Section 3.5- Life Science Foundation Models and More on Protein

2026 Spring

LLM Agents Foundation & Applications

Dr. Yanjun Qi

20260210

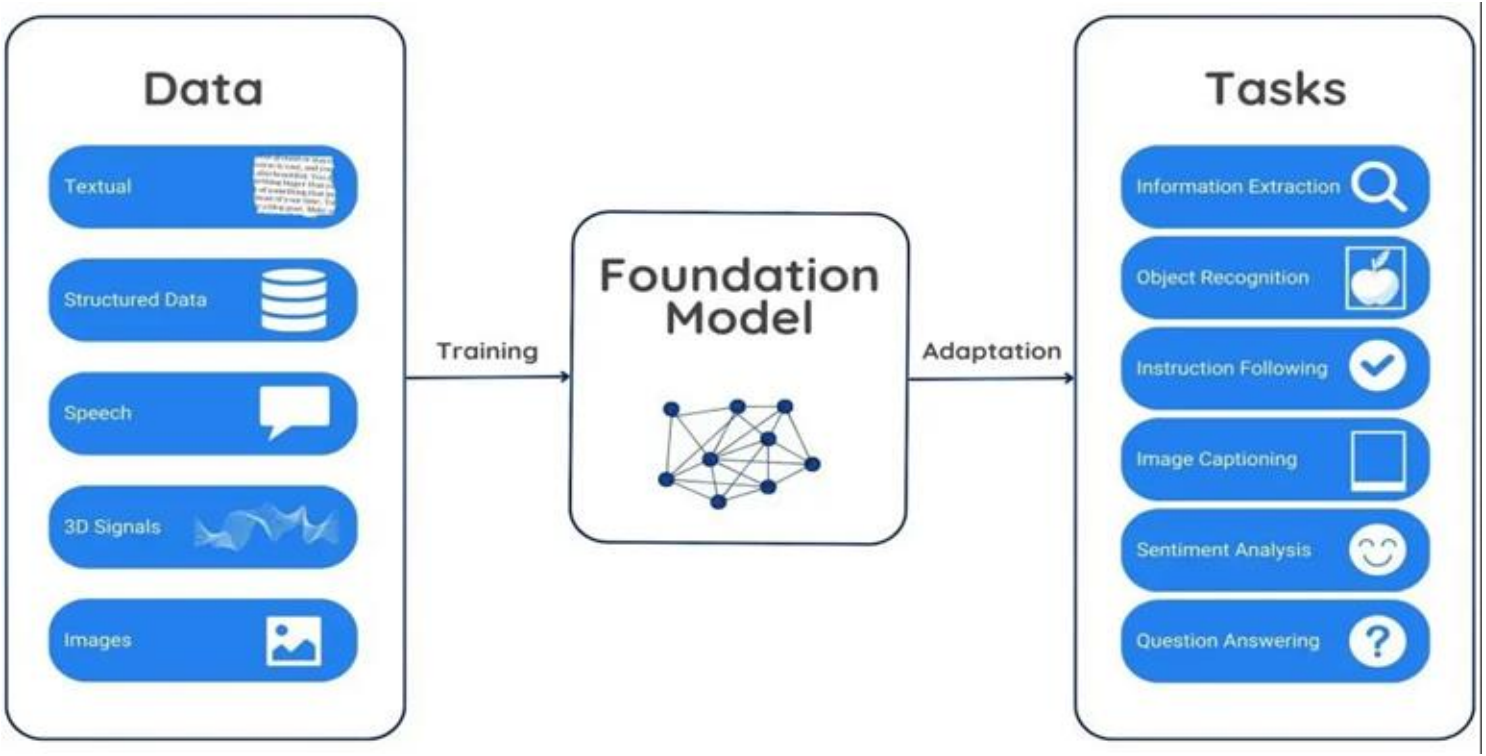
Road Map

- Overview of FMs in Medicine
- Overview of Science FMS
- Overview of Protein LLMs

Overview

Foundation model?

- Very large deep learning models
  - Massive broad datasets
- What is the purpose?
  - Acts a foundation to build upon
  - Can quickly be adapted for a task



Foundation Models: The Core Concept

Traditional ML vs. Foundation Models

Traditional ML
Train one narrow model per task. A tumor detector trained on lung CT scans is useless on pathology slides. Thousands of separate models needed.

Foundation Model
Pre-train one massive model on enormous general data. Fine-tune cheaply for many downstream tasks. The same base model powers radiology reports, drug design, and clinical Q&A.

The Three-Stage Pipeline

① Pre-training

Model learns rich representations from billions of medical texts, images, or DNA sequences at massive scale.

② Fine-tuning / Adaptation

Pre-trained model is adapted to a specific task using far less data — making it accessible.

③ Deployment and Validation

(Where regulatory requirements apply)

Adapted model enters clinical workflows — subject to FDA clearance, hallucination auditing, and bias testing.

https://doi.org/10.1038/s44360-025-00024-7

https://www.nature.com/articles/s44360-025-00024-7

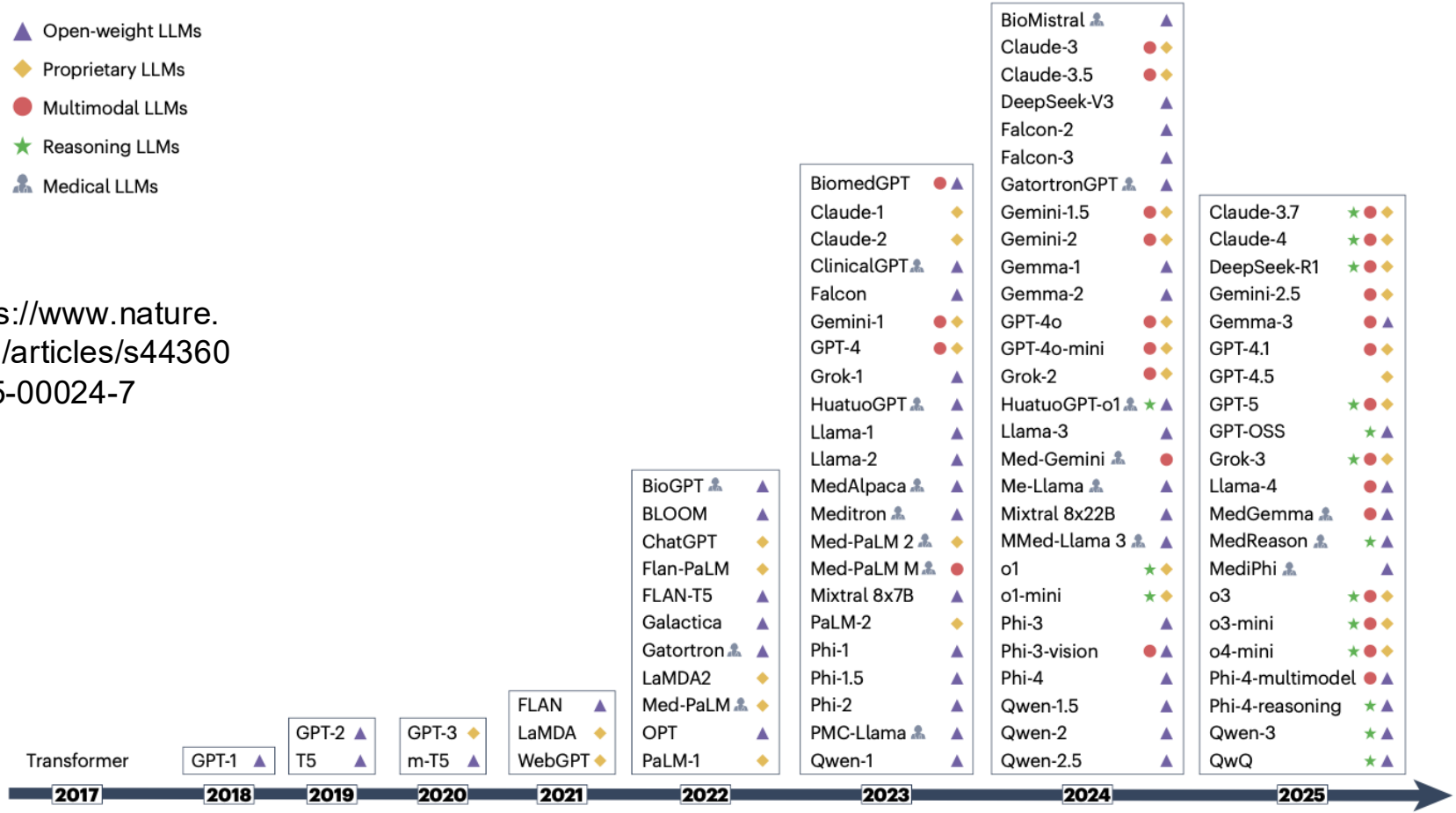
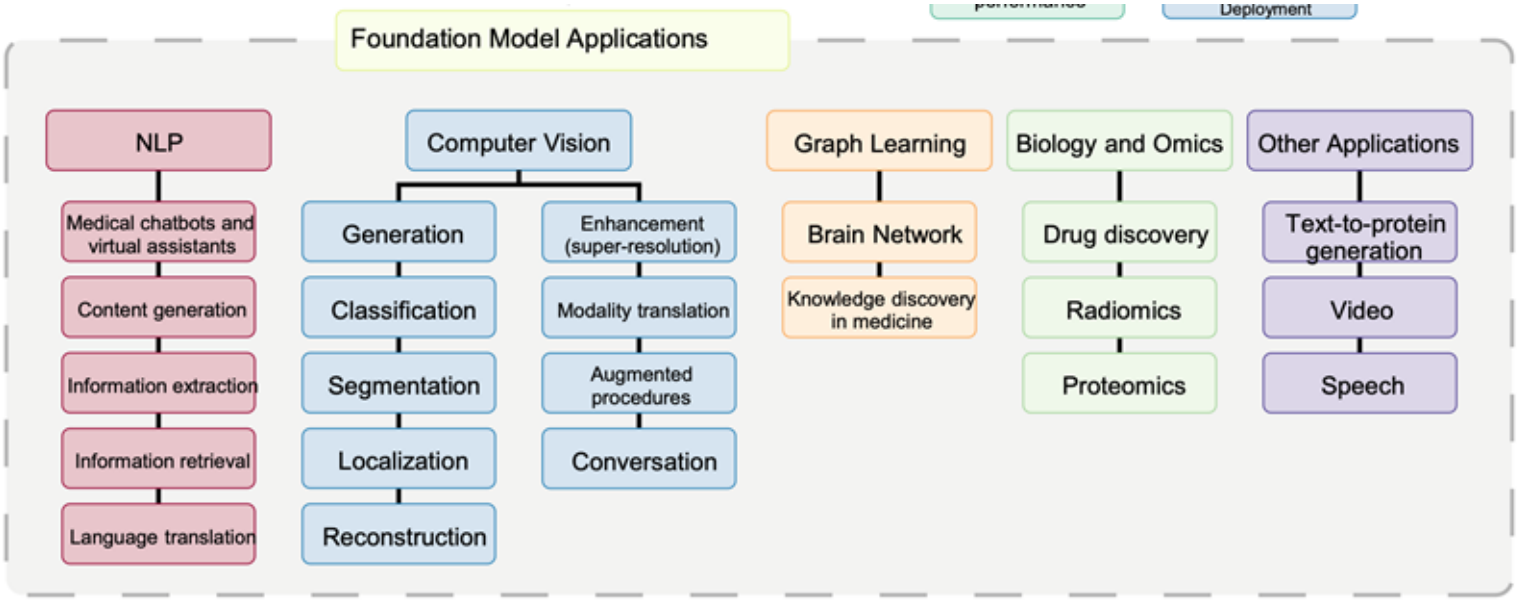


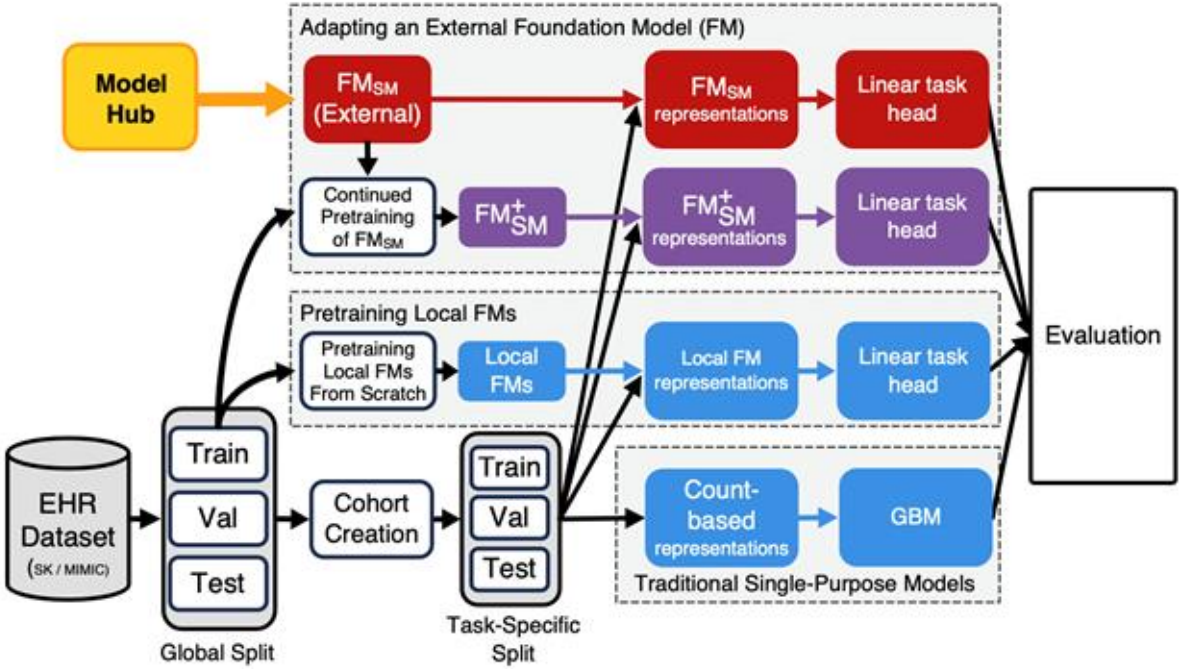
Fig. 1 | Timeline of LLMs released. The timeline begins with the release of the architecture. We categorize LLMs into the following types: open-weight LLMs,

Applications in Healthcare



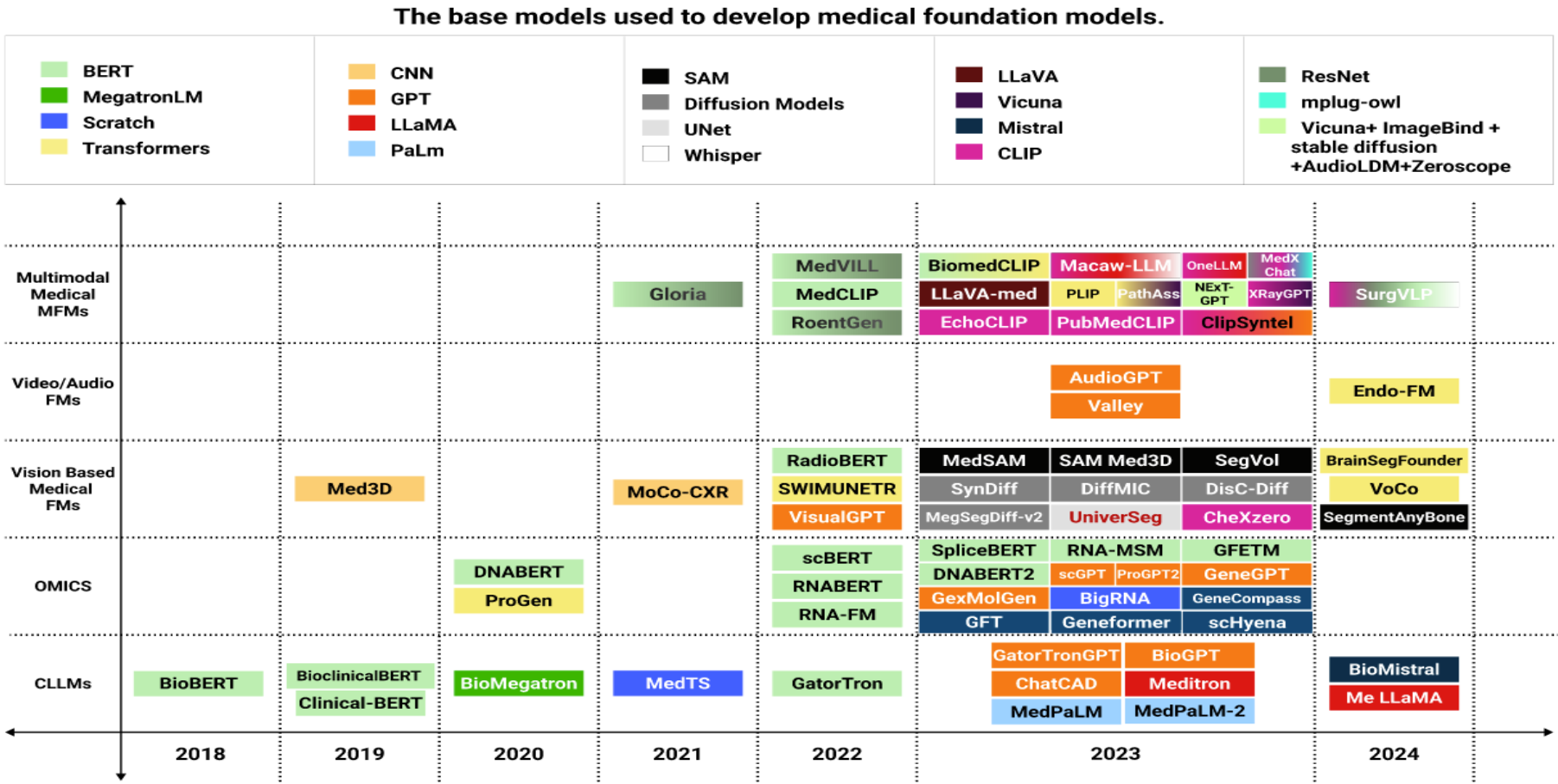
References:
[1] A Comprehensive Survey of Foundation Models in Medicine

One Example of Adapting Model Training and Evaluation



References:

[1] A Multi-center Study on the Adaptability of a Shared Foundation Model for Electronic Health Records



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

A Comprehensive Survey of Foundation Models in Medicine, 2025

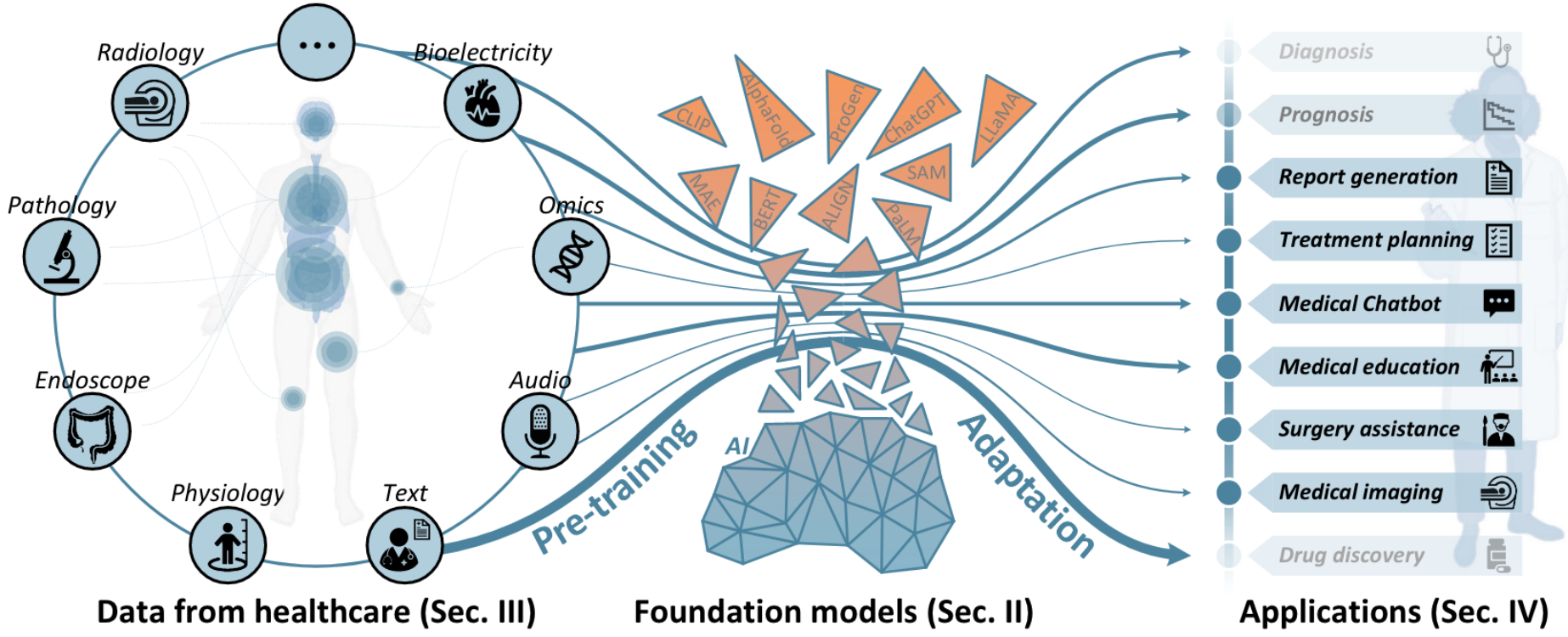


Fig. 1. The pipeline of the healthcare foundation models (HFMs) including the methods (Sec.II), datasets (Sec.III), and applications (Sec.IV).

