, encoded in the hidden representations of the language model**

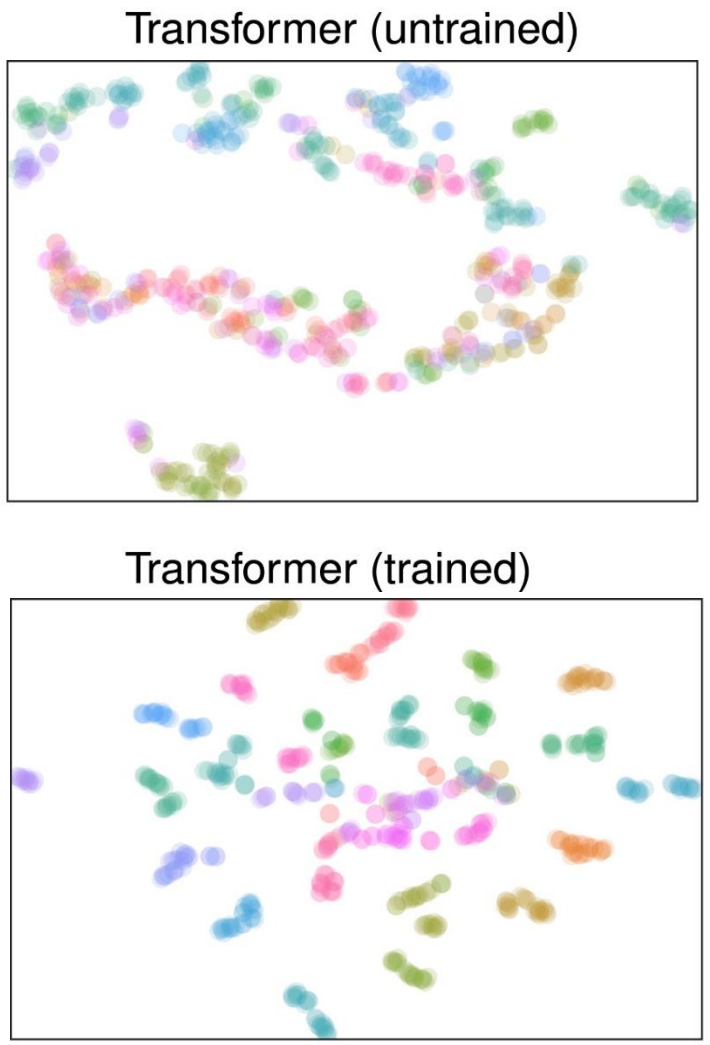


Figure: Projection of hidden representations of trained and untrained transformer model to 2D with t-distributed stochastic neighbor embedding (t-SNE). Each point represents a gene and is colored by the orthologous group it belongs to. Orthologous groups of genes are densely clustered in the trained representation space. By contrast, the untrained representation space does not reflect strong organization by evolutionary relationships.

Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**, e2016239118 (2021).

Downstream Applications & Transfer Learning

Transfer Learning – Applying a model trained on one task to a related but slightly different task can improve performance, especially when annotated data for the second task are scarce.

Fine-tuning protein language models on specific downstream tasks
Using supervised learning with (potentially small) labelled datasets to fine-tune the language model for specific applications

- Protein function prediction
- Prediction of mutational effects
- Prediction of protein stability
- Prediction of protein 3D structure

→ **Pre-trained transformer representations are employed in many downstream tasks**

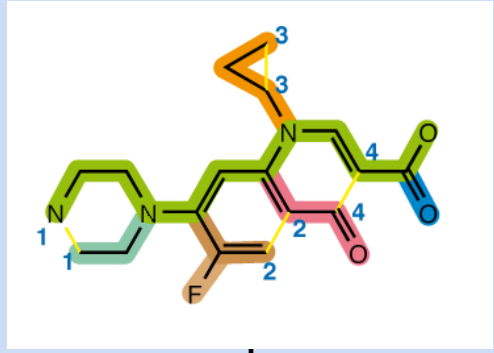
Beyond today's scope

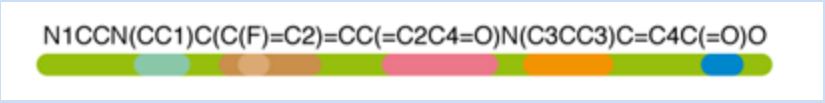
Many chemistry language models have been trained with the MLM-objective on SMILES strings

Up to 1 billion unlabelled molecules from the PubChem and ZINC datasets

Demonstrated chemical learning:

- Recover masked SMILES tokens
- Cluster molecules by chemical features
- Improve downstream molecular property prediction performance





Examples:

MolFormer (IBM Research)

ChemBERTa

Protein Structure Prediction

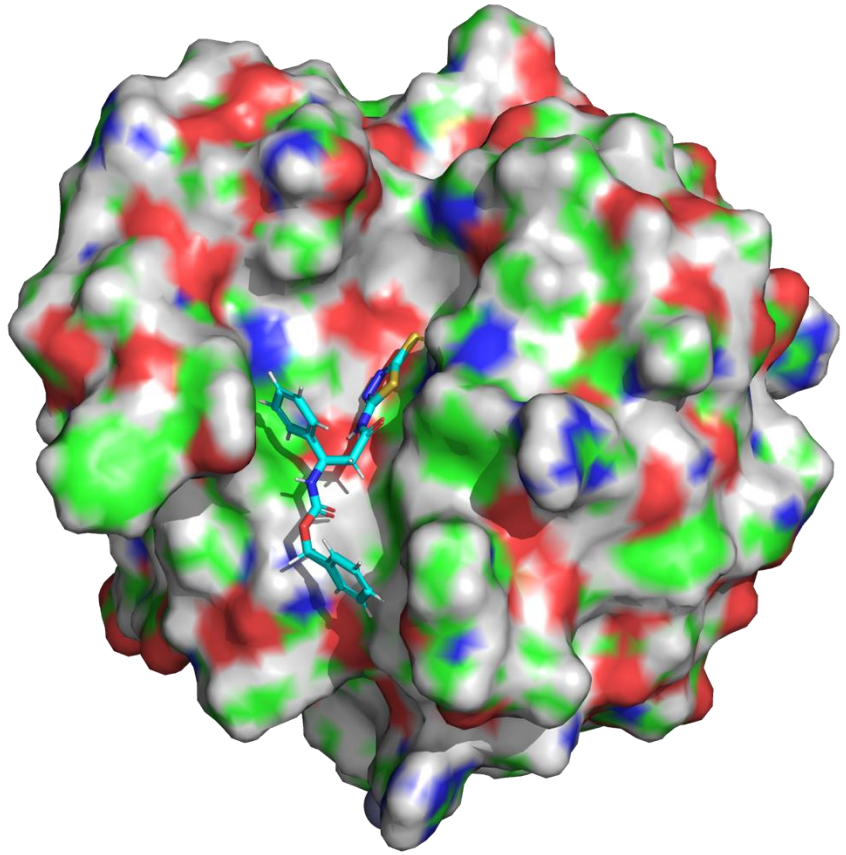
→ AlphaFold2
→ Evolutionary Scale Modelling **2** (ESM**2**)
→ Modelling of Molecular Assemblies

Protein Structure

→ 3D coordinates of atoms of the protein in folded state

Why Protein Structure is important:

- Fundamental understanding:
  The structure of a protein defines its function. Knowing the structure of a protein helps to find its biological role
- Drug discovery:
  Knowing protein structure allows for the identification of active sites and interaction points for potential drug molecules



How to get a protein structure:
For many years, researchers have relied on very time-consuming experimental methods, such as X-ray Crystallography and Nuclear Magnetic Resonance Spectroscopy (NMR) to find the structure of a protein with a given sequence.

→ **Recently, Deep Learning is used to predict protein structures**

The protein folding task

Predicting the 3D fold of a protein from its amino acid sequence is a complex and challenging task

- Amino acid residues can adopt many conformations due to the rotation around its backbone bonds and side chains
- Astronomical number of possible structures even for small protein
- Folding is driven by various interactions between amino acids which are difficult to model (e.g. hydrogen bonds, hydrophobic interactions, van der Waals forces etc.)
- Long-range dependencies: Folding is influenced by interactions between amino acids that are far apart in the sequence but come close in the 3D structure
- Training data is relatively scarce

→ **Critical Assessment of Protein Structure Prediction (CASP)**

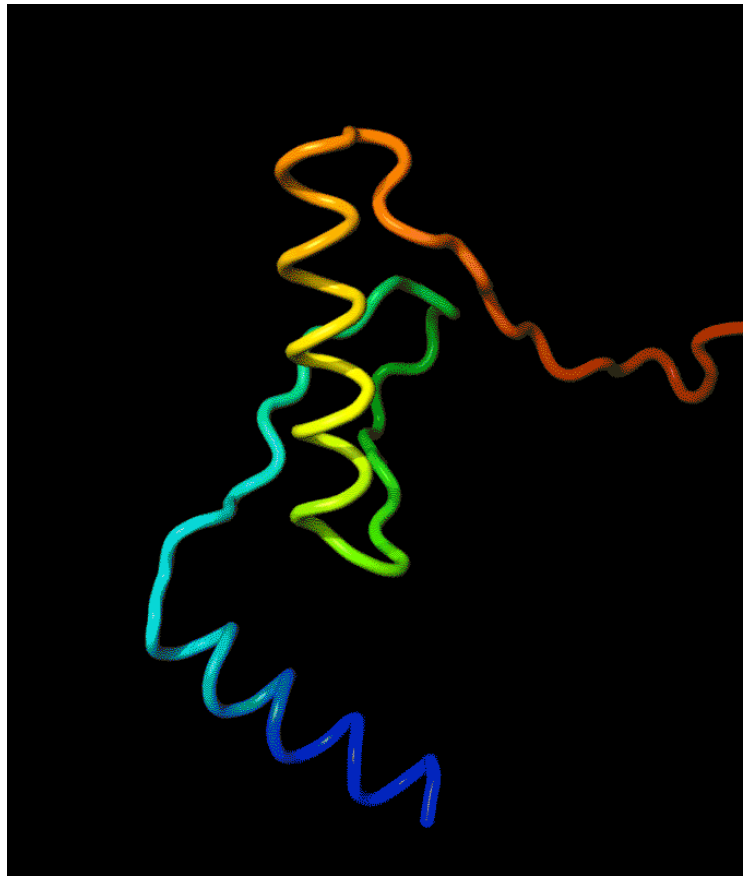


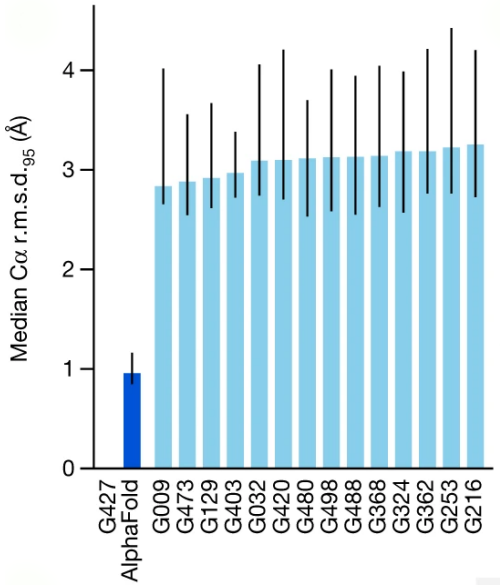
Image Credits: Deepmind

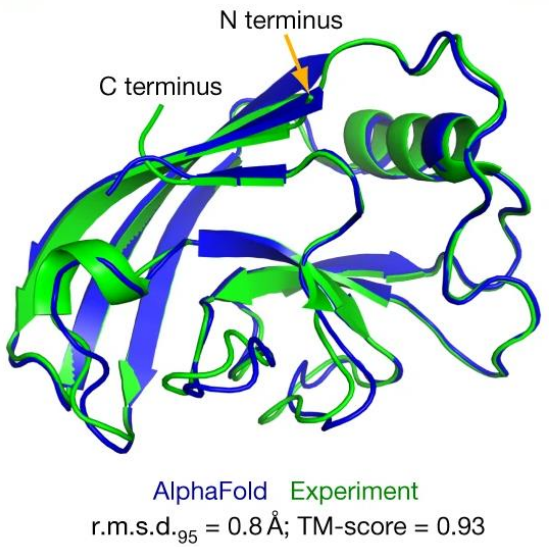
AlphaFold2 for protein structure prediction

First model demonstrating near-experimental accuracy in the task of protein structure prediction

- Winner of the 14th Critical Assessment of Protein Structure Prediction (CASP) challenge
- Predictions showed root-mean-square deviation of Cα-atoms of 0.96Å compared to experimental structures
- To compare: width of carbon atom = 1.4Å

→ **Breakthrough in the field of structural bioinformatics**





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

AlphaFold2 model architecture

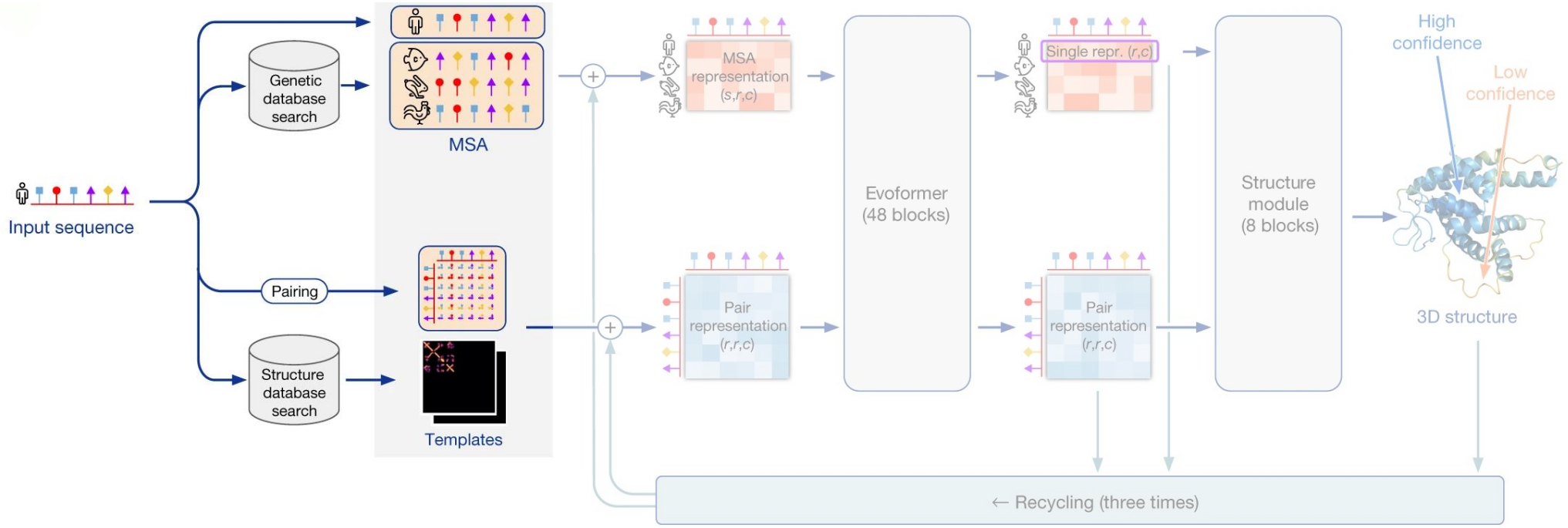


Figure: AlphaFold model architecture. Arrows show the information flow among the various components. Array shapes are shown in parentheses with s, number of sequences, r, number of residues and c, number of channels

