

Week2.2 Bio-LLMs

2025 Spring GenAI

Dr. Yanjun Qi

20250121

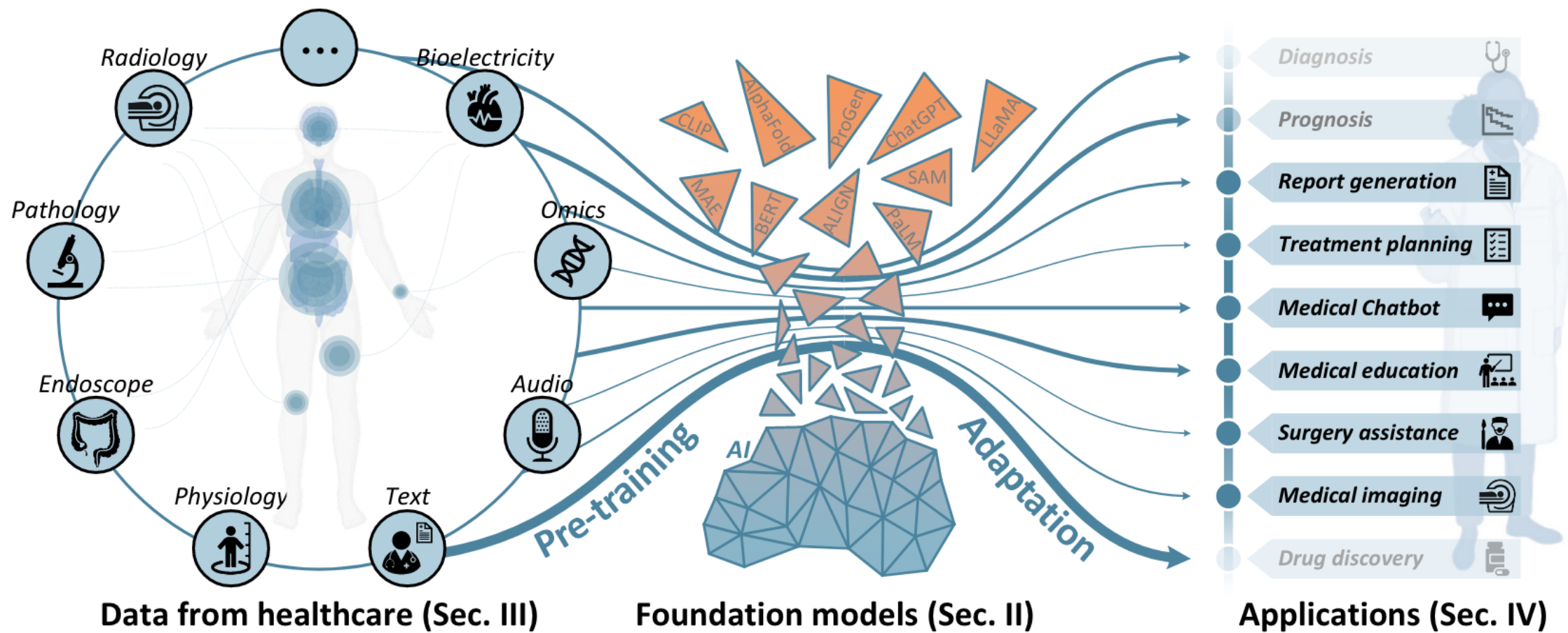
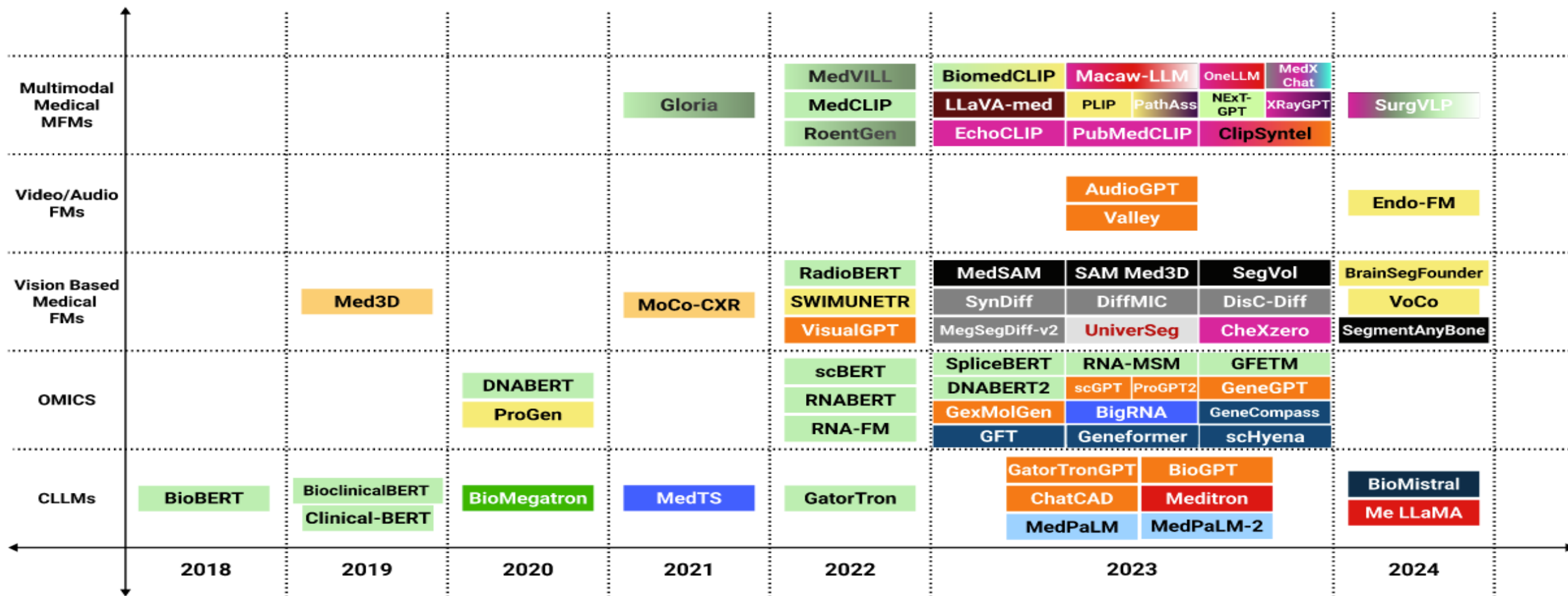


Fig. 1. The pipeline of the healthcare foundation models (HFMs) including the methods (Sec.II), datasets (Sec.III), and applications (Sec.IV).

The base models used to develop medical foundation models.



Development of medical foundation models for multiple healthcare applications (2018-2024)

Scientific Large Language Models: A Survey on Biological & Chemical Domains

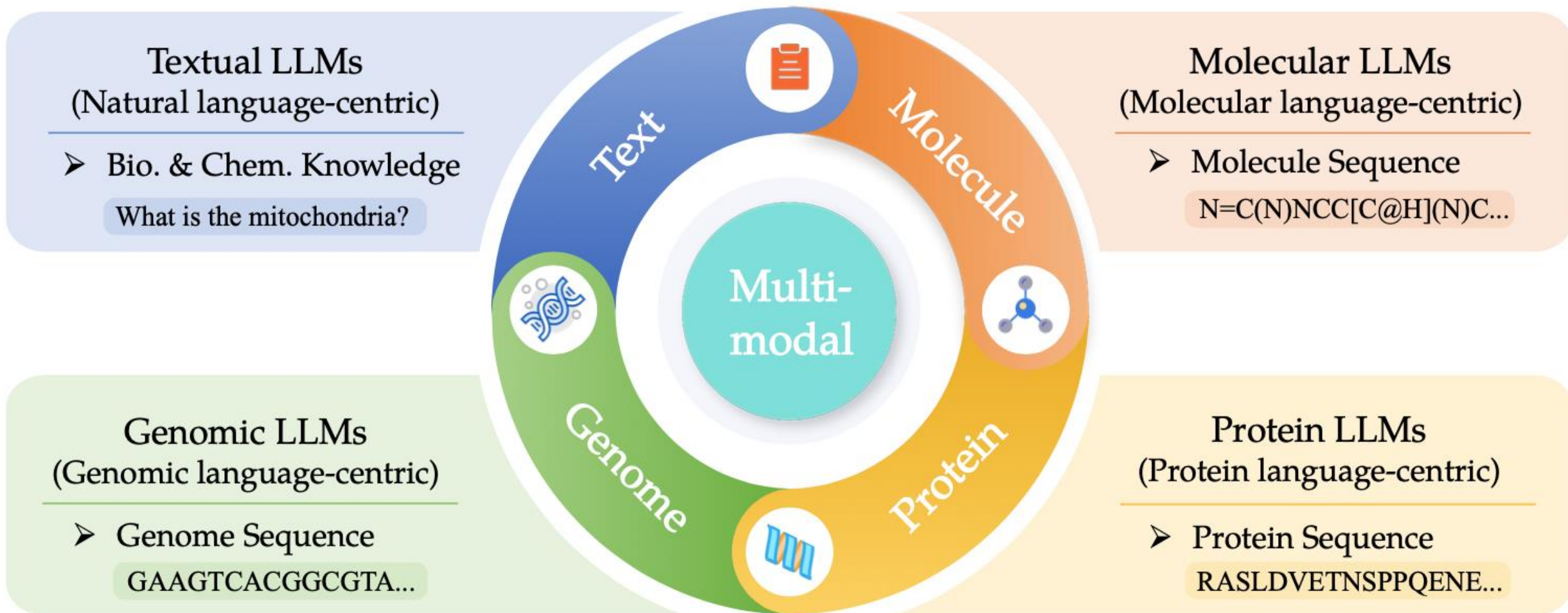
Large Language Models (LLMs) have emerged as a transformative power in enhancing natural language comprehension, representing a significant stride toward artificial general intelligence. The application of LLMs extends beyond conventional linguistic boundaries, encompassing specialized linguistic systems developed within various scientific disciplines. This growing interest has led to the advent of scientific LLMs, a novel subclass specifically engineered for facilitating scientific discovery. As a burgeoning area in the community of AI for Science, scientific LLMs warrant comprehensive exploration. However, a systematic and up-to-date survey introducing them is currently lacking. In this paper, we endeavor to methodically delineate the concept of “scientific language”, whilst providing a thorough review of the latest advancements in scientific LLMs. Given the expansive realm of scientific disciplines, our analysis adopts a focused lens, concentrating on the biological and chemical domains. This includes an in-depth examination of LLMs for textual knowledge, small molecules, macromolecular proteins, genomic sequences, and their combinations, analyzing them in terms of model architectures, capabilities, datasets, and evaluation. Finally, we critically examine the prevailing challenges and point out promising research directions along with the advances of LLMs. By offering a comprehensive overview of technical developments in this field, this survey aspires to be an invaluable resource for researchers navigating the intricate landscape of scientific LLMs.

Additional Key Words and Phrases: Scientific domain, large language models, protein, molecule, genome

ACM Reference Format:

Qiang Zhang, Keyan Ding, Tianwen Lyu, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, Kehua Feng, Xiang Zhuang, Zeyuan Wang, Ming Qin, Mengyao Zhang, Jinlu Zhang, Jiyu Cui, Tao Huang, Pengju Yan, Renjun Xu, Hongyang Chen, Xiaolin Li, Xiaohui Fan, Huabin Xing, and Huajun Chen. 2024. Scientific Large Language Models: A Survey on Biological & Chemical Domains. 1, 1 (July 2024), 90 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

Language Models: A Survey on Biological & Chemical Domains



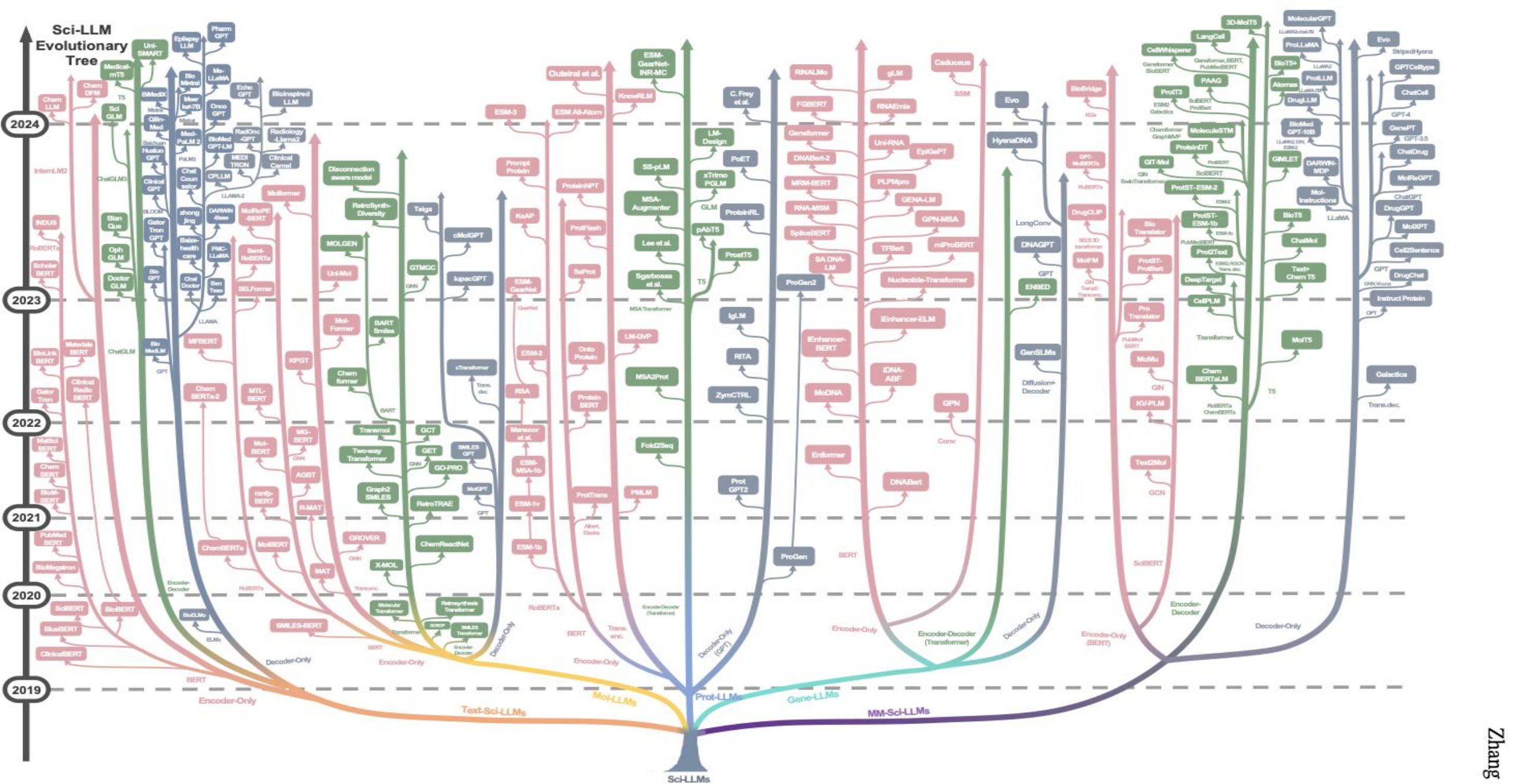


Fig. 3. An evolutionary tree of Sci-LLMs, which consists of five main branches corresponding to the research scopes in this survey. Due to the extensive number of Sci-LLMs, it is not feasible to include all of them in this figure, despite their exceptional quality. For detailed information on the featured models, please refer to Table

- Textual Tokens

<BOS>	aspirin	has	...	?	<EOS>
-------	---------	-----	-----	---	-------

- Protein Tokens

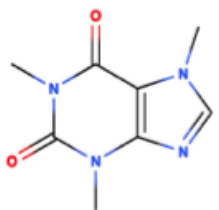


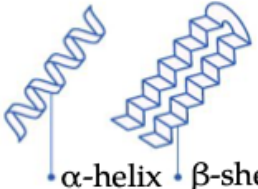




<BOS>	M	E	...	V	<EOS>
-------	---	---	-----	---	-------

- Molecular Tokens

<BOS>	C	NH2	...	(=O)	<EOS>
-------	---	-----	-----	------	-------

- Genomic Tokens

<BOS>	AGT	CG	...	AA	<EOS>
-------	-----	----	-----	----	-------

Molecule	SMILES:	<chem>OC(=O)C1=CC=CC=C1O</chem>		
	SELFIES:	[O][C][=Branch1][C][=O][C][=C][C][=C][C][=C][Ring1][=Branch1][O]		
	InChI:	1S/C7H6O3/c8-6-4-2-1-3-5(6)7(9)10/h1-4,8H,(H,9,10)		
<hr/>				
Protein				
	VDSPQERASLDEN...	α -helix β -sheet		
	Primary Structure (Amino acid sequence)	Secondary Structure	Tertiary Structure	Quaternary Structure
	<hr/>			
Genome	DNA Sequence:	ATCGGTGACTATCG		
	RNA Sequence:	AUCGGUGACUAUCG	Double-stranded DNA Structure	Single-stranded RNA Structure

of molecular, protein and genomic languages. Molecular languages include SMILES, SELFIES and InChI sequences,

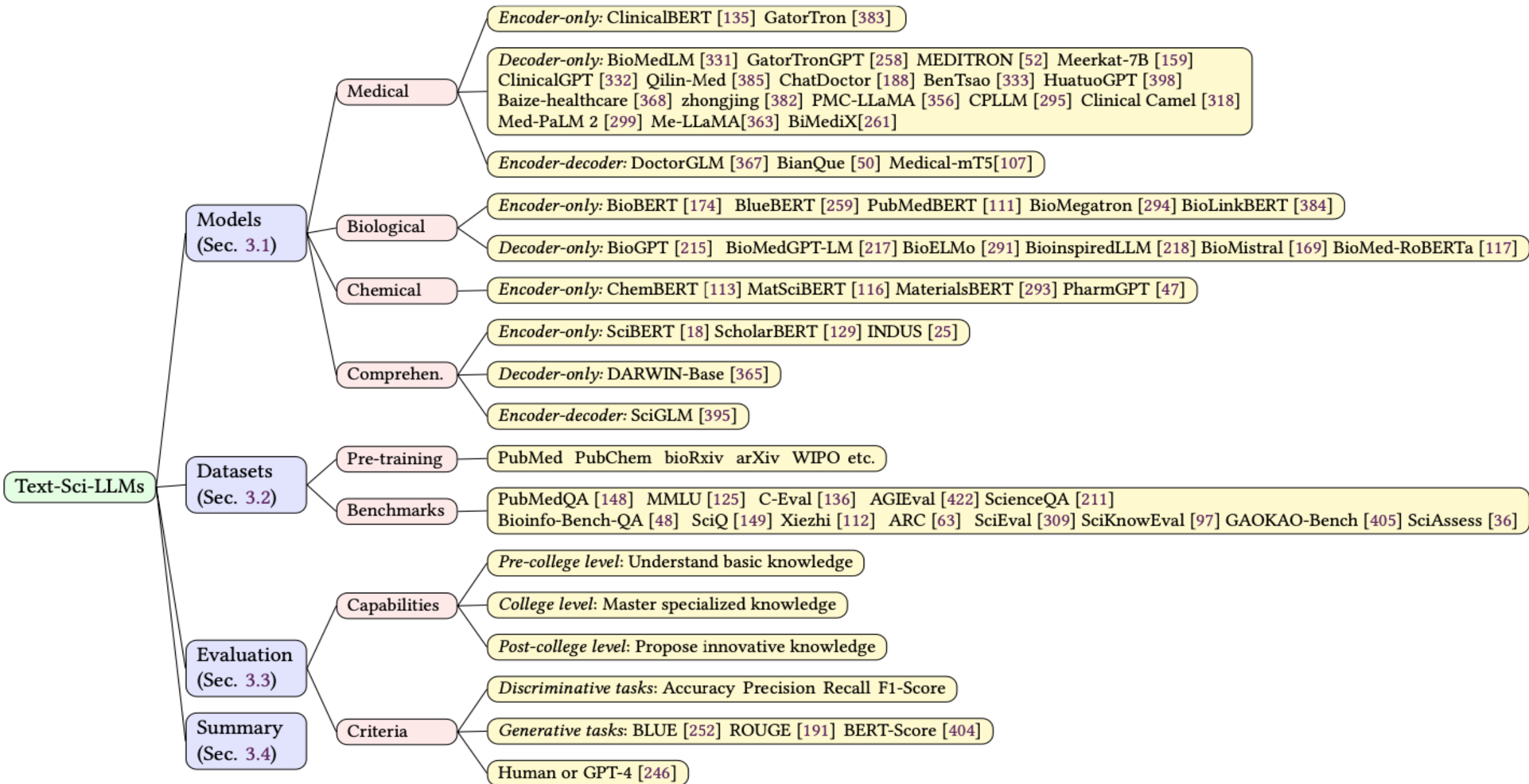


Fig. 6. Chapter overview of Text-Sci-LLMs.