From: Foundation Models for Advancing Healthcare:
Challenges, Opportunities, and Future Directions

Example HC data for healthcare LLM

- Healthcare training data
  - EHR
    - E.g., MIMIC III, MIMIC IV, CPRD
  - Scientific Literature
    - E.g., PubMed, PubMed Central
  - Web Data
    - E.g., COMETA (from Reddit), WebText

A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics, 2025

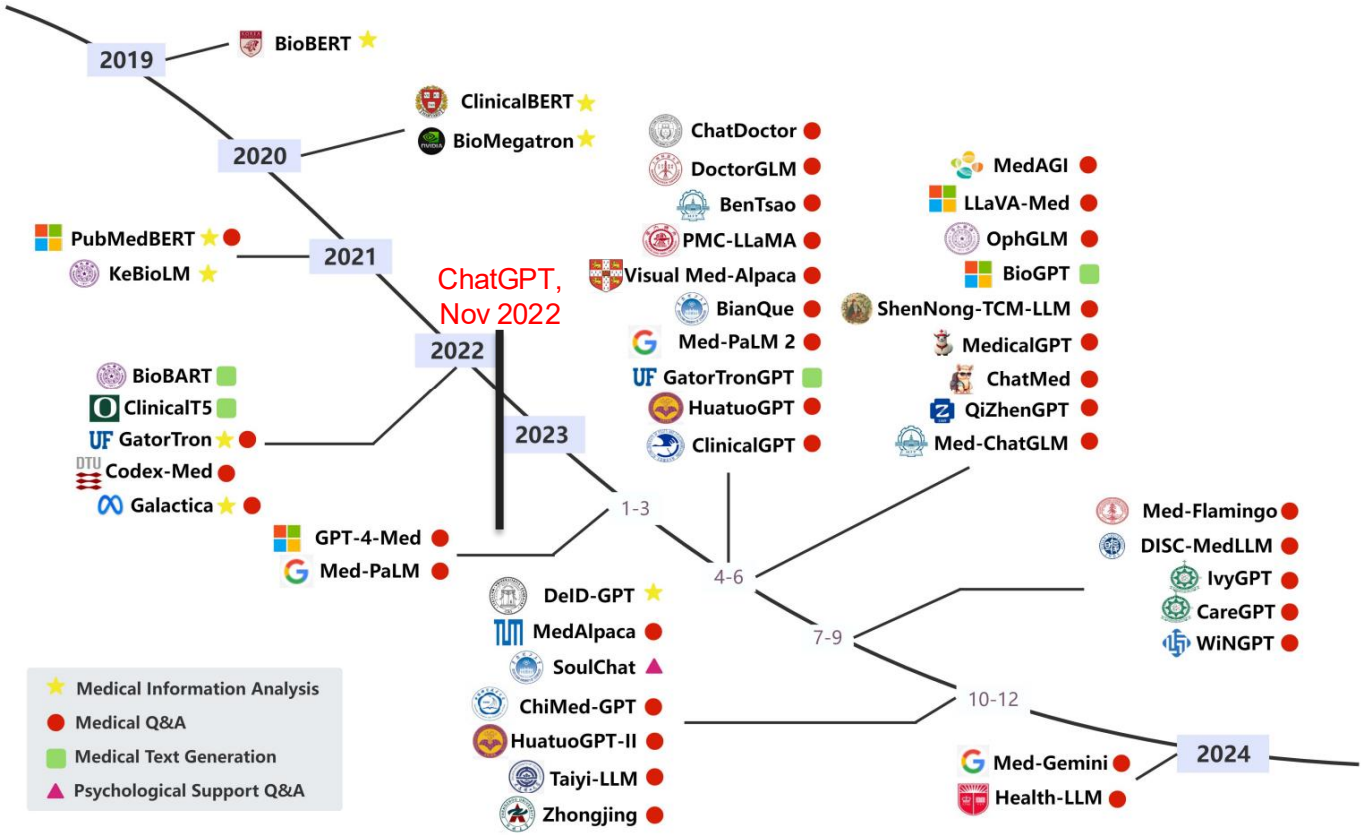
[Submitted on 10 Dec 2025]

Stanford Sleep Bench: Evaluating Polysomnography Pre-training Methods for Sleep Foundation Models

Magnus Ruud Kjaer, Rahul Thapa, Gauri Ganjoo, Hyatt Moore IV, Poul Joergen Jennum, Brandon M. Westover, James Zou, Emmanuel Mignot, Bryan He, Andreas Brink-Kjaer

Polysomnography (PSG), the gold standard test for sleep analysis, generates vast amounts of multimodal clinical data, presenting an opportunity to leverage self-supervised representation learning (SSRL) for pre-training foundation models to enhance sleep analysis. However, progress in sleep foundation models is hindered by two key limitations: (1) the lack of a shared dataset and benchmark with diverse tasks for training and evaluation, and (2) the absence of a systematic evaluation of SSRL approaches across sleep-related tasks. To address these gaps, we introduce Stanford Sleep Bench, a large-scale PSG dataset comprising 17,467 recordings totaling over 163,000 hours from a major sleep clinic, including 13 clinical disease prediction tasks alongside canonical sleep-related tasks such as sleep staging, apnea diagnosis, and age estimation. We systematically evaluate SSRL pre-training methods on Stanford Sleep Bench, assessing downstream performance across four tasks: sleep staging, apnea diagnosis, age estimation, and disease and mortality prediction. Our results show that multiple pretraining methods achieve comparable performance for sleep staging, apnea diagnosis, and age estimation. However, for mortality and

From General to Medical-specific LLMs



A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions

Science FMs

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/nnnnnnn.nnnnnnn

Language Models: A Survey on Biological & Chemical Domains

- Textual Tokens

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

- Molecular Tokens

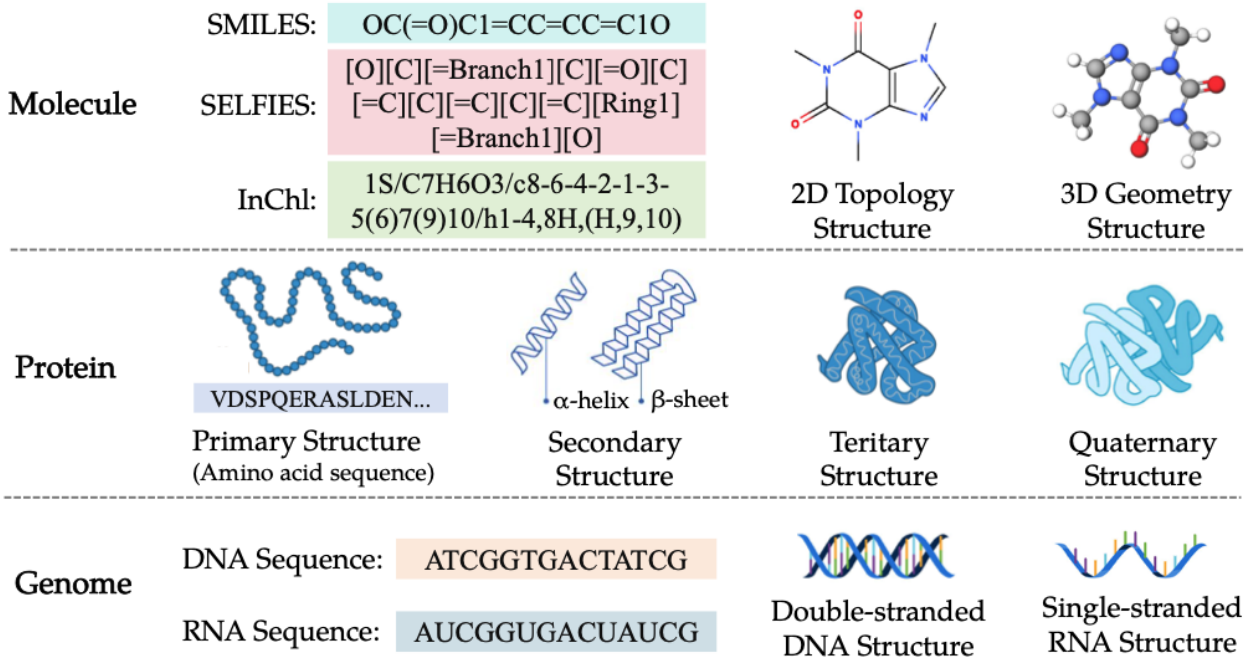
<BOS>	C	NH2	...	(=O)	<EOS>

- Protein Tokens

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

- Genomic Tokens

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



n of molecular, protein and genomic languages. Molecular languages include SMILEs, SELFIES and InChl sequences,

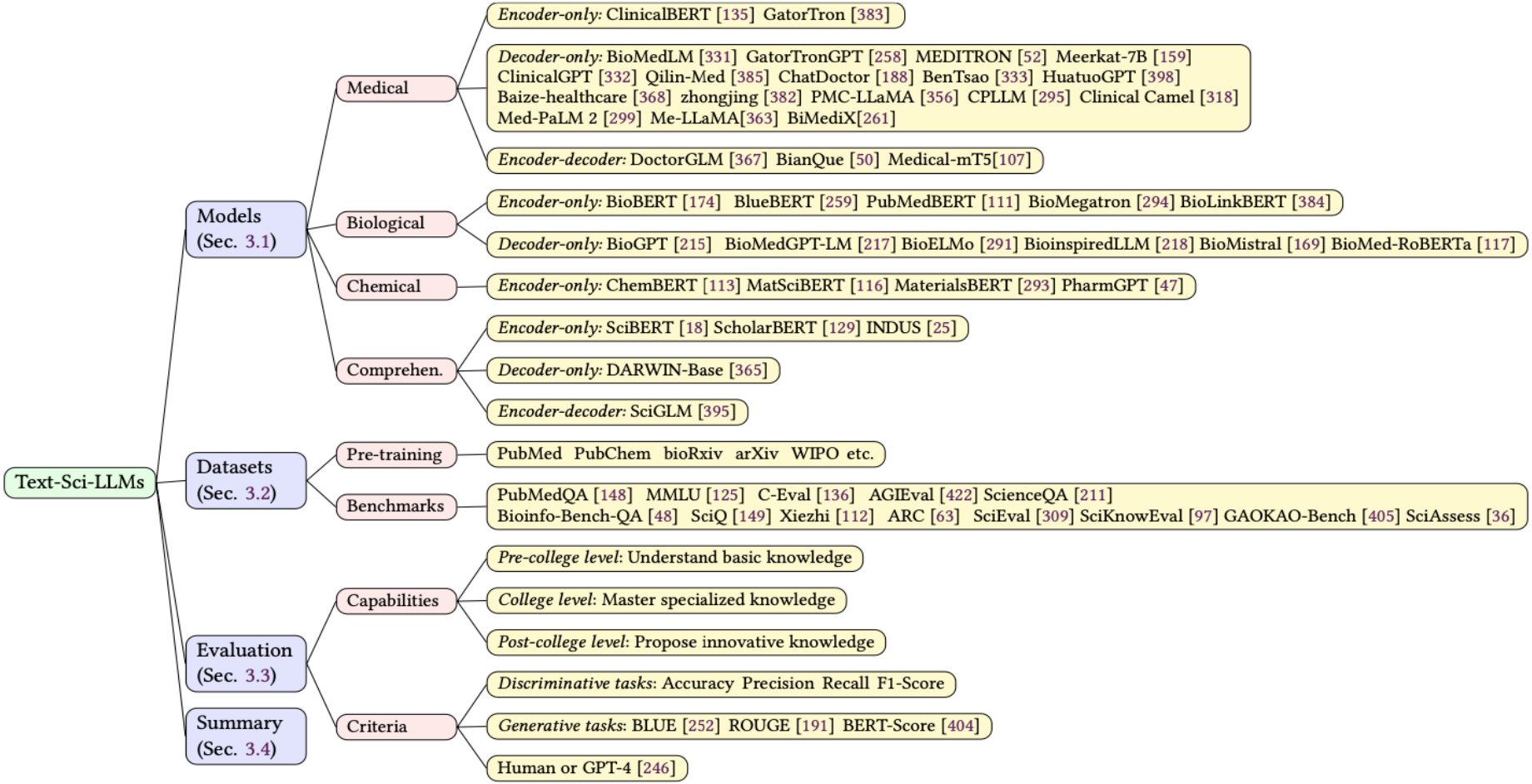


Fig. 6. Chapter overview of Text-Sci-LLMs.

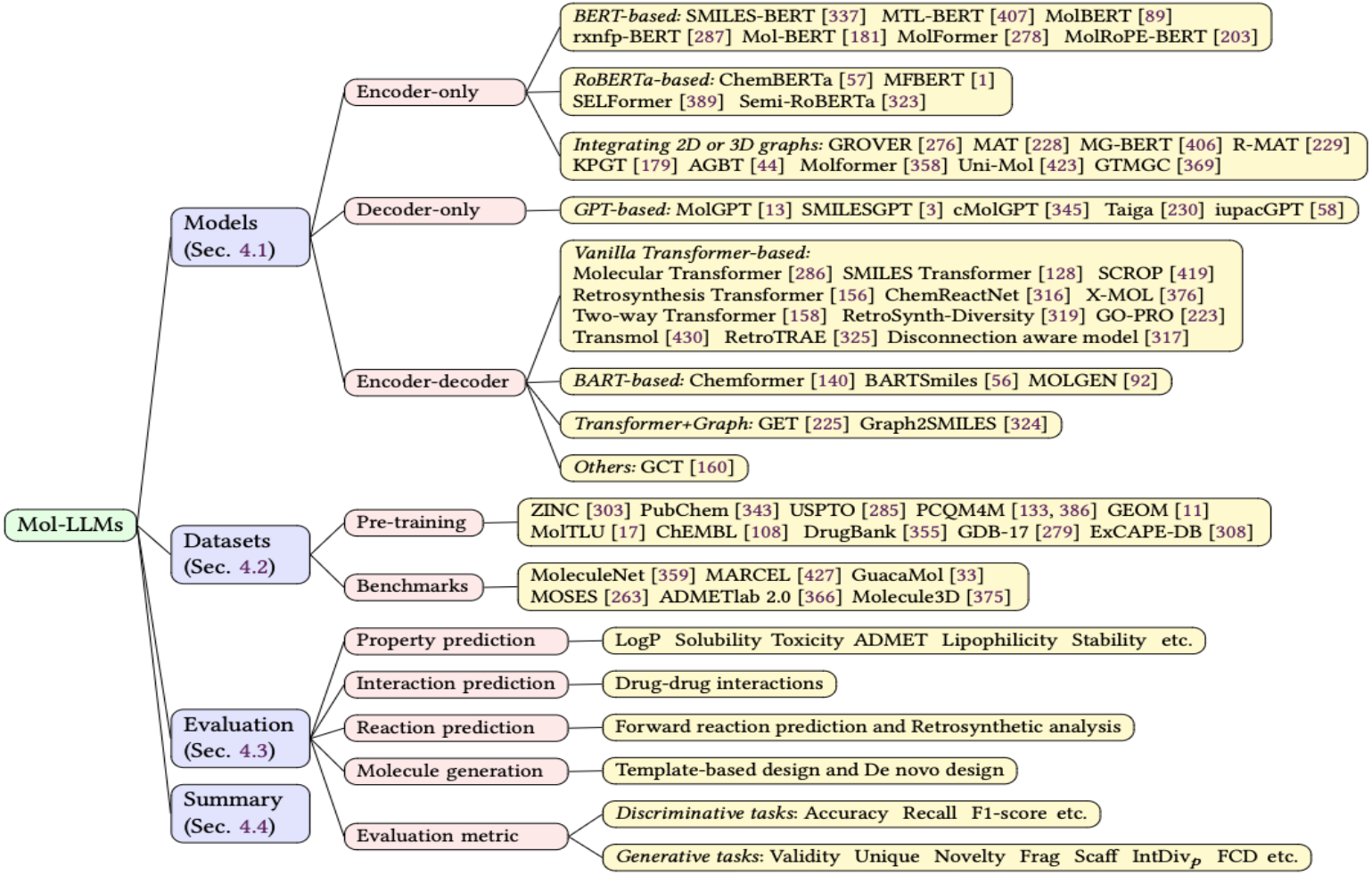


Fig. 7. Chapter overview of Mol-LLMs.

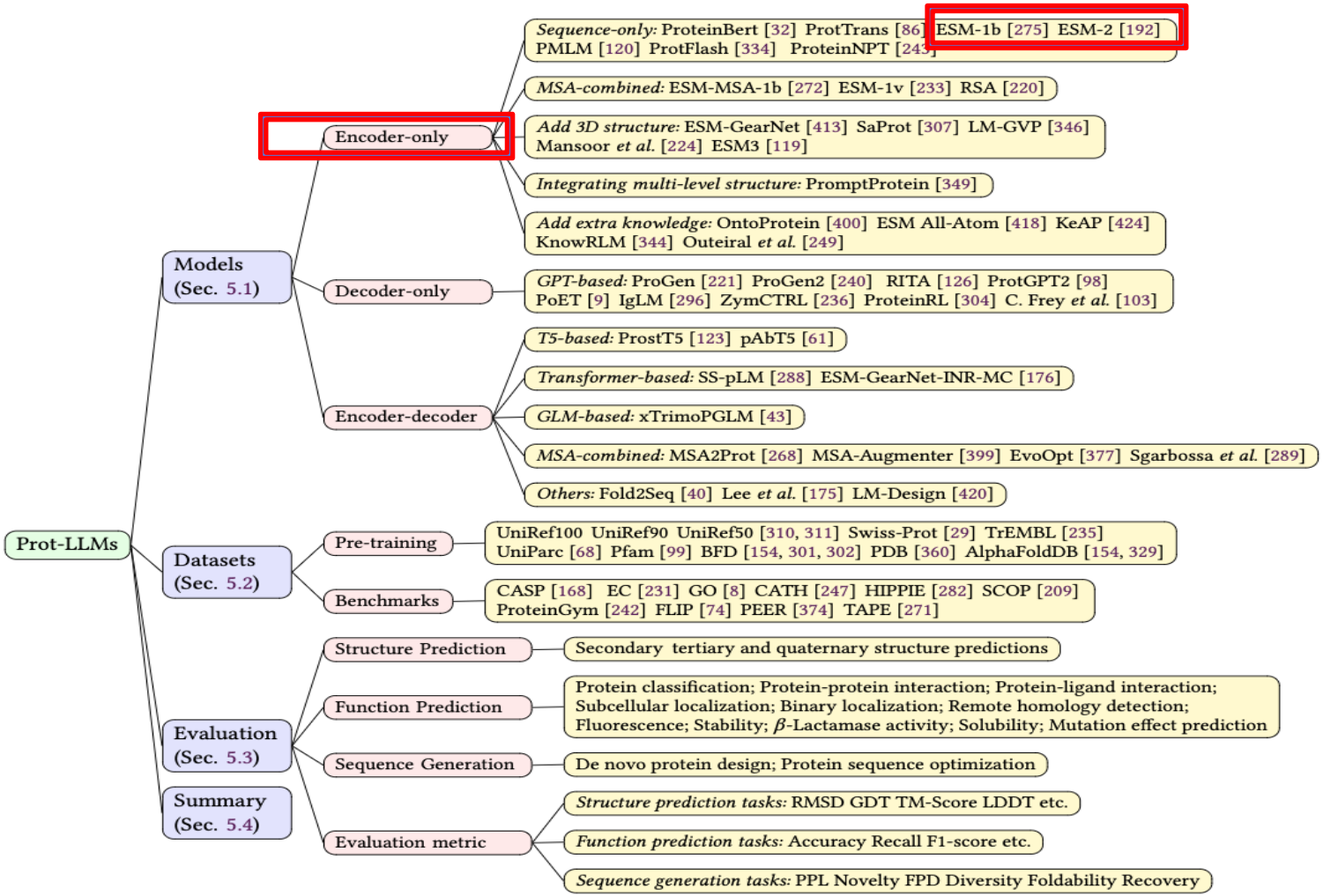


Fig. 8. Chapter overview of Prot-LLMs.