Input to the model

- Input amino acid sequence
- Sequences from evolutionarily related proteins in the form of a multiple sequence alignment
- 3D atom coordinates of homologous structures (templates)

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

AlphaFold2 model architecture

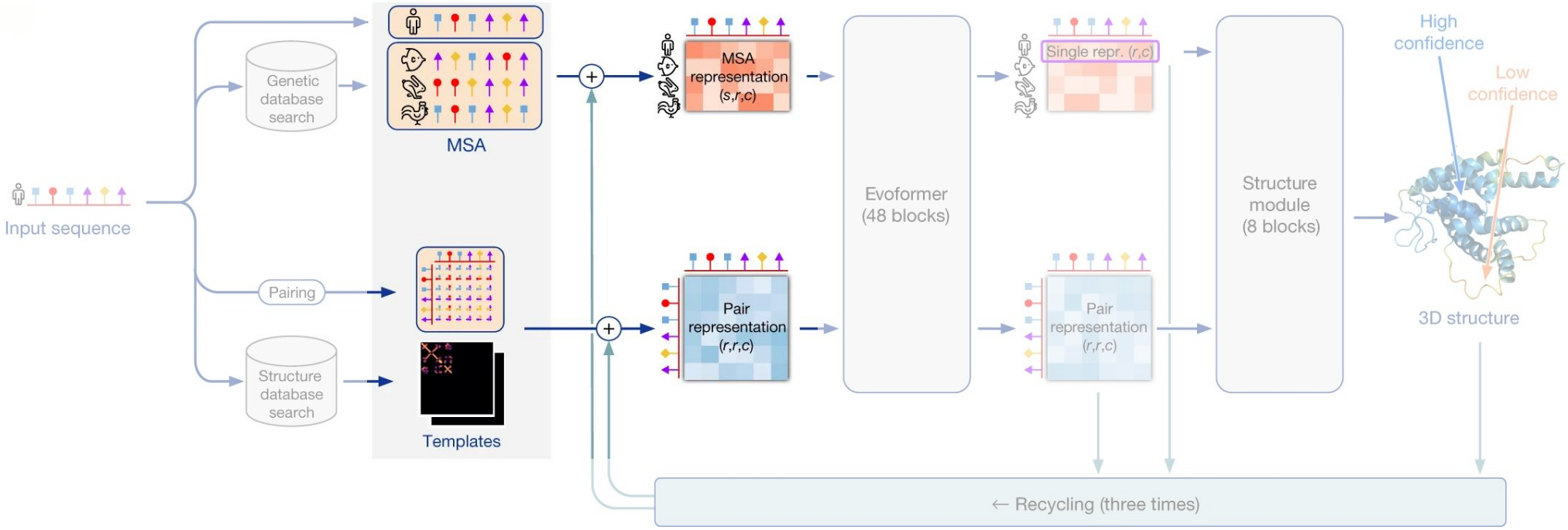


Figure: AlphaFold model architecture. Arrows show the information flow among the various components. Array shapes are shown in parentheses with s, number of sequences, r, number of residues and c, number of channels

Initial representations

- **MSA representation:** The input sequence and the MSA are translated into a MSA representation containing sequence conservation information
- **Pair representation:** Based on analysis of known protein structures (templates), an initial prediction of pairwise amino acid distances is generated

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

AlphaFold2 model architecture

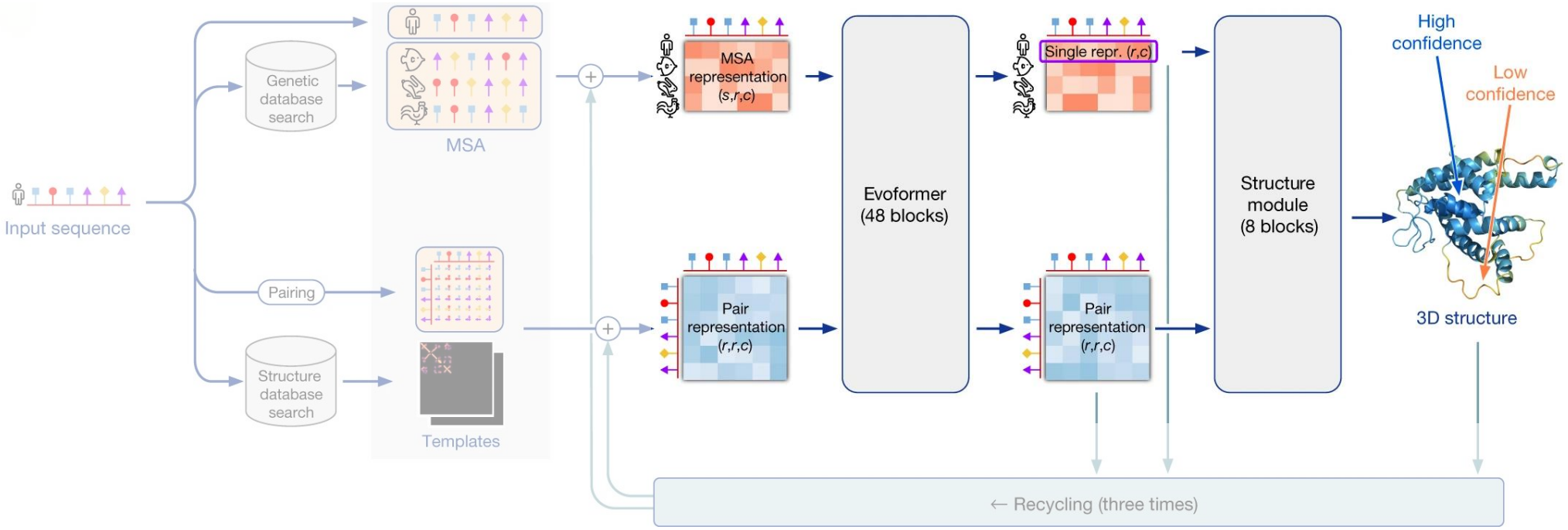


Figure: AlphaFold model architecture. Arrows show the information flow among the various components. Array shapes are shown in parentheses with s, number of sequences, r, number of residues and c, number of channels

Evoformer:

- Core of the architecture
- Updates the two representations through attention mechanisms and convolution-like operations

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

AlphaFold2 model architecture

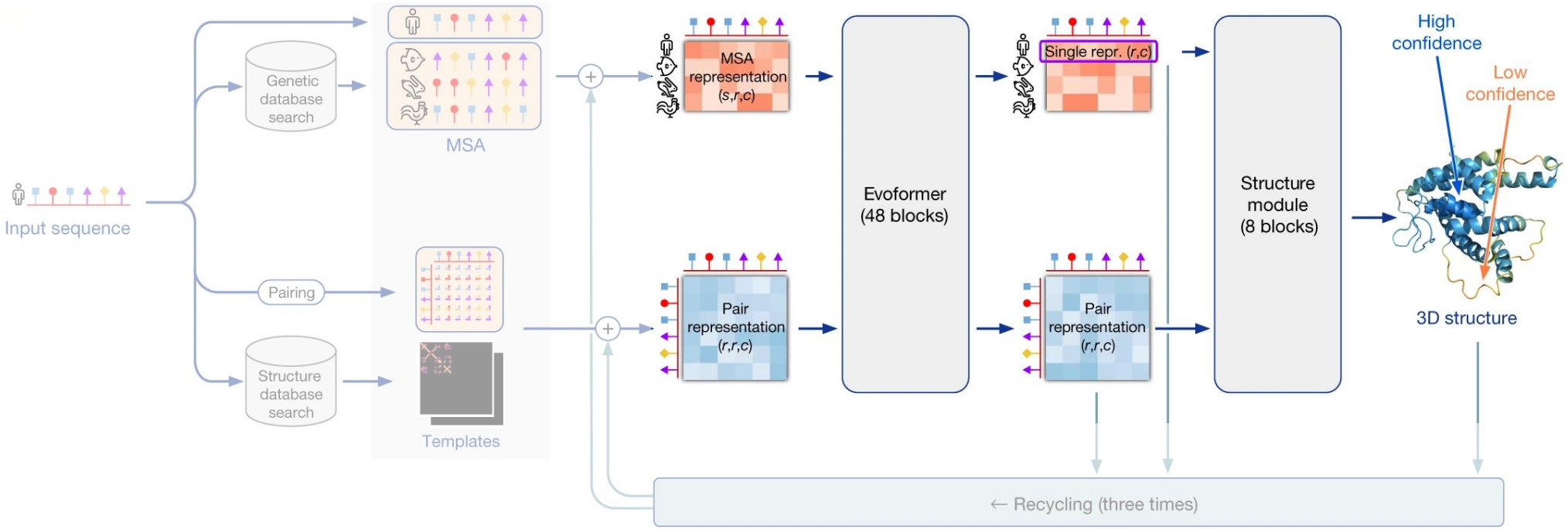


Figure: AlphaFold model architecture. Arrows show the information flow among the various components. Array shapes are shown in parentheses with s, number of sequences, r, number of residues and c, number of channels

Structure Module:

- Converts the processed representations from the Evoformer into spatial coordinates.
- Predicts the distances between residue pairs and the angles of bonds within the protein.
- Constructs a 3D model of the protein.

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

AlphaFold2 model architecture

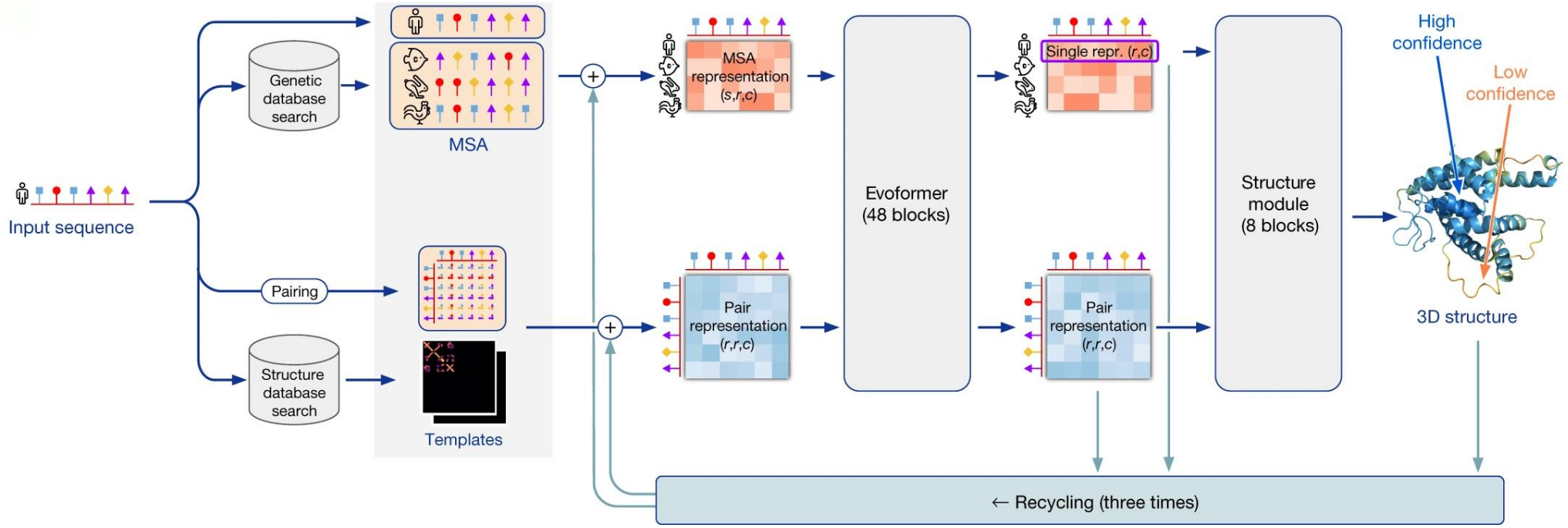


Figure: AlphaFold model architecture. Arrows show the information flow among the various components. Array shapes are shown in parentheses with s, number of sequences, r, number of residues and c, number of channels

Recycling:

- Iterative refinement process where the output of the structure module is fed back multiple times
- Helps refine the predictions by allowing the network to reconsider its earlier outputs
- Benefits of additional contextual information gleaned from subsequent layers.