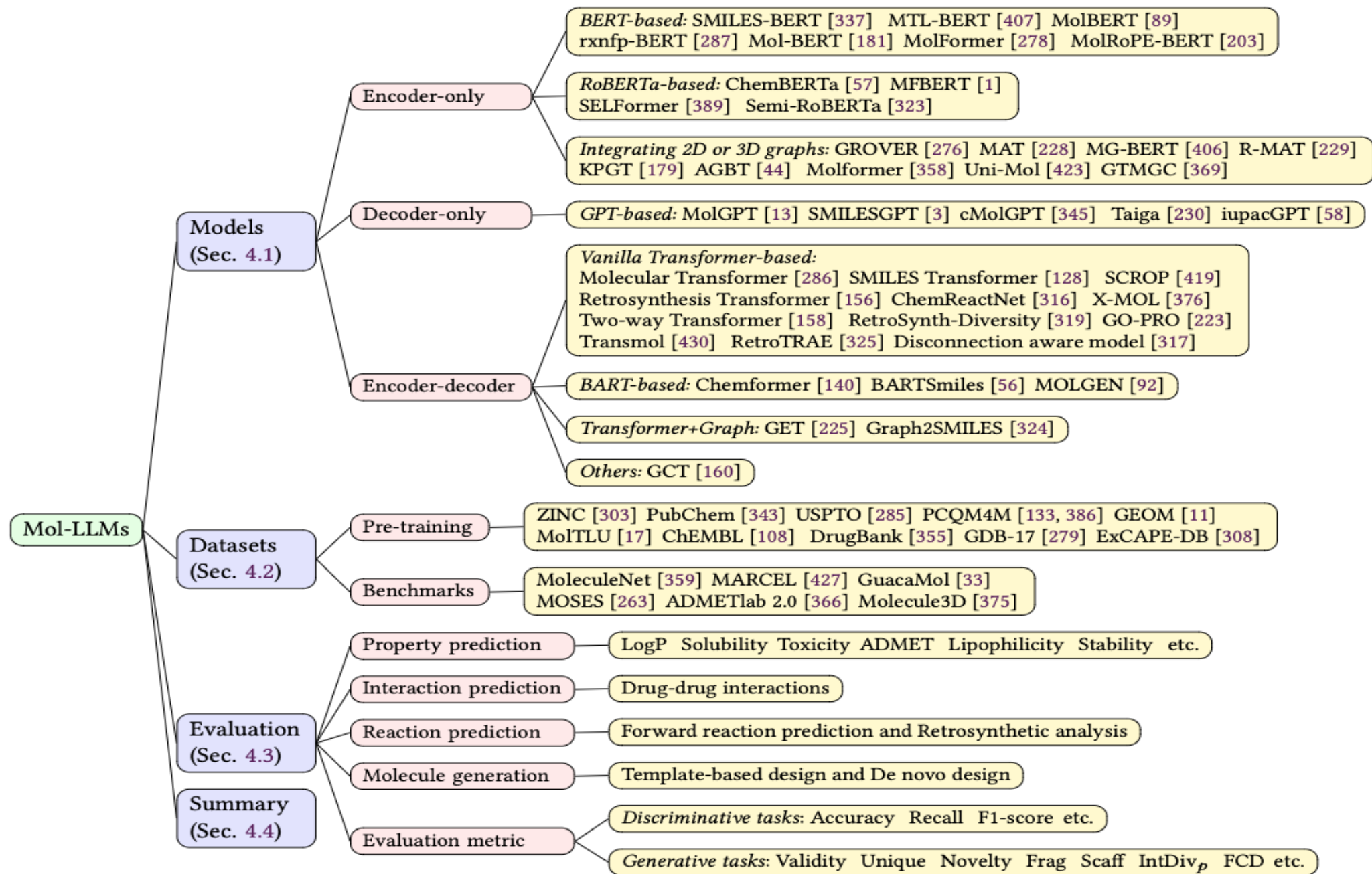


Fig. 7. Chapter overview of Mol-LLMs.

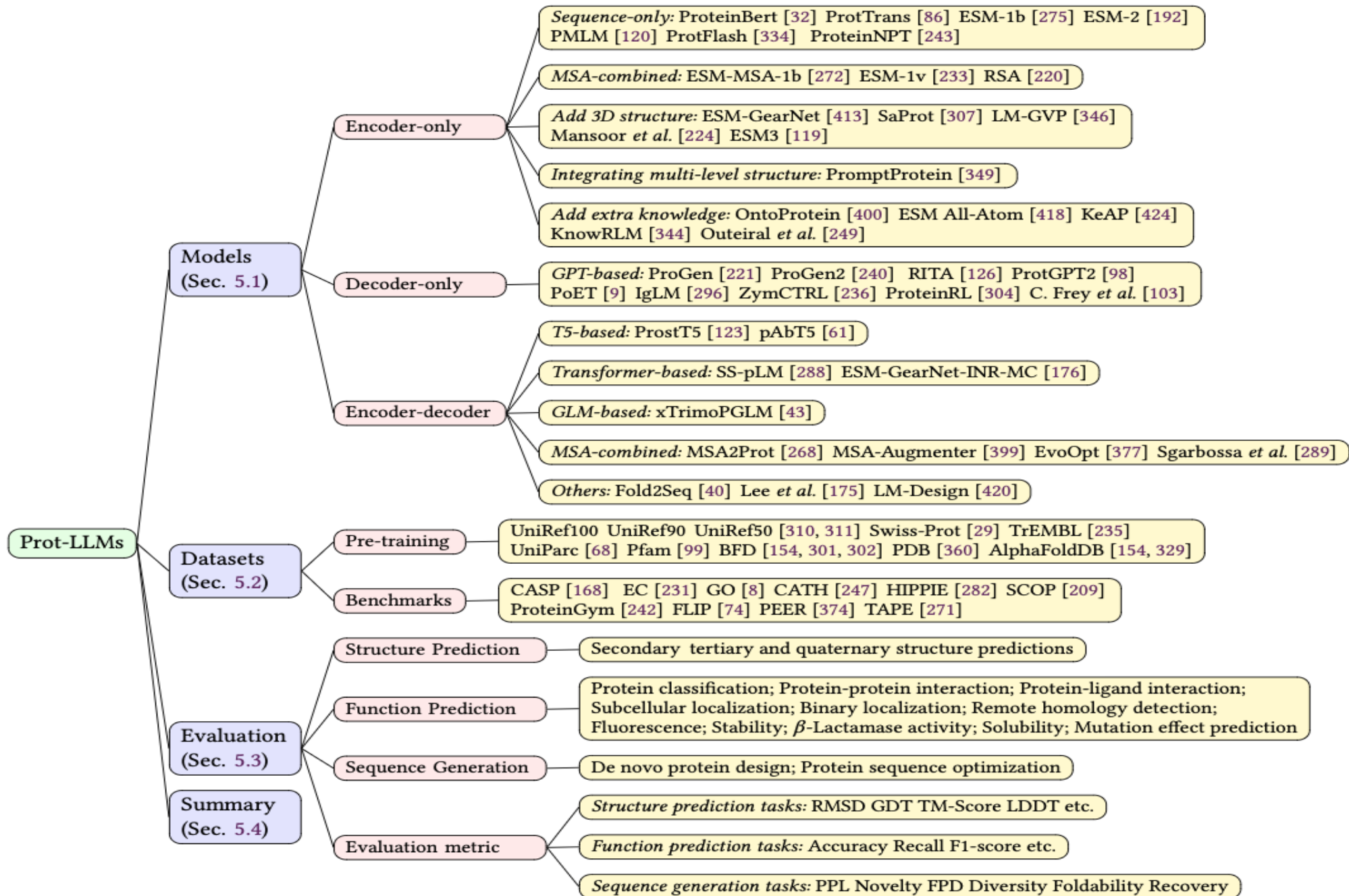


Fig. 8. Chapter overview of Prot-LLMs.

Table 5. Summary of Prot-LLMs

	Model	Time	#Parameters	Base model	Pretraining Dataset	Capability	Open-source
Encoder-only	ESM-1b [275]	2020.02	650M	RoBERTa	UniRef50	Secondary struct. pred., Contact pred., etc.	✓
	ESM-MSA-1b [272]	2021.02	100M	ESM-1b	UniRef50	Secondary struct. pred., Contact pred., etc.	✓
	ESM-1v [233]	2021.02	650M	ESM-1b	UniRef90	Mutation effect pred.	✓
	ProtTrans [86]	2021.07	-	BERT, Albert, Electra	UniRef, BFD	Secondary struct. pred., Func. pred., etc	✓
	PMLM [120]	2021.07	87M - 731M	Trans. enc.	Uniref50/Pfam	Contact pred.	×
	Mansoor <i>et al.</i> [224]	2021.09	100M	ESM-1b	-	Mutation effect pred.	×
	ProteinBERT [32]	2022.02	16M	BERT	UniRef90	Func. pred.	✓
	LM-GVP [346]	2022.04	-	Trans. enc	-	Func. pred.	✓
	RSA [220]	2022.05	-	ESM-1b	-	Func. pred.	✓
	OntoProtein [400]	2022.06	-	BERT	ProteinKG25	Func. pred.	✓
	ESM-2 [192]	2022.07	8M - 15B	RoBERTa	UniRef50	Func. pred., Struct. pred.	✓
	PromptProtein [349]	2023.02	650M	RoBERTa	UniRef50, PDB	Func. pred.	✓
	KeAP [424]	2023.02	-	RoBERTa	ProteinKG25	Func. pred.	✓
	ProtFlash [334]	2023.10	79M/174M	Trans. enc	UniRef50	Func. pred.	✓
	ESM-GearNet [413]	2023.10	-	ESM-1b, GearNet	-	Func. pred.	✓
	SaProt [307]	2023.10	650M	BERT	-	Mutation effect pred.	✓
	ProteinNPT [243]	2023.12	-	Trans. enc.	-	Fitness pred., Redesign	×
	Outeiral <i>et al.</i> [249]	2024.02	10M - 5B	Trans. enc.	European Nucleotide Archive	Protein represent learning	✓
ESM All-Atom [418]	2024.06	35M	RoBERTa	AlphaFold DB	Unified Molecular Modeling	×	
KnowRLM [344]	2024.06	-	Trans. enc.	-	Protein Directed Evolution	×	
ESM3 [119]	2024.06	98B	RoBERTa	PDB	Seq. pred., Func. pred., Struct. pred.	✓	
Decoder-only	ProGen [221]	2020.03	1.2B	GPT	Uniparc SWISS-Prot	Functional prot. gen.	✓
	ProtGPT2 [98]	2021.01	738M	GPT	Uniref50	De novo protein design and engineering	✓
	ZymCTRL [236]	2022.01	738M	GPT	BRENDA	Functional enzymes gen.	✓
	RITA [126]	2022.05	1.2B	GPT	UniRef100	Functional prot. gen.	×
	IgLM [296]	2022.12	13M	GPT	-	Antibody design	✓
	ProGen2 [240]	2023.10	151M - 6.4B	GPT	Uniref90, BFD30, PDB	Functional prot. gen.	✓
	ProteinRL [304]	2023.10	764M	GPT	-	Prot. design	×
	PoET [9]	2023.11	201M	GPT	-	Prot. family. gen.	×
	C. Frey <i>et al.</i> [103]	2024.03	9.87M/1.03M	GPT	hu4D5 antibody mutant	Functional prot. gen.	×
	Fold2Seq [40]	2021.01	-	Transformer	-	Prot. design	✓
MSA2Prot [268]	2022.04	-	Transformer	-	Prot. gen., Variant func. pred.	×	
Sgarbossa <i>et al.</i> [289]	2023.02	-	MSA Transformer	-	Prot. gen.	✓	

Many more
details in the
paper!

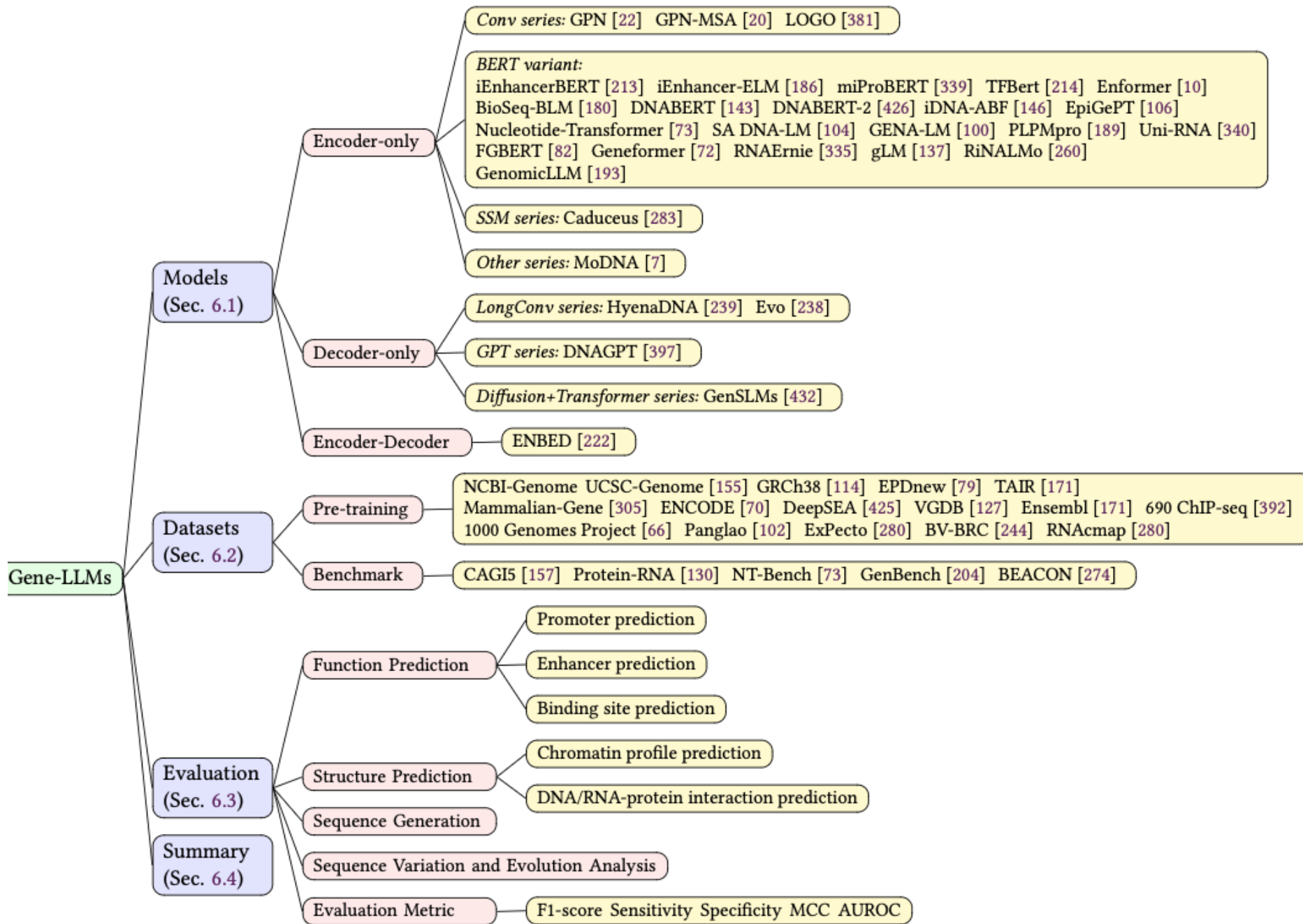


Fig. 9. Chapter overview of Gene-LLMs.