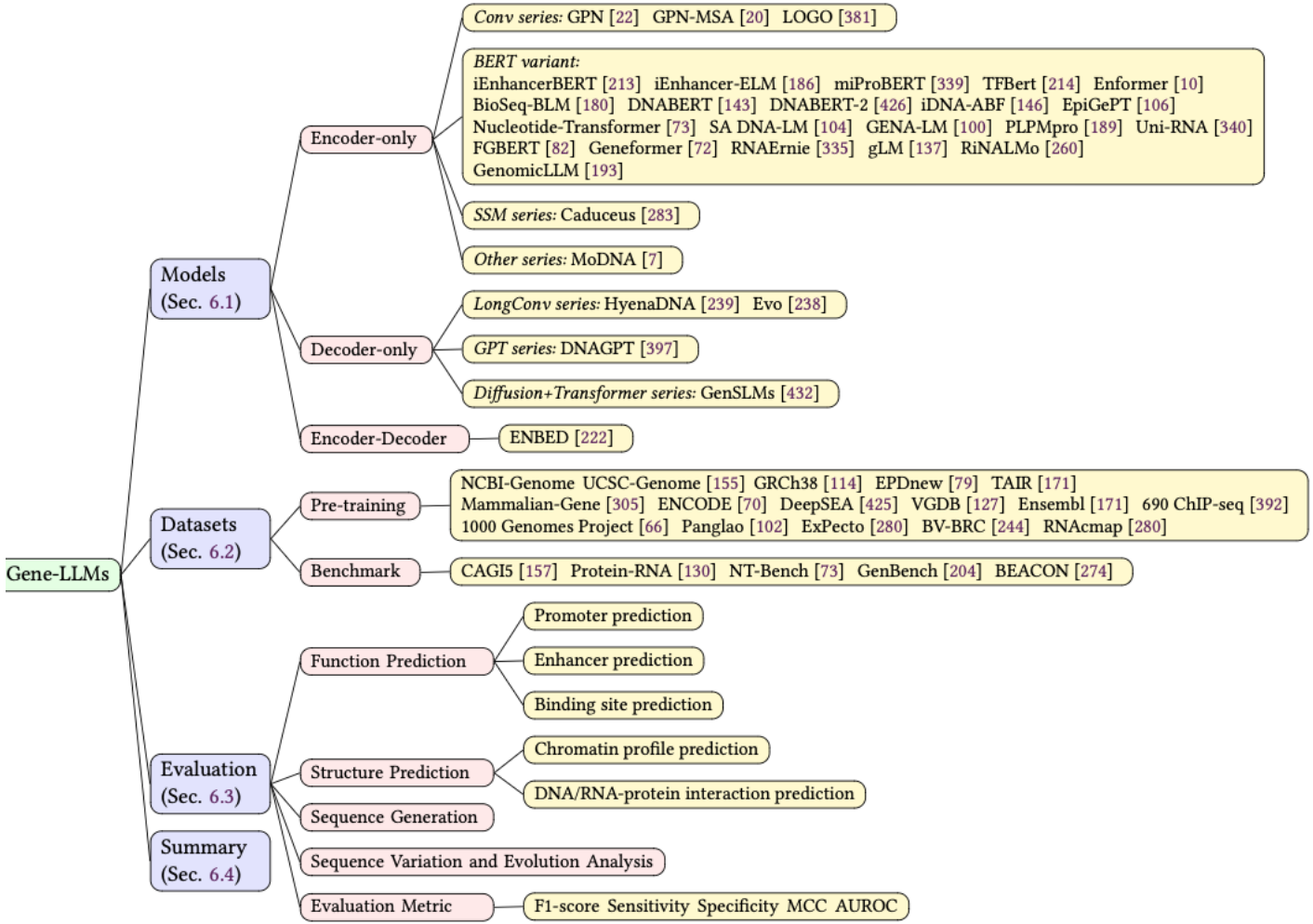


Fig. 9. Chapter overview of Gene-LLMs.

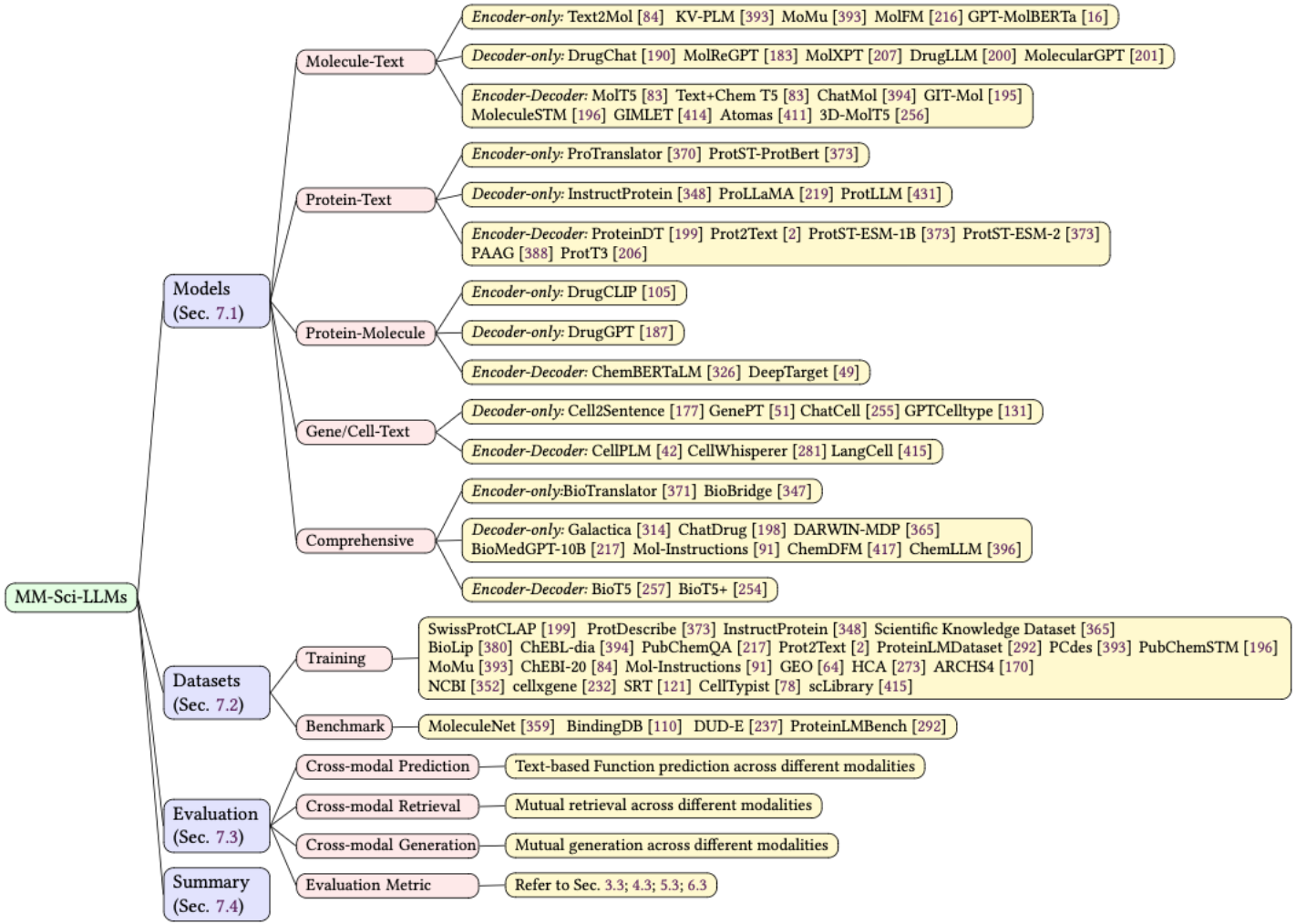
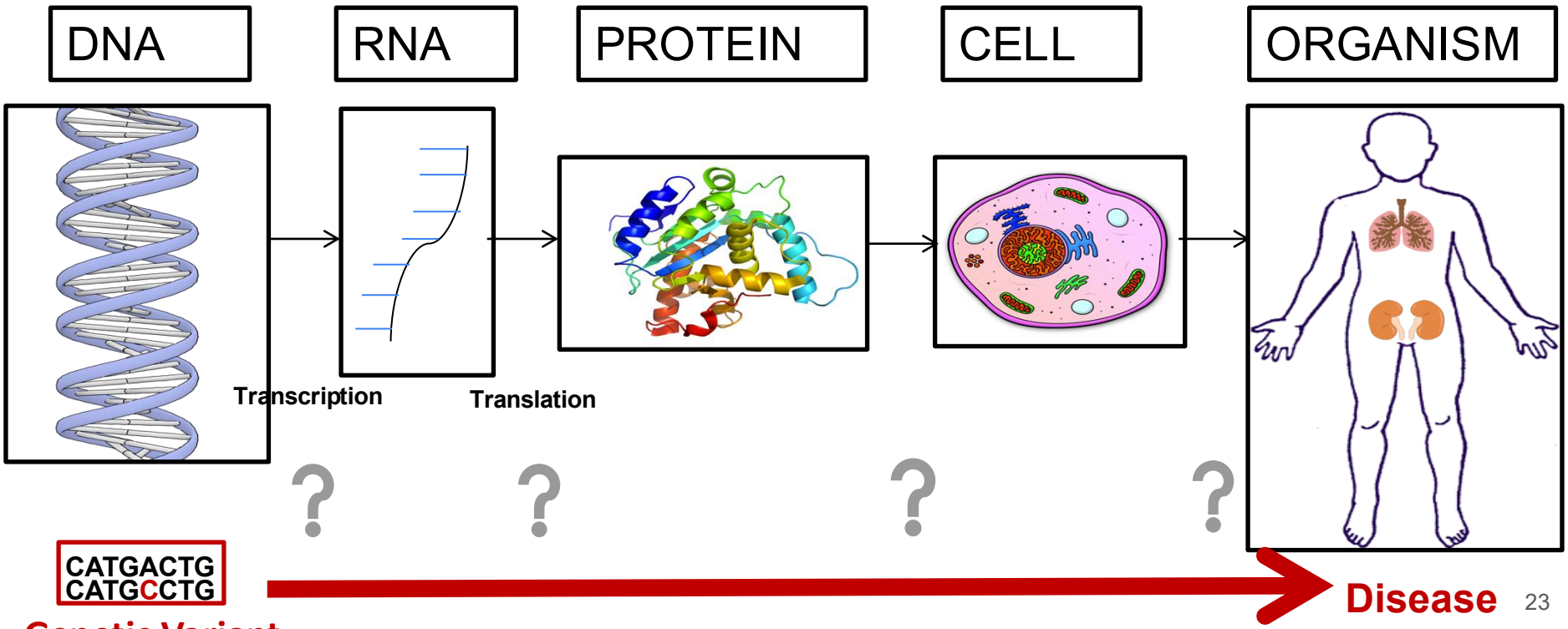


Fig. 10. Chapter overview of MM-Sci-LLMs.

Biology in a Slide:



Protein LLMs

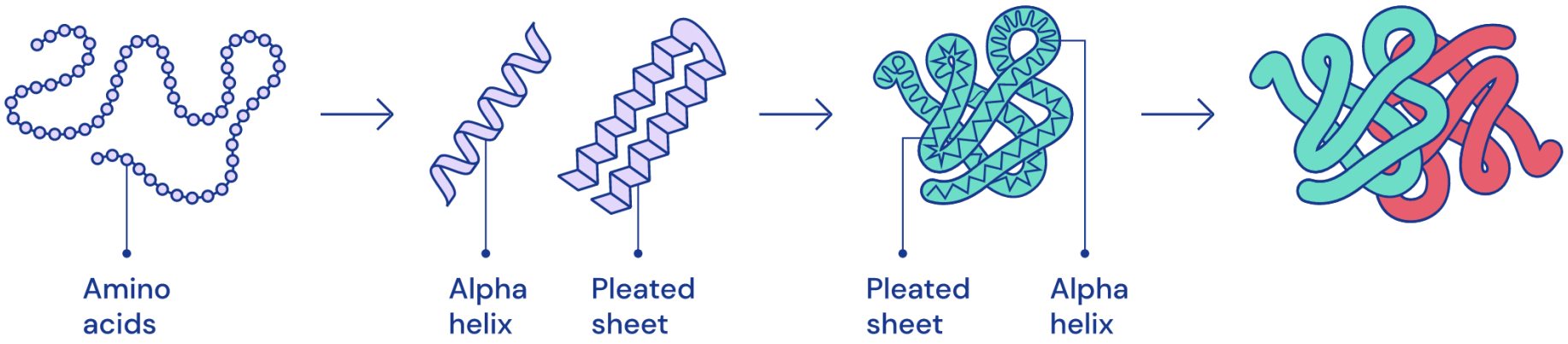
Protein Sequence form and Protein Structure

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

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

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

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



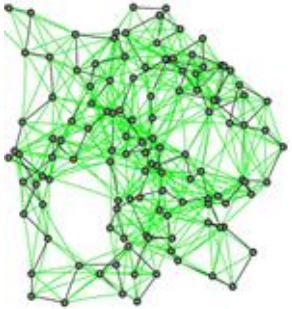
Protein sequence to structure

3D Structure

MHFTEDKATILWGKVNVEGETLGRVYPWQ

Primary Sequence

Protein sequence to structure



3D Structure

MHFTEDKATILWGKVNVEGETLGRVYPWQ

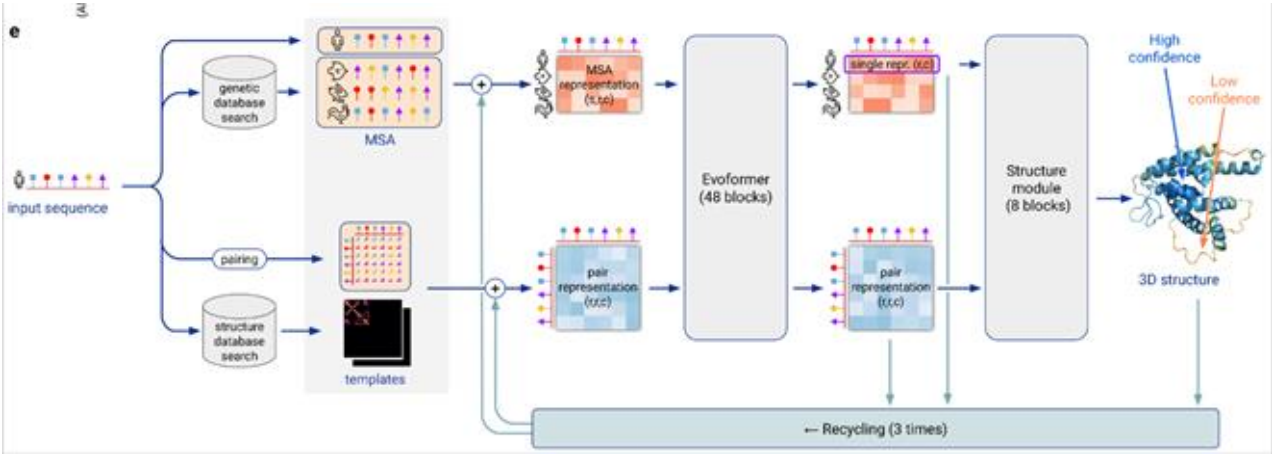
Primary Sequence

nature

https://doi.org/10.1038/s41586-021-03819-2

Accelerated Article Preview

Highly accurate protein structure prediction with AlphaFold



Received: 11 May 2021

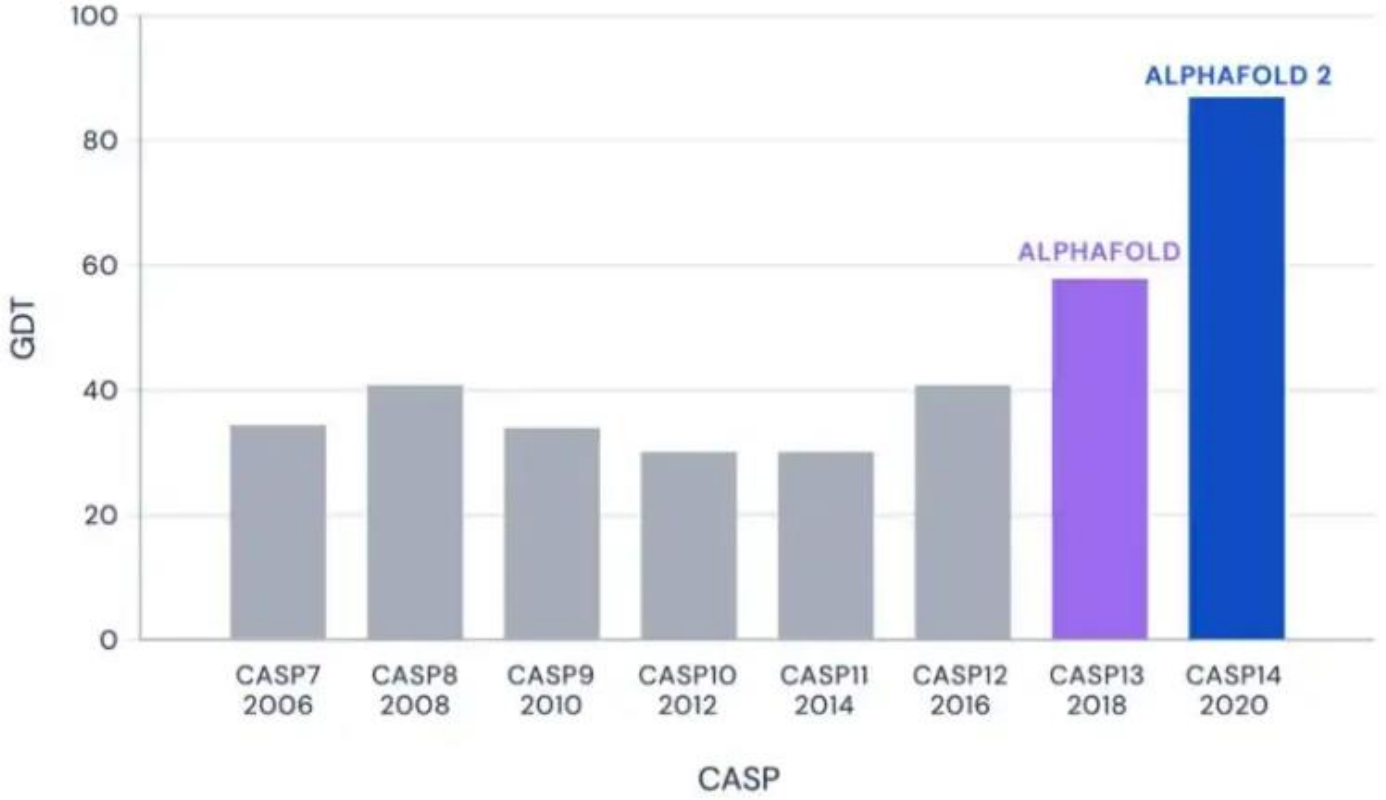
Accepted: 12 July 2021

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 Žídek, 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

Median Free-Modelling Accuracy



Artificial intelligence revolution in structure prediction for entire proteomes

✖ CLOSED SOURCE

➔ Need Open-Source Version (especially in Scientific Domains)

Artificial intelligence revolution in structure prediction for entire proteomes

(A)

AlphaFold2's structure prediction → Searchable AlphaFold DB

(B) FLDNMFGPRDSRVRG

ESMfold

ESM-2
Pretained protein language model
(15 billion parameters)

→ ESMFold's folding trunk →

Predicted structure

FLD MFGP DSR RG

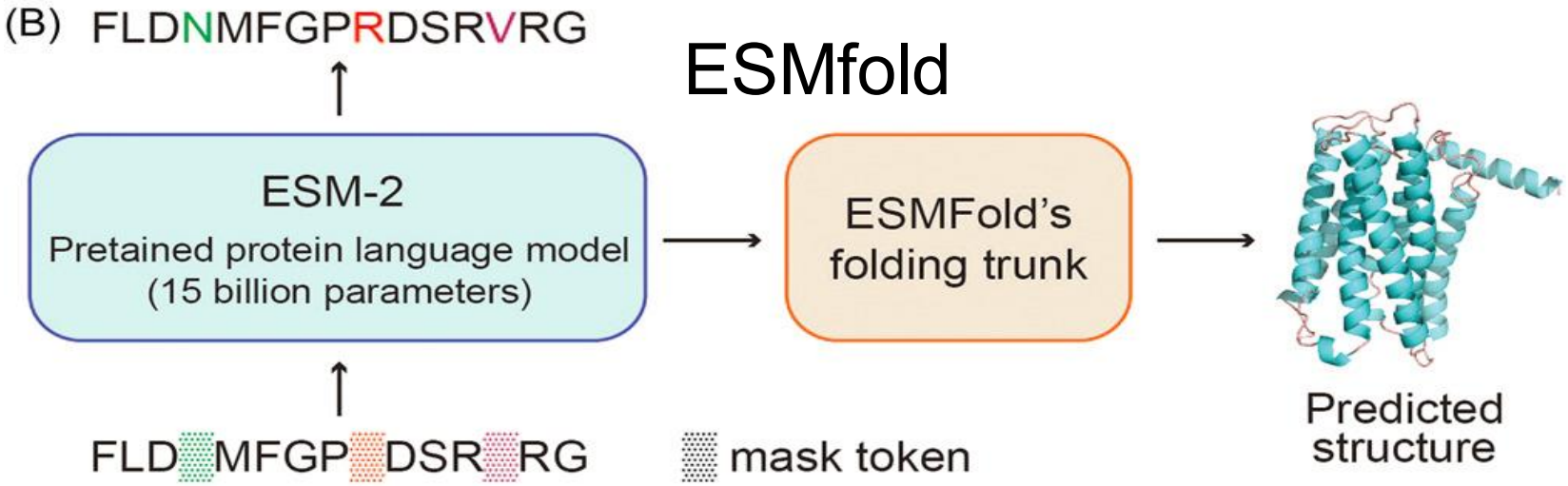
mask token

MedComm – Future Medicine, Volume: 1, Issue: 2, First published: 04 October 2022, DOI: (10.1002/mef2.19)

More:

ESM-2: Open-Source Foundation Protein LLM

Open-Source and Improved Faster Version of AlphaFold



MedComm – Future Medicine, Volume: 1, Issue: 2, First published: 04 October 2022, DOI: (10.1002/mef2.19)

Many more details in the paper!

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 *et al.* [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 *et al.* [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 *et al.* [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 *et al.* [289]	2023.02	-	MSA Transformer	-	Prot. gen.	✓

Why Predicting Protein Folds from Sequence?