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

AlphaFold2 training

Structure objectives: Deviation from true structure is penalized

BERT objective: Random masking is applied on the input MSAs and the network is required to reconstruct the masked regions from the output MSA representation

Training with self-distillation:

1. **Initial supervised training:** Training on a dataset of known protein structures, using conventional supervised learning techniques on all structures in the Protein Data Bank (PDB).
2. **Iterative refinement with self-distillation:** Refine the model's predictions using its own outputs
   - Use the model trained in supervised learning to predict the structure of 350'000 additional sequences
   - Retrain the model from scratch using mix of PDB data and predicted data

→ Makes effective use of unlabelled sequence and improves the accuracy of the model

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

Protein Structure Prediction

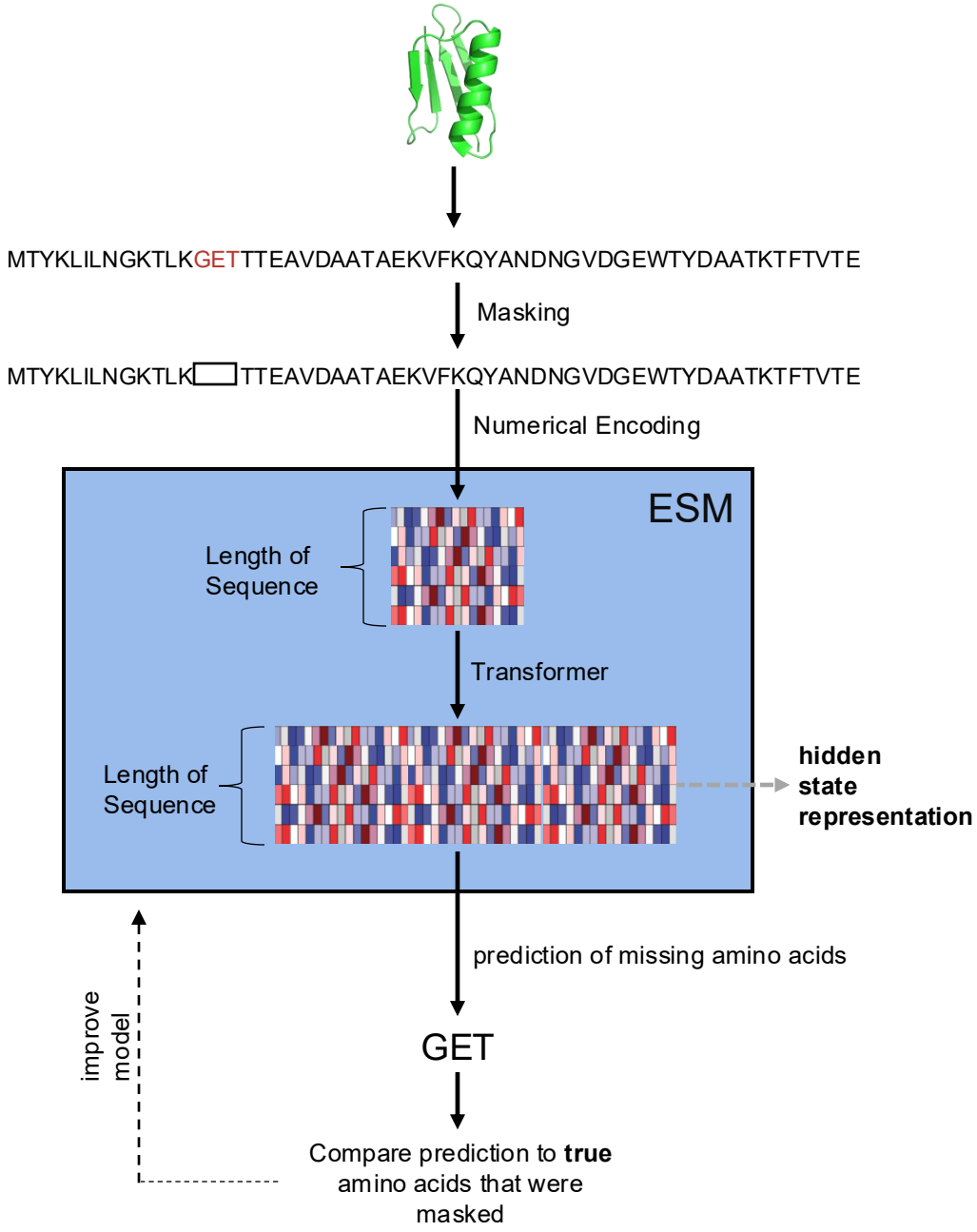
→ AlphaFold2
→ Evolutionary Scale Modelling **2** (ESM**2**)
→ Modelling of Molecular Assemblies

Evolutionary Scale Modelling (ESM)

Self-supervised transformer language model

- Self-supervised learning with unlabelled datasets, far larger amounts of data available
- Trained with the masked language modelling objective on 250 million protein sequences
- Each input sequence is corrupted by replacing a fraction of the amino acids with a mask
- ESM transformer generates prediction for the masked amino acids from the sequence context

→ **Knowledge of intrinsic biological properties of the proteins emerges without supervision, encoded in the hidden representations of the language model**



Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**, e2016239118 (2021).

Evolutionary Scale Modelling (ESM)

Unsupervised pretraining encodes secondary structure into representations

- The model cannot observe protein structure directly. It observes patterns in the sequences that are determined by the structure (a hidden variable)

Dataset of proteins with known structure, each amino acid is part of a helix, a strand or a coil

- Derive a transformer hidden representation for each amino acid of the protein
- Fit logistic regression classifier to predict secondary structure membership for each amino from its hidden representation

→ **High accuracy - self-supervised training generates structural knowledge**

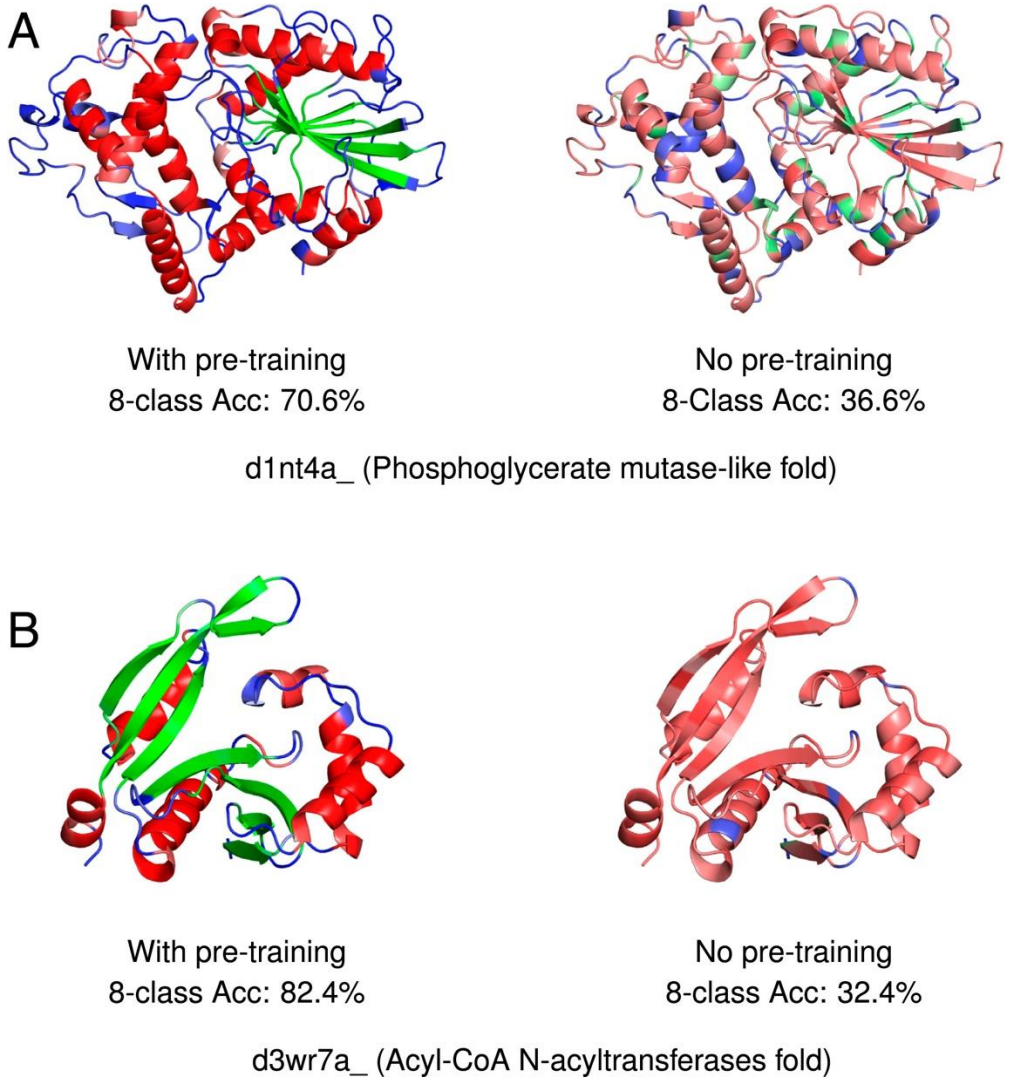


Figure: Unsupervised training encodes secondary structure into representations. Following pretraining, linear projections recover secondary structure (left column). Without pretraining, little information is recovered (right column). Colors indicate secondary structure class identified by the projection: **helix (red), strand (green), and coil (blue).**

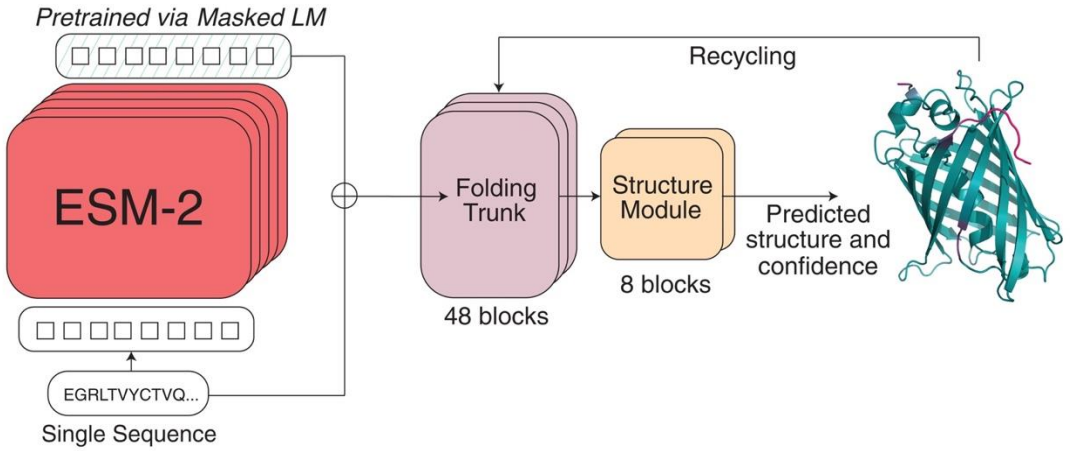
Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**, e2016239118 (2021).

Evolutionary Scale Modelling **2** (ESM**2**)

ESM language model is extended with a **folding head** to predict atomic coordinates

- Pretrained ESM model (MLM objective)
- Folding head extracts atomic coordinates from the hidden representations of ESM
- Fine-tuned end-to-end on 325K experimentally determined structures from the Protein Data Bank (PDB)

→ **Structure prediction accuracy similar to AlphaFold2**



Lin, Z. *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. (2022) doi:10.1101/2022.07.20.500902.

Evolutionary Scale Modelling **2** (ESM**2**)

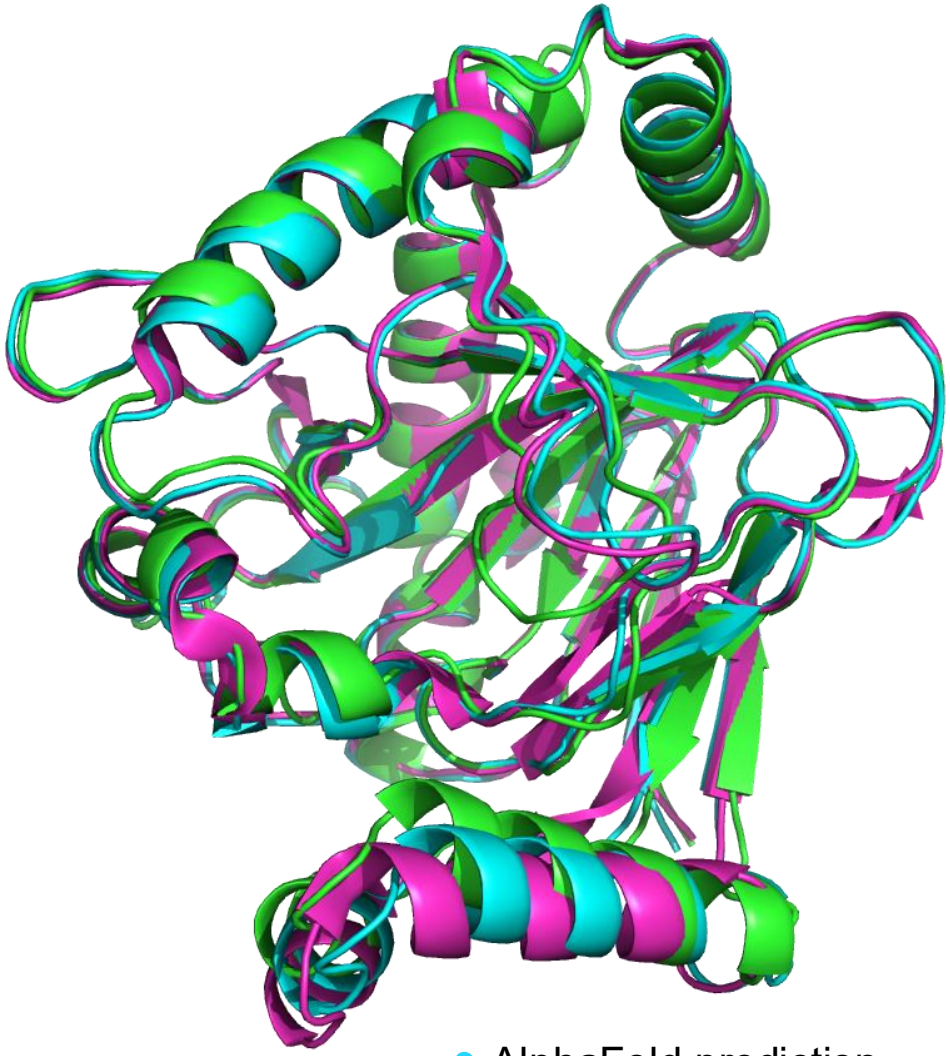
ESM language model is extended with a **folding head** to predict atomic coordinates

- Pretrained ESM model (MLM objective)
- Folding head extracts atomic coordinates from the hidden representations of ESM
- Fine-tuned end-to-end on 325K experimentally determined structures from the Protein Data Bank (PDB)

→ **Structure prediction accuracy similar to AlphaFold2**

→ **Removes costly aspects AlphaFold (multiple sequence alignment), leading to 60x speed-up**

→ **Completely standalone, the only input needed for inference is the protein sequence**



Lin, Z. *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. (2022) doi:10.1101/2022.07.20.500902.