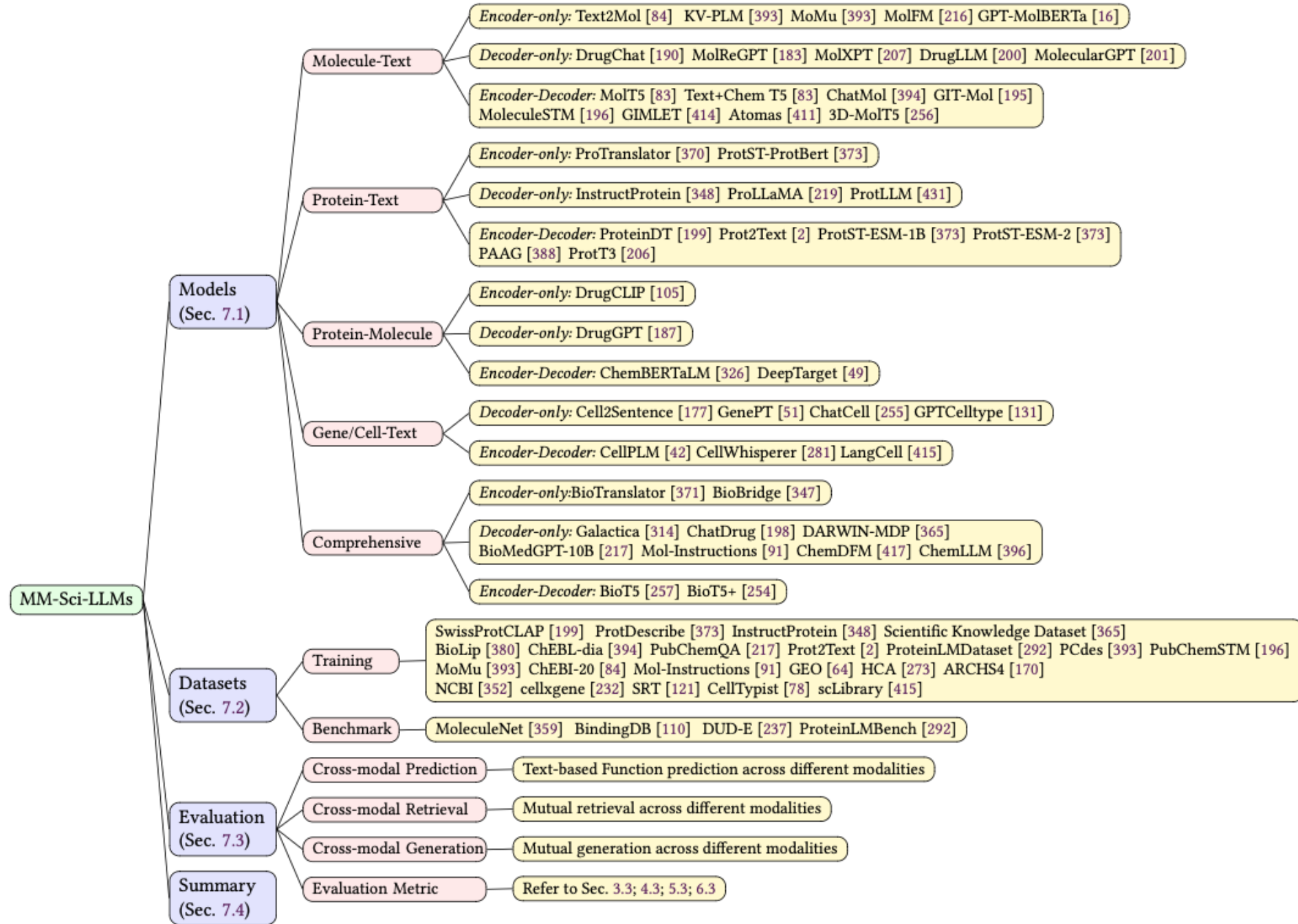


Fig. 10. Chapter overview of MM-Sci-LLMs.

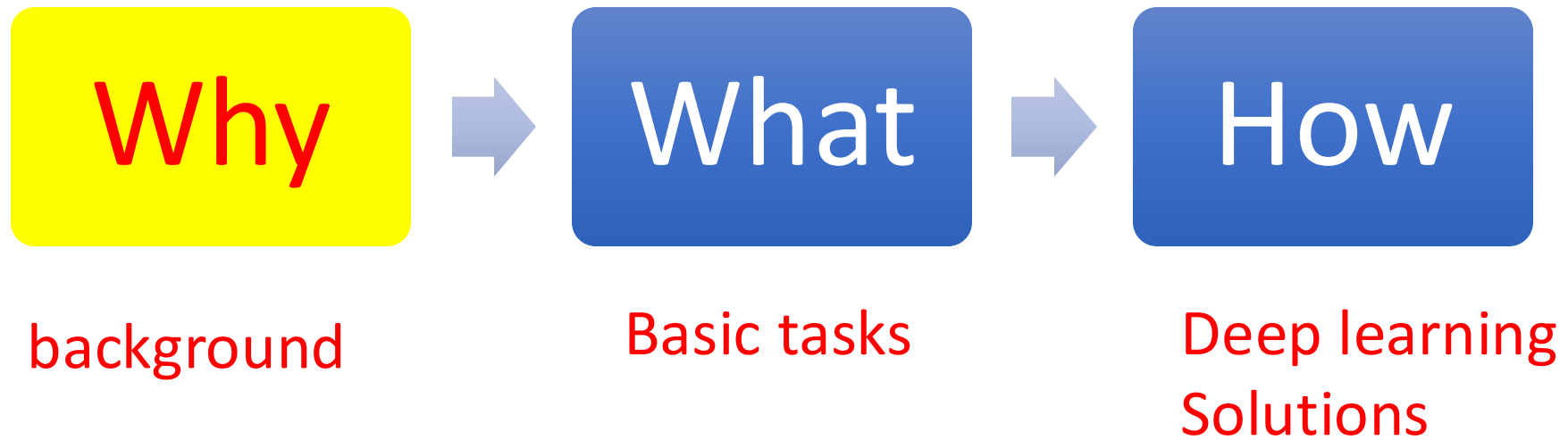
Backup:

A few classic deep learning
papers on Protein
Representation Learning

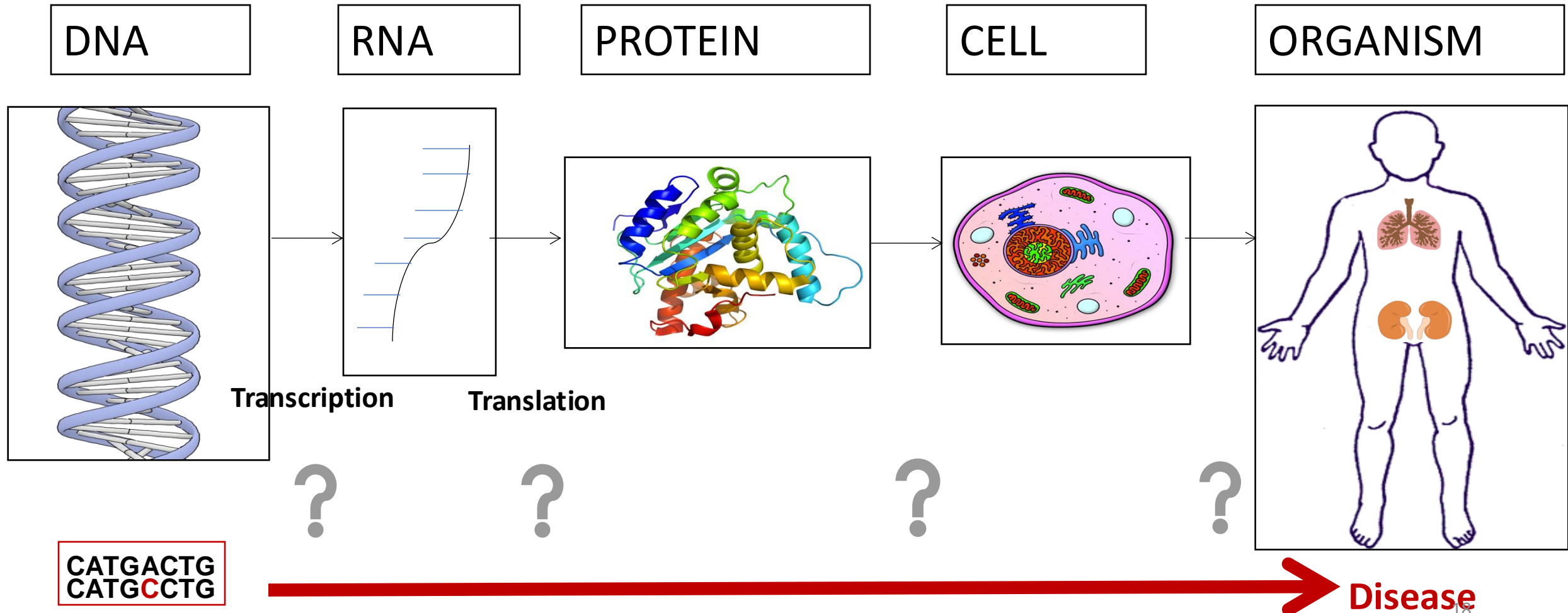
Selected Papers

- ESMfold:
 - Evolutionary-scale prediction of atomic level protein structure with a language model
- Alphafold2:
 - Highly Accurate Protein Structure Prediction with AlphaFold
- RoseTTAfold:
 - Accurate prediction of protein structures and interactions using a three-track neural network
- Related:
 - TRANSFORMER PROTEIN LANGUAGE MODELS ARE UNSUPERVISED STRUCTURE LEARNERS
 - Evfold: Protein 3D structure computed from evolutionary sequence variation

Roadmap

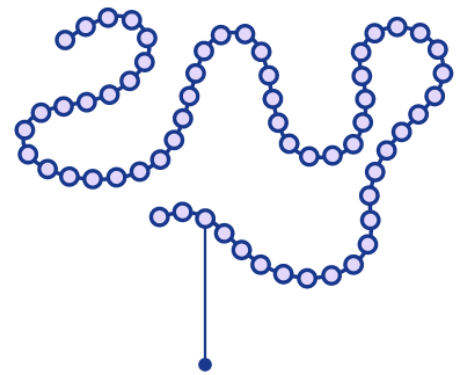


Biology in a Slide:



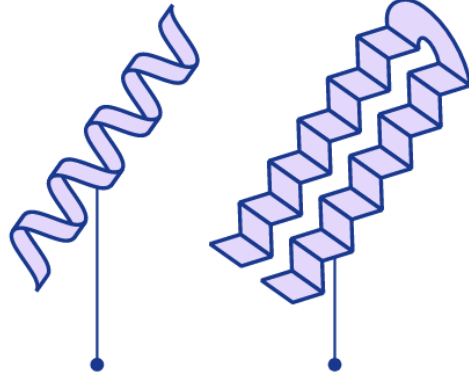
Protein Sequence form and Protein Structure

Every protein is made up of a sequence of amino acids bonded together



Amino acids

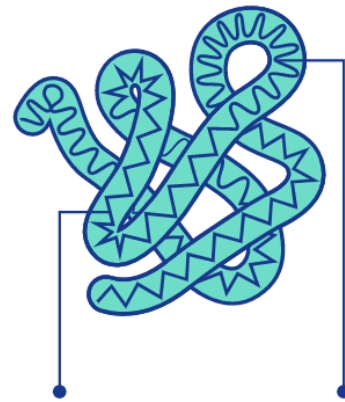
These amino acids interact locally to form shapes like helices and sheets



Alpha helix

Pleated sheet

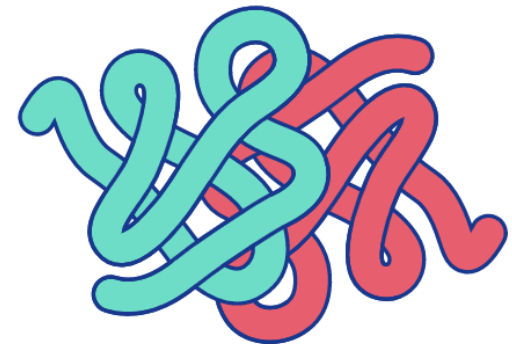
These shapes fold up on larger scales to form the full three-dimensional protein structure



Pleated sheet

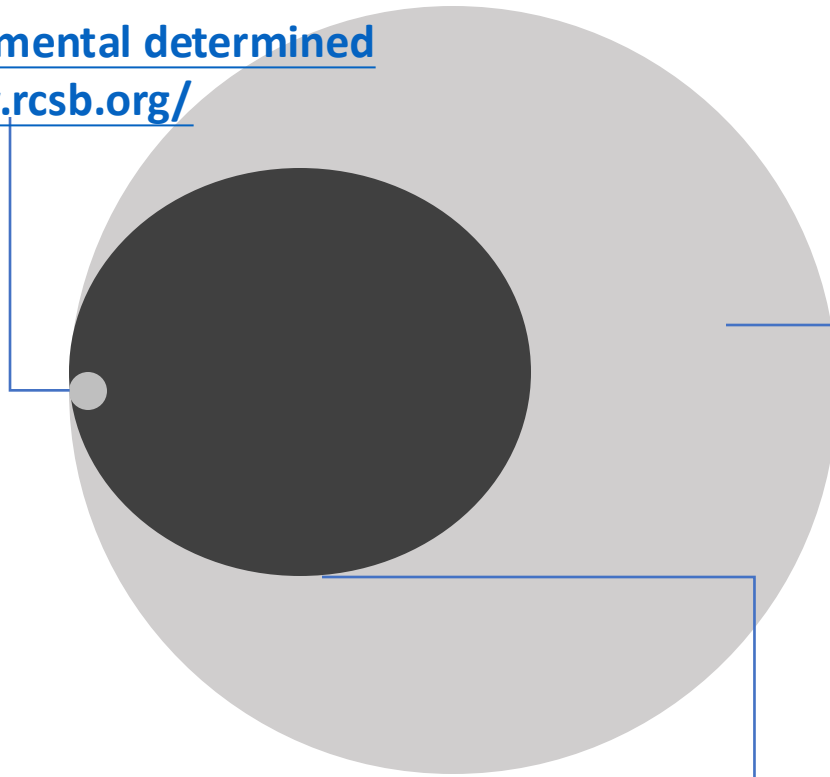
Alpha helix

Proteins can interact with other proteins, performing functions such as signalling and transcribing DNA



Protein Structure landscape

[~200k experimental determined](https://www.rcsb.org/)
<https://www.rcsb.org/>



ESM Metagenomic Atlas (<https://esmatlas.com>): 617M proteins. We are able to complete this characterization in 2 weeks on a heterogeneous cluster of 2,000 GPUs, demonstrating scalability to far larger databases. High confidence predictions are made for over 225M structures

AlphaFold DB 200 million structures in AlphaFold DB, 35% are considered to be highly accurate. Another 45% have reasonable accuracy enough for many studies

Why predicting protein structures?

Structure Prediction
Speed does matter!

Design of entirely new proteins:

- If a designed amino acid sequence could fold into the reliable structure that we desired?

To predict the complex structure of multiple interacting partners

- Proteins work in teams .. what is the interacting team's structure, affinity, function? Team with drug? Ligand? RNA? ...

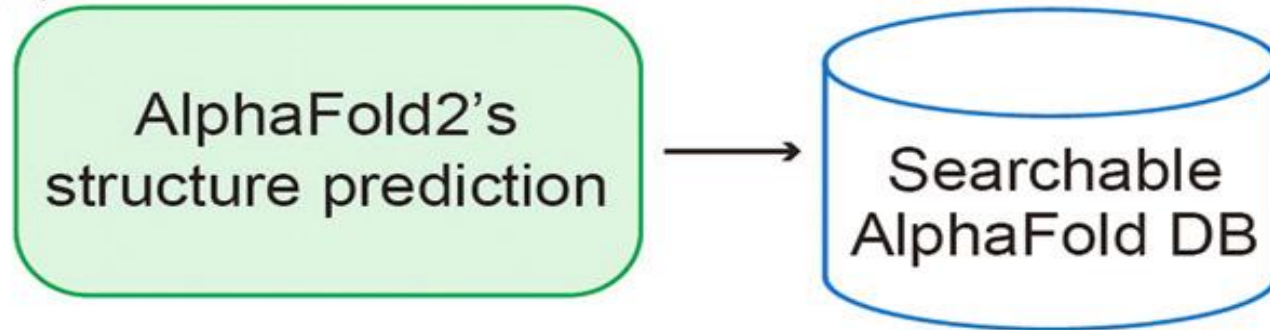
To illustrate the effect of mutations that contribute to rare genetic diseases.

- AlphaFold2 is not specifically designed and is unable to predict how amino acid mutations alter a protein's natural structure

Roadmap

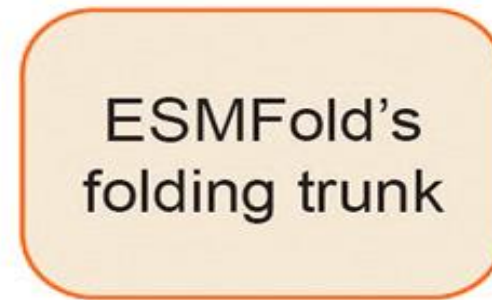
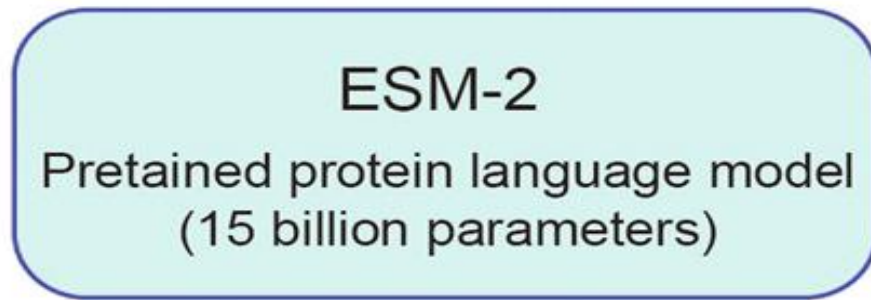


(A)



(B) FLDNMFGP R DSRVRG

ESMfold



FLD MFGP DSR RG

FLD MFGP DSR RG

The sequence "FLD MFGP DSR RG" is shown with colored boxes under each word: green for "FLD", orange for "MFGP", red for "DSR", and purple for "RG".

mask token

mask token

A grid of small squares representing a mask token, followed by the text "mask token".