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

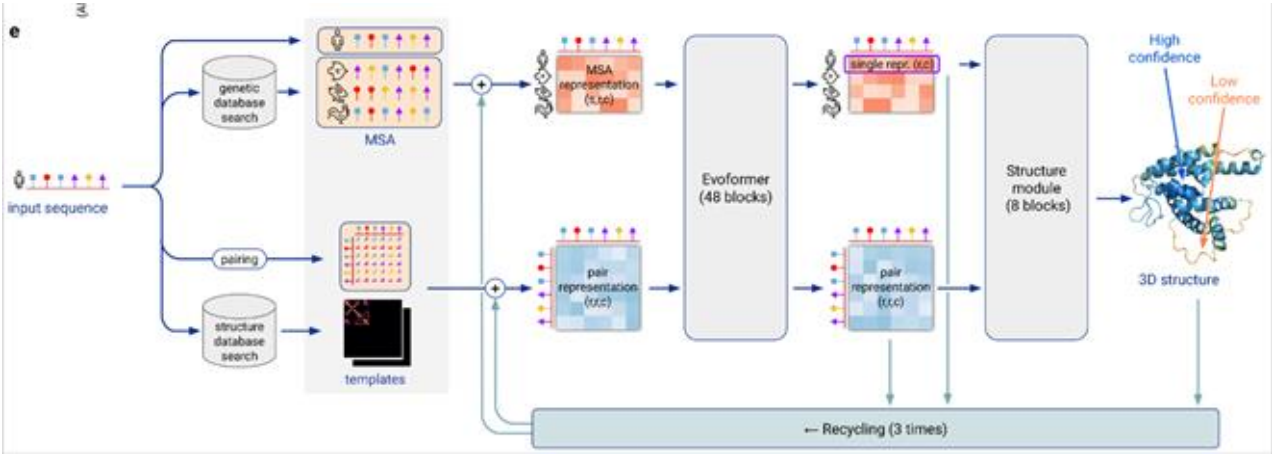
Also the Speed of Protein Structure Prediction also matter!

nature

https://doi.org/10.1038/s41586-021-03819-2

Accelerated Article Preview

Highly accurate protein structure prediction with AlphaFold



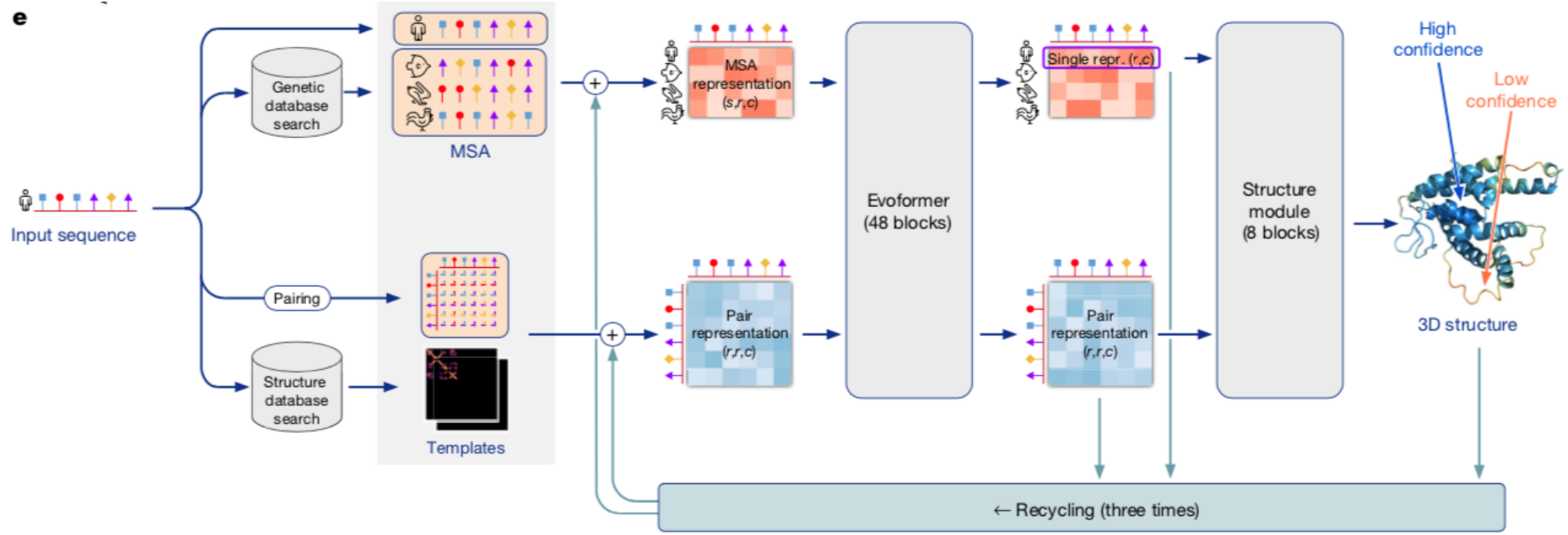
Received: 11 May 2021

Accepted: 12 July 2021

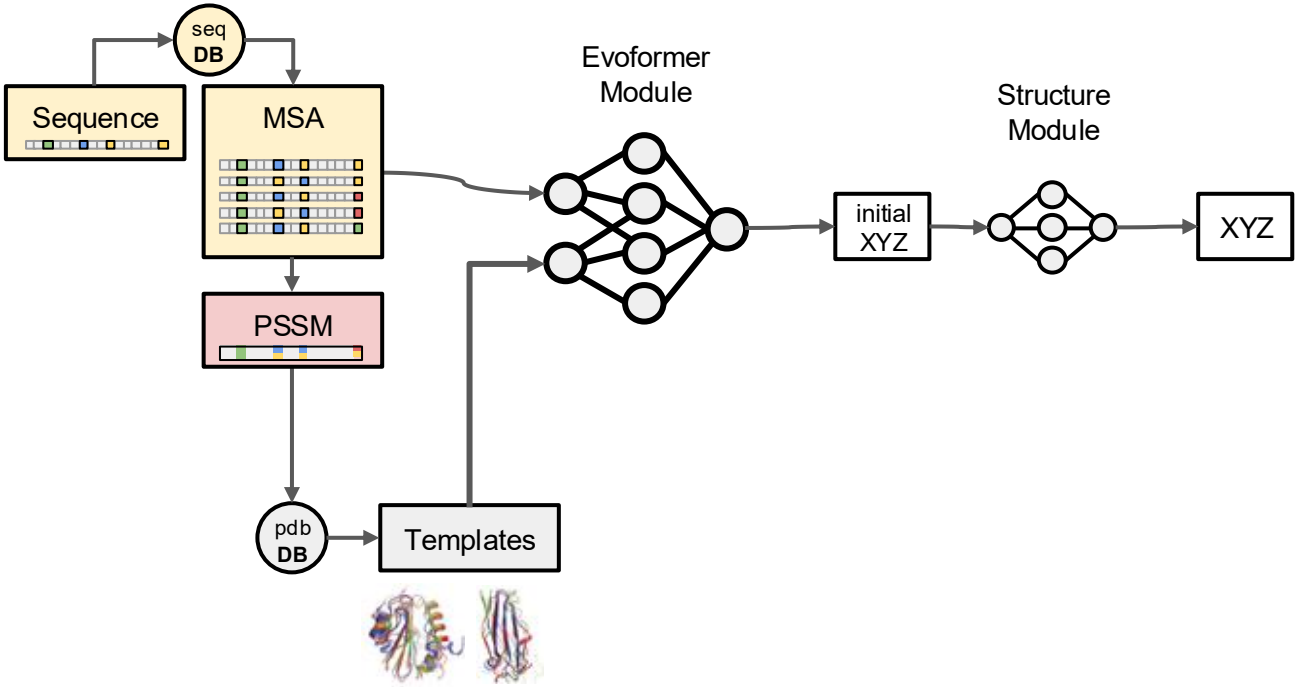
Accelerated Article Preview Published online 15 July 2021

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 Žídek, 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


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

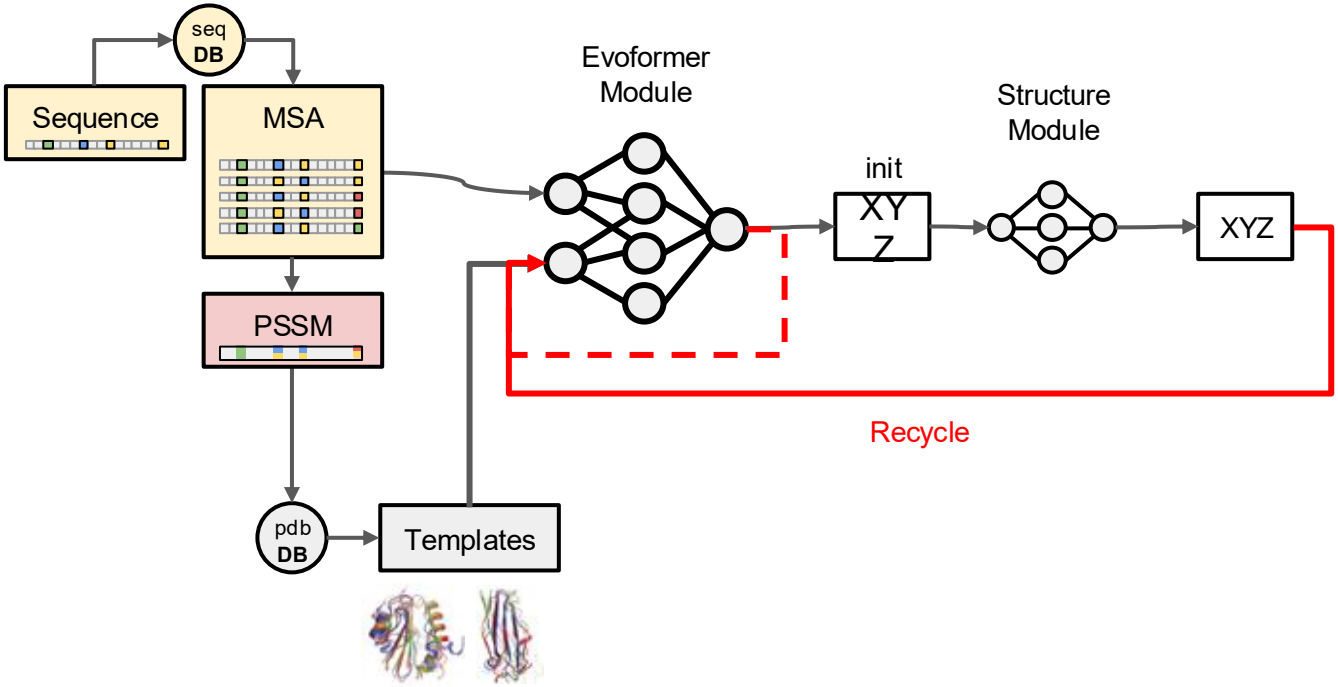
Alphafold2*



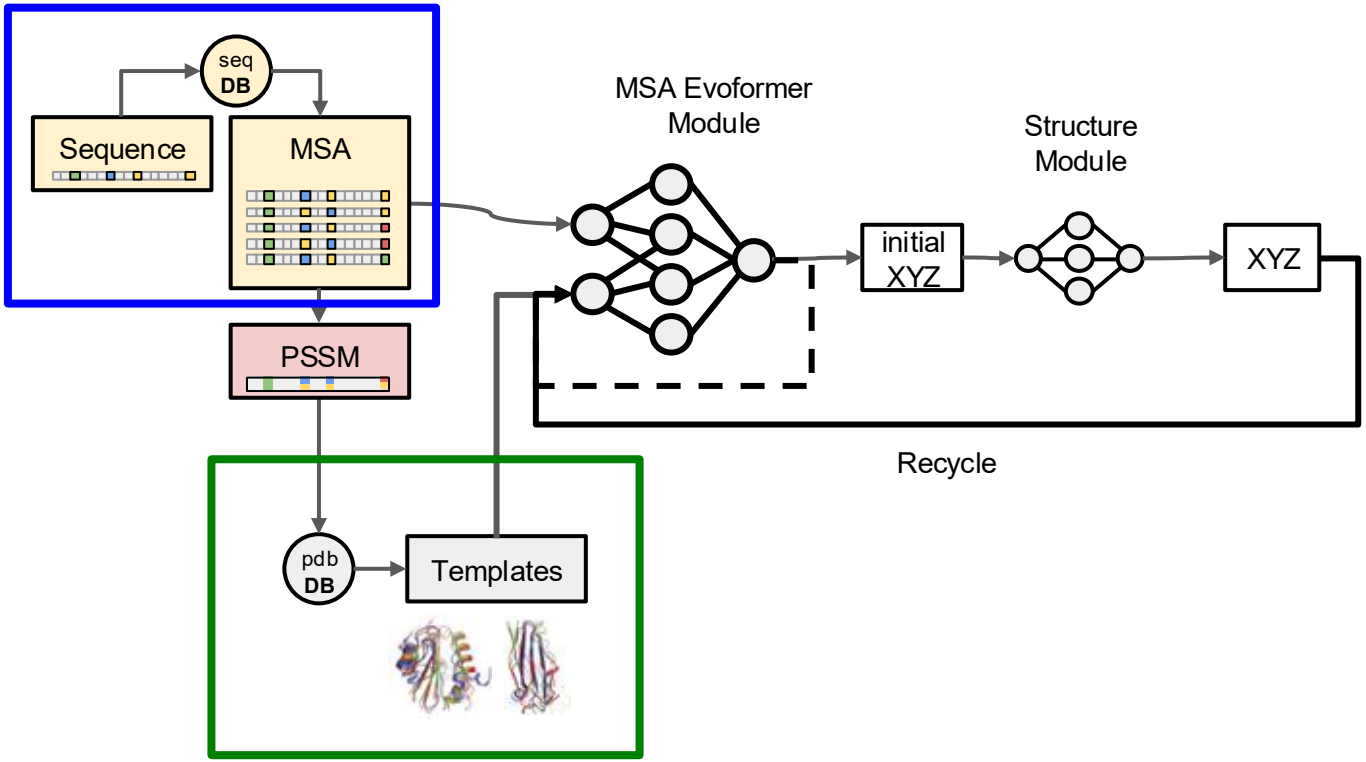
* Past researchers used raw Templates as input and/or did End2End

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.

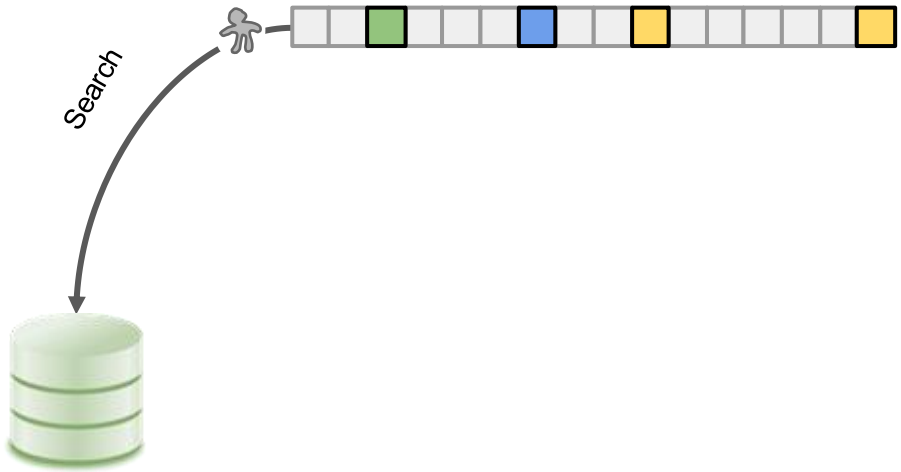
Alphafold2 - New Critical detail Recycling



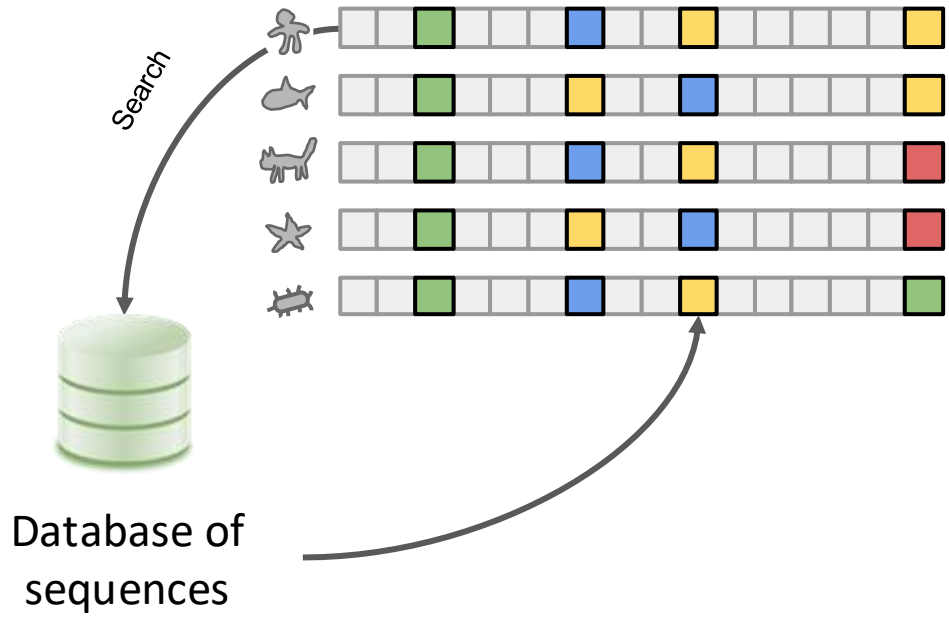
Structure Prediction relies on the input MSA



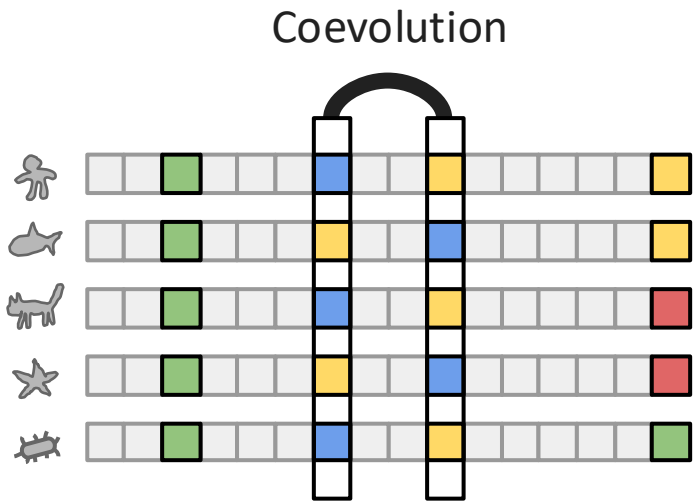
What is a Multiple Sequence Alignment (MSA)?
Search against a database of sequences