AlphaFold and ESM – Performance Comparison

Prediction of backbone atoms:
Both models perform well, with AlphaFold often leading in direct comparisons

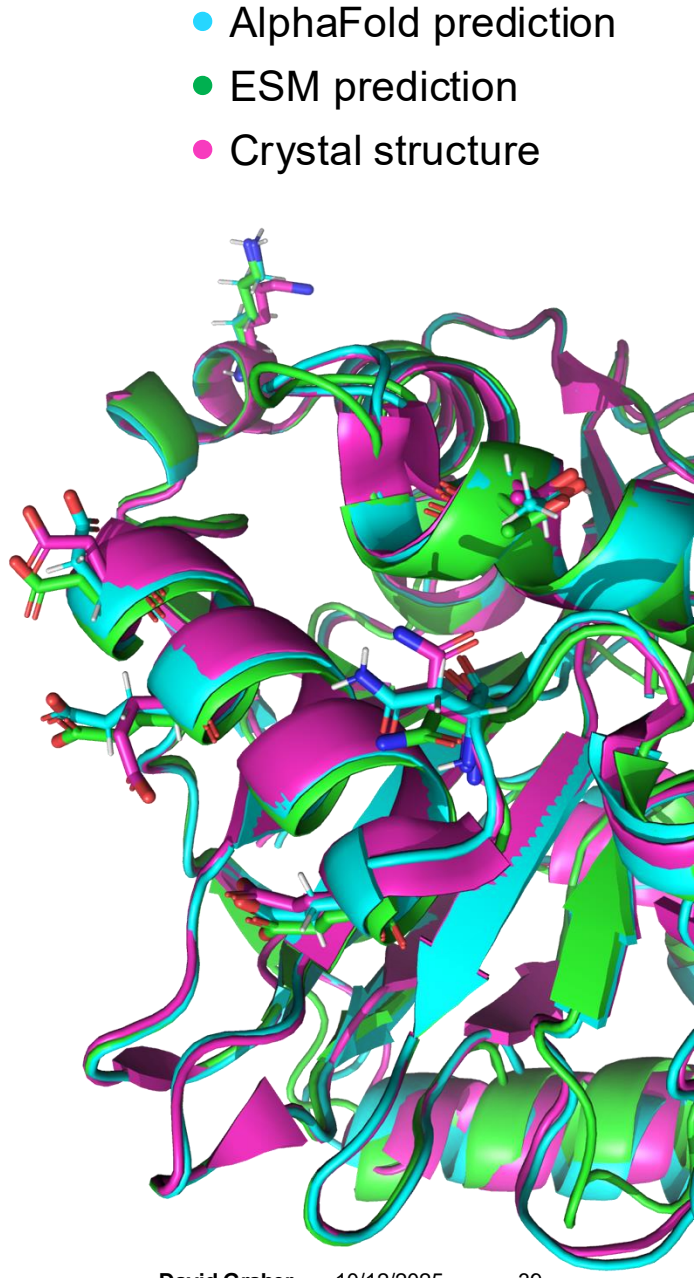
Prediction of side chain atoms:
The precision of both models drops slightly compared to backbone atoms, as side chains are more flexible and harder to predict. AlphaFold generally performs better in modelling side chains.

Which model to use:

→ AlphaFold predictions are often preferred when high accuracy in both backbone and side chains is crucial.

→ ESMFold provides a highly valuable tool, especially when quick predictions are needed for a large number of proteins

Protein Structures are now much easier and faster to obtain!



ESM Metagenomic Atlas

Database containing more than 600 million protein structures predicted with ESM2

- Structure prediction completed in two weeks on a heterogeneous cluster of 2000 GPUs
- Tens of millions of predictions do not have any match to experimentally determined structures, giving a view into previously unexplored protein diversity

All structures can be accessed in the ESM Metagenomic Atlas

https://esmatlas.com/

Zeming Lin et al. Evolutionary-scale prediction of atomic-level protein structure with a language model.Science379,1123-1130(2023).DOI:10.1126/science.ade2574

Protein Structure Prediction

→ AlphaFold2
→ Evolutionary Scale Modelling **2** (ESM**2**)
→ Modelling of Molecular Assemblies

Modelling of biomolecular assemblies

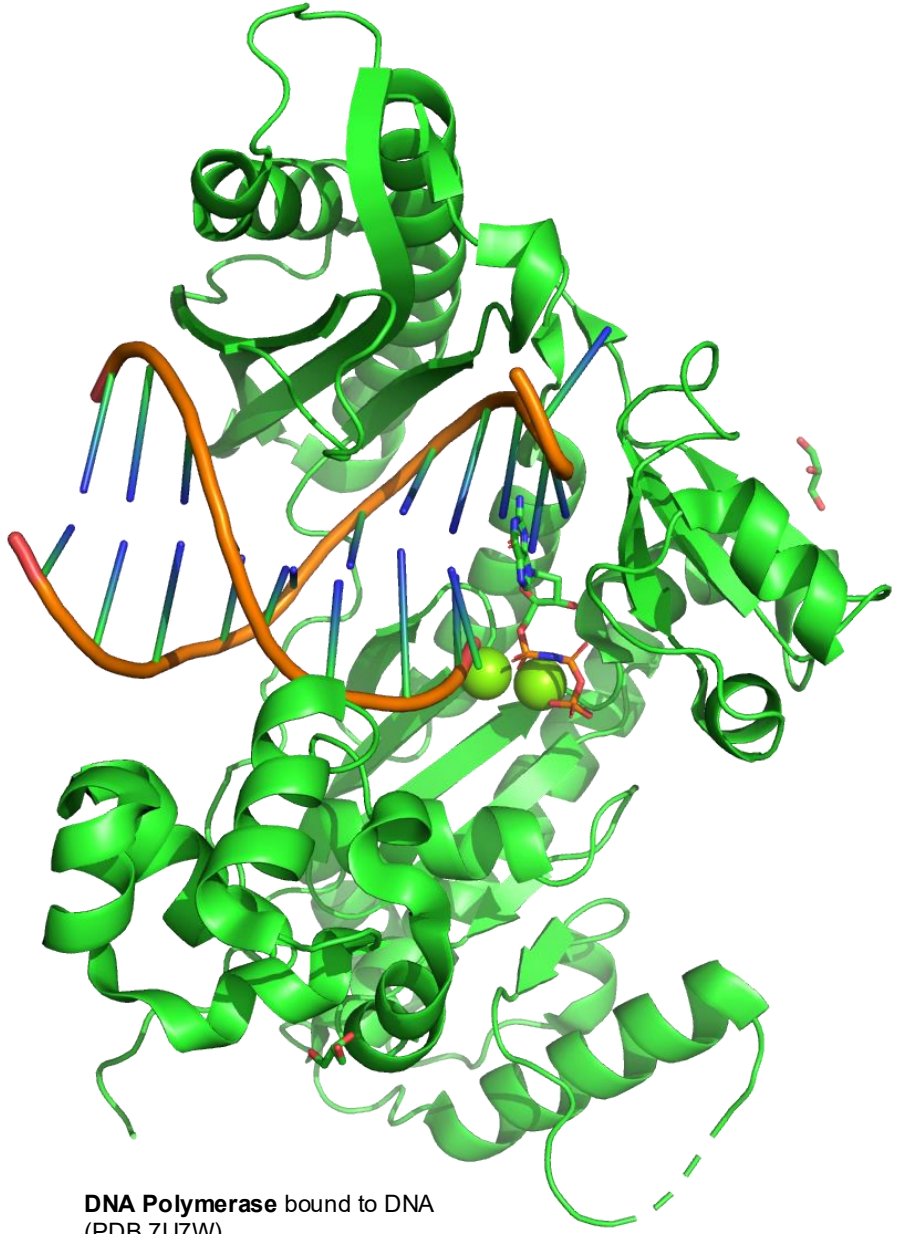
Protein rarely act alone – they form complexes with other proteins, small molecules, DNA and RNA

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

→ RoseTTAFold All-Atom

→ AlphaFold3



DNA Polymerase bound to DNA
(PDB 7U7W)

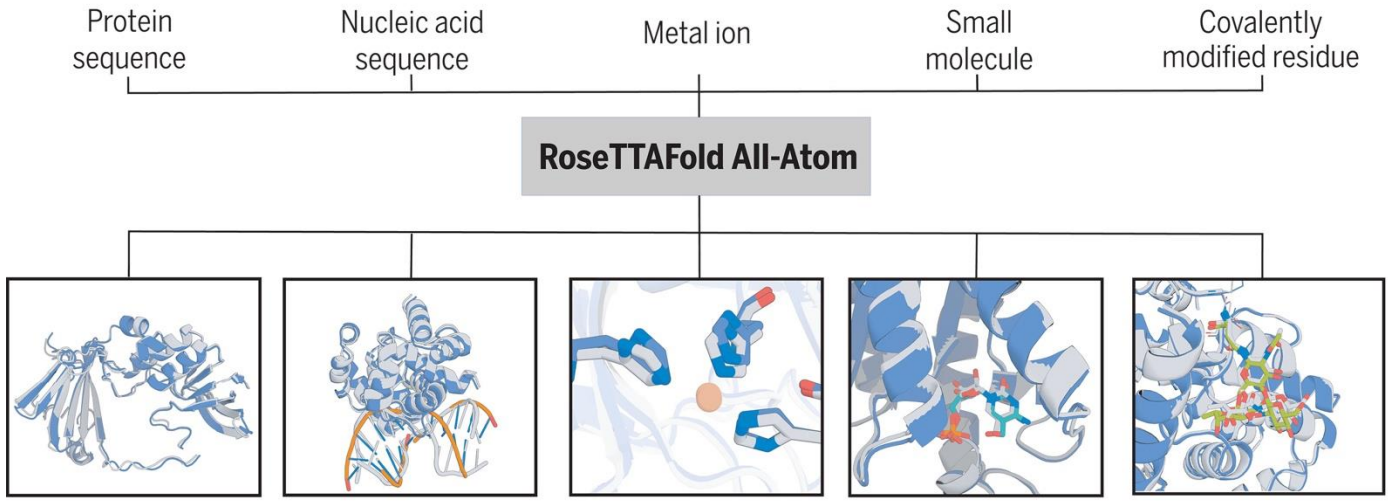
RoseTTAFold All-Atom

Three-track neural network (based on RoseTTAFold protein folding model) that predicts structures of biomolecular assemblies

Trained on a mixed dataset of biomolecules curated from the Protein Data Bank (PDB), including protein-small molecule interactions, protein-metal complexes, proteins with covalently modified amino acids, and protein-nucleic acid complexes.

Input:

- Protein Sequence
- Nucleic Acid Sequence
- graph representation of small molecules
- chemical element type of non-polymer atoms (e.g. metal ions)



Rohith Krishna *et al.*: Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024). DOI: 10.1126/science.adl2528

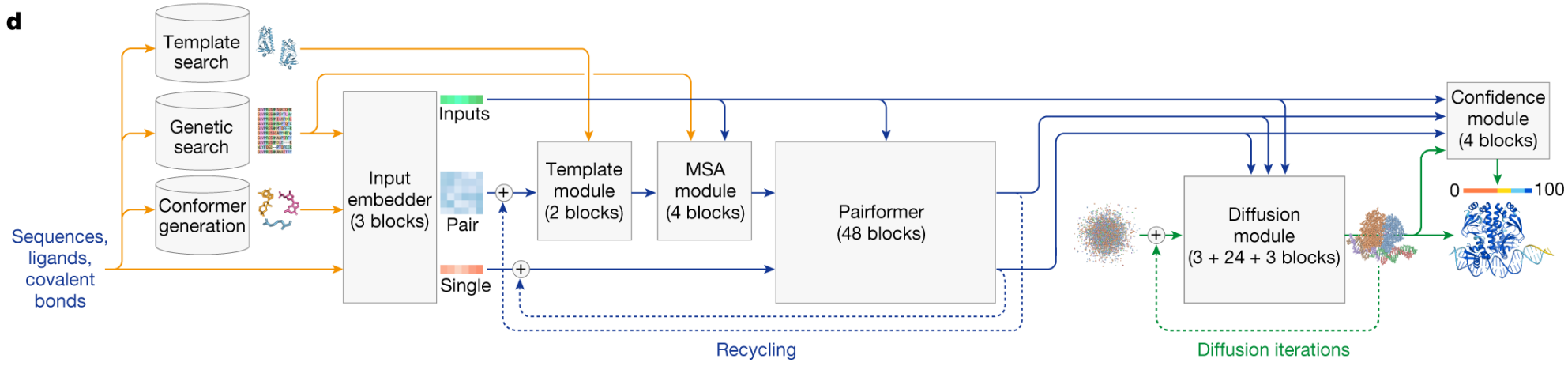
AlphaFold3

Claims to have higher accuracy in biomolecular assembly prediction than RoseTTAFold All-Atom

Differences to AlphaFold2 (AF2):

- Input now includes ligands, nucleic acids and metal ions
- MSA processing is de-emphasized (much smaller MSA module)
- Diffusion module replaces the structure module of AF2.

Diffusion module predicts atom coordinates from pair, single and input representations



Beyond today's scope

Prediction of molecular complexes - open-source alternatives to AlphaFold3:

- **Boltz:**
  Diffusion-based folding models developed by researchers at MIT in collaboration with Recursion, the Boltz models are open-source deep learning tools that rival proprietary models like AlphaFold3 in accuracy.