Protein sequence to structure



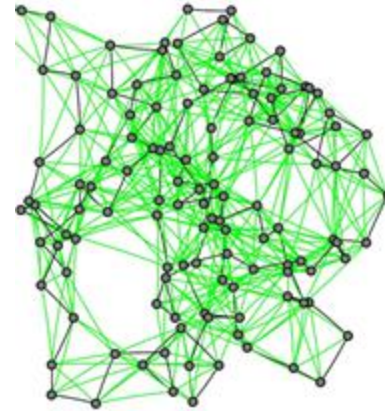
3D Structure



MHFTEDKATILWGKVNVEGETLGRVYPWQ

Primary Sequence

Protein sequence to structure



3D Structure

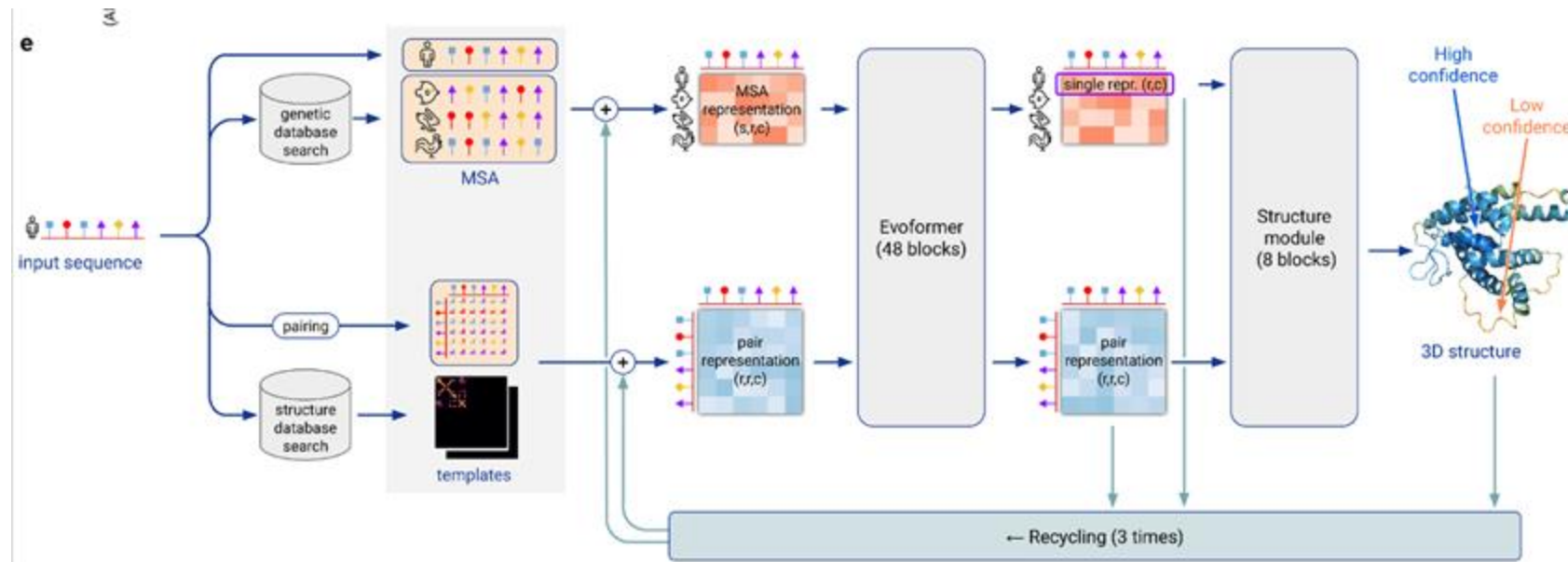


MHFTEDKATILWGKVNVEGETLGRVYPWQ

Primary Sequence

Accelerated Article Preview

Highly accurate protein structure prediction with AlphaFold



Received: 11 May 2021

Accepted: 12 July 2021

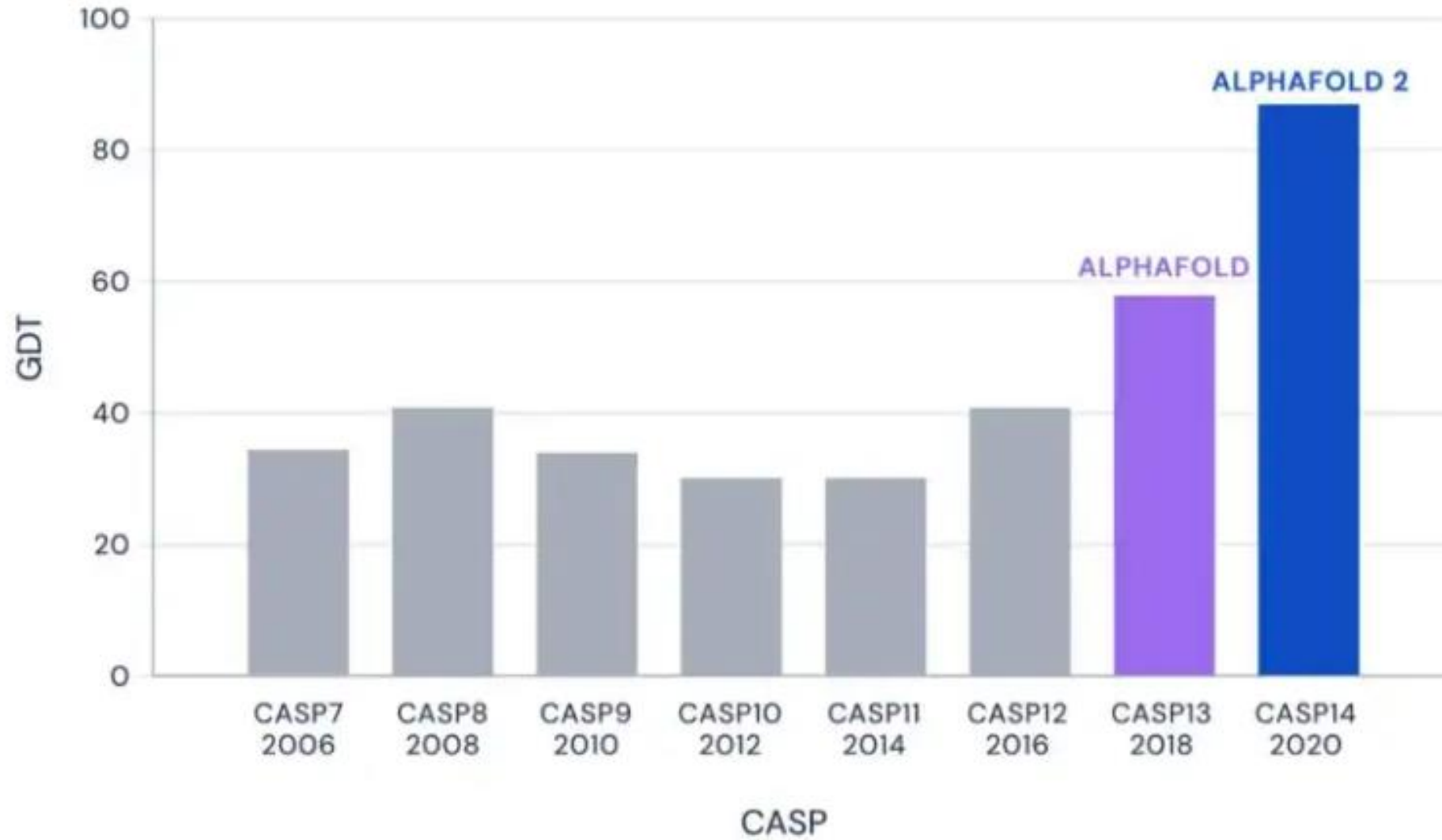
Accelerated Article Preview Published
online 15 July 2021

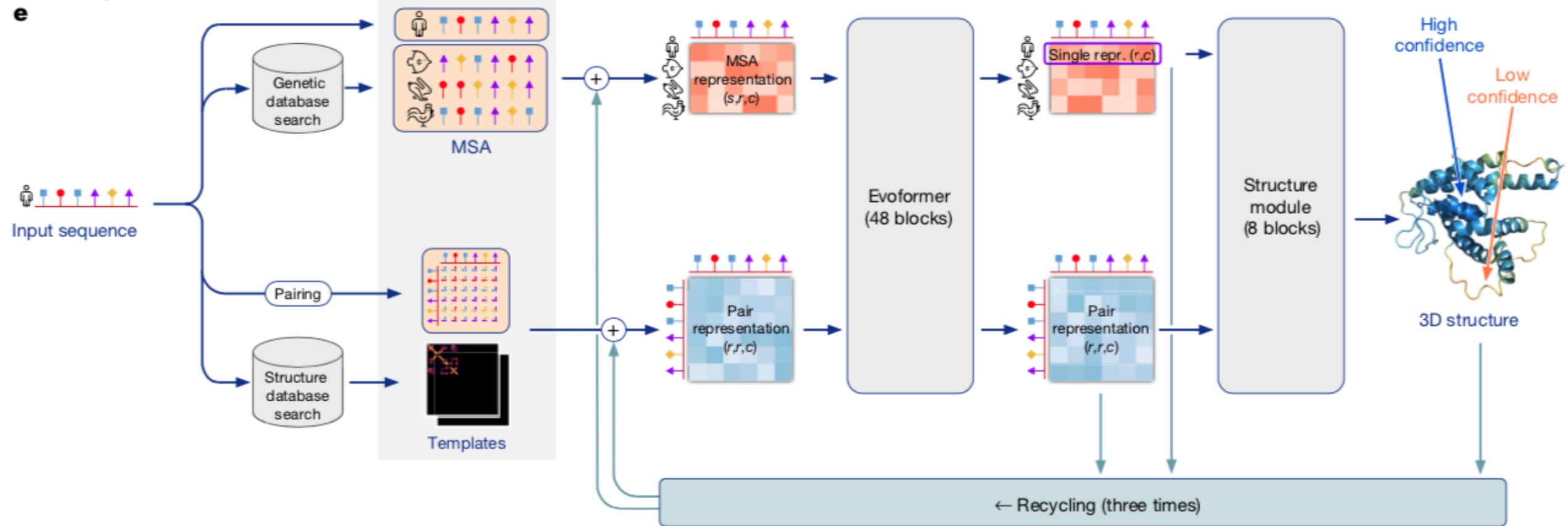
Cite this article as: Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* <https://doi.org/10.1038/s41586-021-03819-2> (2021).

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli & Demis Hassabis

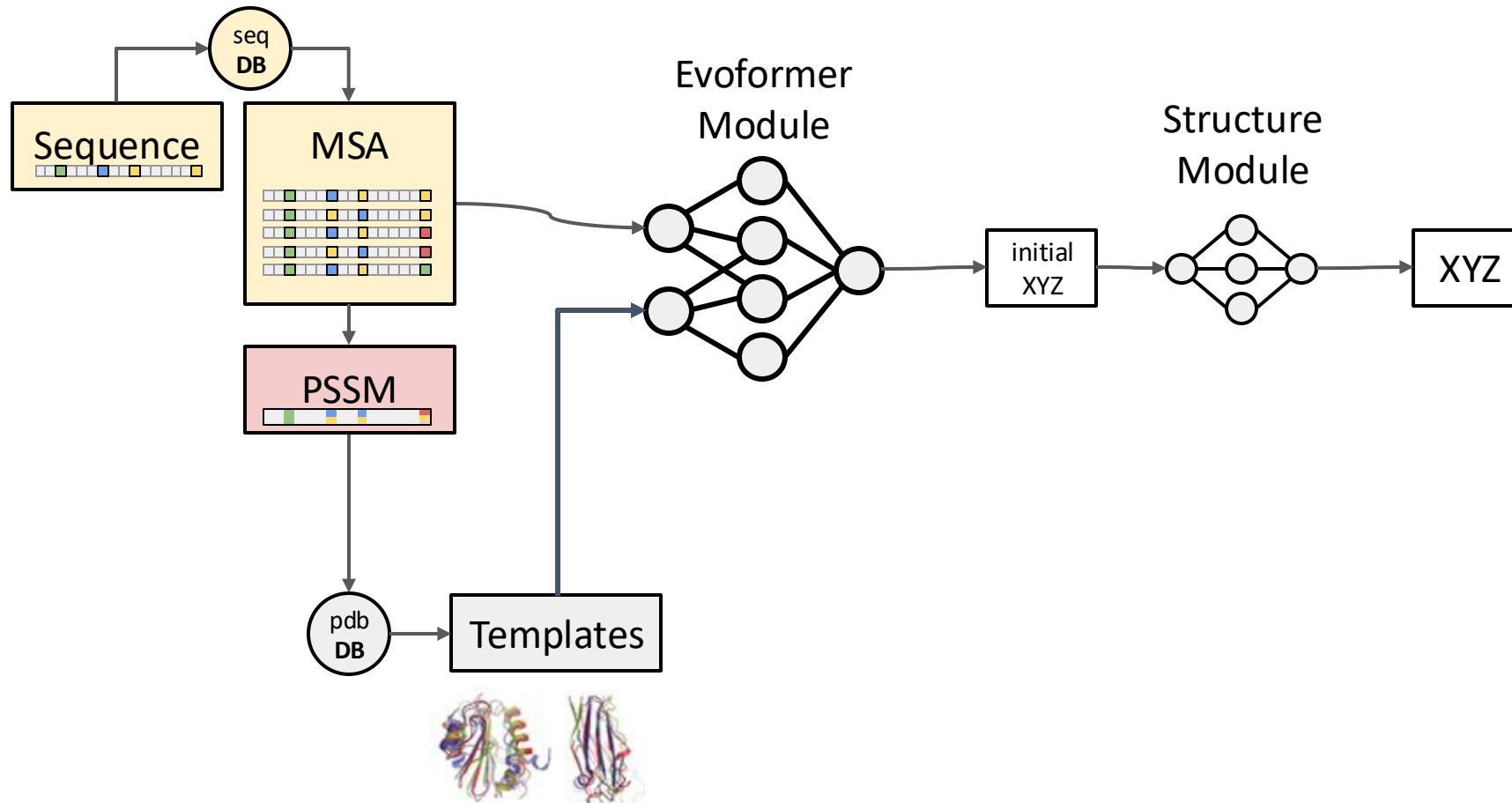
This is a PDF file of a peer-reviewed paper that has been accepted for publication.

Median Free-Modelling Accuracy





AlphaFold2*



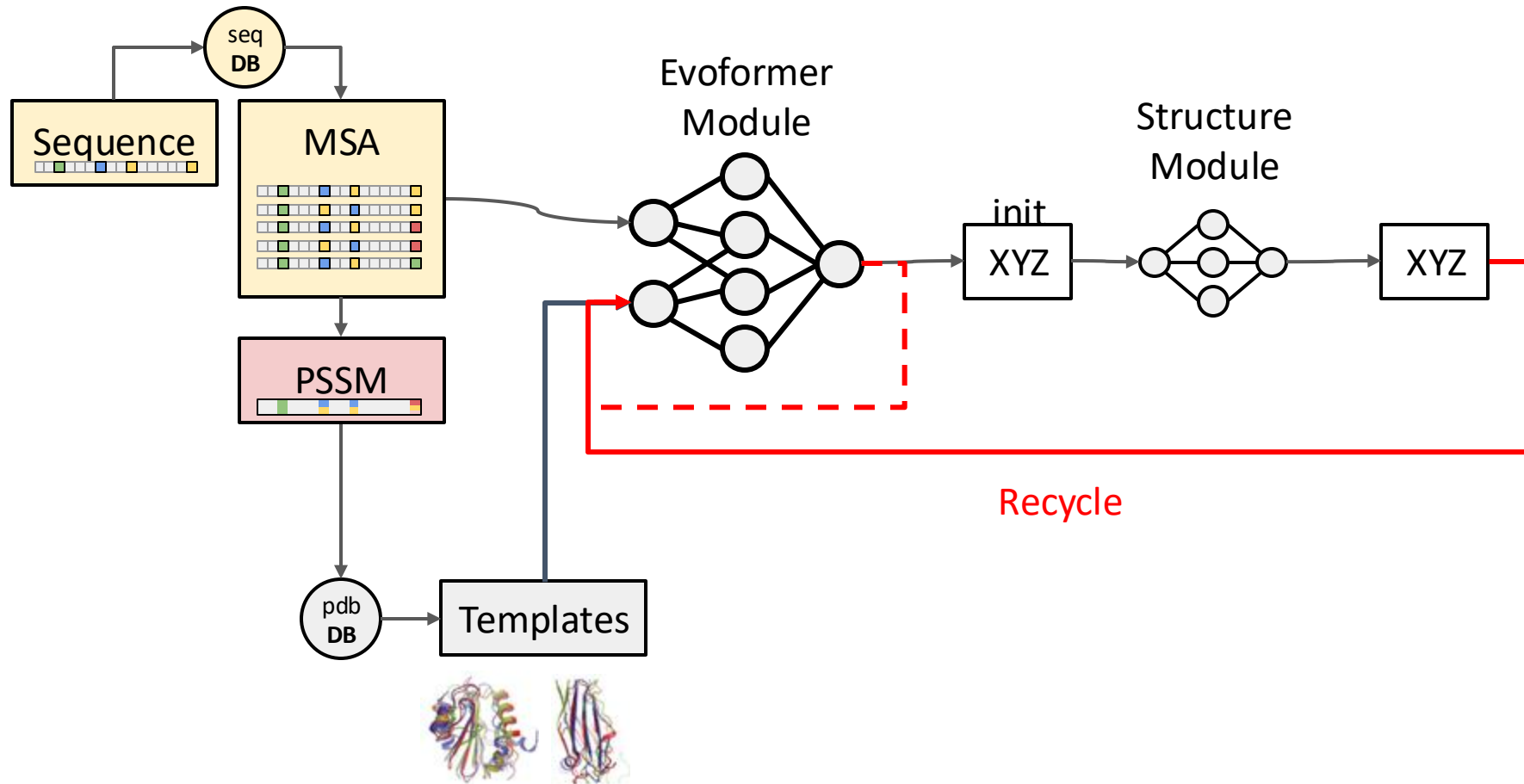
* Past researchers used raw Templates as input and/or did End2End

Analysis of distance-based protein structure prediction by deep learning in CASP13 - Jinbo Xu et al.

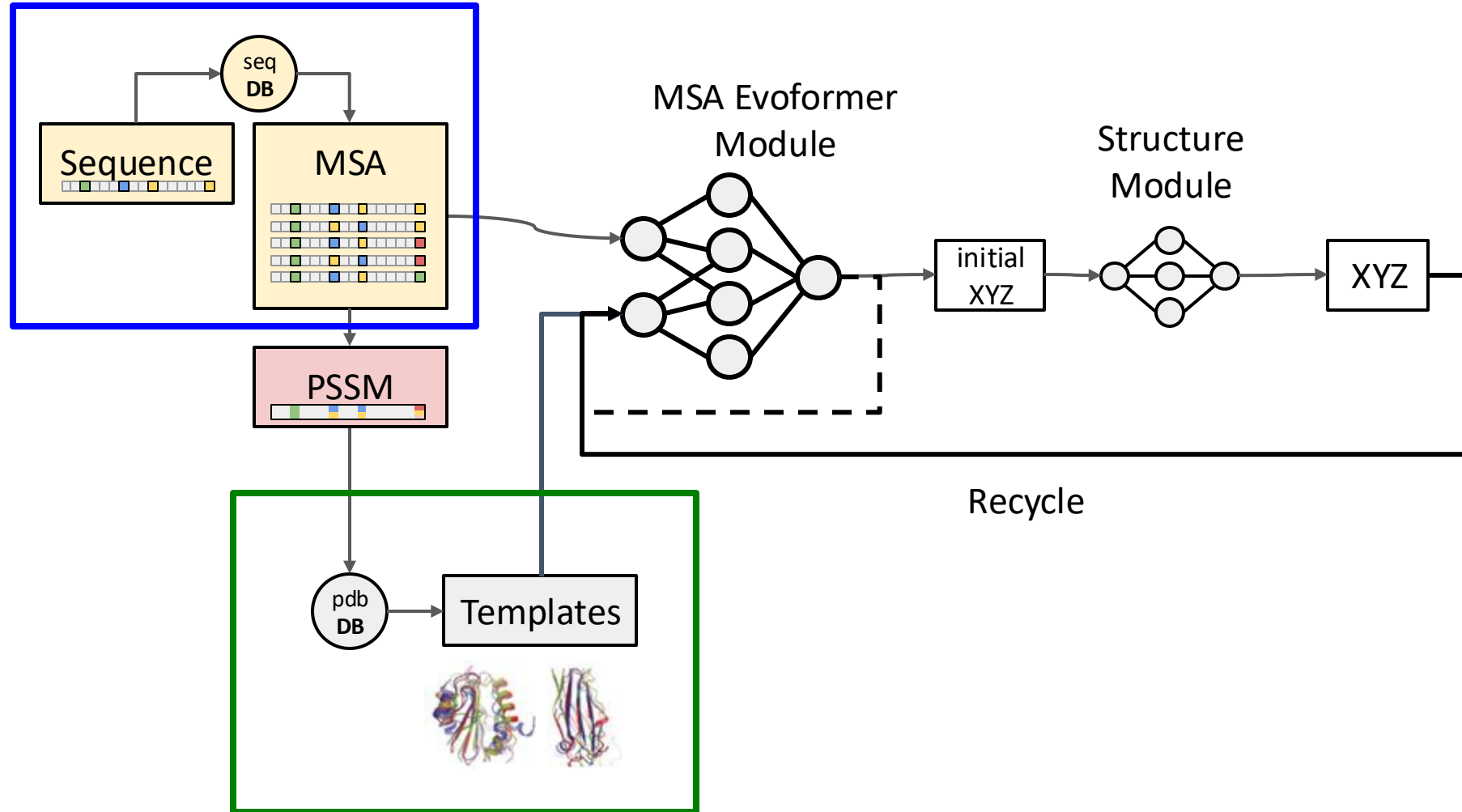
End-to-End Differentiable Learning of Protein Structure - Mohammed AlQuraishi

Learning Protein Structure with a Differentiable Simulator - John Ingraham et al.