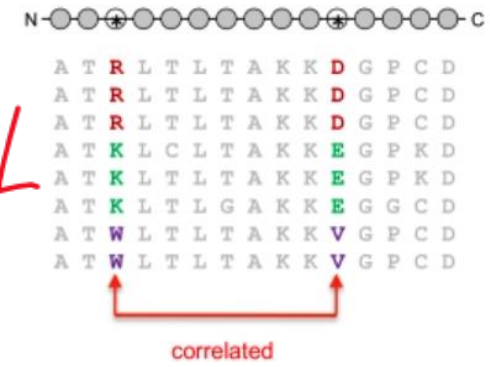


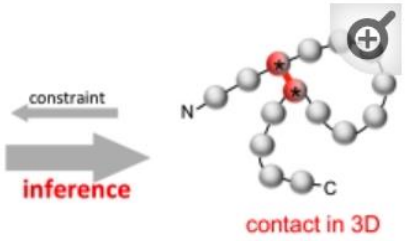
Generate a multiple sequence alignment



Analyze the MSA for coevolution

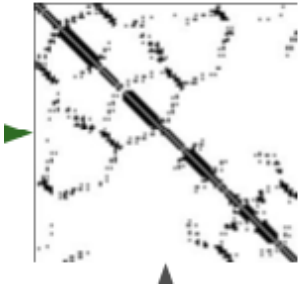


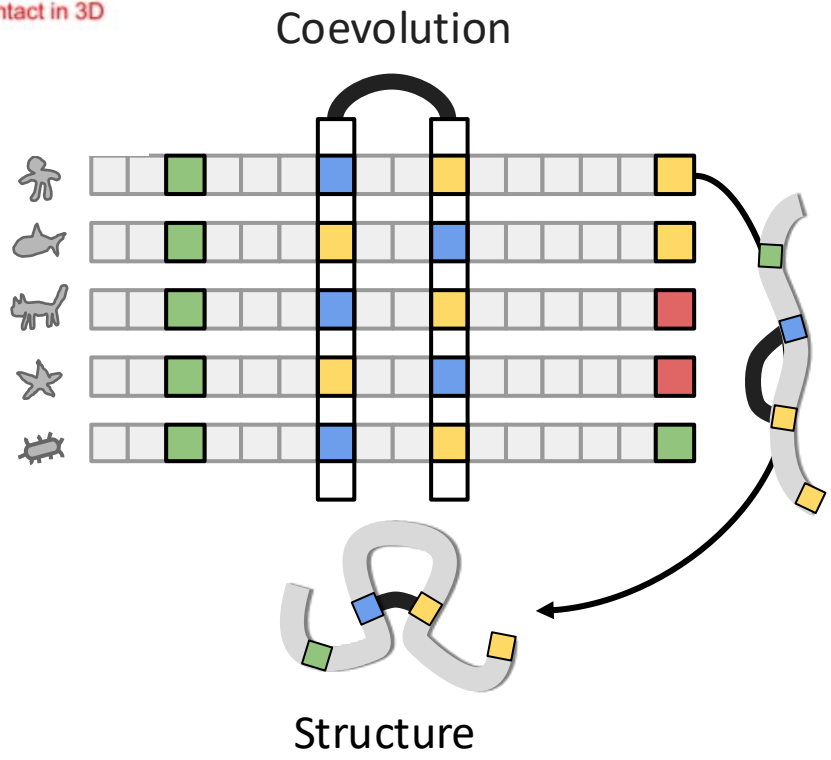




EVfold

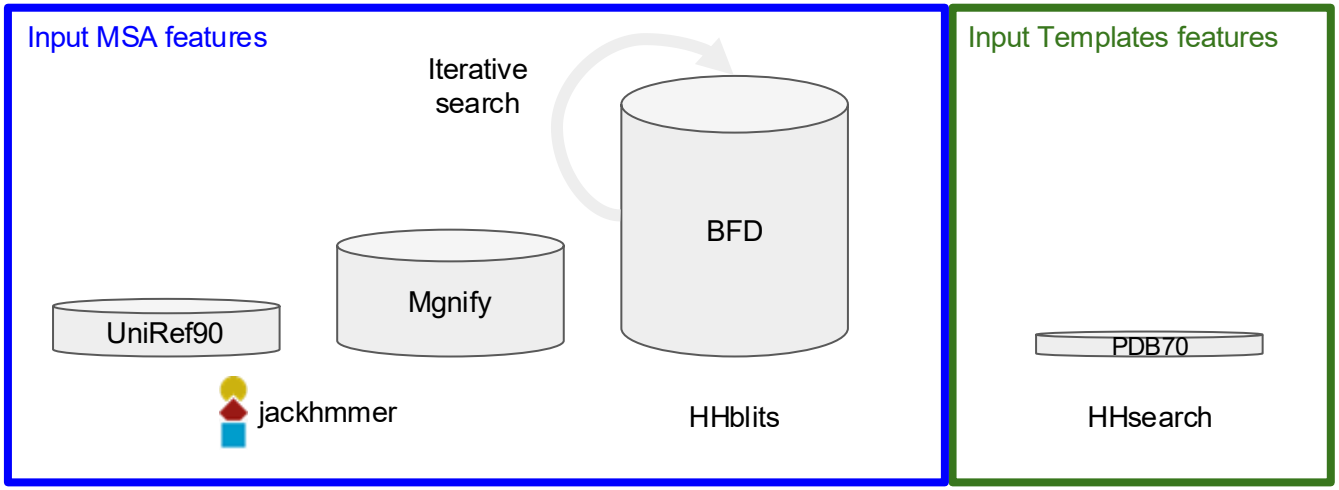
Related Task:
Contact prediction





MSA based
Input feature generation for AlphaFold2

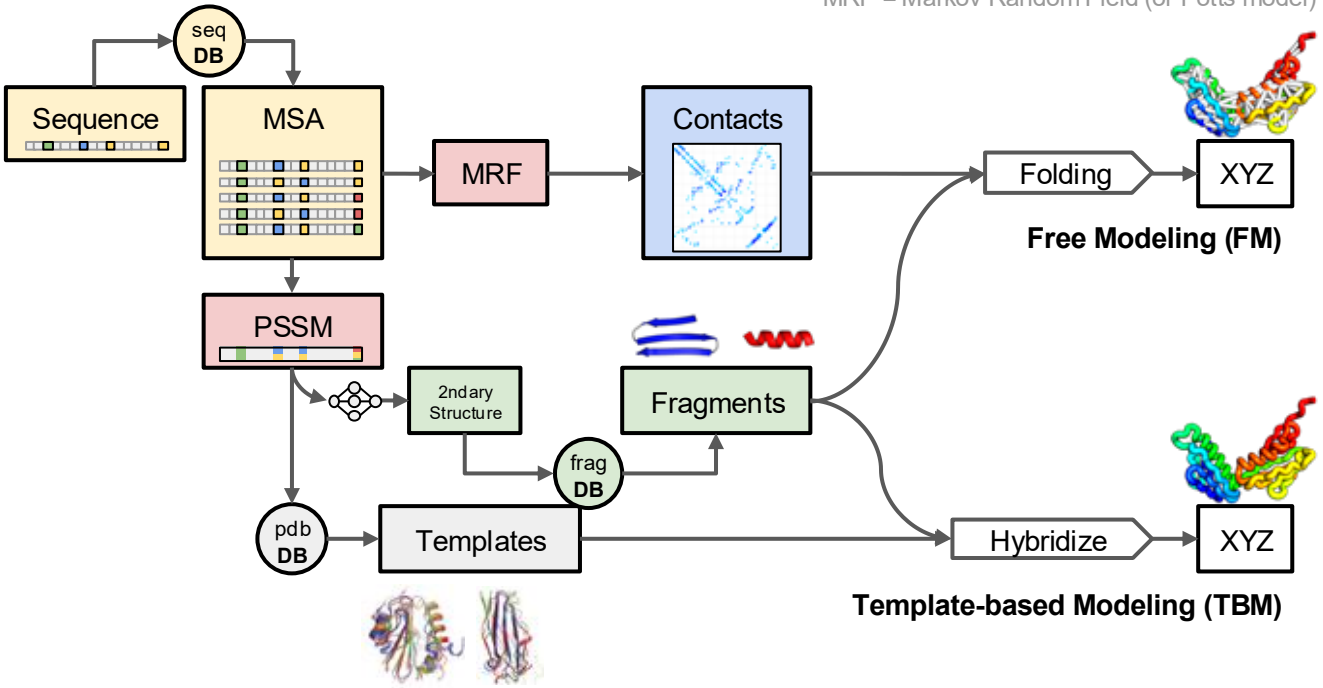
SLOW



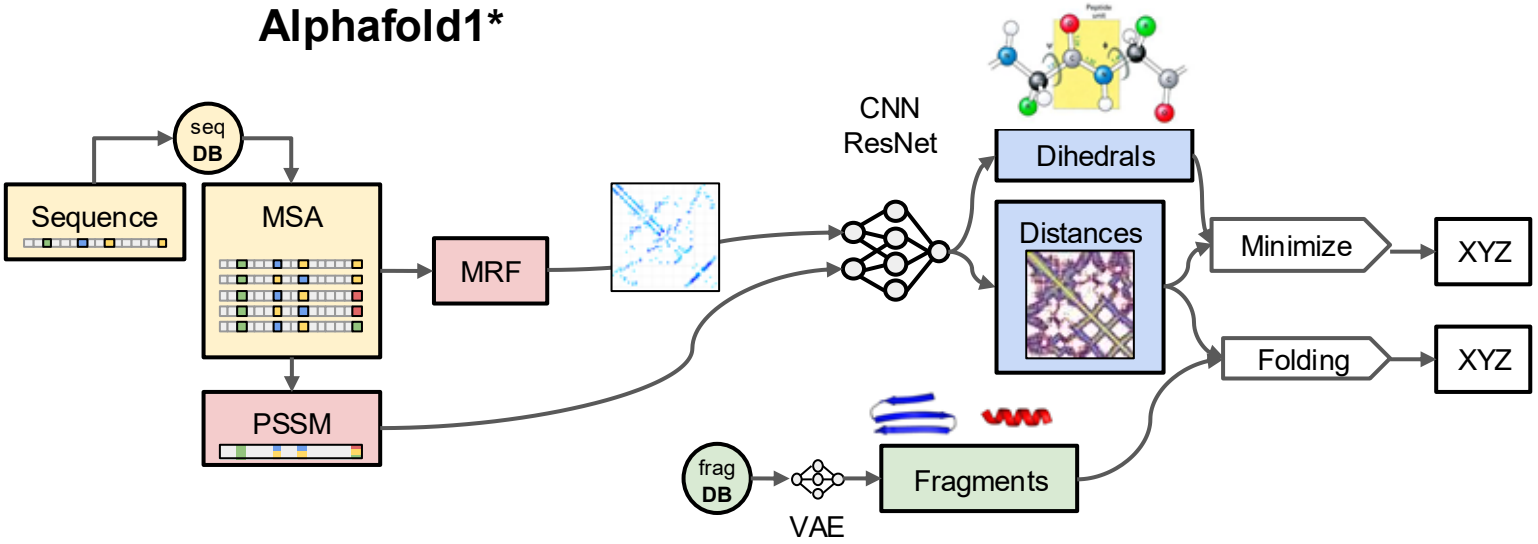
Generation of input MSA features can take **hours** for a single protein on multiple cores

Typical pipeline before Alphafold

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



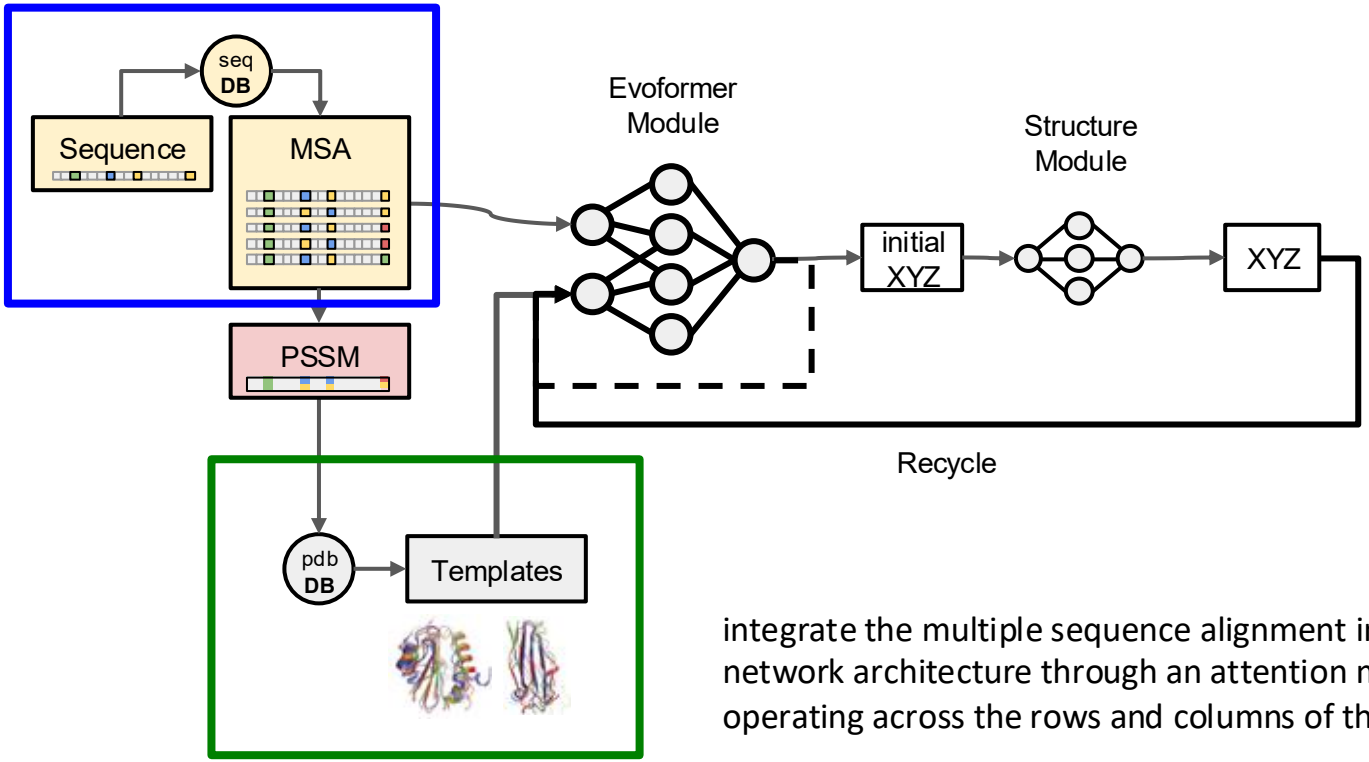
Alphafold1*



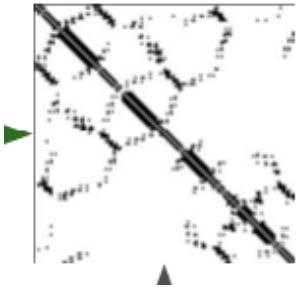
* Past researchers used raw MRF features, and ResNets:

- 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

Related Task:
Contact prediction

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]

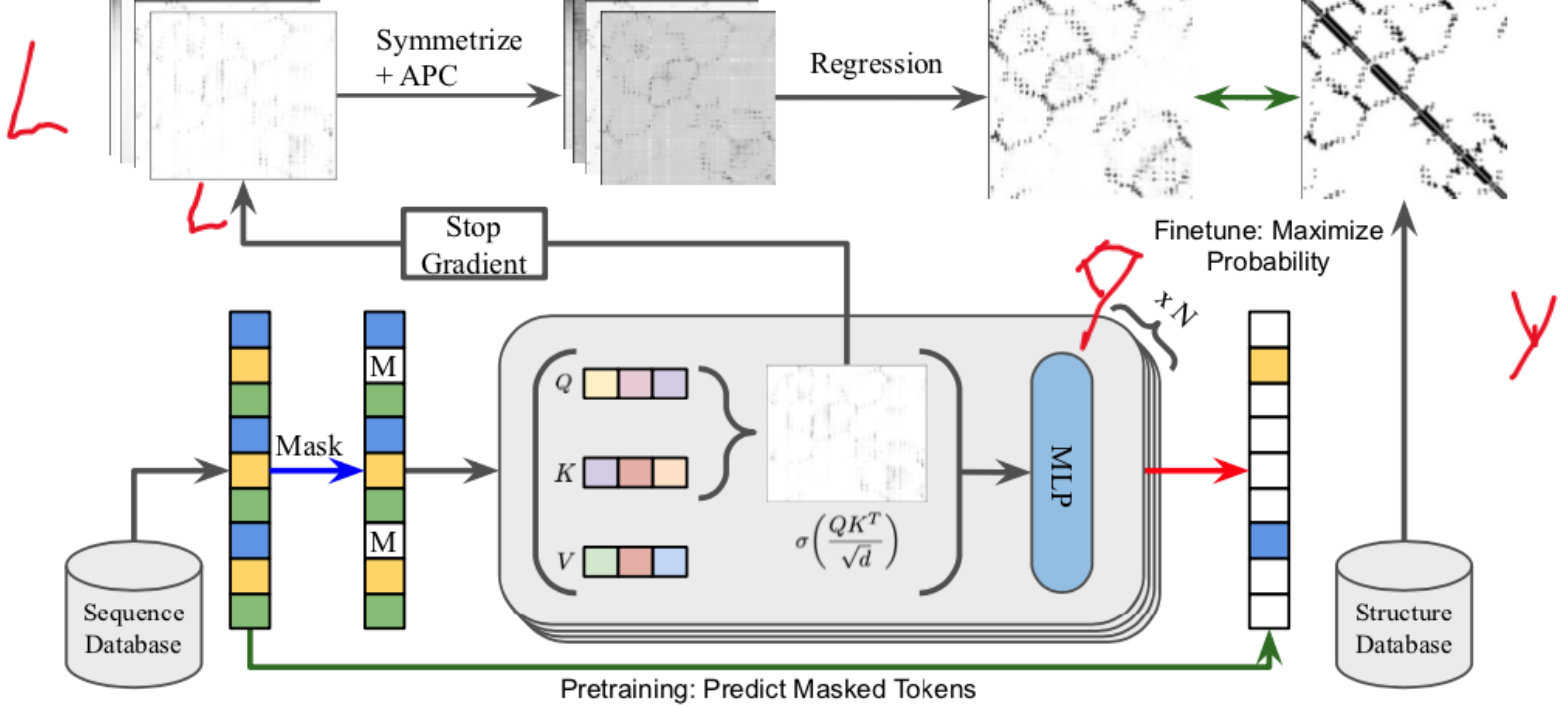


Figure 1: Contact prediction pipeline. The Transformer is first pretrained on sequences from a large database (Uniref50) via Masked Language Modeling. Once finished training, the attention maps are extracted, passed through symmetrization and average product correction, then into a regression. The regression 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 predict (==> 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.

ESM-2

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

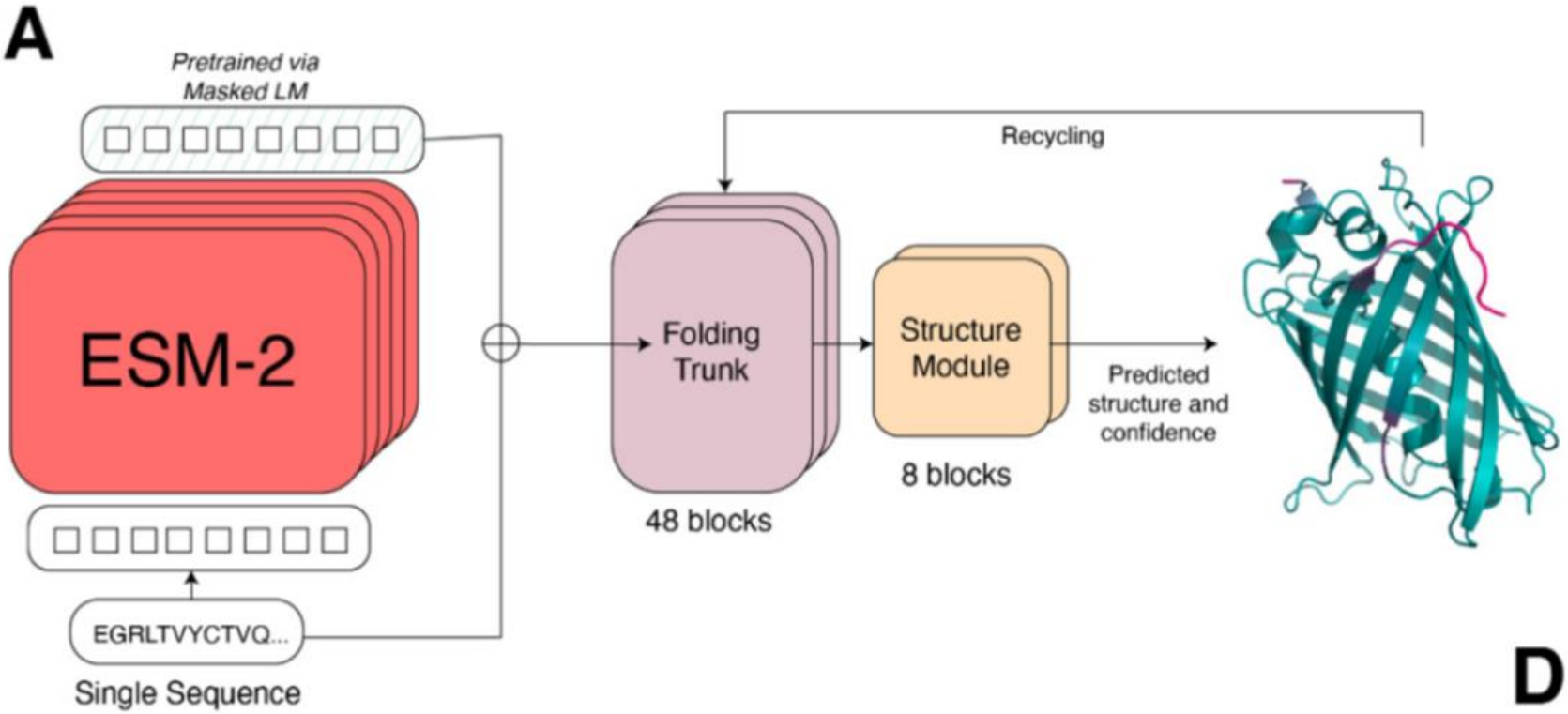
Zeming Lin [1 2 *] Halil Akin [1 *] Roshan Rao [1 *] Brian Hie [1 3 *] Zhongkai Zhu [1] Wenting Lu [1] Nikita Smetanin [1]
Robert Verkuil [1] Ori Kabeli [1] Yaniv Shmueli [1] Allan dos Santos Costa [4] Maryam Fazel-Zarandi [1] Tom Sercu [1 †]
Salvatore Candido [1 †] Alexander Rives [1 † ‡]

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),


A
Pretrained via
Masked LM
ESM-2
EGRLTVYCTVQ...
Single Sequence
Recycling
Folding
Trunk
48 blocks
Structure
Module
8 blocks
Predicted
structure and
confidence
D

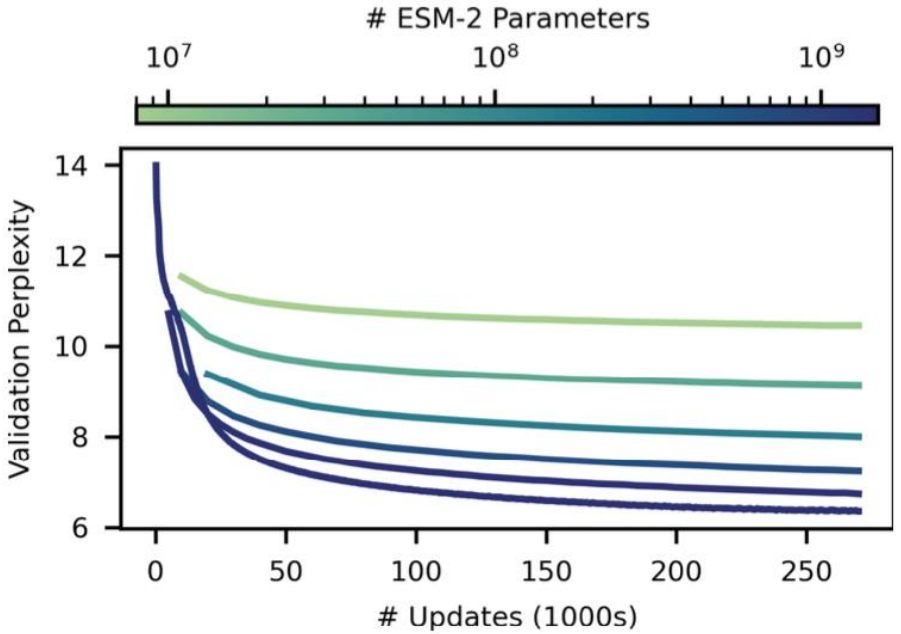
Protein language model (largest to 2022)

- 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.