- **Chai-1:**
  Diffusion-based folding model offering performance comparable to AlphaFold3 while being available for a wide range of commercial applications through a python package and web server

- **OpenFold Consortium:**
  Non-profit AI research and development consortium of academic and industry partners that aims to develop an open ecosystem of AI tools for for biology and drug discovery, including **OpenFold3**

Structure-based models

→ 3D-Convolutional Neural Networks
→ Introduction to Graphs and Graph Neural Networks (GNN)
→ Example GNN models

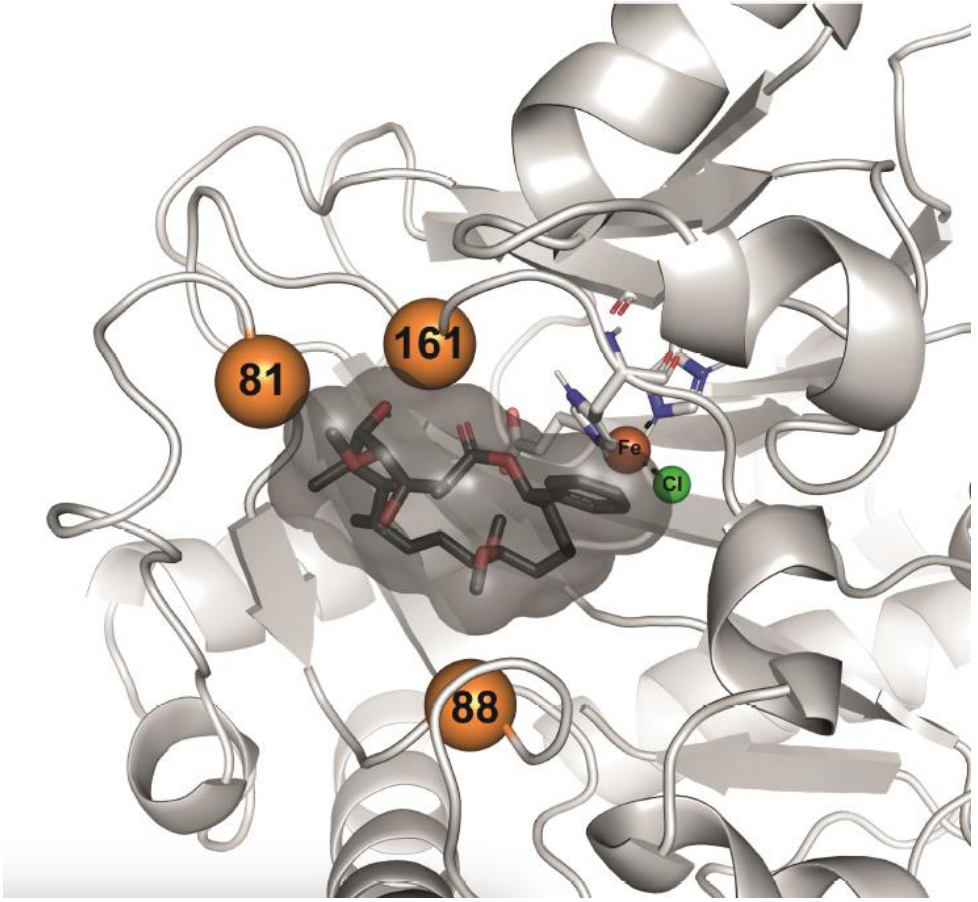
Using structural data as input for AI models

3D structure determines function: Proteins fold into complex shapes that allow them to interact specifically with other with other molecules

- A protein sequence alone provides no information about the actual spatial arrangement of atoms in the 3D space
- In tasks that require an understanding of the physical placement of residues, models that incorporate this structural information **should** be superior
- E.g. interaction prediction, enzyme activity prediction…

Challenges in the use of structural data:

- Availability of structural data is much lower
- Sequence-based methods are generally simpler and more computationally efficient



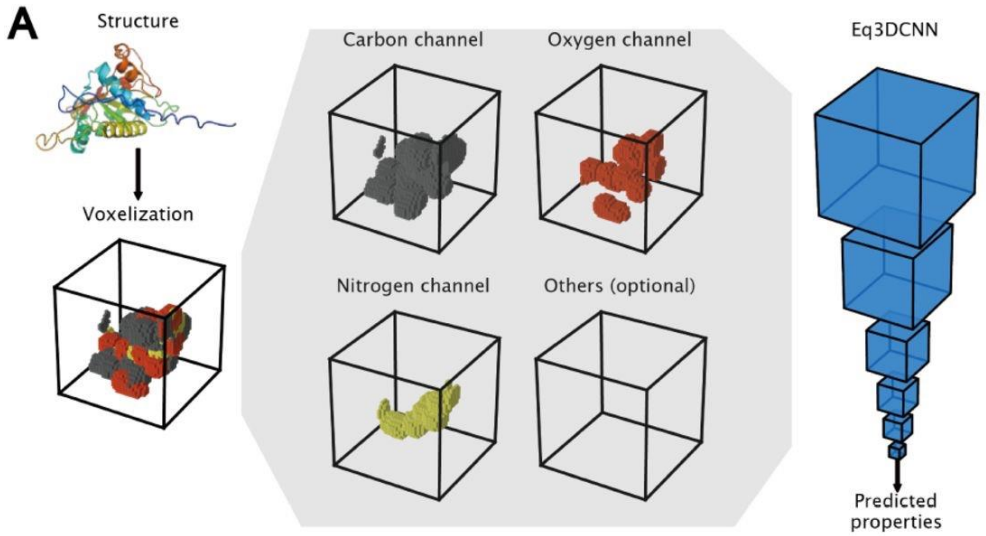
Structure defines function: In the 3-dimensional (3D) fold of this enzyme, different parts of the amino acid chain come together in 3D space to form an active site that accepts a specific substrate. The function of the enzyme is largely defined by the spatial arrangement of the involved residues.

Image credits: Büchler, J. *et al.* Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens. *Nat Commun* **13**, 371 (2022).

3D-Convolutional Neural Networks (3D-CNNs)

3D-CNNs – general procedure

- To convert the atomic coordinates into a format suitable for 3D-CNNs, the space around the protein is divided into a grid of voxels
- Different types of atomic properties can be used to define multiple channels in the input data (e.g. atom types, charges…)
- Convolution and pooling layers to extract relevant features and reduce the spatial dimensions of the input volume
- Fully connected layer derives global prediction



➔ 3D-CNNs have been shown to perform well in tasks such as protein thermostability prediction

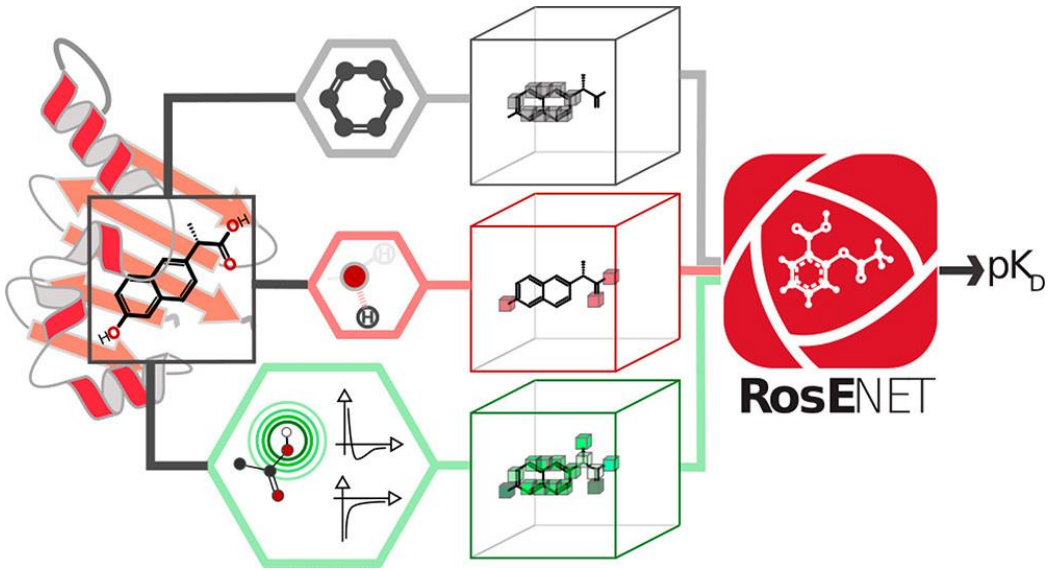
Chen, H. *et al.* Protein property prediction based on local environment by 3D equivariant convolutional neural networks. *bioRxiv* 2024.02.07.579261 (2024) doi:10.1101/2024.02.07.579261.

3D-Convolutional Neural Networks (3D-CNNs)

3D-CNNs for Protein-Ligand Affinity Prediction

- Centering a 25×25×25 Å grid with a spacing of 1Å around the geometric center of the ligand
- The position of each atom within the grid was mapped to a voxel
- Additional chemical properties of the atoms are incorporated as additional channels
- Convolutional Neural Network reduces 3D grid to a single pK_D value representing an affinity prediction

➔ 3D-CNNs were the first structure-based models to outperform sequence-based affinity prediction



RosENet 3D-CNN for binding affinity prediction: A 25x25x25 Å grid around the center of the ligand molecule is defined and featurized with different properties of the underlying protein and ligand atoms (here aromatic carbon atoms, hydrogen bond donors and electrostatic energies). The different property channels are then reduced to a single value with a 3D-convolutional neural network.

Hassan-Harrirou, H., Zhang, C. & Lemmin, T. RosENet: Improving Binding Affinity Prediction by Leveraging Molecular Mechanics Energies with an Ensemble of 3D Convolutional Neural Networks. *J. Chem. Inf. Model.* **60**, 2791–2802 (2020).

3D-Convolutional Neural Networks (3D-CNNs)

Drawbacks

- **Model size:** Significantly larger compared to 2D-CNNs, model parameters grow cubically with the resolution of the grid
- **Sparsity:** Grids often contain a significant portion of voxels representing empty or homogeneous regions that do not contribute meaningful information.
- **Rotational Invariance:** To achieve invariant predictions regardless of the spatial orientation of the objects in the input data, it is necessary to present the objects in multiple orientation

➔ 3D-CNNs usually require large training datasets to perform well and avoid overfitting.

➔ Training and inference require significant memory and processing power

➔ Application to large datasets of structural data is limited (e.g. molecular dynamics simulations)

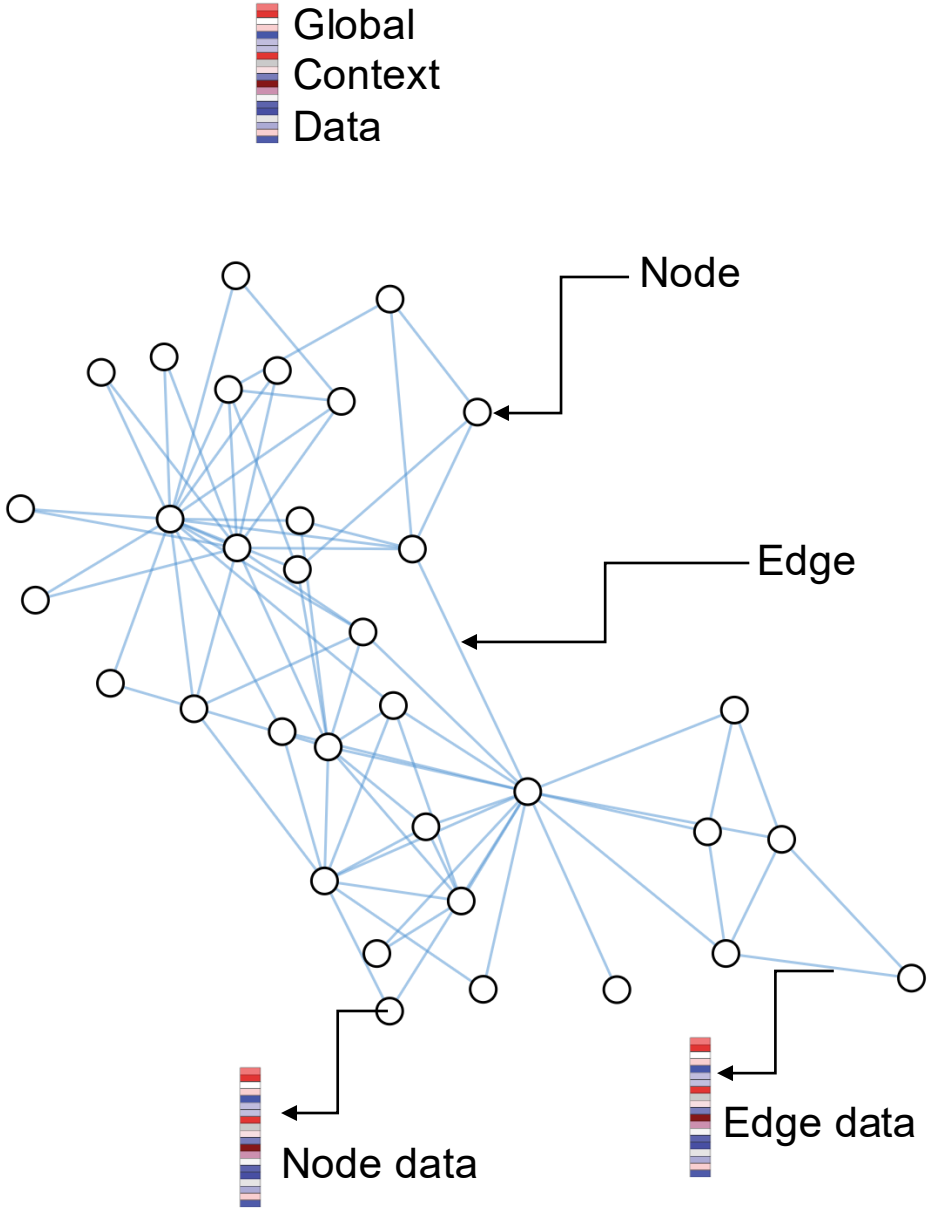
Structure-based models

→ 3D-Convolutional Neural Networks
→ Introduction to Graphs and Graph Neural Networks (GNN)
→ Example GNN models

Introduction to Graphs and GNNs

- Graphs describe a set of objects (nodes) and the connections between them (edges).
- Information can be stored in
  - Nodes
  - Edges
  - Global feature vector
- Powerful representation of data, especially for non-Euclidean and unordered data
  - Social network with friendships
  - Citation networks
  - Atoms in molecules