AlphaFold2 - New Critical detail **Recycling**

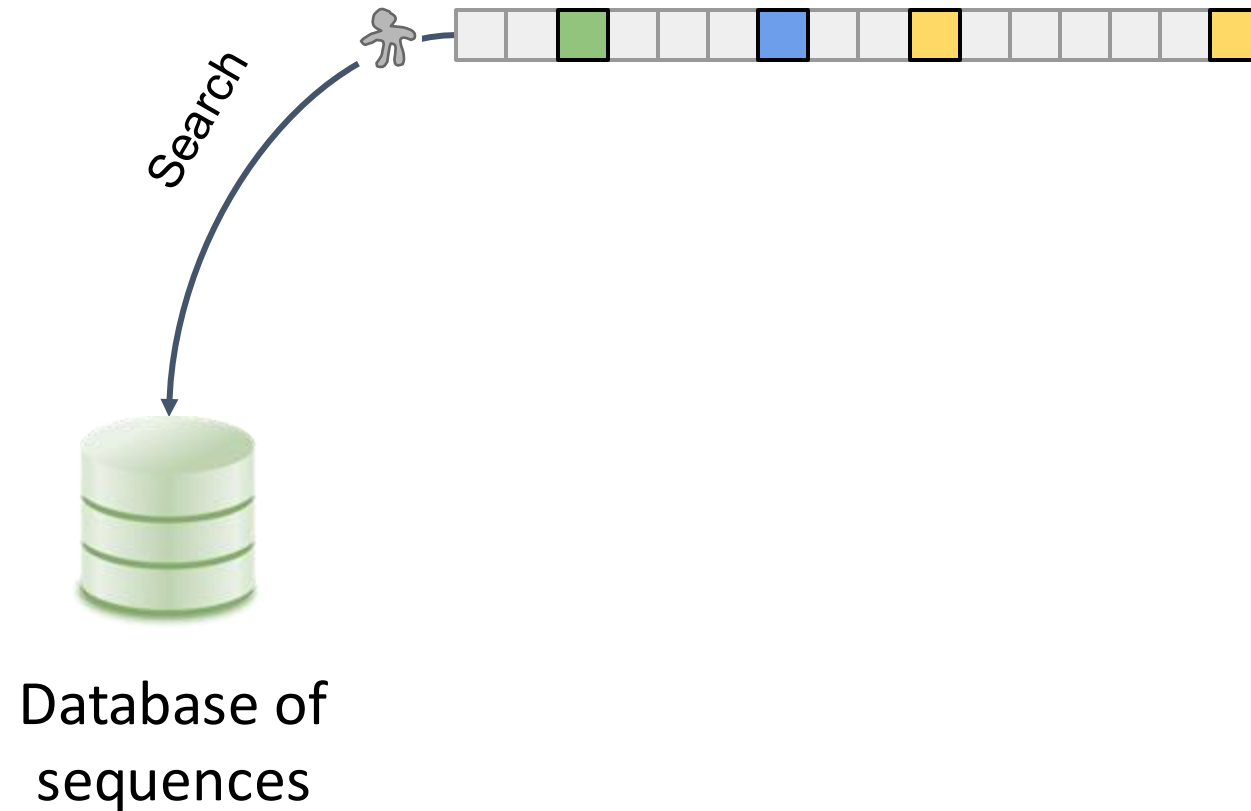


Structure Prediction relies on the input **MSA**

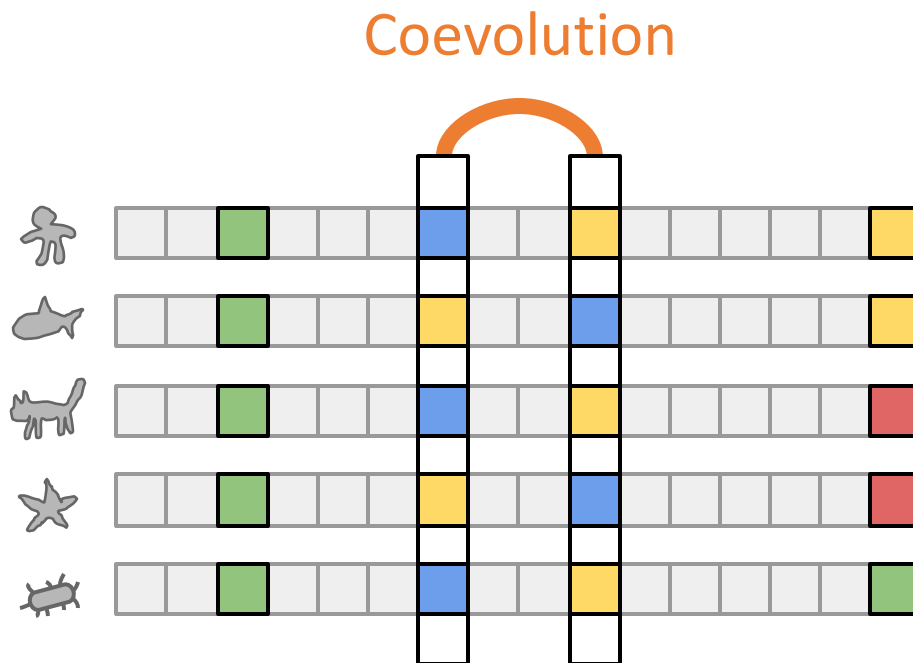


What is a Multiple Sequence Alignment (MSA)?

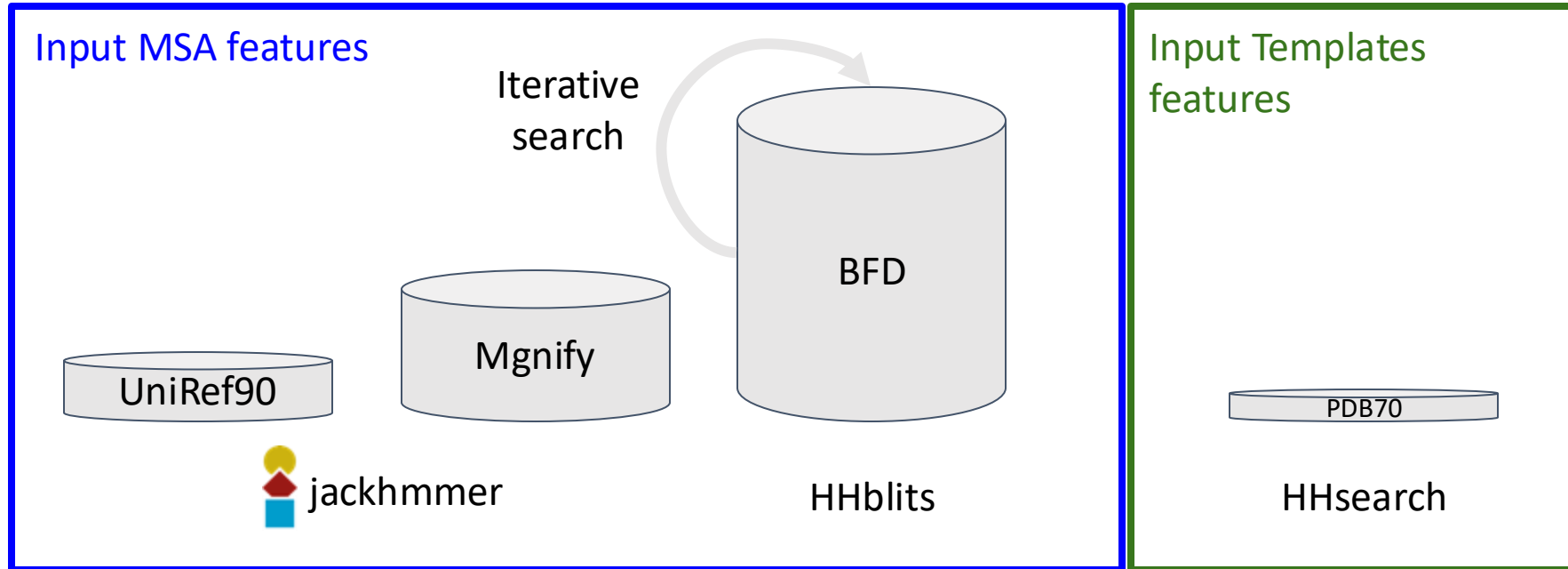
Search against a database of sequences



Analyze the MSA for coevolution



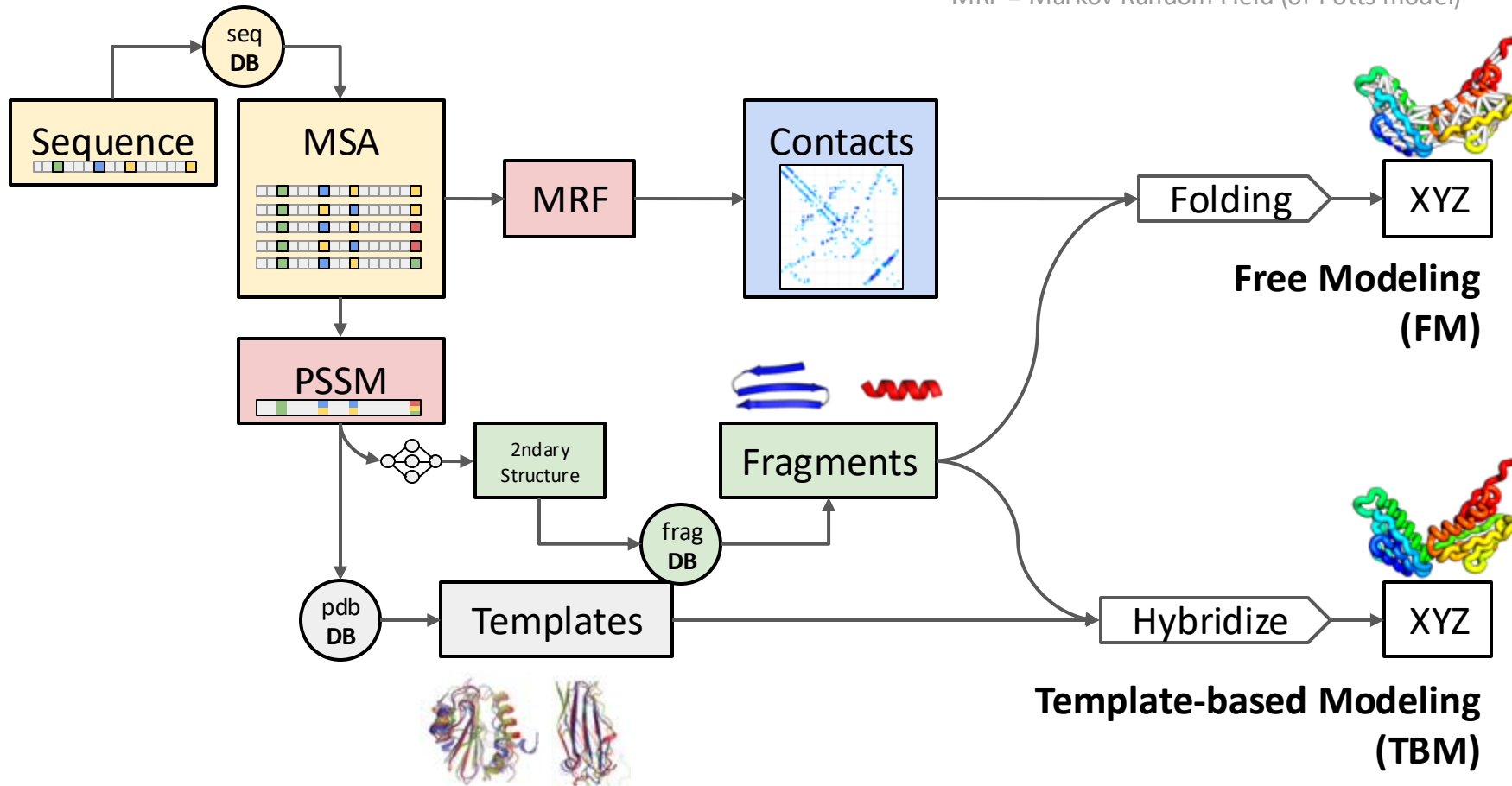
Input feature generation for AlphaFold2



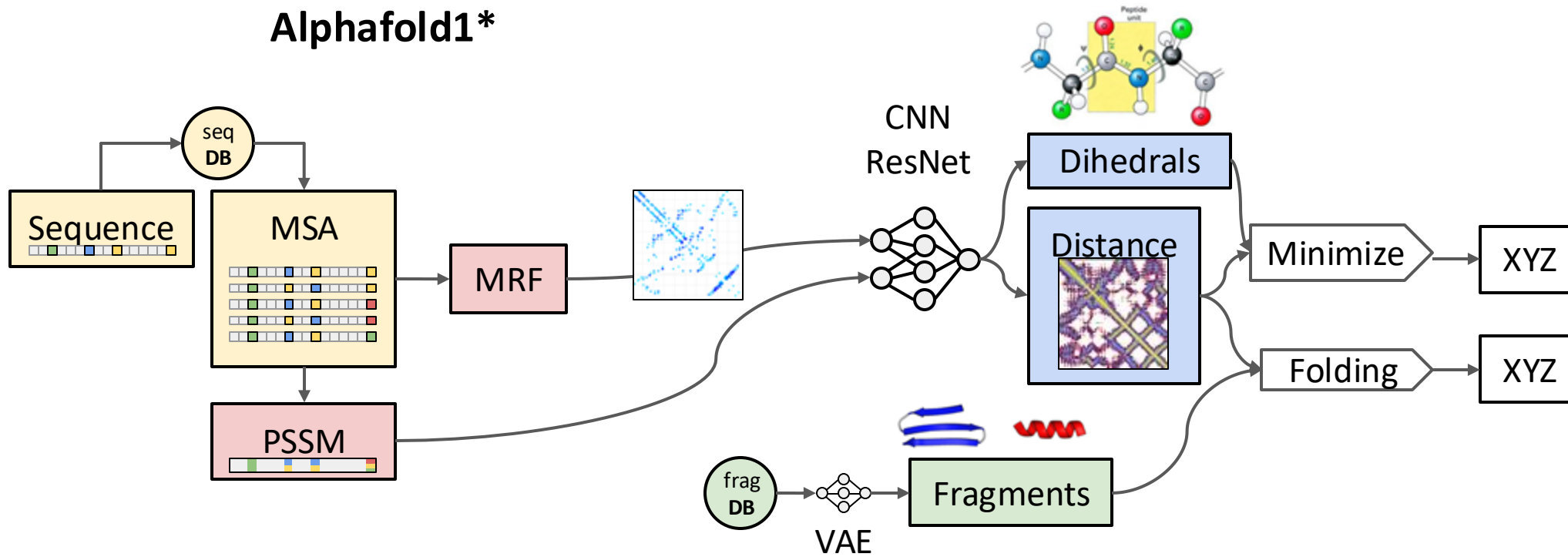
Generation of input features can take **hours** for a single protein on multiple cores

Typical pipeline before Alphafold

MSA = multiple sequence alignment
PSSM = Position-specific-scoring matrix
MRF = Markov Random Field (or Potts model)



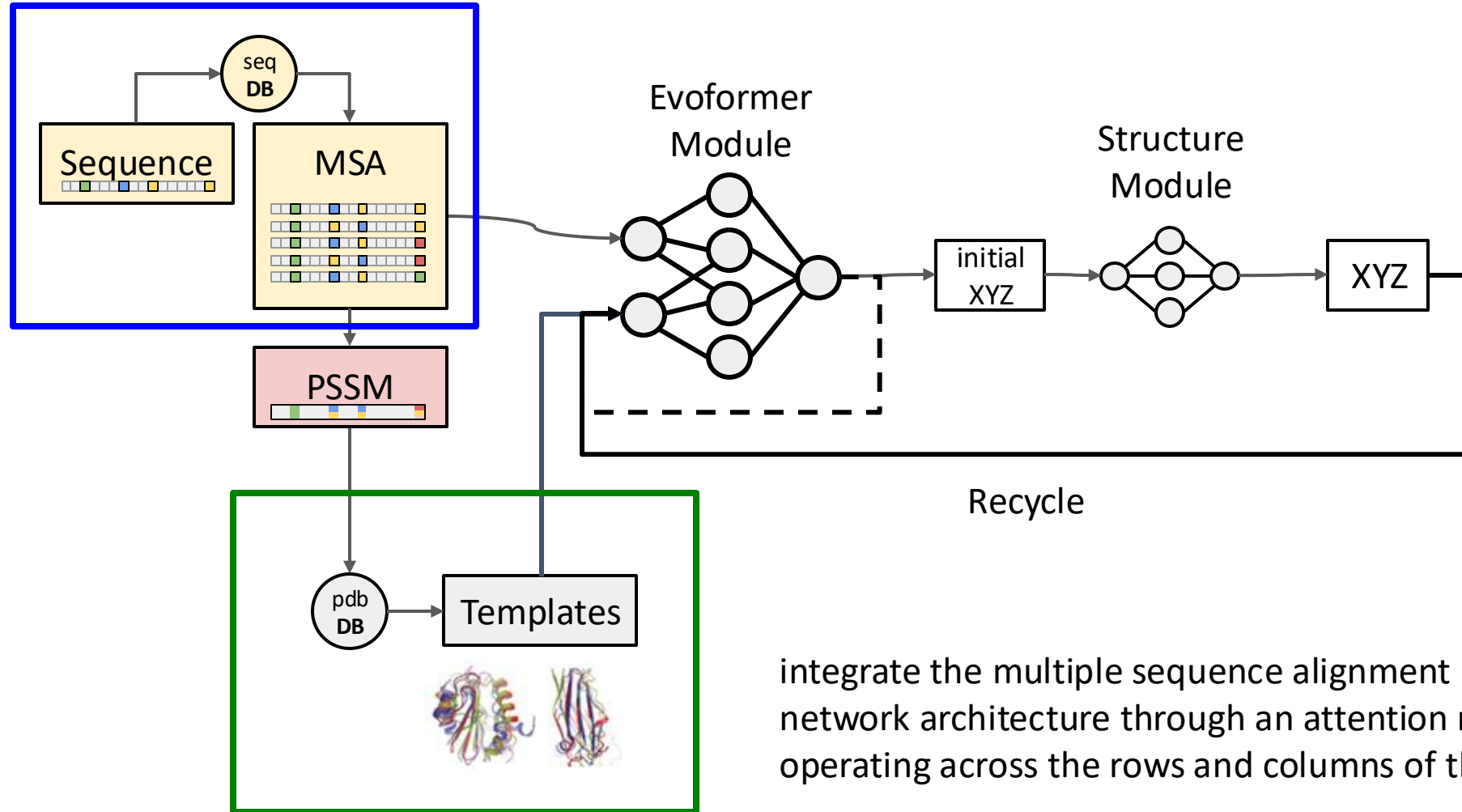
AlphaFold1*



* Past researchers used raw MRF features, and ResNets:

- Golkov, V., Skwark, M.J., Golkov, A., Dosovitskiy, A., Brox, T., Meiler, J. and Cremers, D., 2016, December. **Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images.** In *NIPS* (pp. 4215-4223).
- Wang, S., Sun, S., Li, Z., Zhang, R. and Xu, J., 2017. **Accurate de novo prediction of protein contact map by ultra-deep learning model.** *PLoS computational biology*, 13(1), p.e1005324.