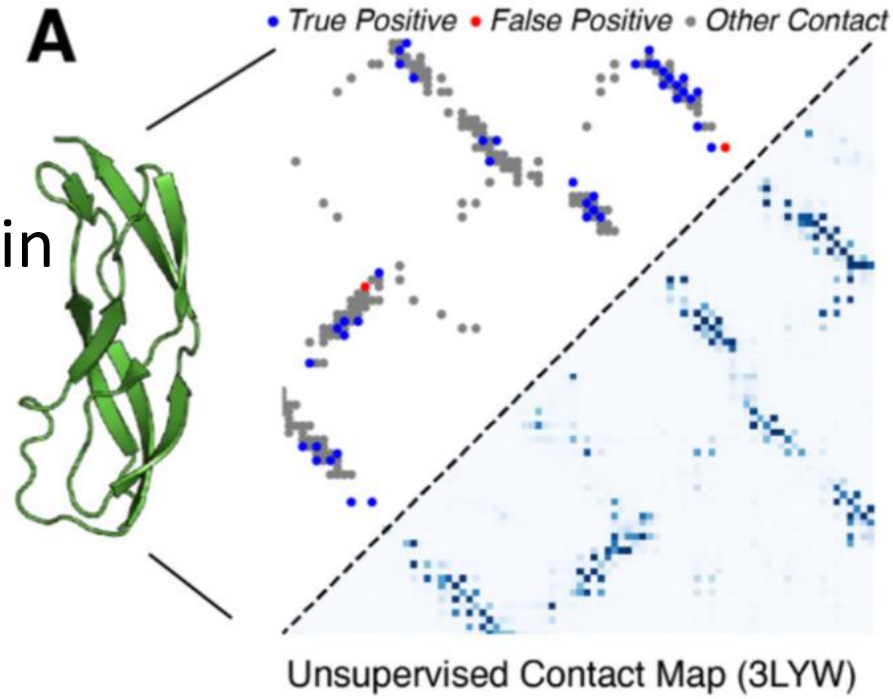


Here perplexity, ranges from 1 for a perfect model to 20 for a model that makes predictions at random.

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

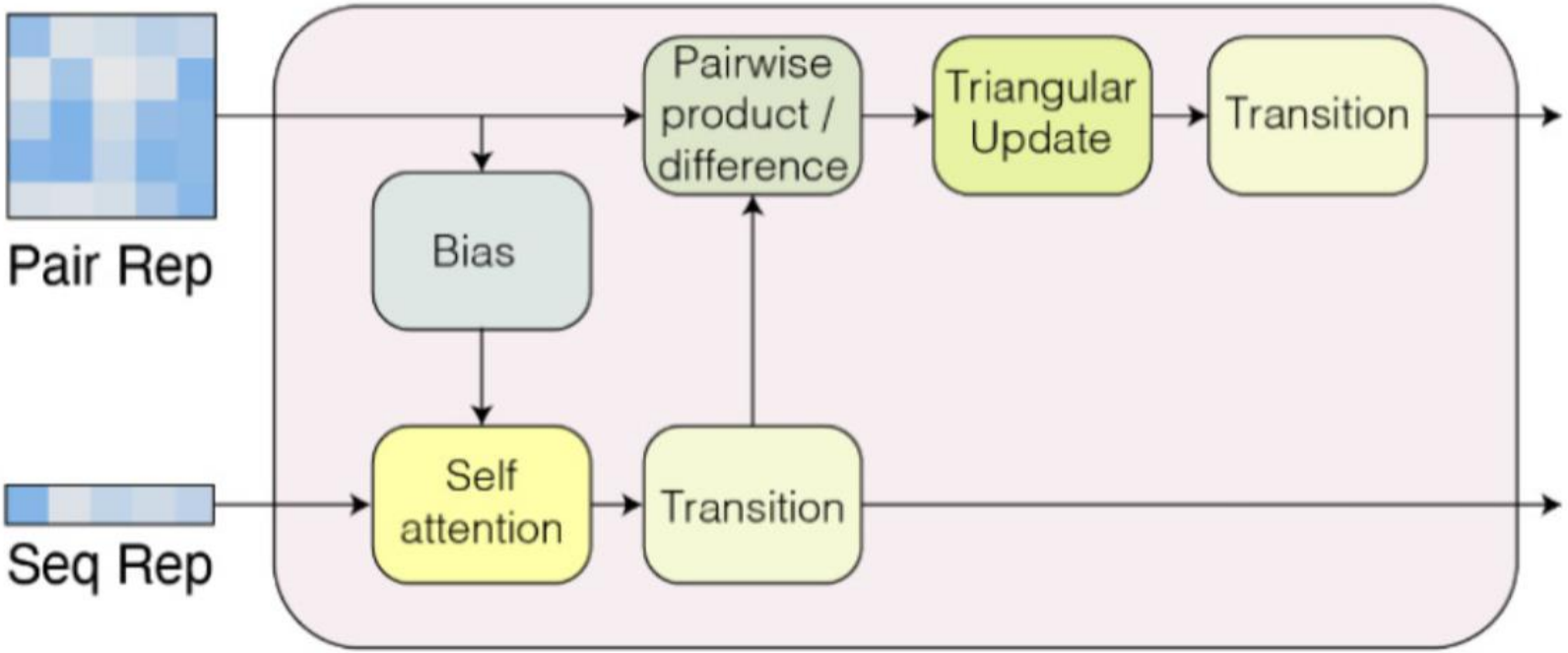
ESM-2 attention patterns correspond to the residue-residue contact map of a protein



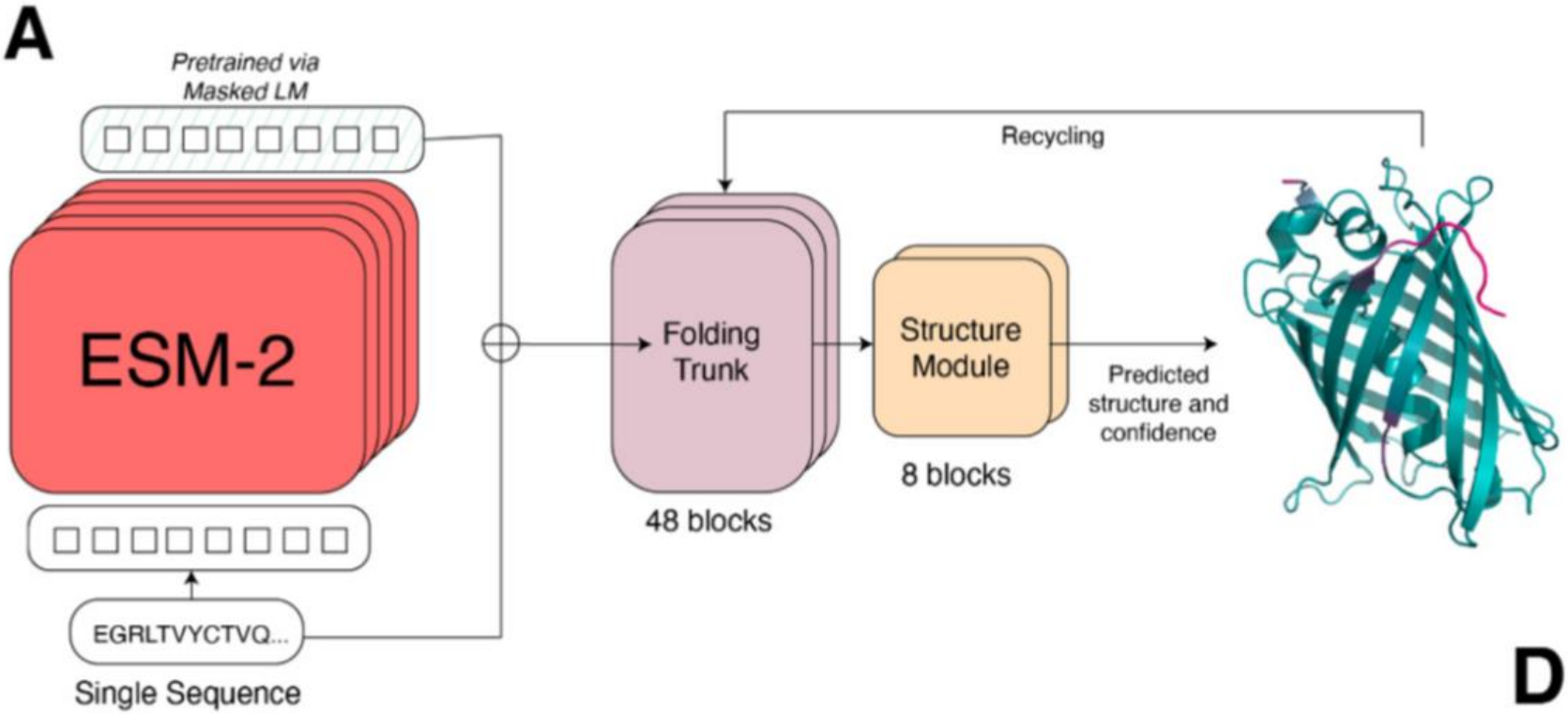
Predicted contact probabilities (bottom right) and actual contact precision (top left) for 3LYW. A contact is a positive prediction if it is within the top-L most likely contacts for a sequence of length L.

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

- 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


A
Pretrained via
Masked LM
ESM-2
EGRLTVYCTVQ...
Single Sequence
Recycling
Folding
Trunk
48 blocks
Structure
Module
8 blocks
Predicted
structure and
confidence
D

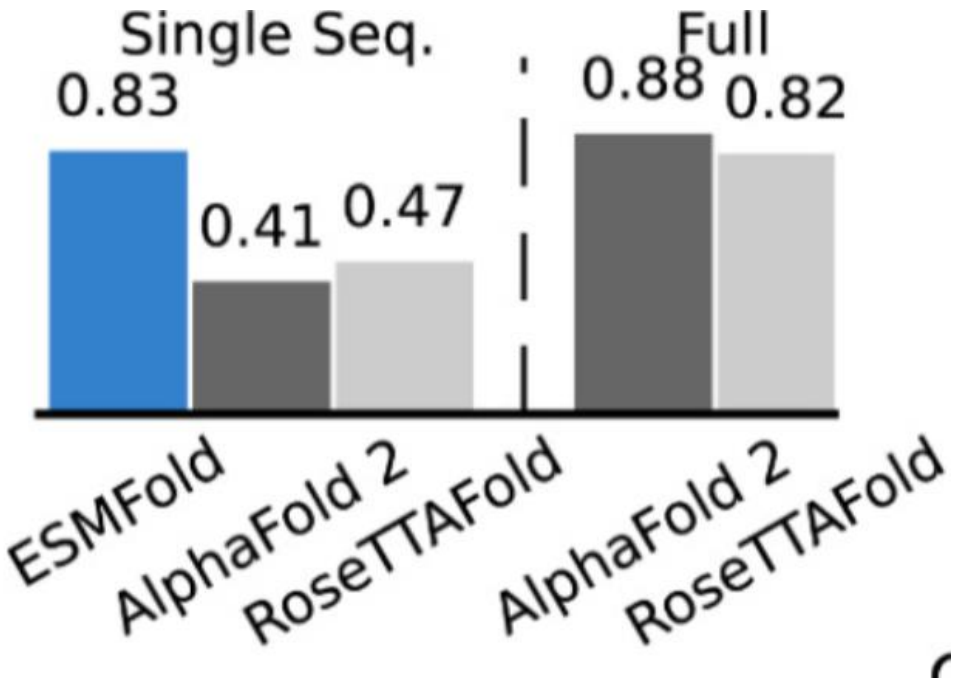
ESMfold architecture

- 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,

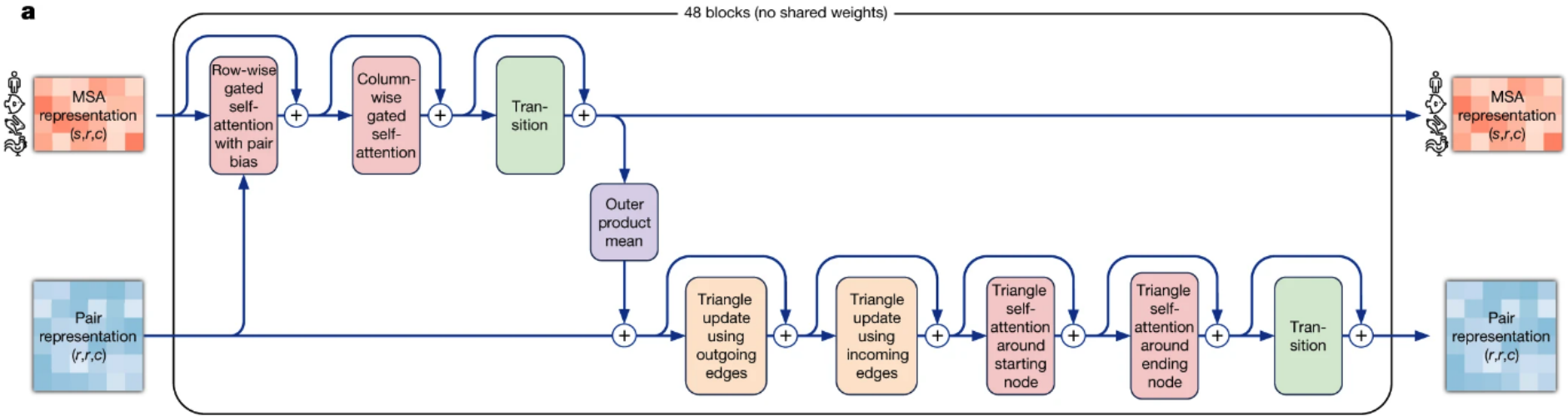
- ESMFold uses the Frame Aligned Point Error (FAPE) and distogram losses introduced in AlphaFold2, as well as heads for predicting LDDT and the pTM score.

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



Vs. 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

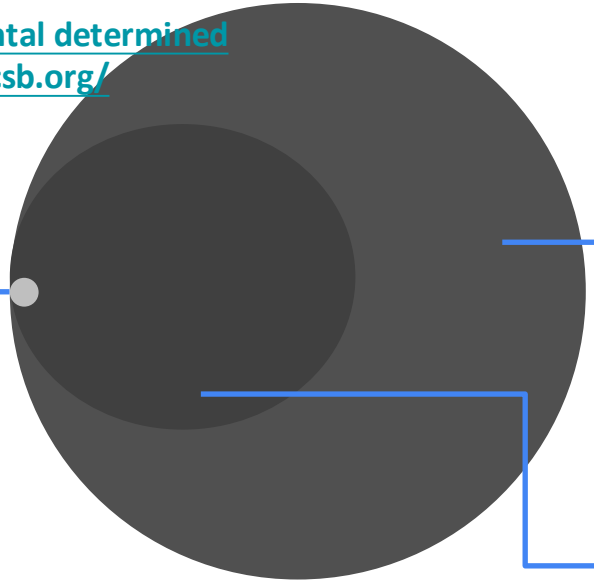


ESMfold more

- 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

Protein Structure landscape

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



ESM Metagenomic Atlas (https://esmatlas.com): 617M proteins. 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

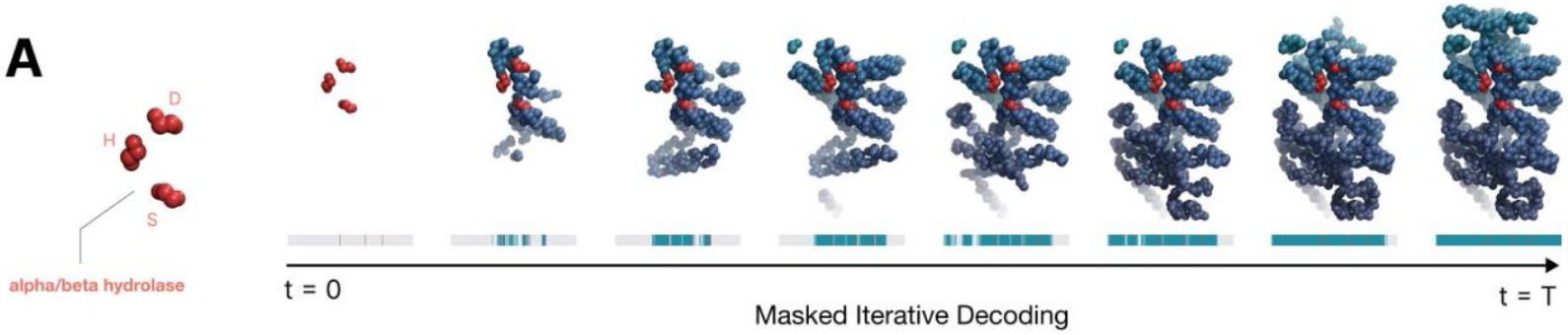
Then/Now ESM-3

Simulating 500 Million Years of Evolution with a Language Model

ESM3: A Multimodal Protein Language Model

Outline

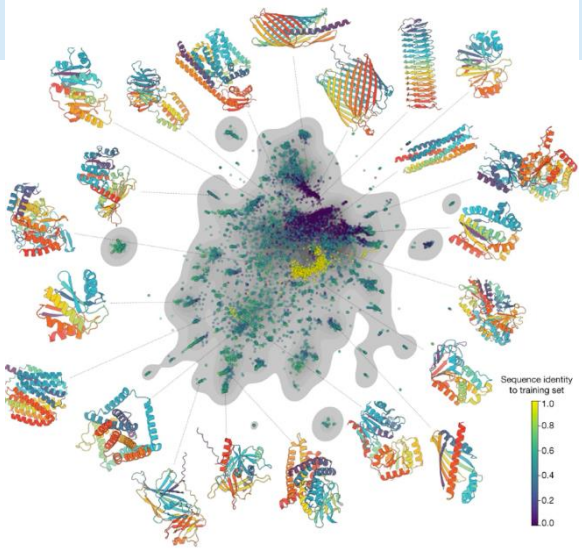
1. **Motivation** — Why protein language models? Why now?
2. **Method Details** — ESM3 architecture & multimodal representation
3. **Explanations** — How tokenization, masking, and generation work
4. **Setup** — Training data, scales, and evaluation framework
5. **Results** — Prompt-following, alignment, and esmGFP
6. **Takeaways** — Implications for protein design and AI biology



The Problem: Exploring Protein Space

- **~3 billion years** of evolution have produced the proteins we observe today
- Natural evolution is a slow, constrained search through sequence space
- Gene sequencing surveys now catalog **billions** of sequences and structures
- Yet the functional protein space remains **vast and mostly unexplored**

Key Question: Can we build a model that learns the deep structure of protein evolution — and use it to generate *functional* proteins far beyond what nature has found?



Why Generative Language Models for Proteins?

Evidence from prior work:

- Representations in protein LMs reflect biological structure & function — *without explicit supervision*
- Performance improves with scale (like NLP)
- Scaling laws predict continued capability gains

What's been missing:

- Multimodal reasoning (sequence + structure + function together)
- Controllable generation
- Ability to reach *distant* functional proteins

Analogy to NLP:

NLP LMs	Protein LMs
Text tokens	Amino acid tokens
Grammar rules	Evolutionary constraints
Sentence completion	Protein generation
GPT/BERT	ESM3

Prior models: ESMFold, ProtGPT2, ProGen2, RFdiffusion — but none unified all three modalities at scale

The Central Claim

500M

years of simulated evolution — in a single model run

> ESM3 generated **esmGFP**: a functional green fluorescent protein with only **58% sequence identity** to its nearest known relative — a distance equivalent to >500 million years of natural evolution.

This is the first time a new GFP this distant from known proteins has been produced *outside* of natural discovery.