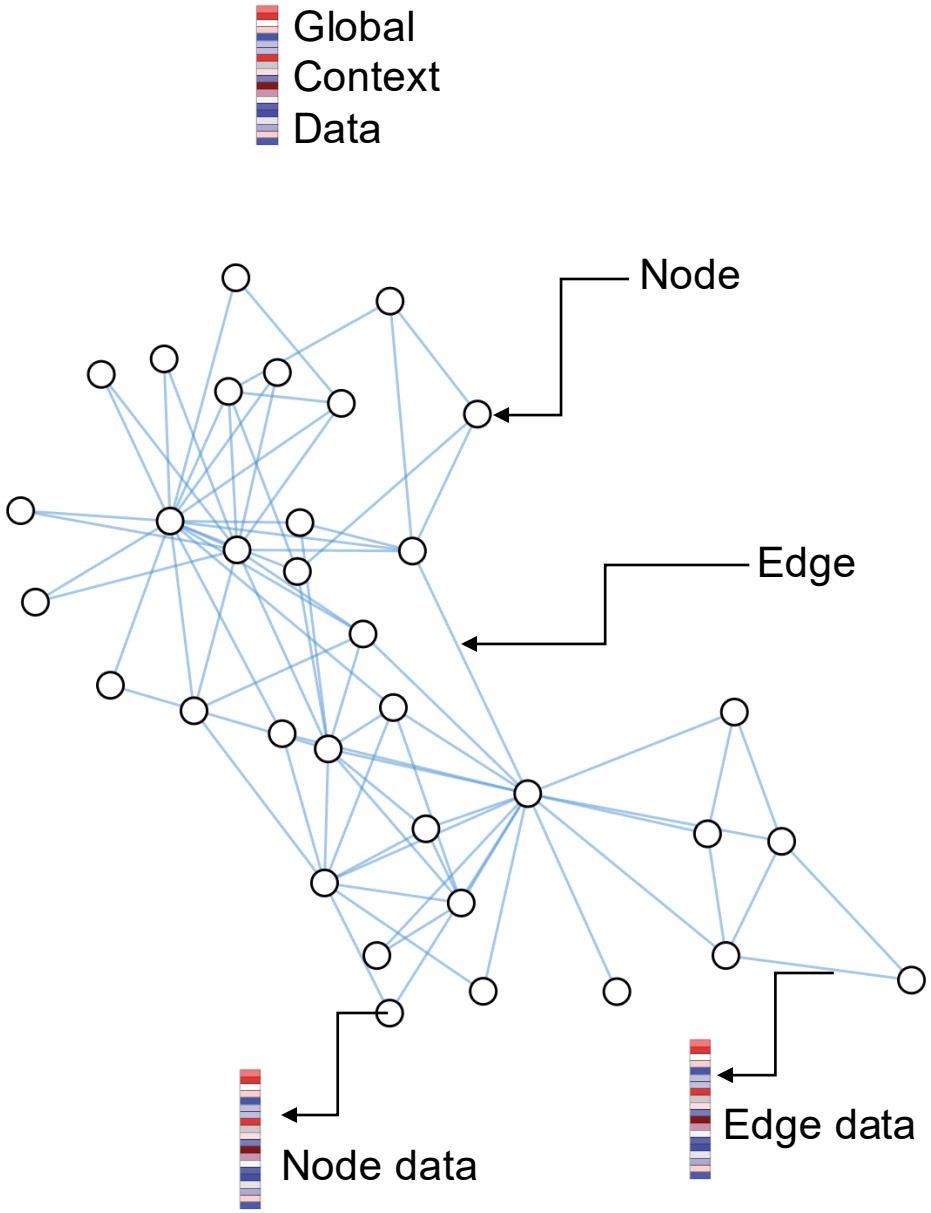
➔ **Graph Neural Networks can process such unordered graphs**



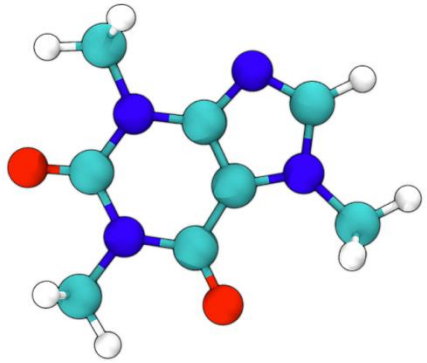
Introduction to Graphs and GNNs

Graph have up to four types of information that we want to use to make predictions:

- Connectivity
- Node features
- Edge features
- Graph-level features / global context



Introduction to Graphs and GNNs

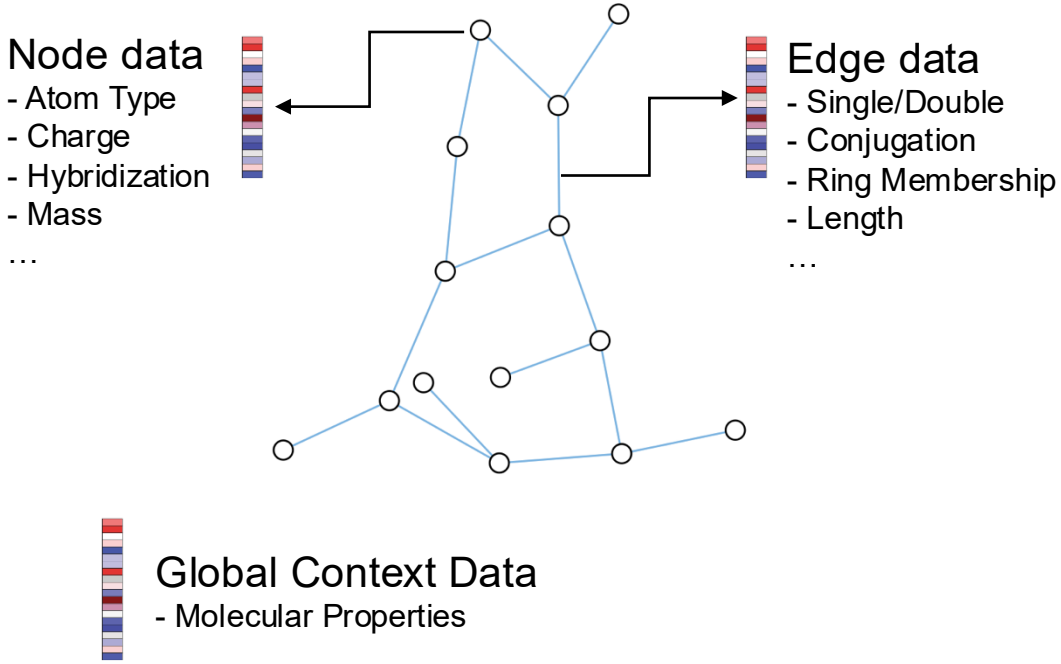


Graph have up to four types of information that we want to use to make predictions:

- Connectivity
- Node features
- Edge features
- Graph-level features / global context

Molecules: It's a very convenient and natural abstraction to describe molecules as graphs, where nodes are atoms and edges are covalent bonds.

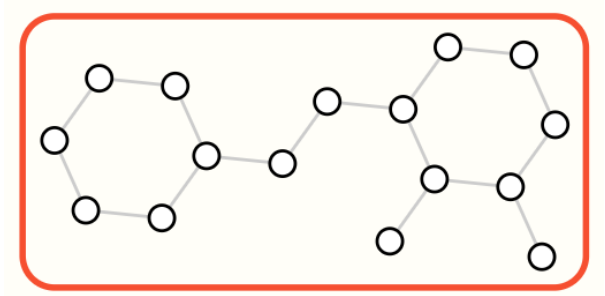
- Allows for integration of chemical properties as node features, edge features and global features



What tasks do we want to perform on this data?

Graph Level Tasks:
We predict a single property for an entire graph

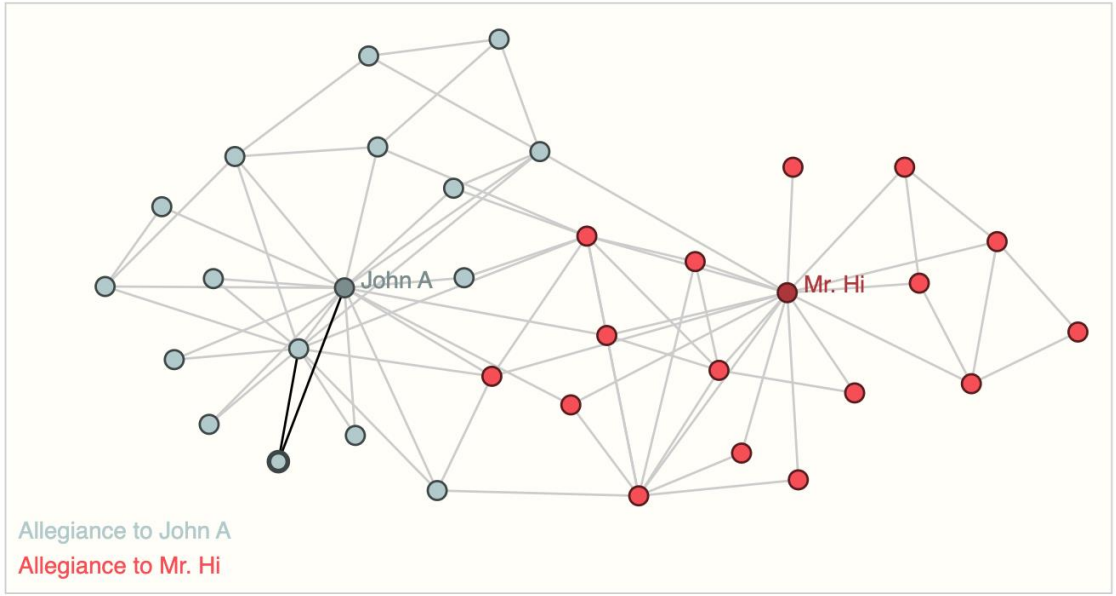
Example: Is this molecule an antibiotic?



Edge-Level Tasks:
We predict the property or presence of edges in a graph

Example: Friendship Suggestions on Facebook

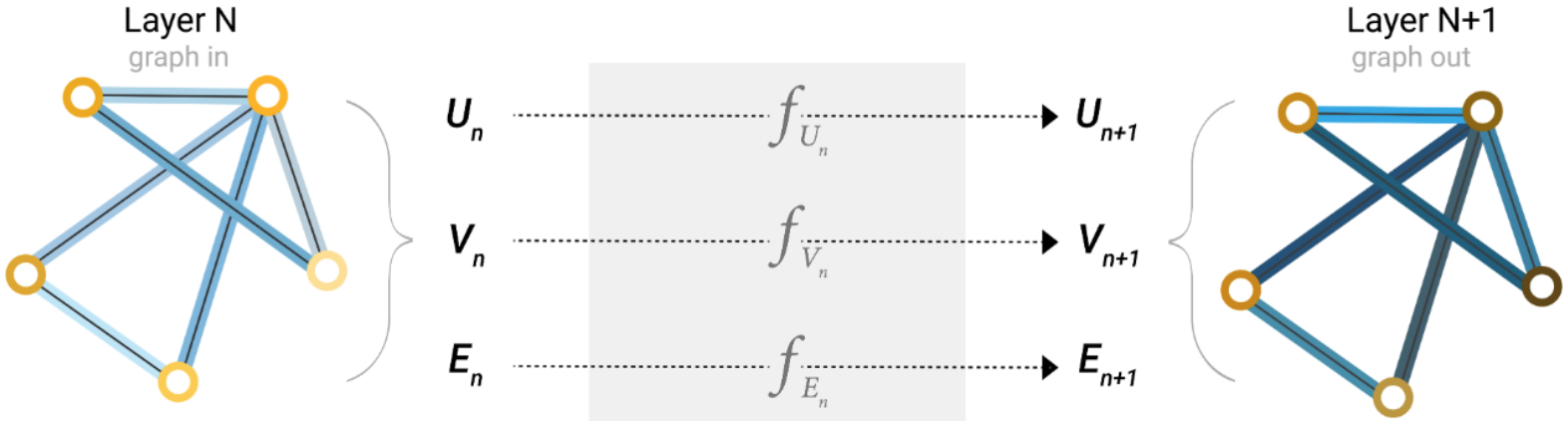
Node-Level Tasks:
We predict some property for each node in a graph



Zach's karate club dataset: A classic example of a node-level prediction problem. The dataset is a single social network graph made up of individuals that have sworn allegiance to one of two karate clubs after a political rift. As the story goes, a feud between Mr. Hi and John H creates a schism in the karate club. The nodes represent individual karate practitioners, and the edges represent interactions between these members outside of karate. The prediction problem is to classify whether a given member becomes loyal to either Mr. Hi or John H, after the feud.

Graph Neural Networks – Basic Principles

A GNN is an optimizable transformation on all attributes of the graph (nodes, edges, global-context) that preserves graph connectivity



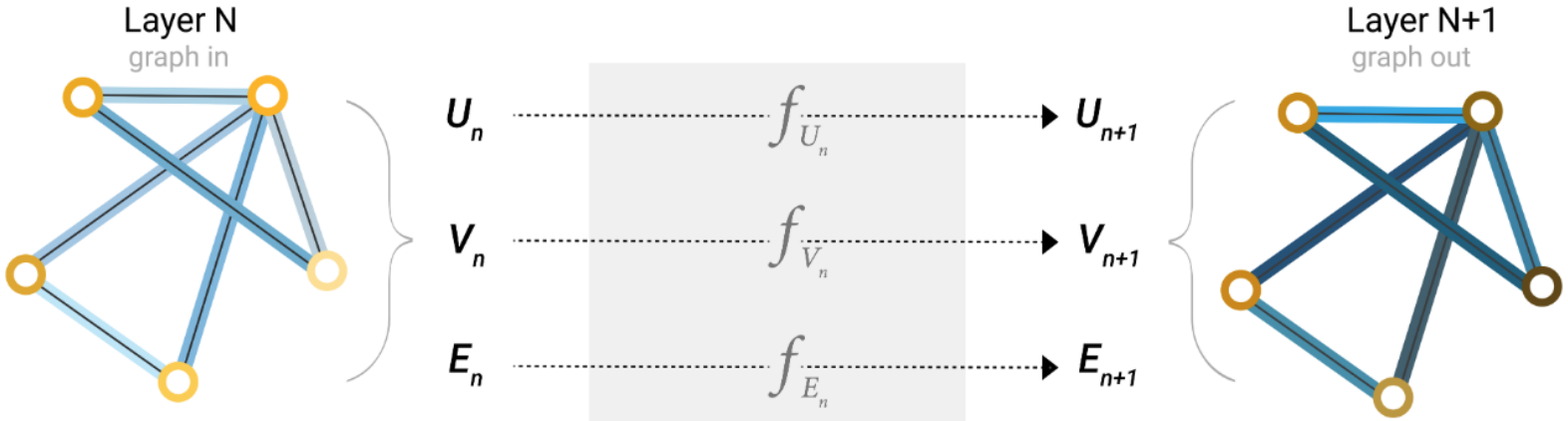
Transformation of graph attributes with update functions f_U, f_V and f_E: Each attribute (global context U, nodes V and edges E) is transformed and updated in a GNN layer. The output of the GNN is the same graph with updated node, edge and global context representations.

GNNs adopt a "graph-in, graph-out" architecture

- Accept a graph as input, with information loaded into its nodes, edges and global-context, and progressively transform this information, without changing the connectivity of the input graph.

Graph Neural Networks – Basic Principles

A GNN is an optimizable transformation on all attributes of the graph (nodes, edges, global-context) that preserves graph connectivity



Transformation of graph attributes with update functions f_U, f_V and f_E: Each attribute (global context U, nodes V and edges E) is transformed and updated in a GNN layer. The output of the GNN is the same graph with updated node, edge and global context representations.