AlphaFold2:



integrate the multiple sequence alignment into the neural network architecture through an attention mechanism operating across the rows and columns of the MSA

TRANSFORMER PROTEIN LANGUAGE MODELS ARE UNSUPERVISED STRUCTURE LEARNERS

Roshan Rao*
UC Berkeley
rmrao@berkeley.edu

Joshua Meier
Facebook AI Research
jmeier@fb.com

Tom Sercu
Facebook AI Research
tsercu@fb.com

Sergey Ovchinnikov
Harvard University
so@g.harvard.edu

Alexander Rives
Facebook AI Research & New York University
arives@cs.nyu.edu

ABSTRACT

Unsupervised contact prediction is central to uncovering physical, structural, and functional constraints for protein structure determination and design. For decades, the predominant approach has been to infer evolutionary constraints from a set of related sequences. In the past year, protein language models have emerged as a potential alternative, but performance has fallen short of state-of-the-art approaches in bioinformatics. In this paper we demonstrate that Transformer attention maps learn contacts from the unsupervised language modeling objective. We find the highest capacity models that have been trained to date already outperform a state-of-the-art unsupervised contact prediction pipeline, suggesting these pipelines can be replaced with a single forward pass of an end-to-end model.¹

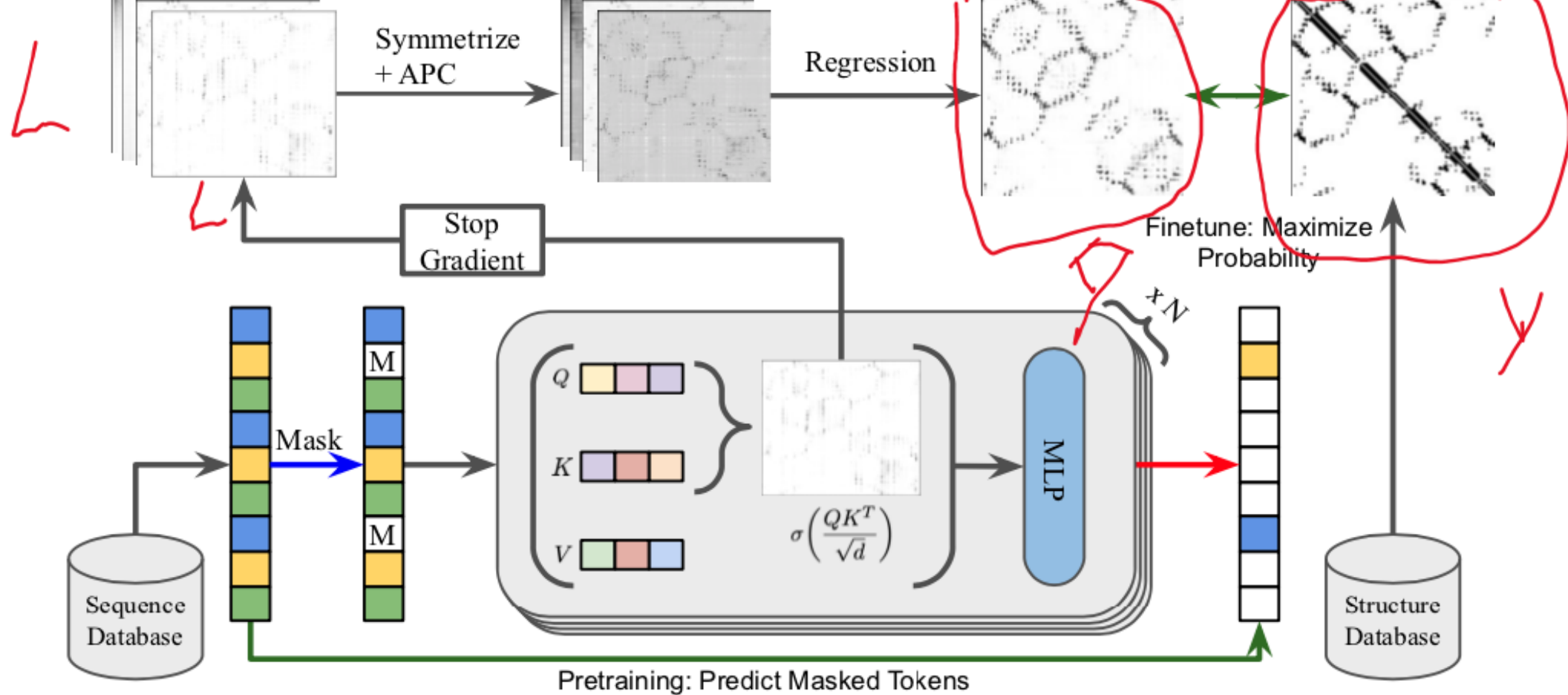


Figure 1: Contact prediction pipeline. The Transformer is first pretrained on sequences from a large database (Uniref50) via Masked Language Modeling. Once finished training, the attention maps are extracted, passed through symmetrization and average product correction, then into a regression. The regression is trained on a small number ($n \leq 20$) of proteins to determine which attention heads are informative. At test time, contact prediction from an input sequence can be done entirely on GPU in a single forward pass.



From Language model to contact (==> no MSA!!!)

Structure prediction from contacts While we do not perform structure prediction in this work, many methods have been proposed to extend contact prediction to structure prediction. For example, EVFold (Marks et al., 2011) and DCAFold (Sulkowska et al., 2012) predict co-evolving couplings using a Potts Model and then generate 3D conformations by directly folding an initial conformation with simulated annealing, using the predicted residue-residue contacts as constraints. Similarly, FragFold (Kosciolek & Jones, 2014) and Rosetta (Ovchinnikov et al., 2016) incorporate constraints from a Potts Model into a fragment assembly based pipeline. Senior et al. (2019), use features from a Potts model fit with pseudolikelihood maximization to predict pairwise distances with a deep residual network and optimize the final structure using Rosetta. All of these works build directly upon the unsupervised contact prediction pipeline.

Supervised contact prediction Recently, supervised methods using deep learning have resulted in breakthrough results in *supervised* contact prediction (Wang et al., 2017; Jones & Kandathil, 2018; Yang et al., 2019; Senior et al., 2020; Adhikari & Elofsson, 2020). State-of-the-art methods use deep residual networks trained with supervision from many protein structures. Inputs are typically covariance statistics (Jones & Kandathil, 2018; Adhikari & Elofsson, 2020), or inferred coevolutionary parameters (Wang et al., 2017; Liu et al., 2018; Senior et al., 2020; Yang et al., 2019). Other recent work with deep learning uses sequences or evolutionary features as inputs (AlQuraishi, 2018; Ingraham et al., 2019). Xu et al. (2020) demonstrates the incorporation of coevolutionary features is critical to performance of current state-of-the-art methods.

Unsupervised contact prediction In contrast to supervised methods, unsupervised contact prediction models are trained on sequences *without information from protein structures*. In principle this allows them to take advantage of large sequence databases that include information from many sequences where no structural knowledge is available. The main approach has been to learn evolutionary constraints among a set of similar sequences by fitting a Markov Random Field (Potts model) to the underlying MSA, a technique known as Direct Coupling Analysis (DCA). This was proposed by Lapedes et al. (1999) and reintroduced by Thomas et al. (2008) and Weigt et al. (2009).

Contact prediction from protein language models Since the introduction of large scale language models for natural language processing (Vaswani et al., 2017; Devlin et al., 2019), there has been considerable interest in developing similar models for proteins (Alley et al., 2019; Rives et al., 2019; Heinzinger et al., 2019; Rao et al., 2019; Elnaggar et al., 2020; Lu et al., 2020; Madani et al., 2020; Shen et al., 2021). Rives et al. (2019) were the first to study protein Transformer language models, demonstrating that information about residue-residue contacts could be recovered from the learned representations by linear projections supervised with protein structures. Recently Vig et al. (2020) performed an extensive analysis of Transformer attention, identifying correspondences to biologically relevant features, and also found that different layers of the model are responsible for learning different features. In particular Vig et al. (2020) discovered a correlation between self-attention maps and contact patterns, suggesting they could be used for contact prediction.

Prior work benchmarking contact prediction with protein language models has focused on the supervised problem. Bepler & Berger (2019) were the first to fine-tune an LSTM pretrained on protein sequences to fit contacts. Rao et al. (2019) and Rives et al. (2020) perform benchmarking of multiple protein language models using a deep residual network fit with supervised learning on top of pretrained language modeling features.

Evolutionary-scale prediction of atomic level protein structure with a language model

Zeming Lin^{1 2 *} Halil Akin^{1 *} Roshan Rao^{1 *} Brian Hie^{1 3 *} Zhongkai Zhu¹ Wenting Lu¹ Nikita Smetanin¹
Robert Verkuil¹ Ori Kabeli¹ Yaniv Shmueli¹ Allan dos Santos Costa⁴ Maryam Fazel-Zarandi¹ Tom Sercu^{1 †}
Salvatore Candido^{1 †} Alexander Rives^{1 † ‡}

Abstract

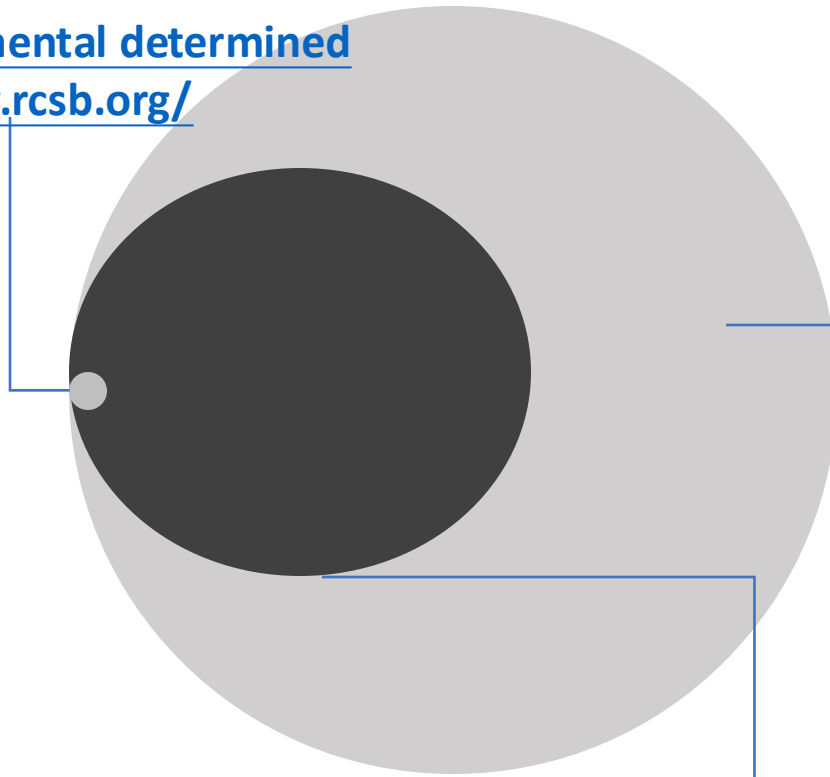
Artificial intelligence has the potential to open insight into the structure of proteins at the scale of evolution. It has only recently been possible to extend protein structure prediction to two hundred million cataloged proteins. Characterizing the structures of the exponentially growing billions of protein sequences revealed by large scale gene sequencing experiments would necessitate a breakthrough in the speed of folding. Here we show

1. Introduction

The sequences of proteins at the scale of evolution contain an image of biological structure and function. This is because the biological properties of a protein act as constraints on the mutations to its sequence that are selected through evolution, recording structure and function into evolutionary patterns (1–3). Within a protein family, structure and function can be inferred from the patterns in sequences (4, 5). This insight has been central to progress in computational structure prediction starting from classical methods (6, 7),

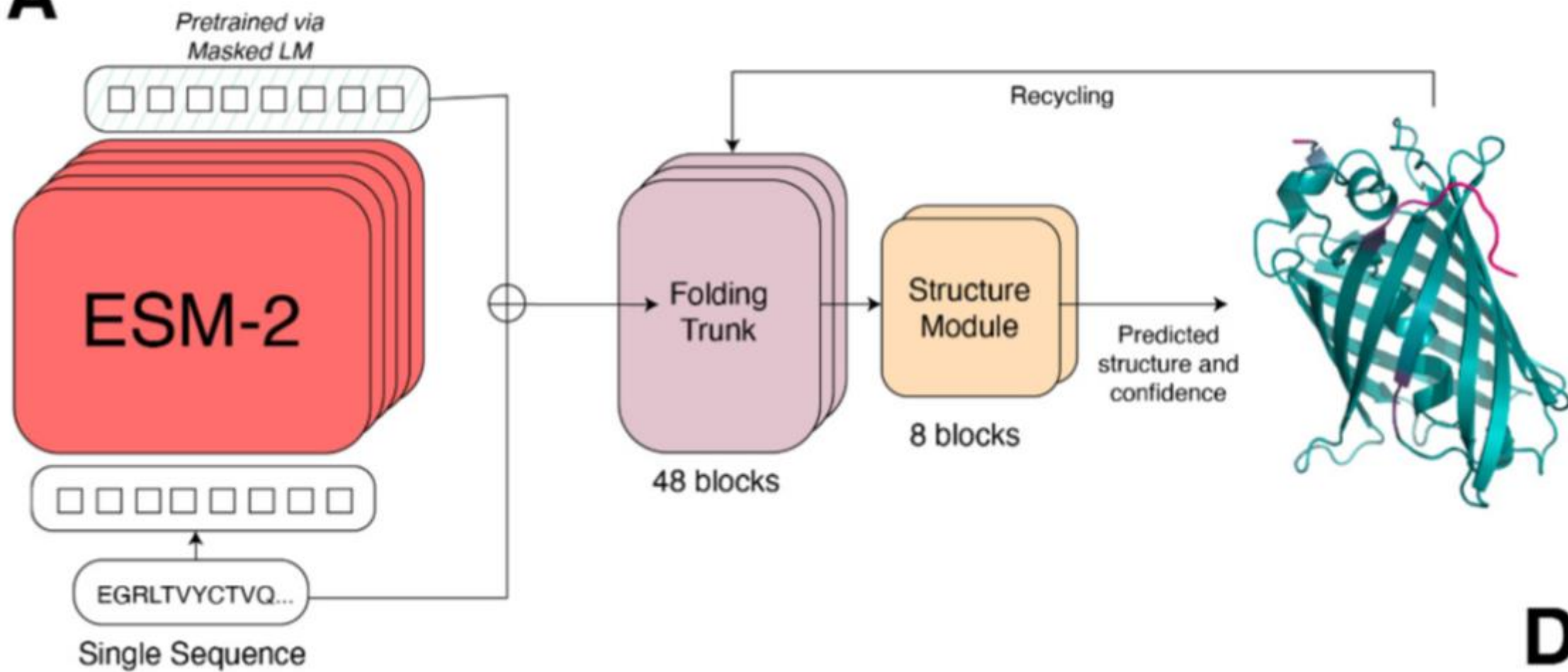
Protein Structure landscape

190k experimental determined
<https://www.rcsb.org/>



ESM Metagenomic Atlas (<https://esmatlas.com>): 617M proteins. We are able to complete this characterization **in 2 weeks on a heterogeneous cluster of 2,000 GPUs**, demonstrating scalability to far larger databases. High confidence predictions are made for over 225M structures

AlphaFold DB 200 million structures in AlphaFold DB, 35% are considered to be highly accurate. Another 45% have reasonable accuracy enough for many studies

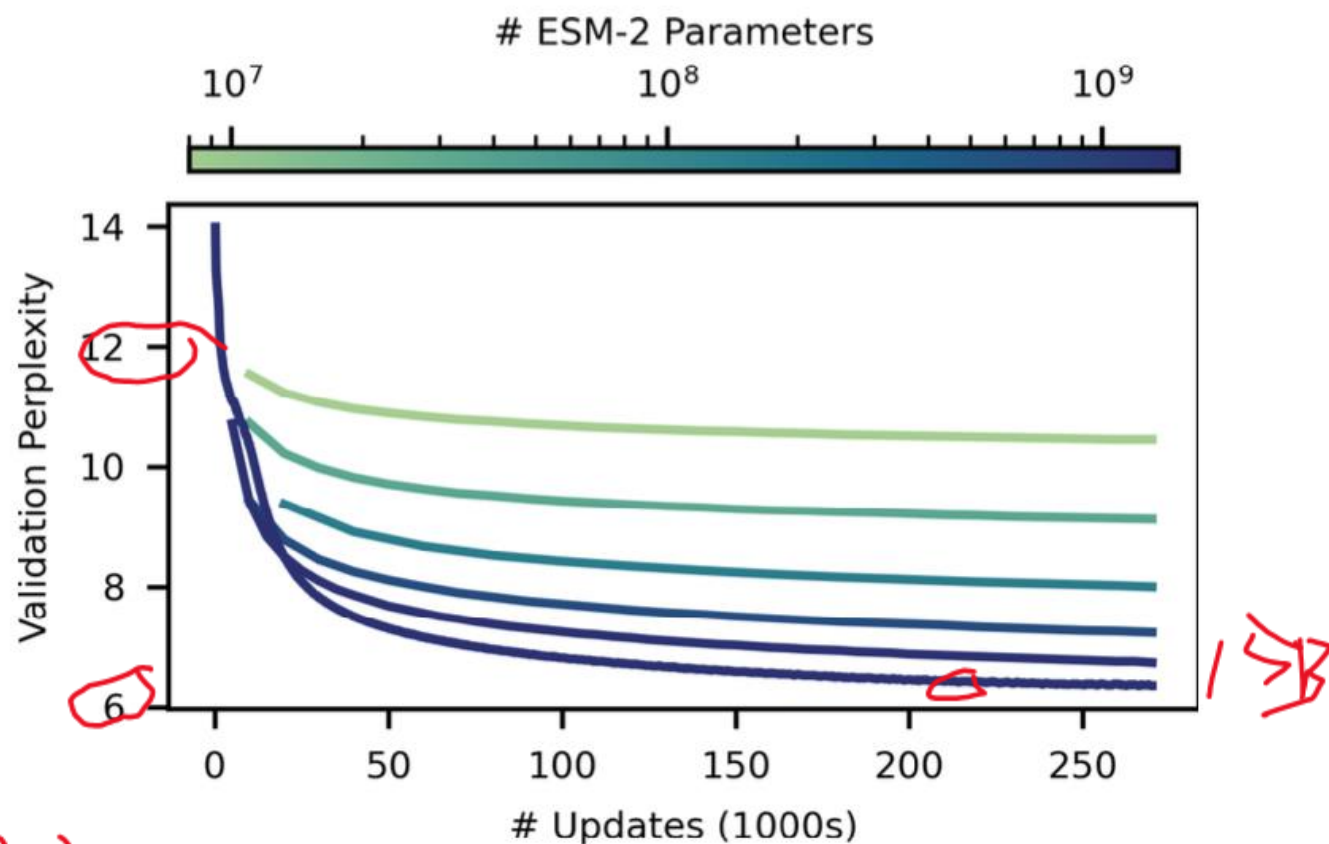
A**D**

Protein language model (largest to date)

- ESM-2, at scales from 8 million parameters up to 15 billion parameters.
- Relative to previous generation model ESM-1b, ESM-2 introduces improvements in architecture, training parameters, and increases computational resources and data
- Enabling the structure prediction from primary sequence,
 - On a single NVIDIA V100 GPU, ESMFold makes a prediction on a protein with 384 residues in 14.2 seconds, 6x faster than a single AlphaFold2 model. On shorter sequences the improvement increases up to ~60x

ESM-2

- During training sequences are sampled with even weighting across ~43 million UniRef50 training clusters from ~138 million UniRef90 sequences so that over the course of training the model sees ~65 million unique sequences.
- Training curves for ESM-2 models from 8M (highest curve, light) to 15B parameters (lowest curve, dark). Models are trained to 270K updates. Validation perplexity is measured on a 0.5% random-split holdout of UniRef50. After 270K updates the 8M parameter model has a perplexity of 10.45, and the 15B model reaches a perplexity of 6.37.