ESM3 Architecture Overview

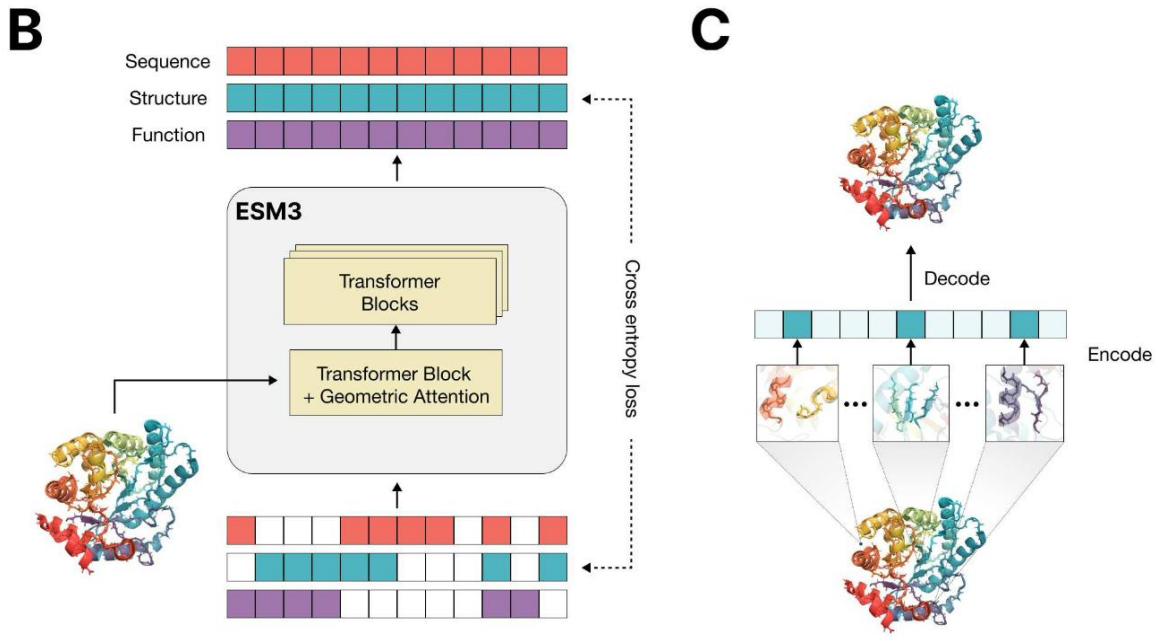
ESM3 is a **bidirectional transformer** that reasons jointly over three modalities:

Inputs (all optional/maskable):

- Amino acid sequence tokens
- 3D structure tokens (via VQ-VAE)
- Secondary structure (SS8) tokens
- Solvent Accessible Surface Area (SASA)
- Function keyword tokens
- Residue-level annotations

Output:

- Token probability distributions for each track
- Decoded full atomic structure



The Three Modalities in Detail

Modality	Representation	Token Vocab	Key Role
Sequence	20 canonical amino acids + special	29 tokens	Primary chain identity
Structure	VQ-VAE discrete codes of local 3D atomic neighborhoods	4096 + 4 special	Fold & binding geometry
Function	LSH-quantized TF-IDF of keyword annotations (InterPro)	255 × 8 per residue	Biological activity
SS8	8-class secondary structure	10 tokens	Coarse topology
SASA	16-bin discretized surface exposure	18 tokens	Solvent accessibility

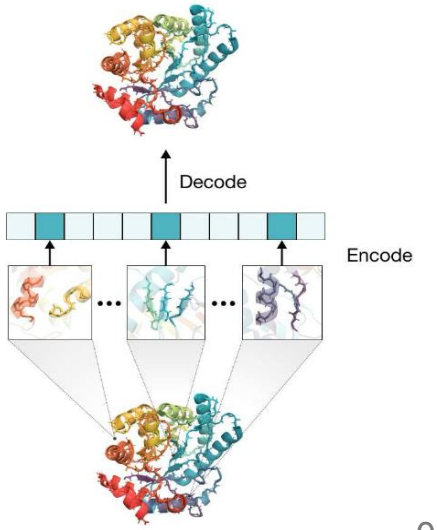
All modalities share a single latent space — the model learns cross-modal relationships entirely from data

Structure Tokenization: The VQ-VAE

Converting 3D structure to discrete tokens is non-trivial. ESM3 uses a **discrete autoencoder**:

1. **Encode:** Local atomic neighborhood around each amino acid → discrete token (codebook size 4,096)
2. **Invariant geometric attention** operates in local reference frames (bond geometry at each residue), then transforms to global frame
3. **Decode:** Structure tokens → full atomic coordinates via 700M parameter transformer decoder

> **Quality:** Near-perfect reconstruction — **<0.5 Å RMSD** on CAMEO test set



Structure tokenization schematic (Fig. 1C) — local neighborhood → VQ-VAE → discrete token → decode → coordinates]

Training Objective: Generative Masked Language Modeling

$$\mathcal{L} = -\mathbb{E}_{x,m}\left[\frac{1}{|m|}\sum_{i\in m}\log p(x_i|x_{\setminus m})\right]$$

- A random mask **m** is applied to tokens across all modalities
- The model predicts **masked tokens** from unmasked context
- Key innovation: mask fraction varies over a **noise schedule** (not fixed like BERT)

> This means the model learns to generate any modality from any other — sequence from structure, structure from function keywords, etc. — enabling flexible **all-to-all** generation.

Higher masking rates → better generative capability; lower masking rates → better representation learning. The noise schedule balances both.

Chain-of-Thought Generation

For complex design tasks, ESM3 uses a **multi-step generation protocol**:

Step 1: Generate structure tokens (backbone)
→ Filter for active site coordination quality
Step 2: Condition on new structure + original prompts
→ Generate sequence tokens
Step 3: Iterative joint optimization
→ Alternate sequence ↔ structure optimization
Step 4: Rank designs by multiple metrics
→ Select top candidates per sequence-identity bucket

> This is analogous to chain-of-thought reasoning in LLMs — intermediate "thoughts" (structures) guide the final output (functional sequence).

Biological Alignment (Fine-tuning)

Base ESM3 is further improved via **preference optimization** (similar to RLHF for LLMs):

1. Generate many proteins for backbone coordinate prompts
2. Score each generation: **pTM** (structure confidence) + **backbone cRMSD** (prompt adherence)
3. Pair **high-quality vs. low-quality** generations for the same prompt → preference dataset
4. Fine-tune with **Direct Preference Optimization (DPO)**-style loss

> Result: Model learns to put higher likelihood on good scaffolds. Larger models benefit *much more* from alignment — revealing latent capability.

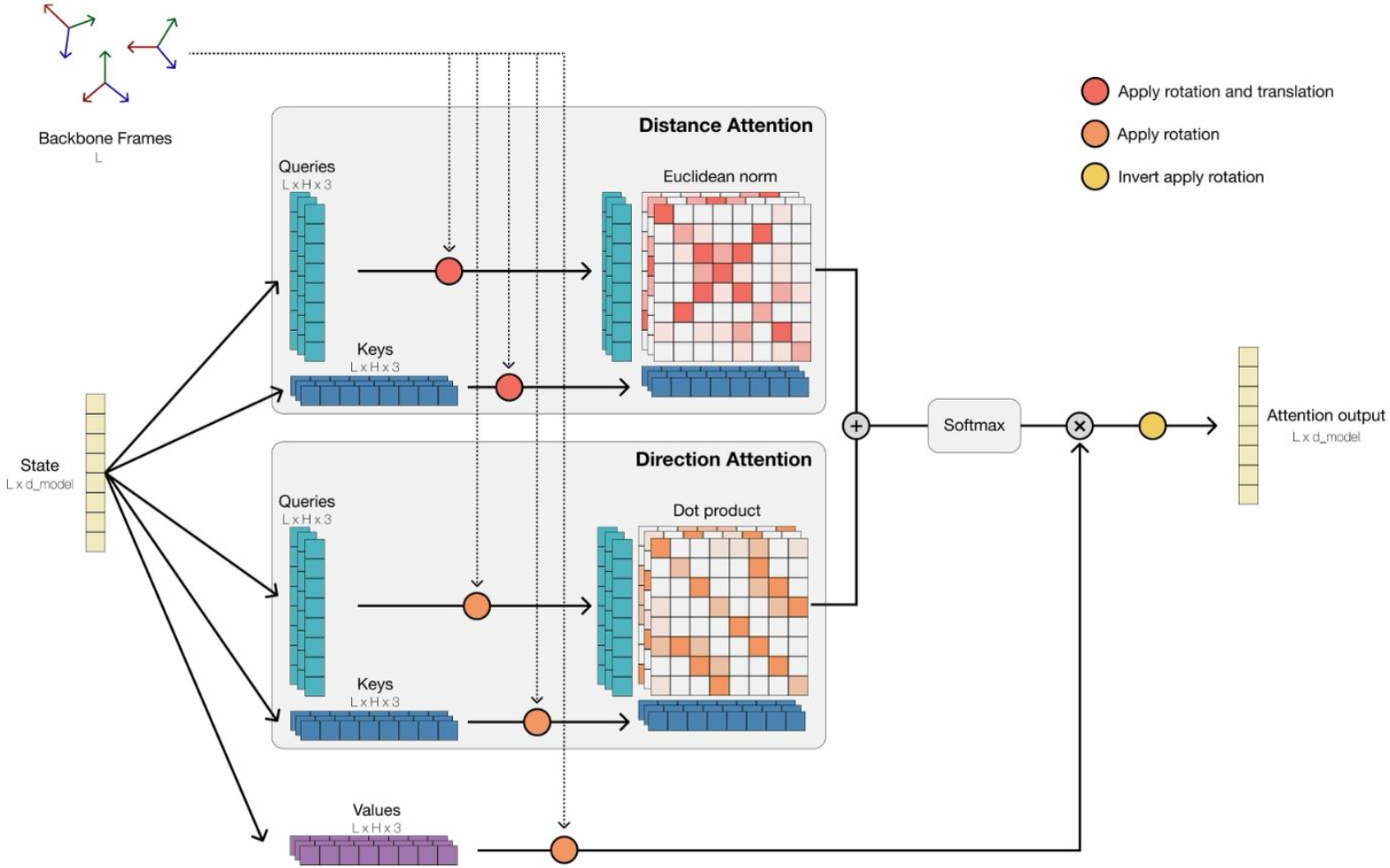
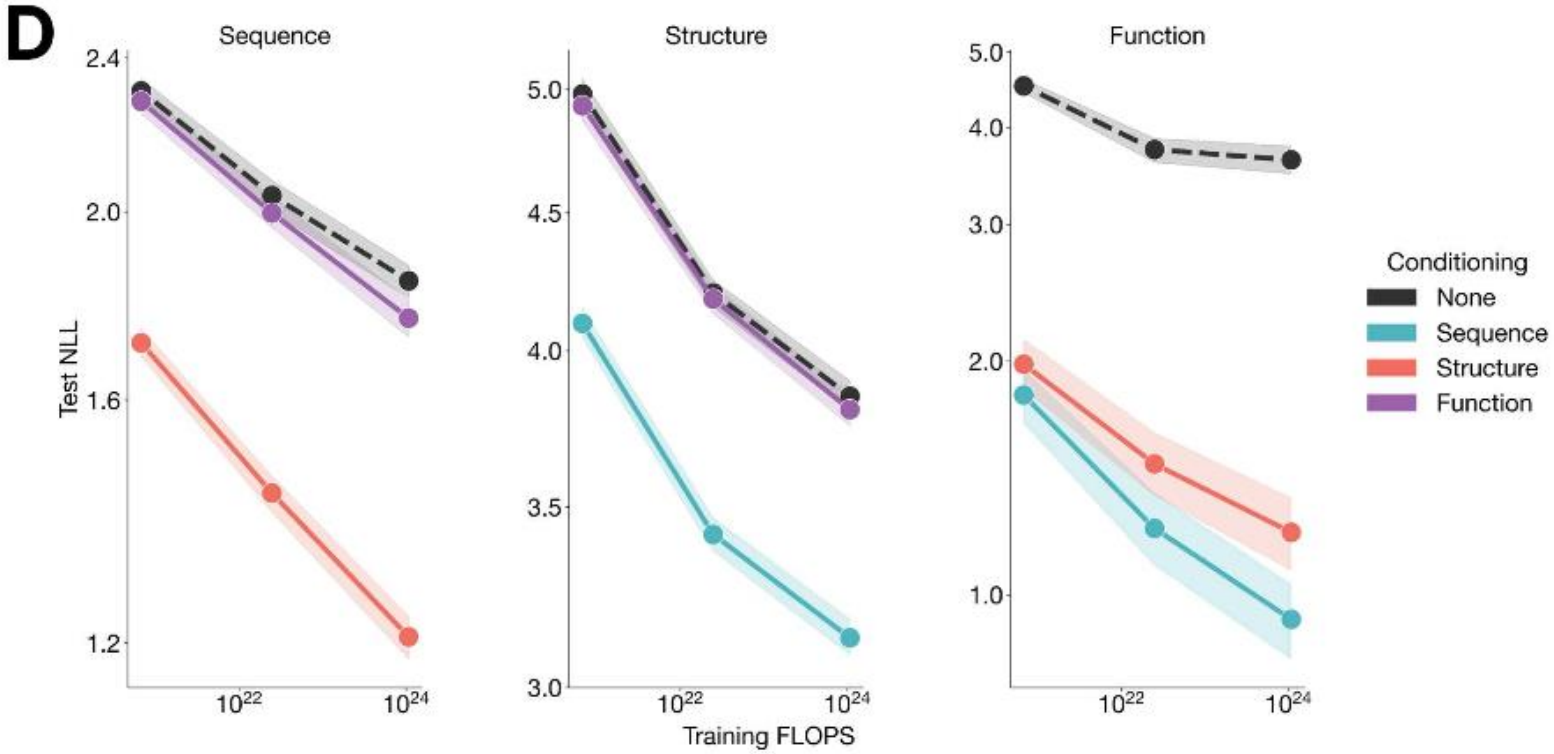


Figure S2. Geometric attention. Geometric attention is an SE(3) invariant all-to-all attention mechanism where the attention score matrix

(D) Models are trained at three scales: 1.4B, 7B, and 98B parameters. Negative log likelihood on test set as a function of training FLOPs shows response to conditioning on each of the input tracks, improving with increasing FLOPs.



Training Data

Data Type	Count
Natural protein sequences	2.78 billion
Protein structures (experimental + predicted)	236 million
Proteins with function annotations	539 million
Total unique tokens	**771 billion**

Data sources: UniRef, JGI, MGnify, OAS (antibodies), PDB, AlphaFold DB, ESMAtlas

Augmentation: Synthetic sequences generated via inverse folding (ESM-IF1) for all structures, including predicted ones

→ expands training to **3.15 billion** total sequences

Model Scales & Compute

Model	Parameters	Transformer Blocks
ESM3-Small	1.4B	—
ESM3-Medium	7B	—
ESM3-Large	**98B**	**216**

Compute: Largest model trained with **1.07×10^{24} FLOPs**

Architecture choice: Deeper (more layers) > wider — larger depth response found in architecture search

Evaluation Framework

Prompt-following metrics (per track):

- **cRMSD** — backbone atom RMSD between prompt coordinates and generation
- **SS3 accuracy** — 3-class secondary structure match fraction
- **SASA Spearman ρ** — correlation between prompted and generated SASA
- **Keyword recovery** — fraction of function keywords recovered by InterProScan

Generative quality:

- **pTM / pLDDT** — ESMFold-predicted structure confidence scores
- **scTM** — self-consistency TM-score (inverse fold → refold)

Alignment evaluation:

- **Pass@128** — fraction of 46 ligand binding motif tasks solved in 128 attempts

Result 1 — ESM3 Follows Complex Prompts Faithfully

Across all individual tracks, ESM3 7B achieves **high prompt consistency AND high structural confidence**:

> Key finding: The model generalizes to **out-of-distribution folds** (TM < 0.7 to training set) while maintaining coherent structures — mean pTM 0.85 ± 0.03

Result 2 — Creative Motif Scaffolding

ESM3 composes **atomic-level motif prompts** with **high-level fold/keyword prompts** — finding novel scaffolds:

Result 2 (cont.) — Protein Compression

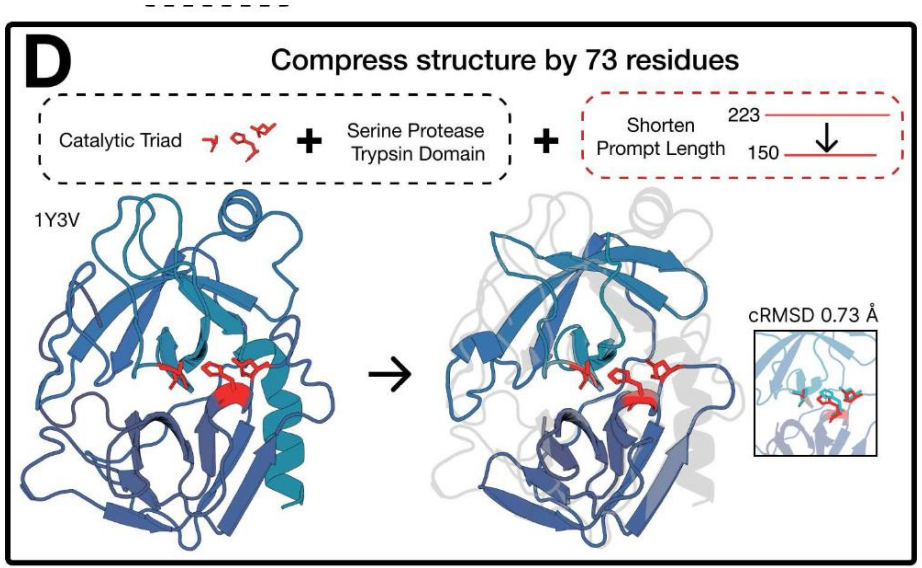
A striking example of creative generation: ESM3 **compresses trypsin by 33%** while preserving function:

Prompt:

- Catalytic triad atomic coordinates (from PDB 1Y3V)
- Trypsin function keywords
- **Target length: 150 residues** (vs. 223 natural)

Result:

- Active site RMSD: **0.73 Å**
- Structure confidence: pTM **0.84**
- Self-consistency: scTM mean **0.97**

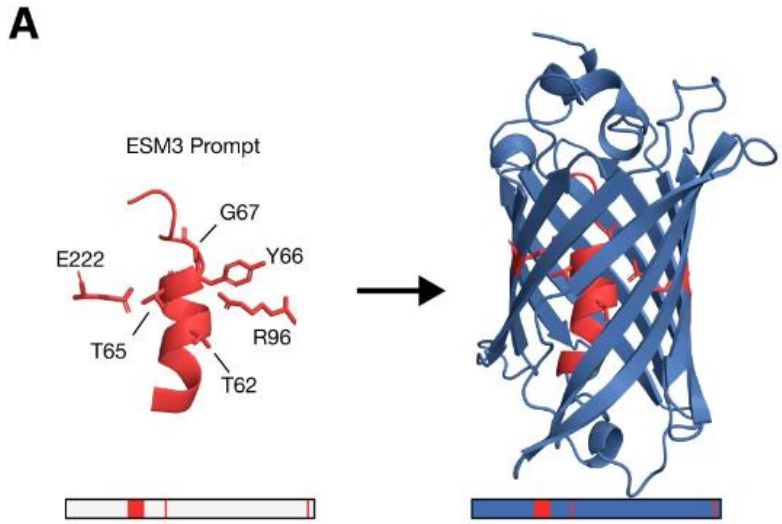


Result 3 — Alignment Unlocks Latent Capability

Model	Base	Aligned	Gain
1.4B	9.5%	18.8%	+2×
7B	19.0%	37.4%	+2×
98B	**26.8%**	**65.5%**	**+2.4×**

Pass@128 on tertiary coordination tasks

> The **largest model** benefits most from alignment — revealing capabilities not apparent in the base model



Result 4 — Generating esmGFP

Goal: Create a functional GFP with low sequence identity to all known GFPs

Prompt (base pretrained 7B model):

- Sequence + structure tokens + backbone coords of residues **Thr62, Thr65, Tyr66, Gly67, Arg96, Glu222** (chromophore-forming residues)
- Structure of residues 58–71 from PDB 1QY3 (energetically important helix region)
- Target length: **229 residues**

Using a time-calibrated phylogenetic analysis of anthozoans:

- esmGFP similarity to known FPs mirrors **cross-order diversity** within the Anthozoa class
- E.g., comparable to divergence between Scleractinia (stony corals) and Actiniaria (sea anemones)

Estimated equivalent evolutionary distance: **>500 million years**

Key Takeaways

1. Proteins as a language — and evolution as its "corpus"
ESM3 frames protein biology as a token prediction problem. By learning to predict evolutionarily generated tokens, it implicitly learns the deep constraints that govern what proteins can exist.

2. Tokenization enables scalable multimodal modeling
Encoding structure as discrete tokens — rather than diffusion in 3D space — makes the model highly scalable and allows all-to-all conditioning across modalities.

3. Scale + alignment = emergent capability
Larger models have greater latent capability, but this is only revealed through preference alignment. The 98B aligned model solves 65.5% of tertiary coordination tasks vs. 9.5% for 1.4B base.

4. Evolutionary simulation beyond natural constraints
ESM3 can reach functional proteins that natural evolution hasn't found — not by mimicking evolution's path, but by learning the underlying fitness landscape structure.

Broader Implications

For protein engineering:

- Rational design at multiple abstraction levels (atoms → keywords)
- New scaffolds for binding sites with no natural precedent
- Protein compression / miniaturization

For AI & biology:

- Protein LMs as *simulators* of evolutionary possibility
- Representation space encodes fundamental biology
- Scaling laws apply in the biological domain

Limitations & open questions:

- Experimental validation cost remains high
- Chromophore maturation of esmGFP is slow (2 days vs. hours for natural GFPs)
- Does "simulating evolution" require understanding physics, or just statistics?
- How does this scale to more complex multi-domain proteins?

Model weights (ESM3-open) and esmGFP sequence are publicly available for academic research.