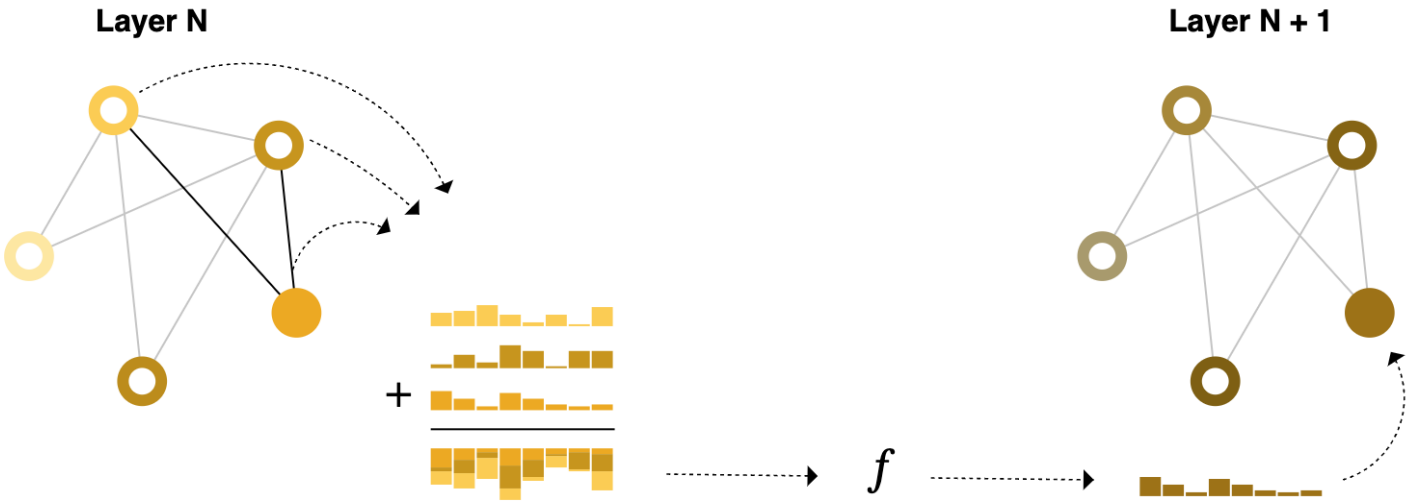
To update the graph features, GNNs apply the "message passing neural network" framework:

➔ Neighboring nodes or edges exchange information and influence each other's updated embeddings

➔ Makes the learned embeddings aware of graph connectivity

Graph Neural Networks – Basic Principles

A GNN is an optimizable transformation on all attributes of the graph (nodes, edges, global-context) that preserves graph connectivity



The three steps of message-passing in graphs:

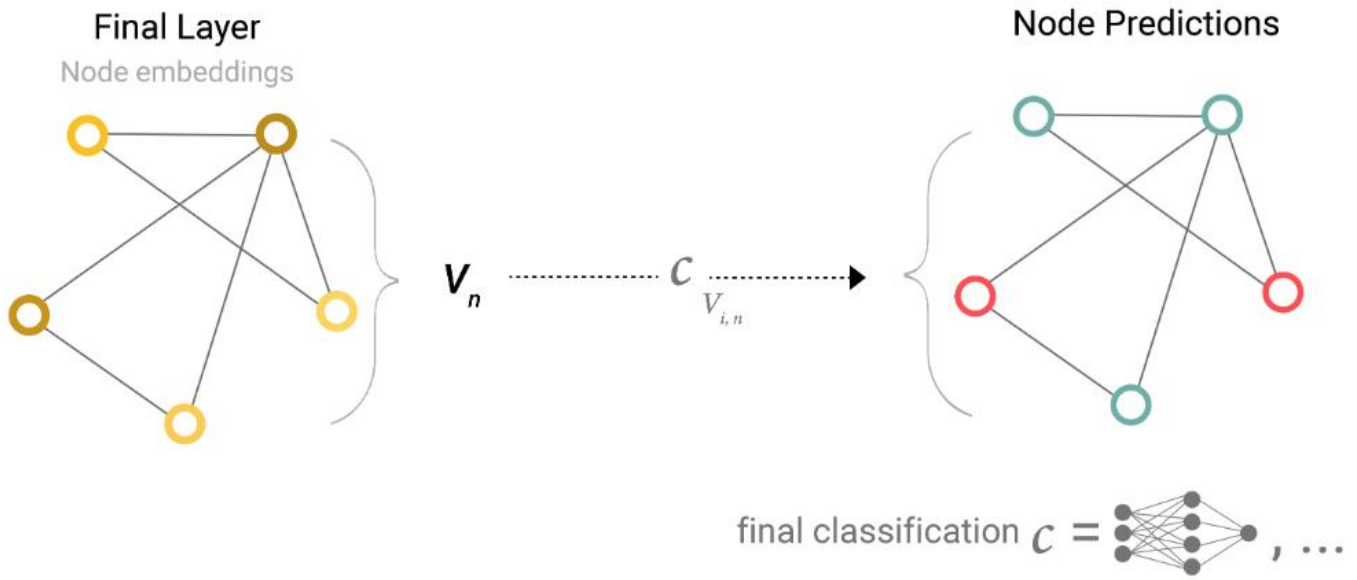
1. Aggregate information from adjacent nodes
2. Transform information
3. Update graph with new information

Message-passing works in three steps

1. For each node in the graph, gather all the neighboring node embeddings (or messages)
2. Aggregate all messages using an aggregation function (including the node's own embedding)
3. Pass the pooled messages through an update function, usually a learned neural network

Graph Neural Networks – Basic Principles

A GNN is an optimizable transformation on all attributes of the graph (nodes, edges, global-context) that preserves graph connectivity



Node classification:
Given a graph with transformed node embeddings (V_n), apply a fully connected neural network C to derive a prediction for each node.

Predictions:

- Node Level: For each **node** in the graph, process transformed **node** embeddings with fully-connected NN
- Edge Level: For each **edge** in the graph, process transformed **edge** features with a fully-connected NN
- Graph Level: Perform global pooling (or process transformed global features) with a fully-connected NN

Graph Neural Networks – Basic Principles

Message-passing vs. **Image convolution.**

- Both are operations to aggregate and process the information of an element's neighbors in order to update the element's value
- In graphs the elements are nodes, and in images the elements are pixels
- The number of neighboring nodes in a graph can be variable, unlike in an image where each pixel has a fixed number of neighboring elements

➔ **The message-passing process in graphs is also called graph convolution**

As in 2D images processed with CNNs, we can stack graph convolutional layers. After **three** layers, a node contains information about all nodes that are up to **three** steps away from it

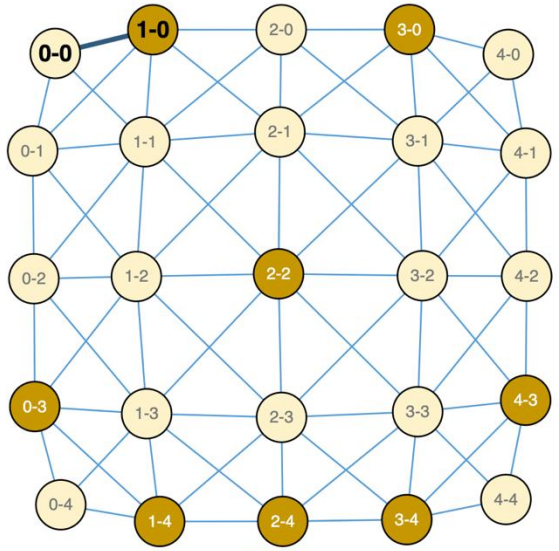


Image Convolution can be thought of as graph convolution on a graph with regular structure, where each node represents a pixel and is connected to adjacent pixels by and edge.

Graph Convolutional Layers

Several types of graph convolutional layers are commonly used, each with its own method of aggregating information from neighboring nodes.

Graph Convolutional Network Layer (GCNConv):

The updated node features $x_i^{(l+1)}$ of node i at layer $(l+1)$ are computed as

$$x_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{\sqrt{d_j d_i}} W^{(l)} x_j^{(l)}\right)$$

$\mathcal{N}(i)$	neighboring nodes of node i
$x_j^{(l)}$	node features of neighbor j layer l
d_i	the degree of node i
$W^{(l)}$	learnable weight matrix at layer l,
σ	activation function (e.g. ReLU)

Graph Attention Network Layer (GATConv)

The updated node features $x_i^{(l+1)}$ of node i at layer $(l+1)$ are computed as

$$x_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} W^{(l)} x_j^{(l)}\right)$$

$a_{ij}^{(l)}$	attention score between nodes i and j
$\mathcal{N}(i)$	neighboring nodes of node i
$x_j^{(l)}$	node features of neighbor j layer l
$W^{(l)}$	learnable weight matrix at layer l
σ	activation function (e.g. ReLU)

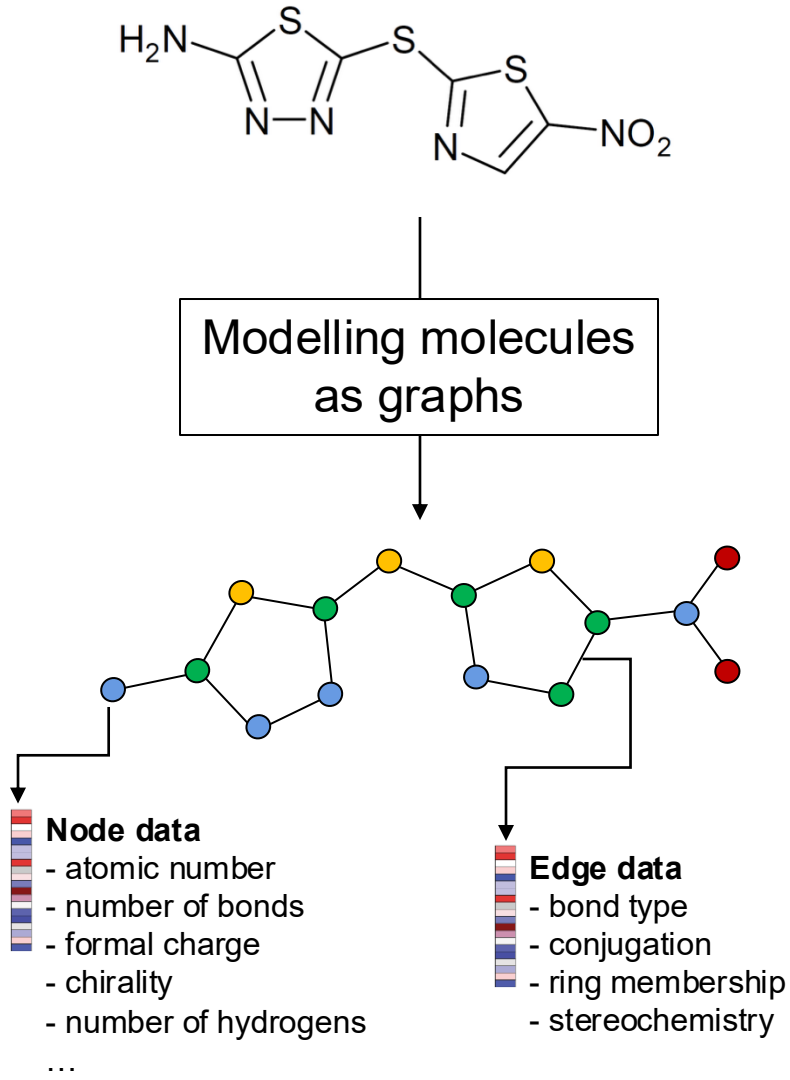
Structure-based models

→ 3D-Convolutional Neural Networks
→ Introduction to Graphs and Graph Neural Networks (GNN)
→ Example GNN models

Antibiotic Discovery with GNNs (Stokes et al. 2020)

Graph Neural Network capable of predicting molecules with antibacterial activity

- **Graph-level binary classification problem**
- **Training set:** Library of 2560 drug molecules of diverse structure and function containing 120 molecules with known growth inhibition against E.Coli
- **Graph Modelling:** Generate graph representation of molecules, include chemical properties as edge and node features
- **Graph Convolutional Neural Network**
  - GCNConv convolution applied to edges and nodes, followed by global add pooling
  - Feed-forward neural network that outputs a predicted probability of growth-inhibition of E.Coli



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

Antibiotic Discovery with GNNs

Make predictions on a large molecular library
(Drug Repurposing Hub molecule library n=6111)

- Inference with an ensemble of models trained on twenty random folds of the training data
- Among the **99** molecules with **highest** model prediction, **51** molecules displayed growth inhibition against E.Coli
- Among the **63** molecules with **lowest** model prediction, **two** molecules displayed growth inhibition against E.Coli

➔ **Discovery of Halicin, which displays bactericidal activity against a wide range of bacteria**

➔ **Very low structural similarity to its nearest neighbor antibiotic shows that the model was capable of generalization, accessing new antibiotic chemistry**

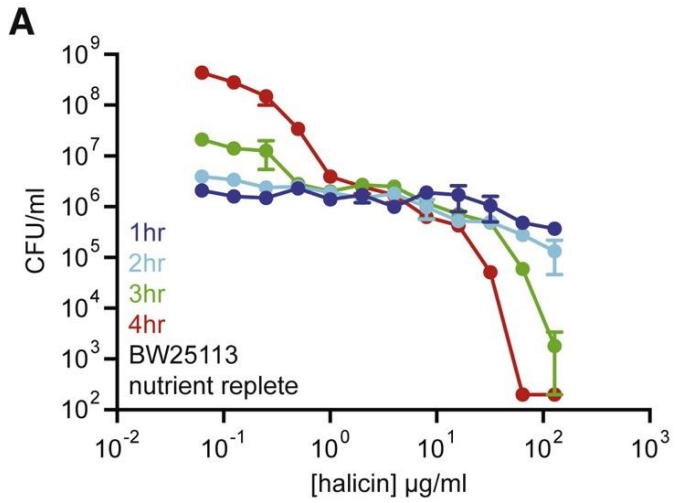


Figure A: Killing of E. coli in LB media in the presence of varying concentrations of halicin after 1 h (blue), 2 h (cyan), 3 h (green), and 4 h (red).