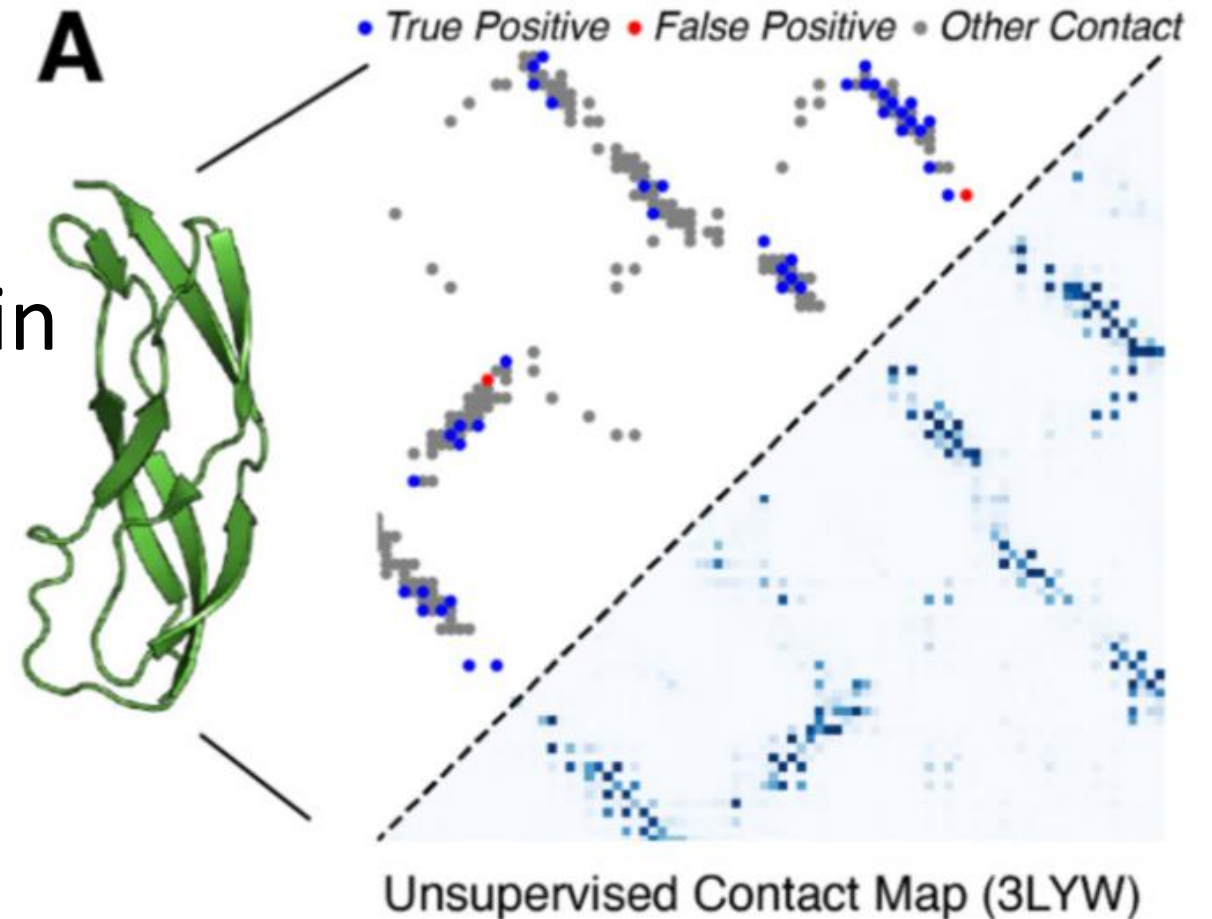


Here perplexity, ranges from 1 for a perfect model to 20 for a model that makes predictions at random.

ESM-2

- BERT encoder only transformer
- Rotary Position Embedding (RoPE) to allow the model extrapolate beyond the context window it is trained on
- Absolute plus, Learned positional encodings

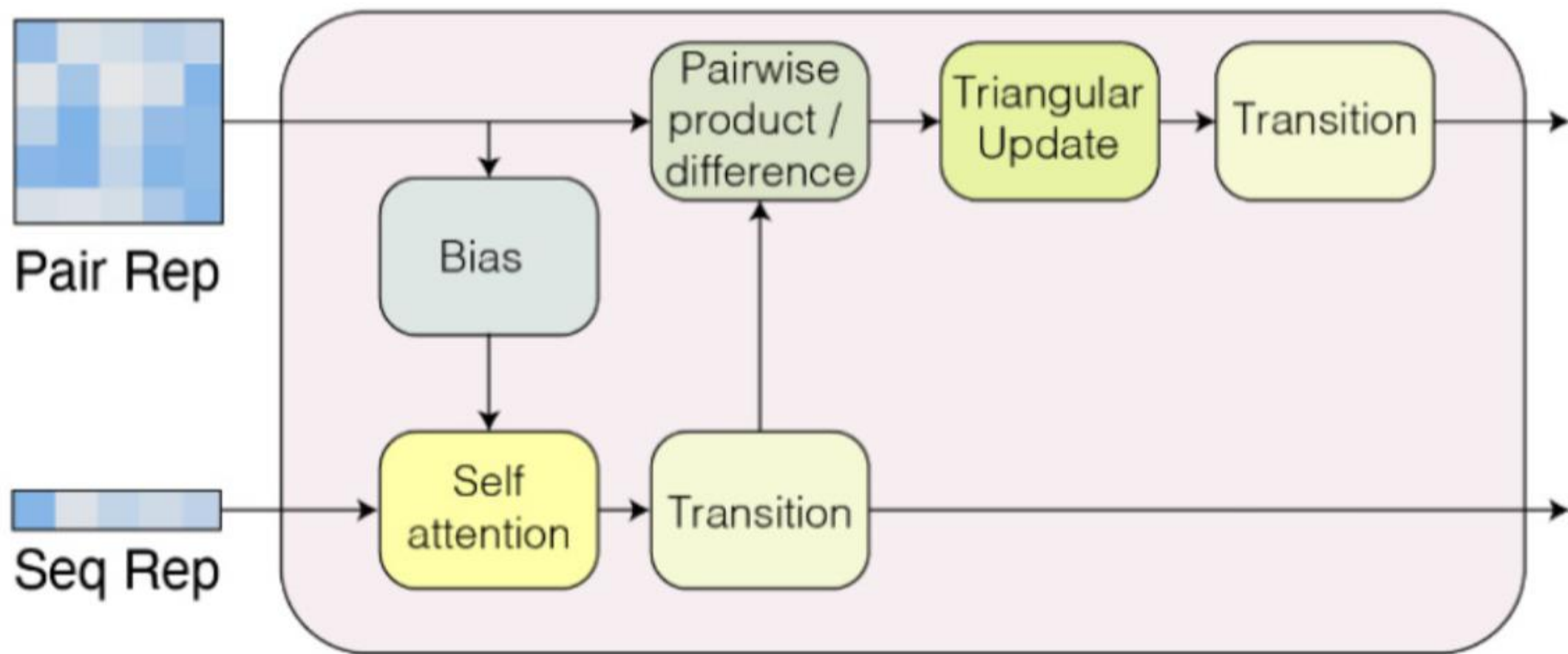
ESM-2 attention patterns correspond to the residue-residue contact map of a protein



Predicted contact probabilities (bottom right) and actual contact precision (top left) for 3LYW. A contact is a positive prediction if it is within the top-L most likely contacts for a sequence of length L.

ESMfold

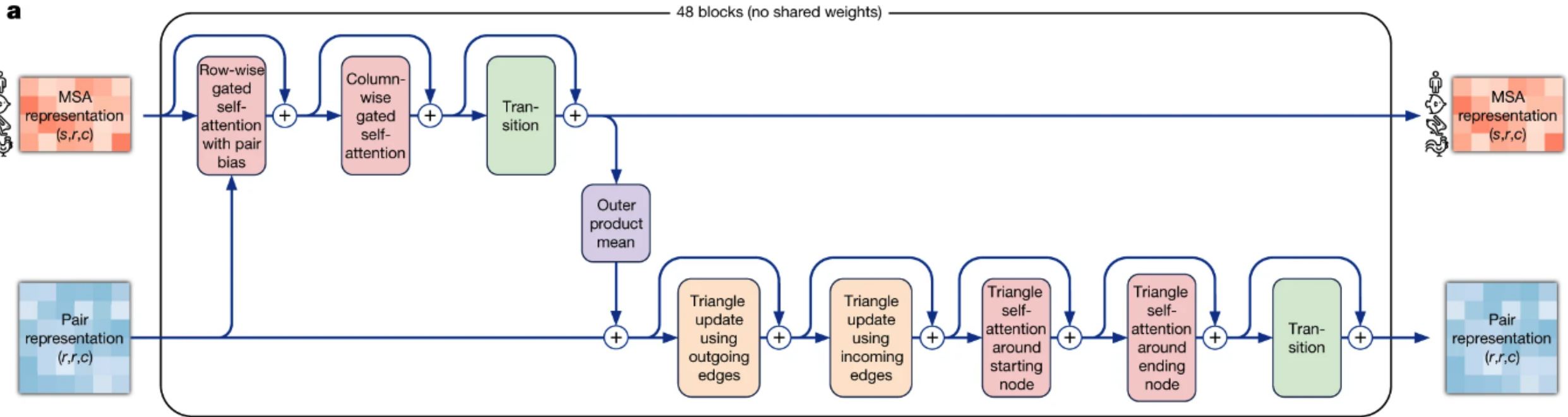
- Eliminating the need for external evolutionary databases, multiple sequence alignments, and templates.
- Each folding block alternates between updating a sequence representation and a pairwise representation.
- The output of these blocks is passed to an equivariant transformer structure module, and three steps of recycling are performed before outputting a final atomic-level structure and predicted confidences



Folding Block

AlphaFold2's Evoformer

integrate the multiple sequence alignment into the neural network architecture through an attention mechanism operating across the rows and columns of the MSA



ESMFold architecture

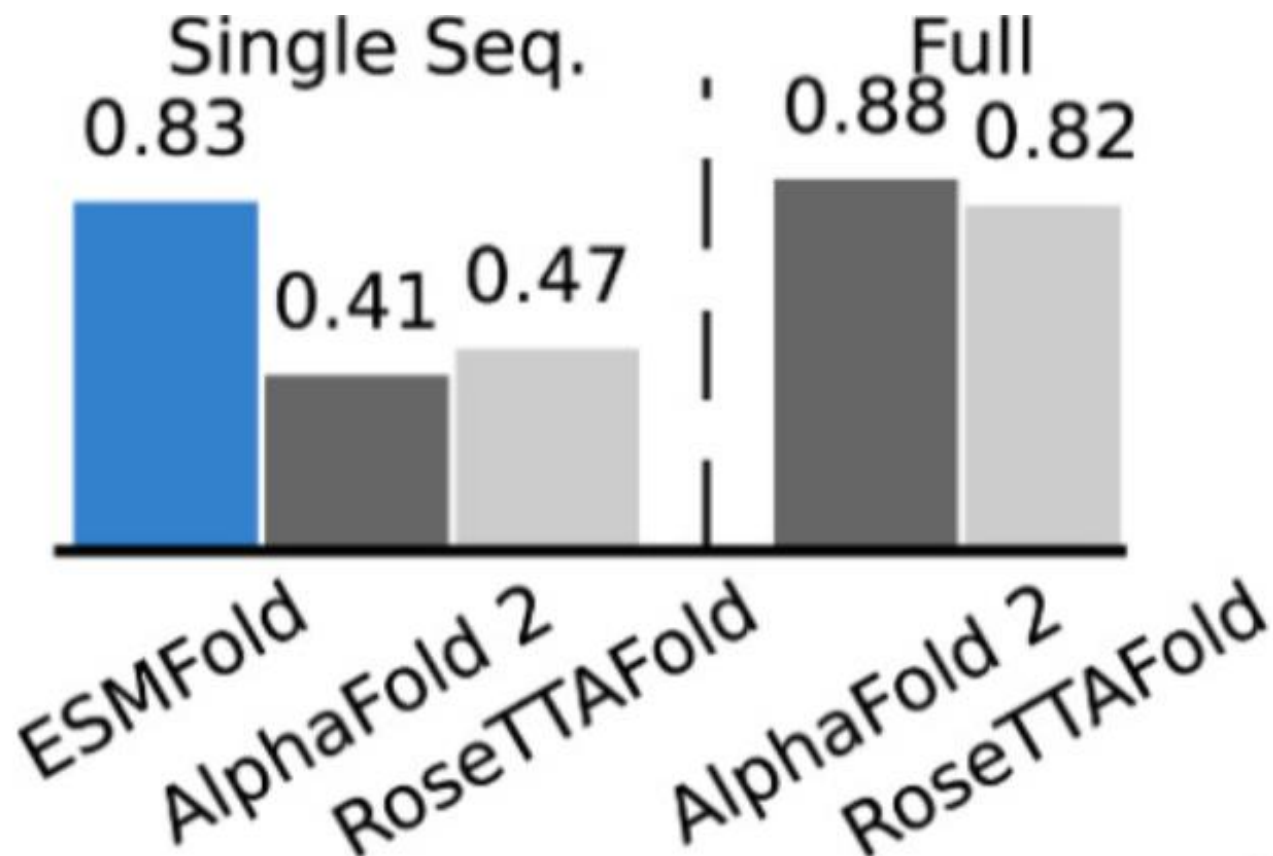
- replace the axial attention with a standard attention. All other operations are the same as in the Evoformer block. Call this simplified architecture the Folding block.
- the removal of templates. Template information is passed to the Alphafold2 model as pairwise distances, input to the residue-pairwise embedding. ESMFold simply omit this information, passing instead the attention maps from the language model,
- ESMFold uses the Frame Aligned Point Error (FAPE) and distogram losses introduced in AlphaFold2, as well as heads for predicting LDDT and the pTM score.

ESMfold more

- ESMfold train the folding head on ~25K clusters covering a total of ~325K experimentally determined structures from the PDB, further augmented with a dataset of ~12M structures we predicted with AlphaFold2

??

- ESMFold produces accurate atomic resolution predictions, with similar accuracy to RosettaFold on CAMEO.



Roadmap



What is next?

Structure Prediction
Speed does matter!

Design of entirely new proteins:

- If a designed amino acid sequence could fold into the reliable structure that we desired?

To predict the complex structure of multiple interacting partners

- Proteins work in teams .. what is the interacting team's structure, affinity, function? Team with drug? Ligand? RNA? ...

To illustrate the effect of mutations that contribute to rare genetic diseases.

- AlphaFold2 is not specifically designed and is unable to predict how amino acid mutations alter a protein's natural structure

Backup related

Highly accurate protein structure prediction with AlphaFold

•[John Jumper](#), et al

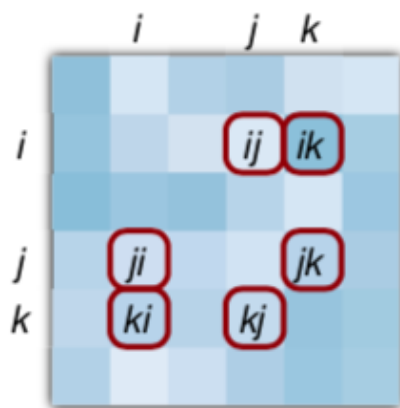
•[Demis Hassabis](#)

Show authors [Nature](#) volume 596, pages 583–589 (2021) [Cite this article](#)

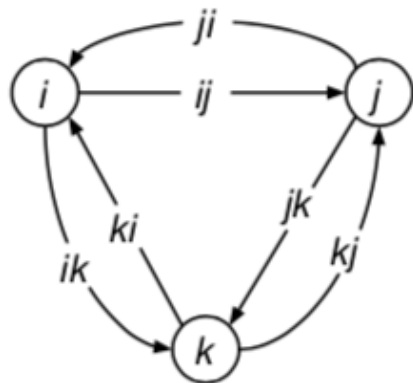
Abstract

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort^{1,2,3,4}, the structures of around 100,000 unique proteins have been determined⁵, but this represents a small fraction of the billions of known protein sequences^{6,7}. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the ‘protein folding problem’⁸—has been an important open research problem for more than 50 years⁹. Despite recent progress^{10,11,12,13,14}, existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)¹⁵, demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm

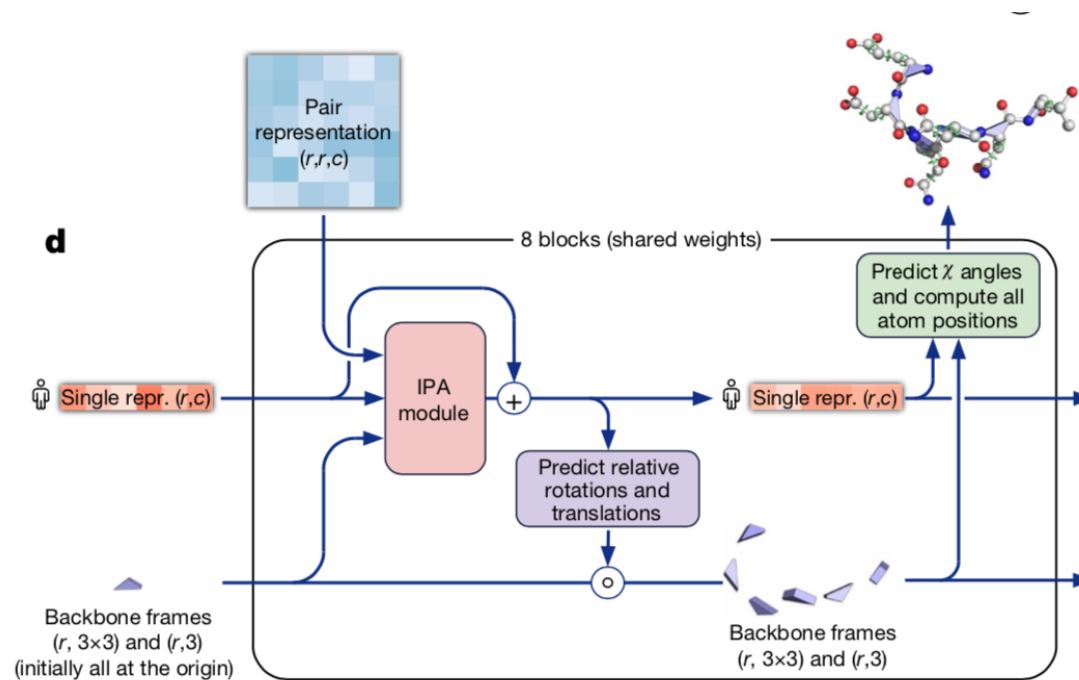
b Pair representation (r,r,c)



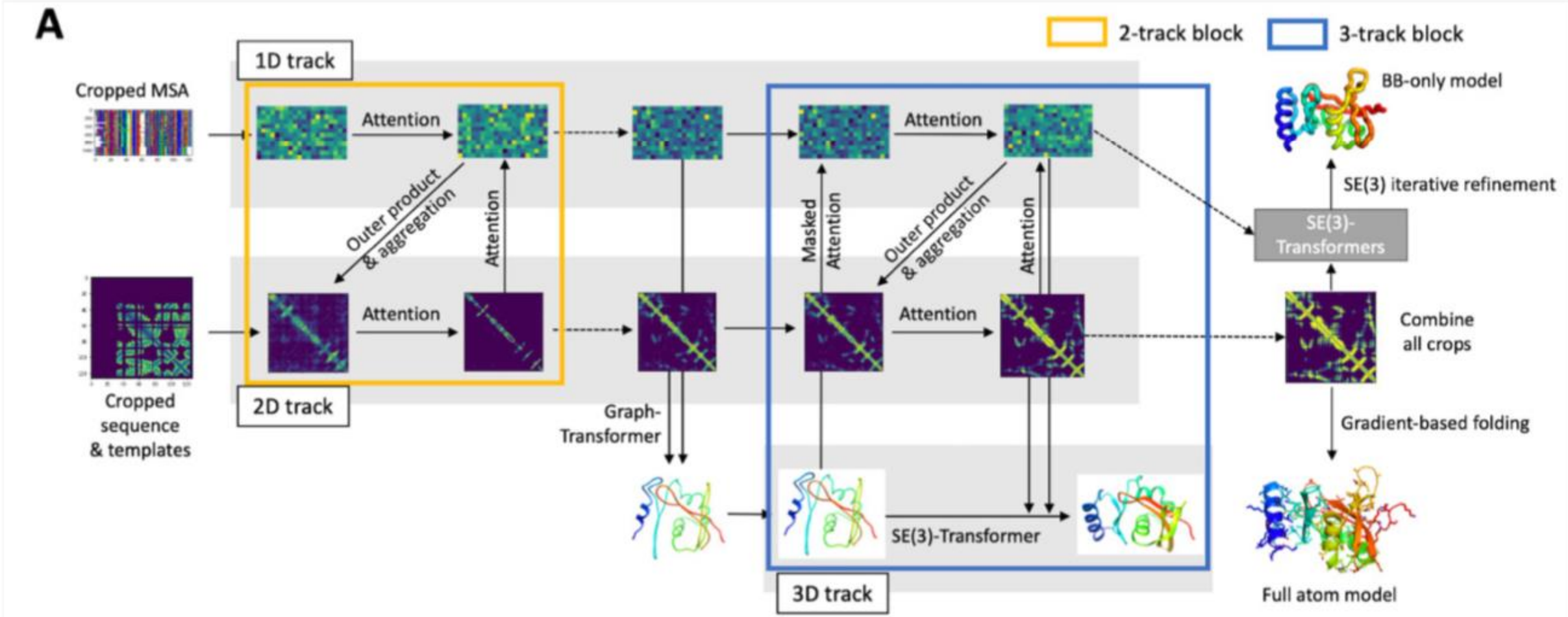
Corresponding edges in a graph



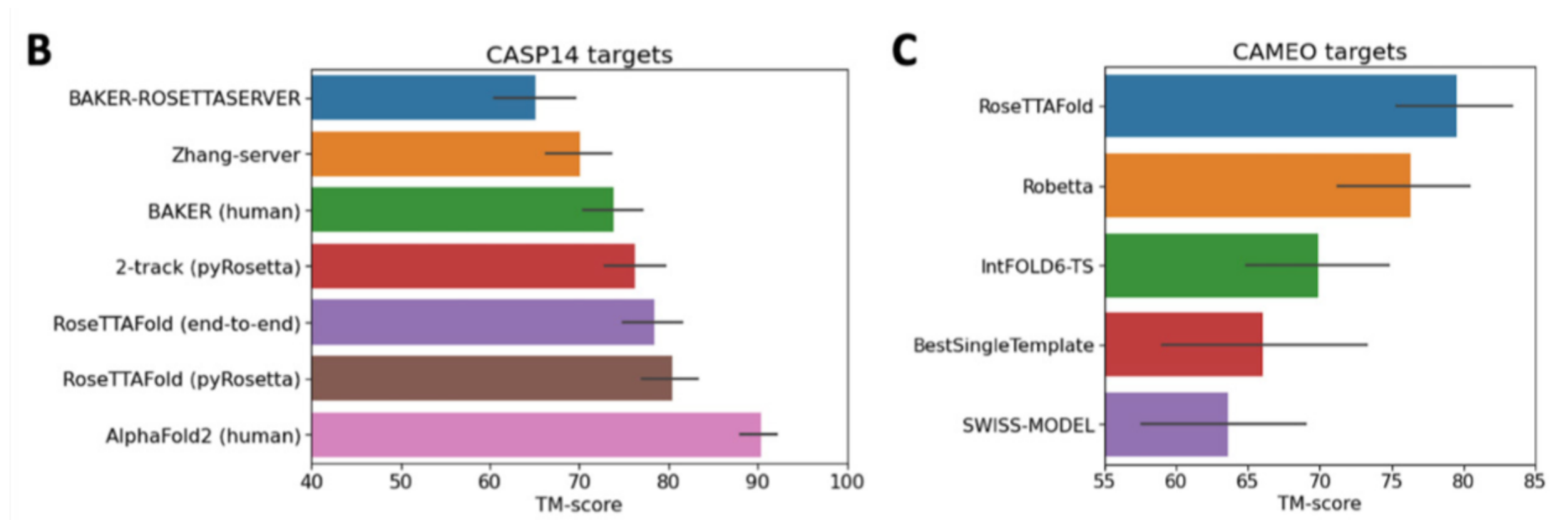
Structure module including Invariant point attention (IPA) module.



RoseTTAFold



RoseTTAFold



Protein 3D structure computed from evolutionary sequence variation

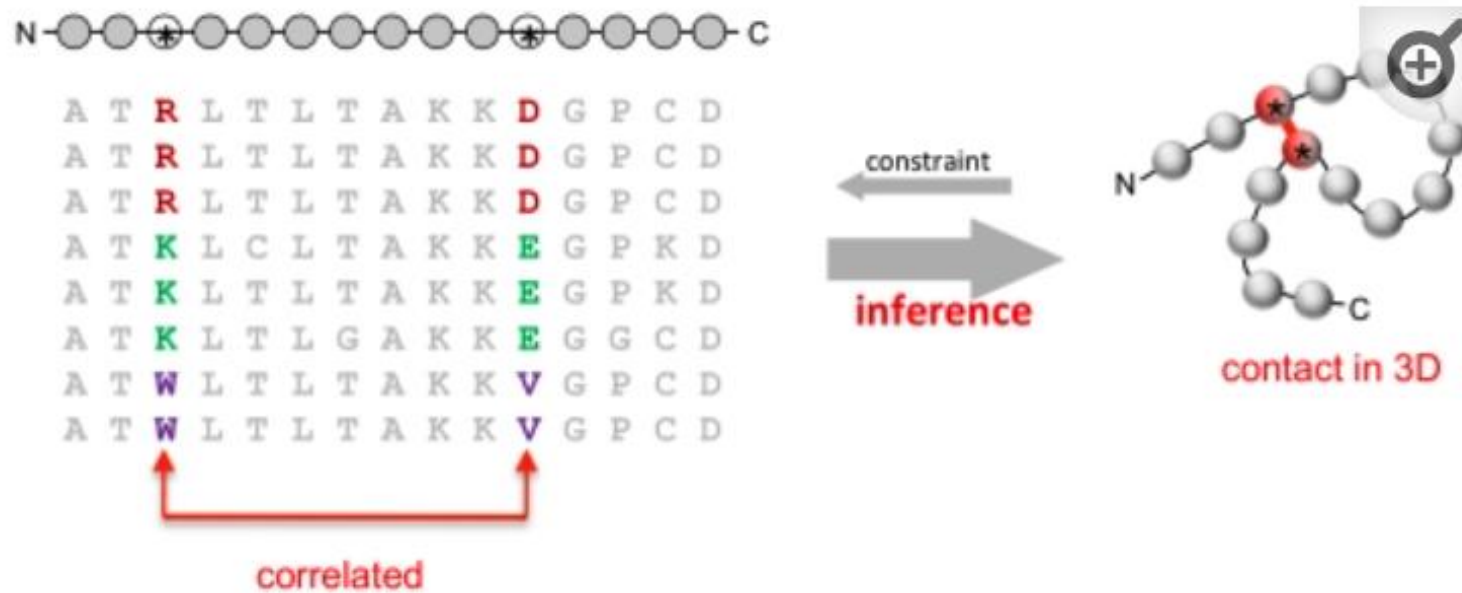
[Debora S Marks](#)¹, [Lucy J Colwell](#), [Robert Sheridan](#), [Thomas A Hopf](#), [Andrea Pagnani](#), [Riccardo Zecchina](#), [Chris Sander](#)

Affiliations expand

Abstract

The evolutionary trajectory of a protein through sequence space is constrained by its function. Collections of sequence homologs record the outcomes of millions of evolutionary experiments in which the protein evolves according to these constraints. Deciphering the evolutionary record held in these sequences and exploiting it for predictive and engineering purposes presents a formidable challenge. The potential benefit of solving this challenge is amplified by the advent of inexpensive high-throughput genomic sequencing. In this paper we ask whether we can infer evolutionary constraints from a set of sequence homologs of a protein. The challenge is to distinguish true co-evolution couplings from the noisy set of observed correlations. We address this challenge using a maximum entropy model of the protein sequence, constrained by the statistics of the multiple sequence alignment, to infer residue pair couplings. Surprisingly, we find that the strength of these inferred couplings is an excellent predictor of residue-residue proximity in folded structures. Indeed, the top-scoring residue couplings are sufficiently accurate and well-distributed to define the 3D protein fold with remarkable accuracy. We quantify this observation by computing, from sequence alone, all-atom 3D structures of fifteen test proteins from different fold classes, ranging in size from 50 to 260 residues, including a G-protein coupled receptor. These blinded inferences are de novo, i.e., they do not use homology modeling or sequence-similar fragments from known structures. The co-evolution signals provide sufficient information to determine accurate 3D protein structure to 2.7-4.8 Å C(α)-RMSD error relative to the observed structure, over at least two-thirds of the protein (method called EVfold, details at <http://EVfold.org>). This discovery provides insight into essential interactions constraining protein evolution and will facilitate a comprehensive survey of the universe of protein structures, new strategies in protein and drug design, and the identification of functional genetic variants in normal and disease genomes.

Correlated mutations carry information about distance relationships in protein structure.



Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences (ESM)

[Alexander Rives](https://orcid.org/0000-0003-2208-0796) arives@cs.nyu.edu, [Joshua Meier](https://orcid.org/0000-0003-2947-6064), [Tom Sercu](https://orcid.org/0000-0003-2947-6064) <https://orcid.org/0000-0003-2947-6064>, +7, and [Rob Fergus](#) [Authors](#)

Edited by David T. Jones, University College London, London, United Kingdom, and accepted by Editorial Board Member William H. Press December 16, 2020 (received for review August 6, 2020)

April 5, 2021

118 (15) e2016239118

<https://doi.org/10.1073/pnas.2016239118>

Significance

Learning biological properties from sequence data is a logical step toward generative and predictive artificial intelligence for biology. **Here, we propose scaling a deep contextual language model with unsupervised learning to sequences spanning evolutionary diversity.** We find that without prior knowledge, information emerges in the learned representations on fundamental properties of proteins such as secondary structure, contacts, and biological activity. We show the learned representations are useful across benchmarks for remote homology detection, prediction of secondary structure, long-range residue–residue contacts, and mutational effect. Unsupervised representation learning enables state-of-the-art supervised prediction of mutational effect and secondary structure and improves state-of-the-art features for long-range contact prediction.

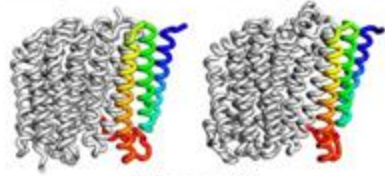
Abstract

In the field of artificial intelligence, a combination of scale in data and model capacity enabled by unsupervised learning has led to major advances in representation learning and statistical generation. In the life sciences, the anticipated growth of sequencing promises unprecedented data on natural sequence diversity. Protein language modeling at the scale of evolution is a logical step toward predictive and generative artificial intelligence for biology. **To this end, we use unsupervised learning to train a deep contextual language model on 86 billion amino acids across 250 million protein sequences spanning evolutionary diversity.** The resulting model contains information about biological properties in its representations. The representations are learned from sequence data alone. The learned representation space has a multiscale organization reflecting structure from the level of biochemical properties of amino acids to remote homology of proteins. Information about secondary and tertiary structure is encoded in the representations and can be identified by linear projections. Representation learning produces features that generalize across a range of applications, enabling state-of-the-art supervised prediction of mutational effect and secondary structure and improving state-of-the-art features for long-range contact prediction.

Also works in RoseTTAFold

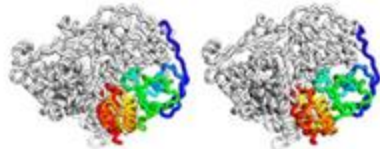
A

NADH-quinone oxidoreductase (nuoJ/nuoK)



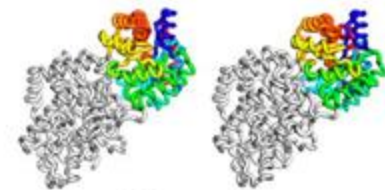
TM-score: 97

Aldehyde oxidoreductase (paoA/paoB)



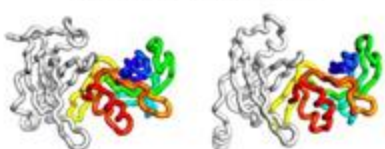
TM-score: 95

Tryptophan synthase (trpA/trpB)



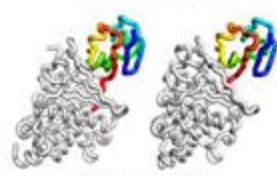
TM-score: 92

50S ribosome (rplS/rplN)



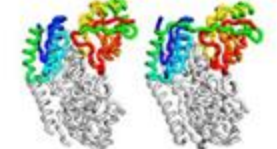
TM-score: 93

thiS/thiF complex



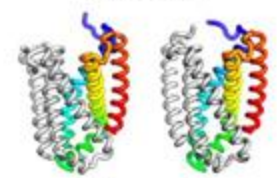
TM-score: 96

5-oxoprolinase (pxpB/pxpC)



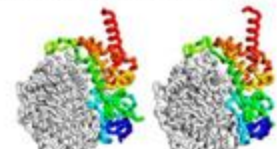
TM-score: 95

Succinate dehydrogenase (sdhD/sdhC)



TM-score: 90

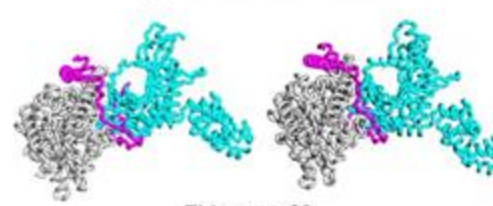
Fumarate reductase (frdA/frdC)



TM-score: 95

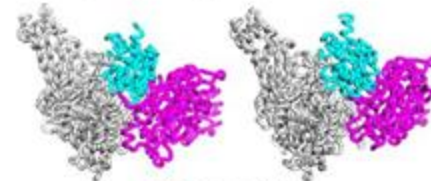
B

tRNA-dependent amidotransferase



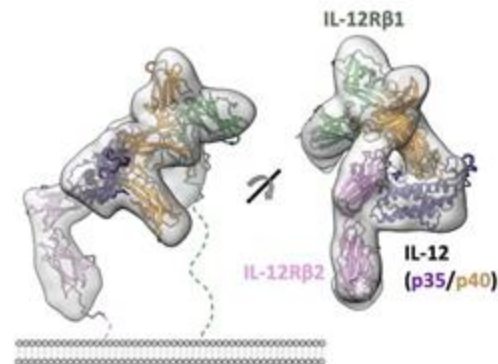
TM-score: 89

4-hydroxybenzoyl-CoA reductase



TM-score: 90

C



Accurate prediction of protein structures and interactions using a three-track neural network

<https://science.sciencemag.org/content/early/2021/07/19/science.abj8754>