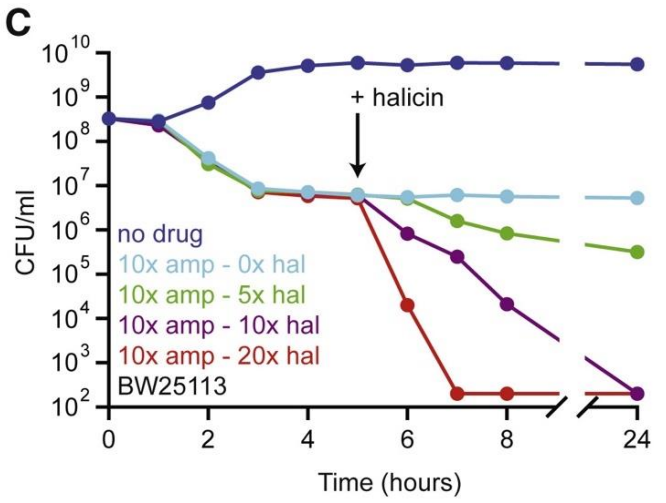


Figure B: Killing of E. coli persisters by halicin after treatment with 10 µg/mL of ampicillin. Light blue is no halicin. Green is 5× MIC halicin. Purple is 10× MIC halicin. Red is 20× MIC halicin.

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

Graph Neural Networks for Protein-Ligand Interactions

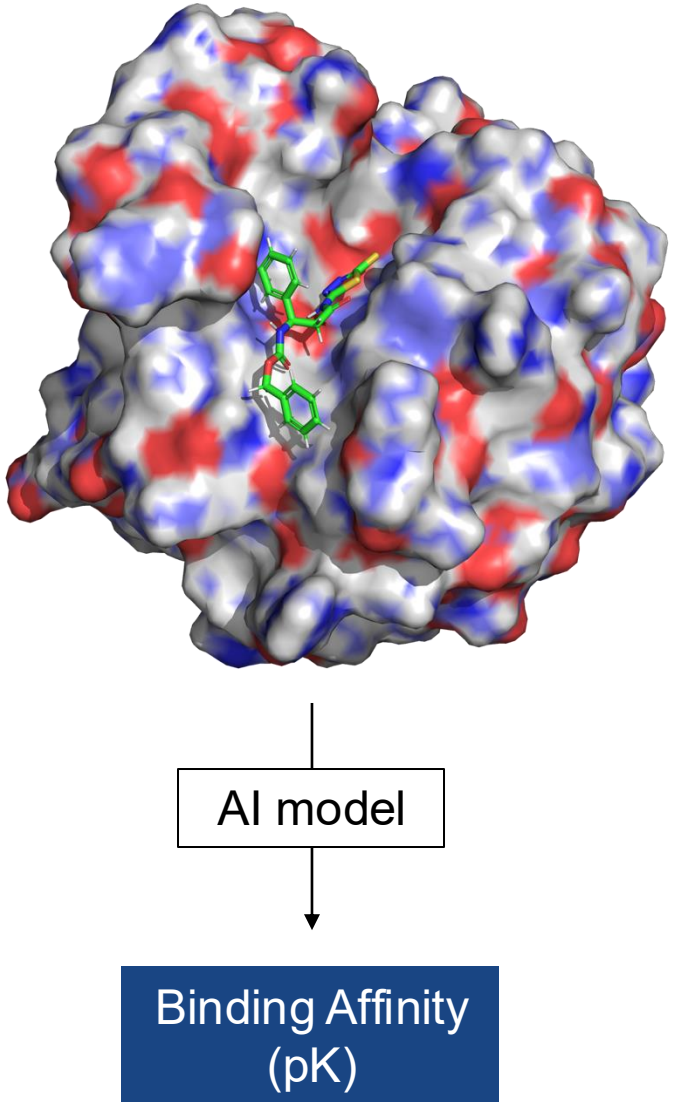
Aim: Predict binding affinities (interaction strength) for protein-ligand interactions from their 3D-interaction structure

- Important for computational drug design
- Small-molecule drugs should bind with high affinity to specific protein targets

➔ Scoring Functions

➔ Classical Scoring functions are force-field-based, empirical, or knowledge-based and show limited accuracy in binding affinity prediction

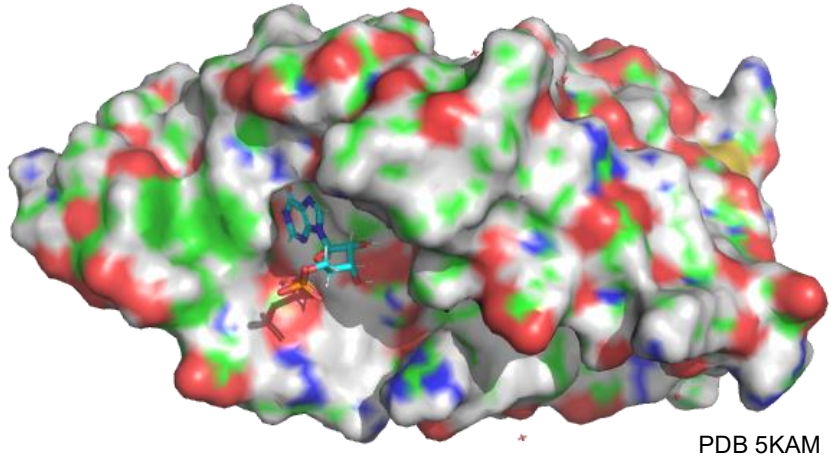
➔ **AI models should improve this**



Graph Neural Networks for Protein-Ligand Interactions

GEMS: Graph convolutional neural network for predicting binding affinities

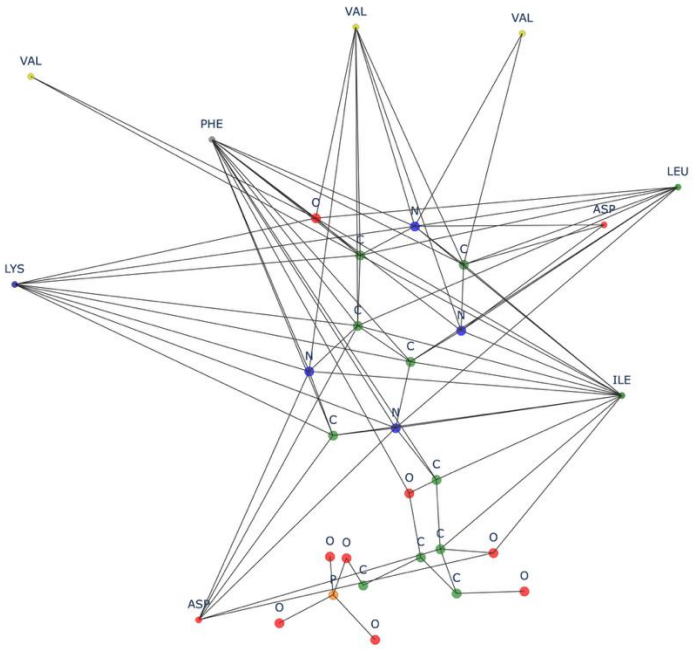
Trained on PDBbind: Database of affinity-labelled protein-ligand complexes



PDB 5KAM

Protein-ligand interactions are modelled as graphs

- Atom-level molecular graph of the ligand
- Surrounding amino acids as additional nodes
- Node Features: Protein Language Model Embeddings



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

Graph Neural Networks for Protein-Ligand Interactions

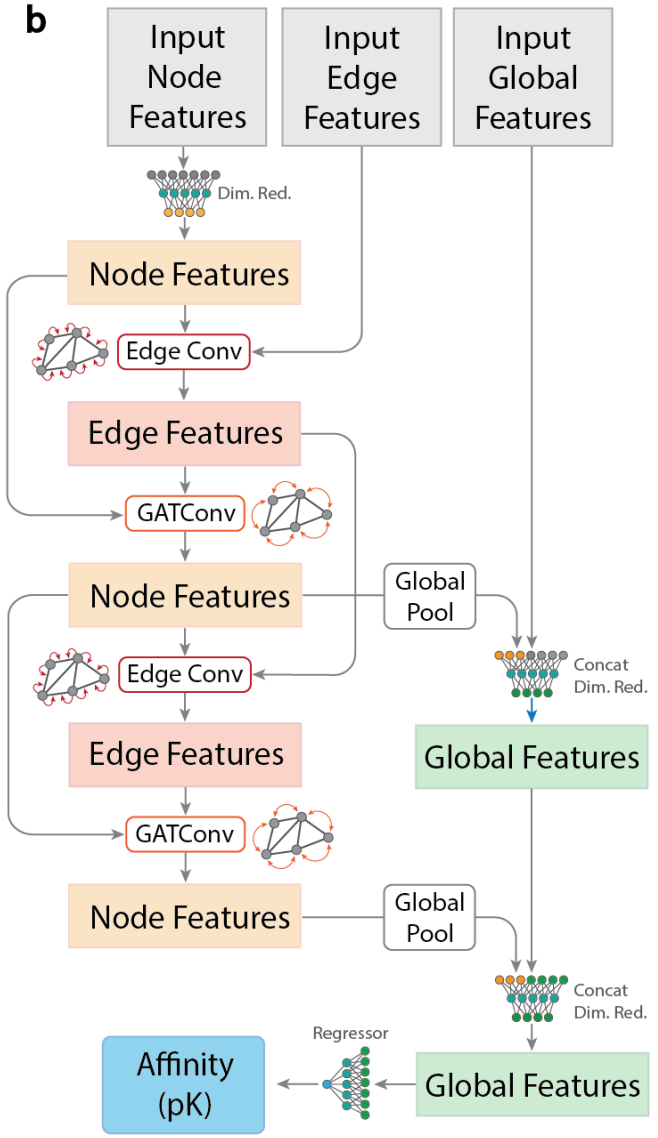
GEMS: Graph convolutional neural network for predicting binding affinities

Input:

- Node, edge and global features
- Connectivity

Model Architecture:

- Initial feature transformation
- Alternating sequence of node and edge convolutions
- Global graph features are updated after each convolution
- Final prediction with fully-connected NN



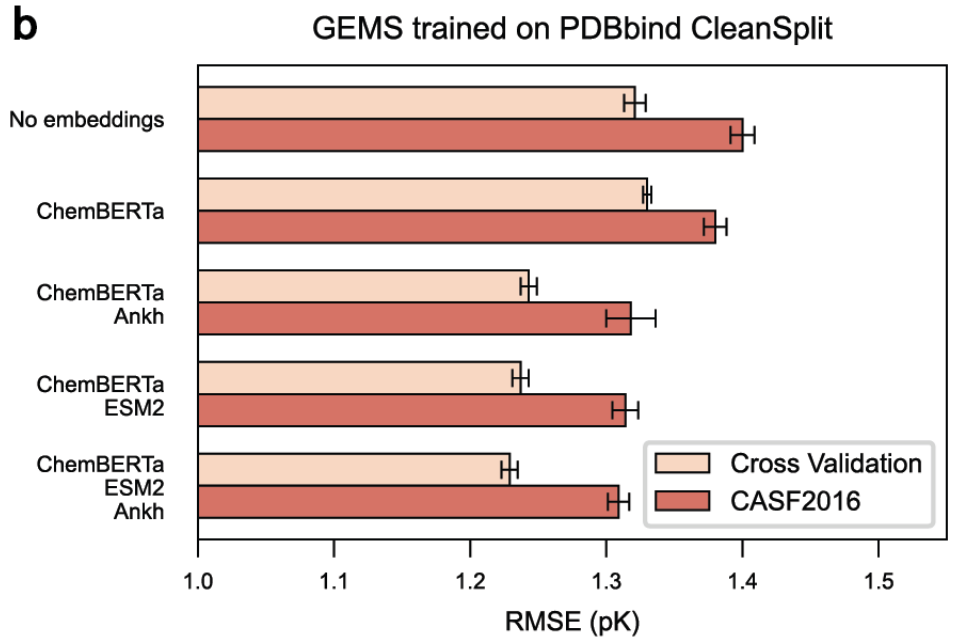
Graber, D. *et al.* Resolving data bias improves generalization in binding affinity prediction. *Nat. Mach. Intell.* **7**, 1713–1725 (2025).

Graph Neural Networks for Protein-Ligand Interactions

GEMS: Graph convolutional neural network
for predicting binding affinities

Prediction performance on independent benchmark:

- Significantly more accurate (RMSE = 1.3) than classical scoring functions (RMSE = 1.73)
- Adding protein language model embeddings as node features boosts performance
- Incorporation of embeddings of several language models leads to synergistic improvement



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

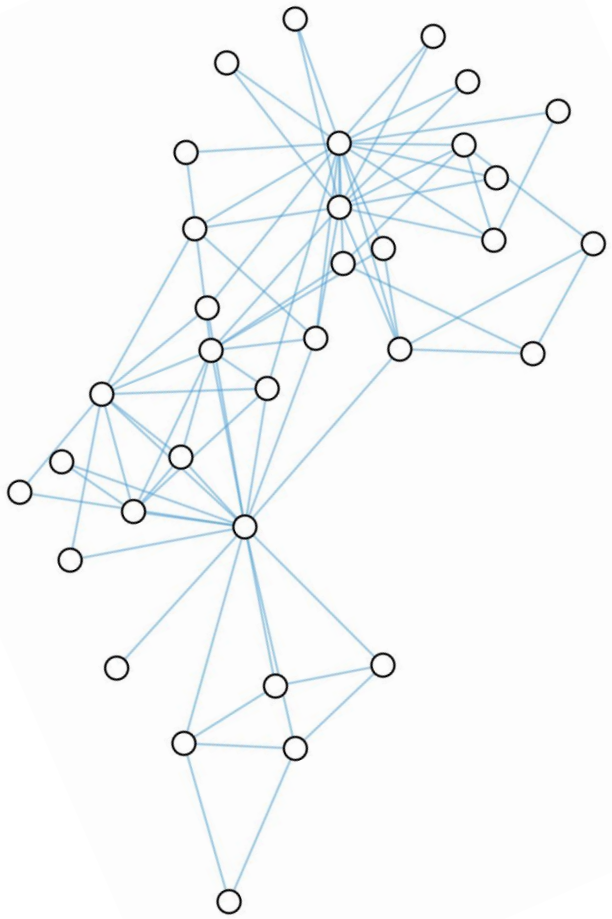
Graph Neural Networks - Conclusions

Graph Neural Networks offer many advantages for modelling and processing 3D objects, such as molecules

- Can handle data defined in irregular domains, without mapping onto a regular grid
- Graph representations are invariant to translation and rotation
- Much sparser representation of 3D objects than voxel grids
- Many possibilities for data integration
- Efficient update mechanisms with graph convolution and weight sharing

➔ **GCNs perform well in tasks where an understanding of the structure and the interconnectivity of data is important**

➔ **Models usually perform well with relatively few parameters**



Thank you for your attention