Summary: ESM3 at a Glance

Aspect	Details
Model type	Bidirectional transformer, generative masked LM
Modalities	Sequence + Structure (VQ-VAE) + Function keywords
Largest model	98B parameters, 216 transformer blocks
Training data	771B tokens from 3.15B proteins
Key capability	All-to-all conditional generation via discrete tokens
Alignment method	Preference optimization (DPO-style) on prompted generations
Flagship result	esmGFP — functional GFP at 58% identity, ~500M years distant
Availability	ESM3-open weights + esmGFP sequence in public domain

Appendix: Architecture Details

Transformer block design:

- Standard transformer (Vaswani et al.) with pre-norm (RMSNorm)
- Rotary position embeddings (RoPE) for sequence position
- SwiGLU activation in feed-forward layers
- Geometric attention in **first block only** for direct coordinate conditioning

Structure tokenizer:

- VQ-VAE with invariant geometric attention
- Local reference frames defined by bond geometry at each amino acid
- Loss: pairwise distances + relative orientations of bond vectors/normals

Function tokenizer:

- TF-IDF over InterPro keyword annotations → 8 LSH hash tokens per residue (vocab size 255)
- Decoded post-hoc with 3-layer transformer inverting the LSH quantization

Appendix: GFP Generation Protocol

1. Prompt ESM3 7B with:
   - Sequence + structure + backbone coords of 6 chromophore residues
   - Structure of helix residues 58–71 from PDB 1QY3
   - Target length: 229 residues; all other positions masked

2. Chain-of-thought (Stage 1 — Structure):
   - Generate structure tokens (backbone)
   - Filter: good active site coordination + structure differentiated from 1QY3

3. Chain-of-thought (Stage 2 — Sequence):
   - Add generated structure to prompt → generate sequence

4. Iterative joint optimization:
   - Alternate: optimize sequence | fix structure
     optimize structure | fix sequence
   - Reject if active site atomic coordination lost

5. Rank pool of ~10,000 designs:
   - Bucket by sequence identity to known FPs
   - Rank by pLDDT, pTM, active site cRMSD, chromophore geometry

6. Synthesize top designs per bucket → express in E. coli → measure fluorescence

Backups

Back up on Protein ESM LLM

References Selected Papers on Protein LLM for Fold

- 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

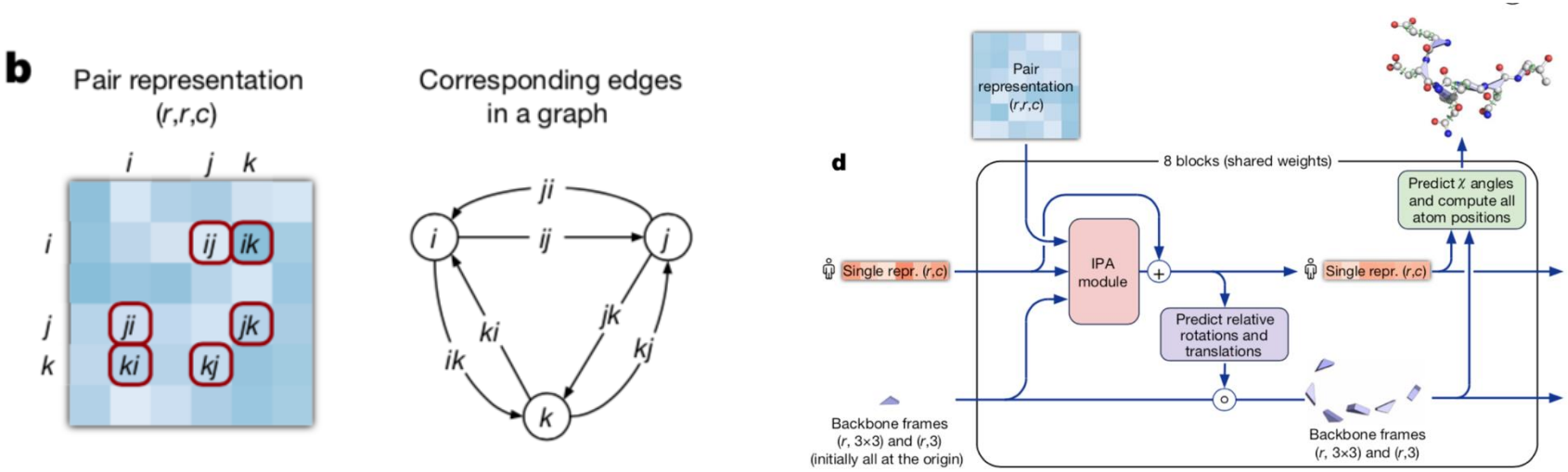
Highly accurate protein structure prediction with AlphaFold

•John Jumper, et al
•Demis Hassabis
Show authors*Nature* **volume 596**, pages583–589 (2021)Cite this article

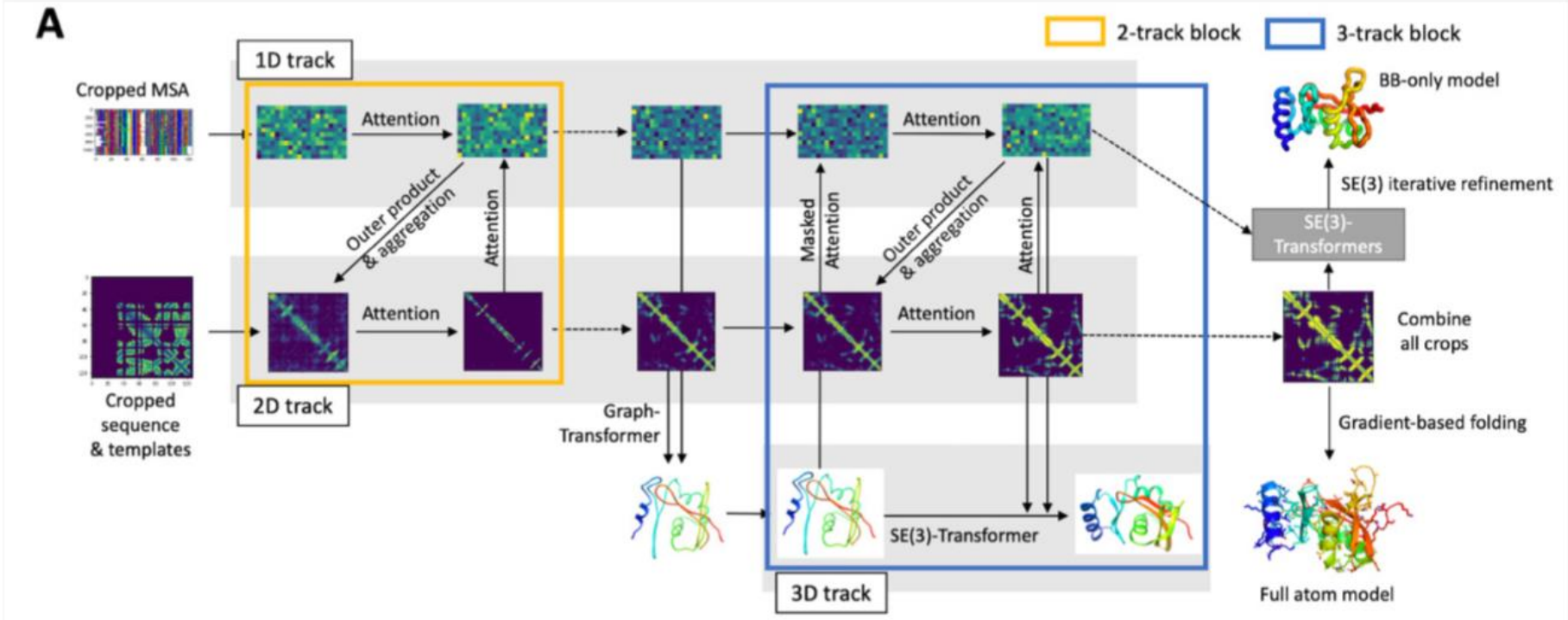
Abstract

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 determined[5], 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 years[9]. 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

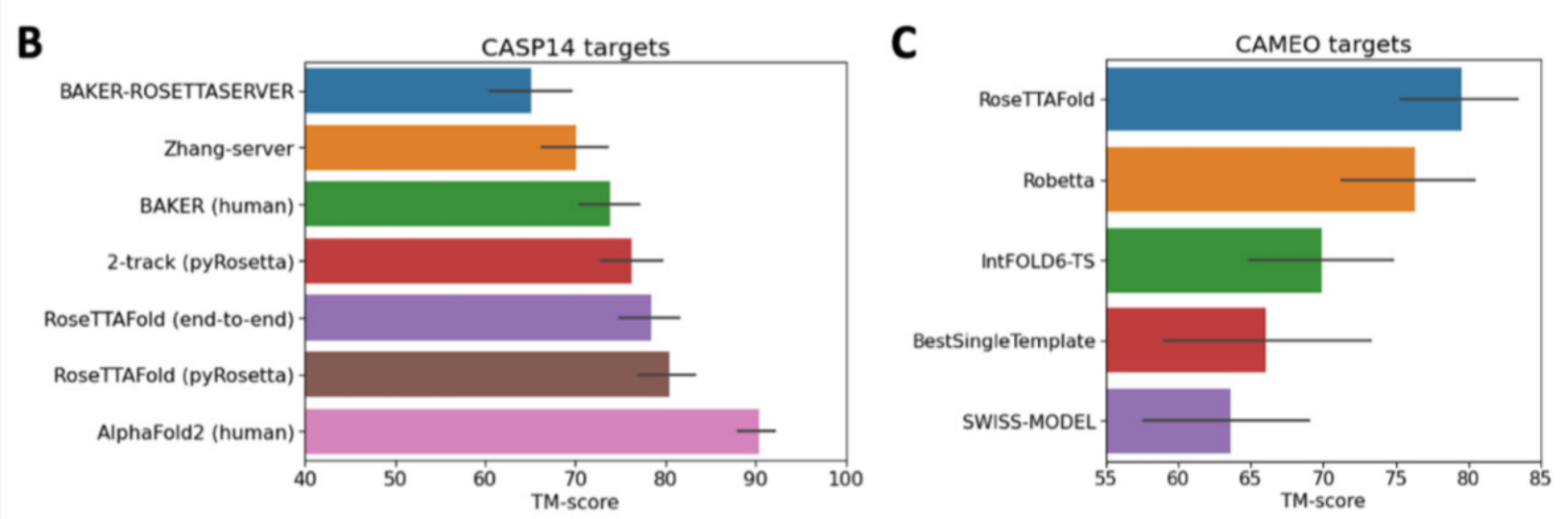
Structure module including Invariant point attention (IPA) module.



RoseTTAFold



RoseTTAFold



EVFold

PLoS One
. 2011;6(12):e28766.
doi: 10.1371/journal.pone.0028766. Epub 2011 Dec 7.

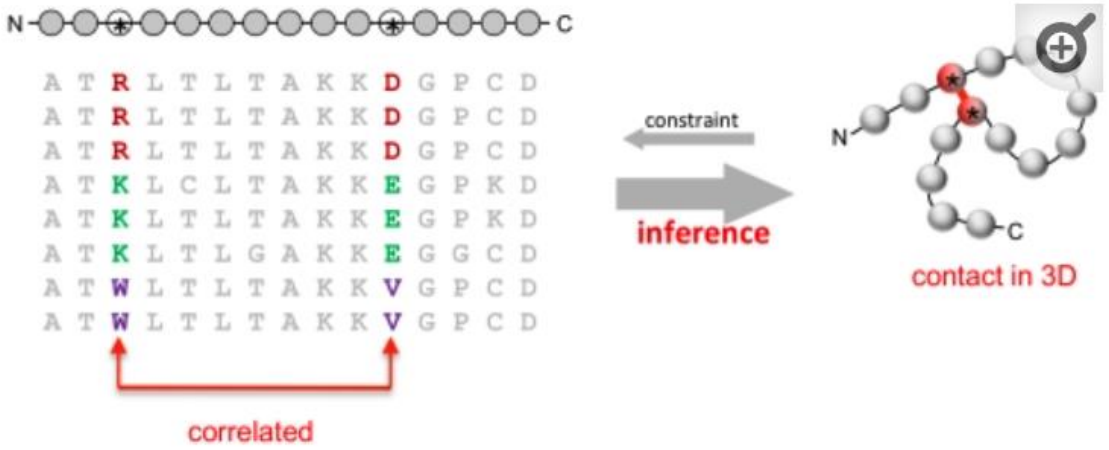
Protein 3D structure computed from evolutionary sequence variation

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, Tom Sercu https://orcid.org/0000-0003-2947-6064, **+7**, and Rob FergusAuthors

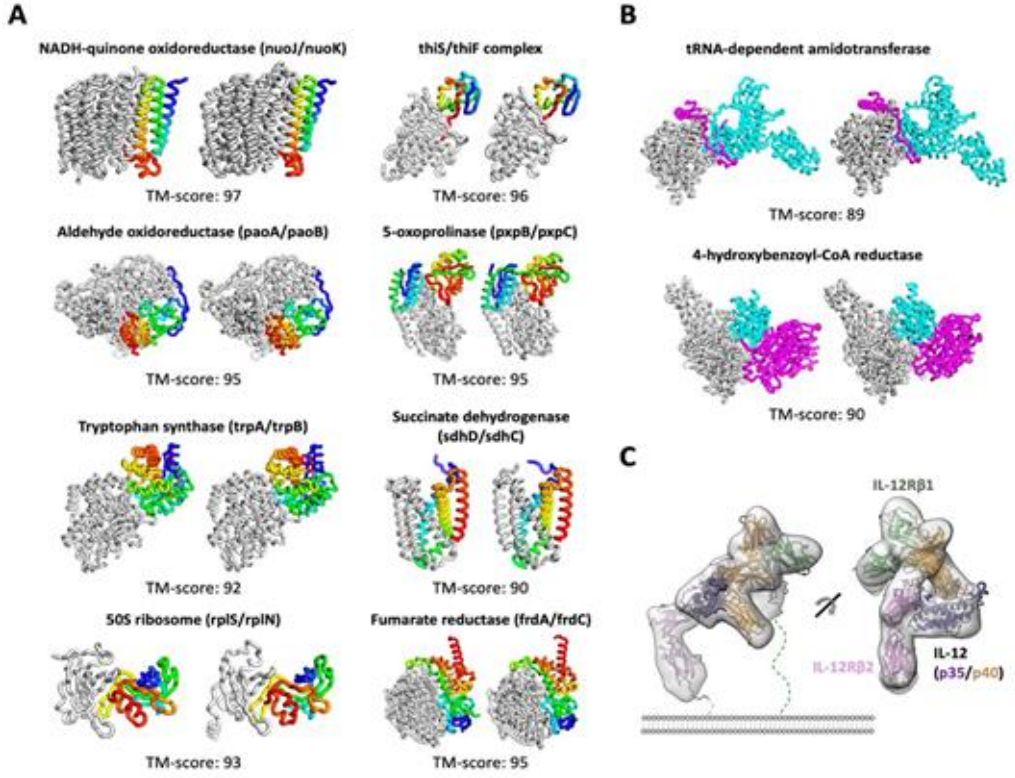
Edited by David T. Jones, University College London, London, United Kingdom, and accepted by Editorial Board Member William H. Press December 16, 2020 (received for review August 6, 2020)

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.

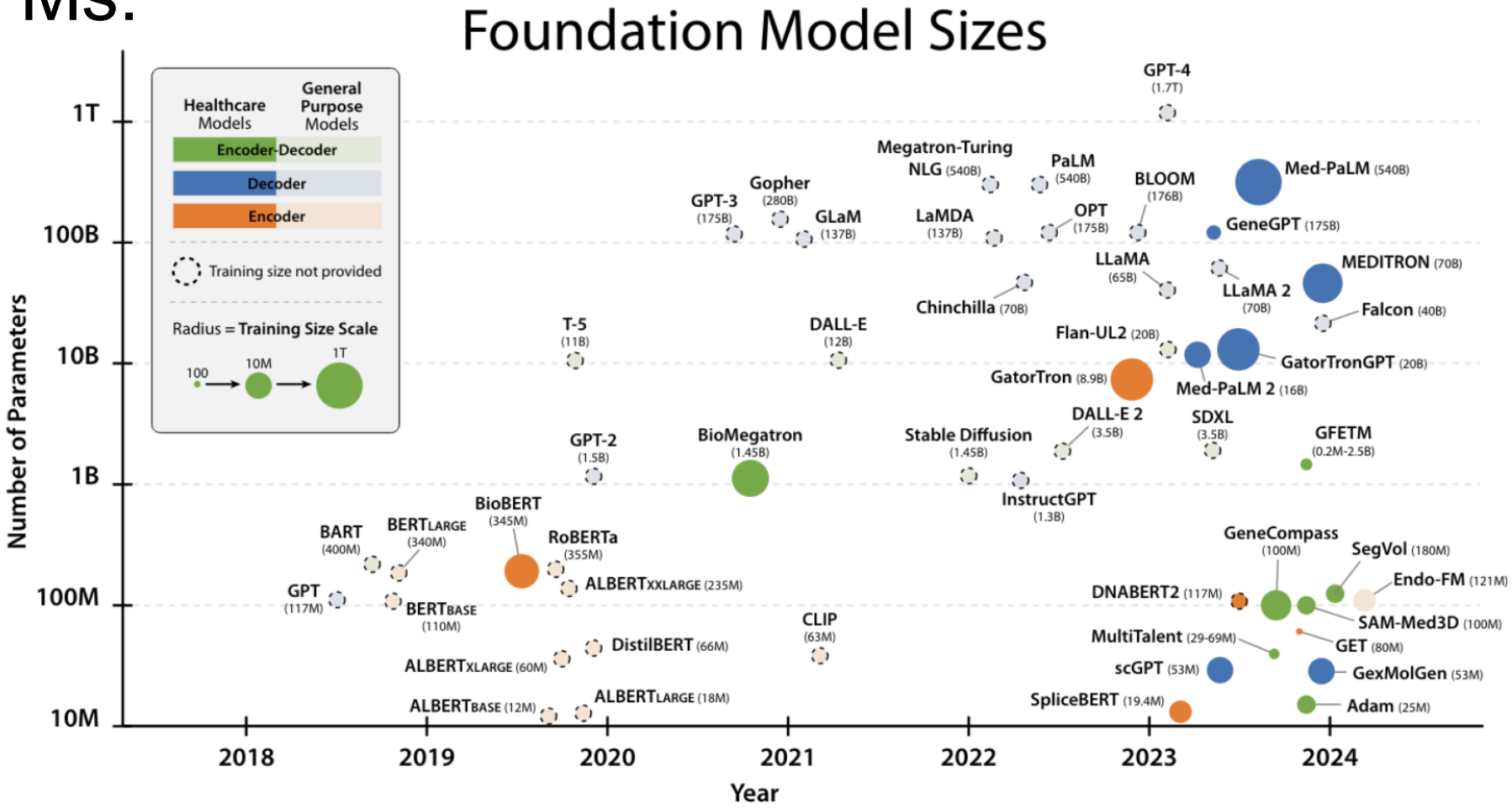
Also works in RoseTTAFold



Accurate prediction of protein structures and interactions using a three-track neural network
https://science.sciencemag.org/content/early/2021/07/19/science.abj8754

Back up on BioFM Survey

FMs:



A Comprehensive Survey of Foundation Models in Medicine, 2025

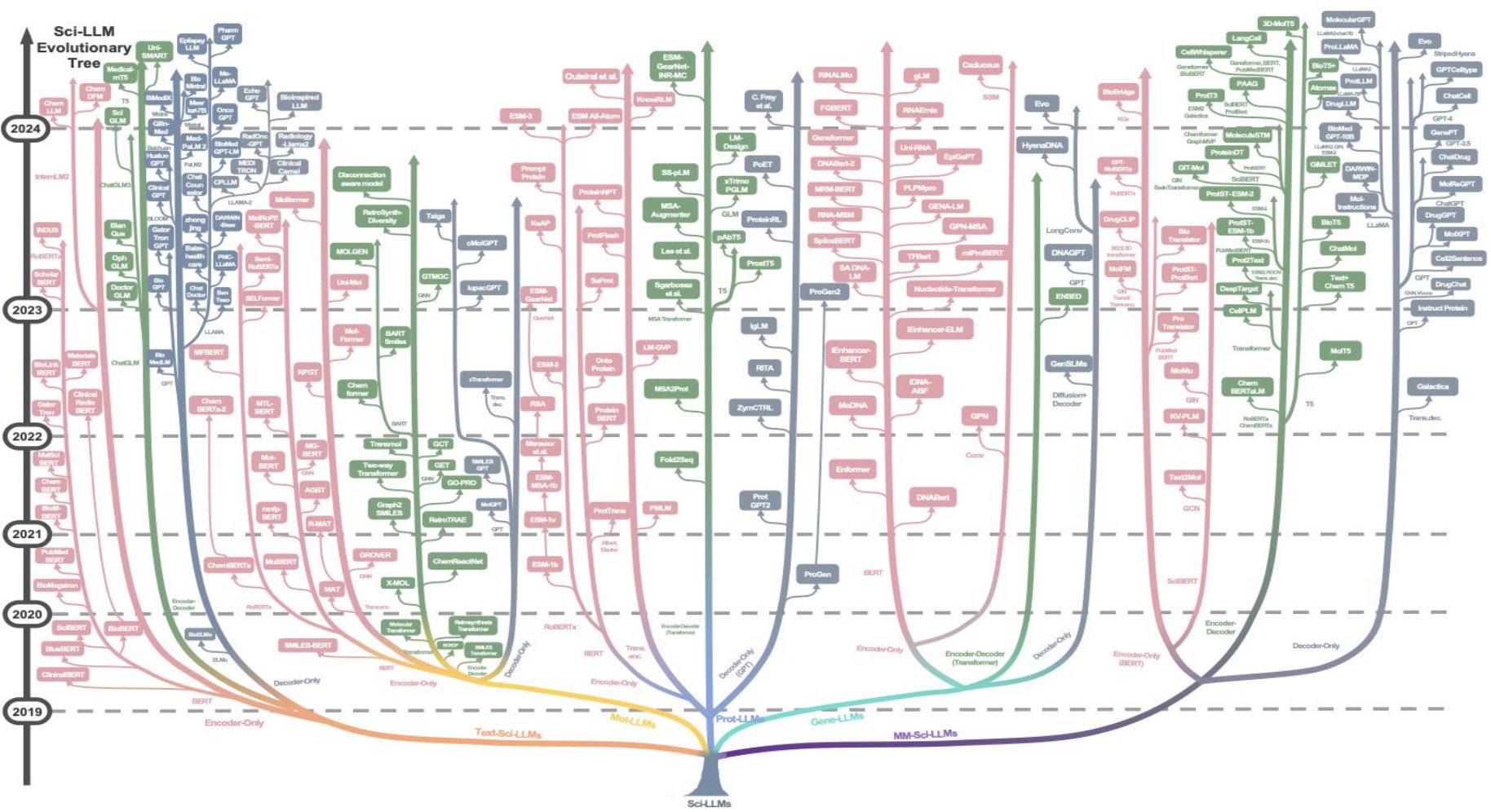


Fig. 3. An evolutionary tree of Sci-LLMs, which consists of five main branches corresponding to the research scopes in this survey. Due to the extensive number of Sci-LLMs, it is not feasible to include all of them in this figure, despite their exceptional quality. For detailed information on the featured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

Example HC data for healthcare LLM

- Healthcare training data
  - EHR
    - E.g., MIMIC III, MIMIC IV, CPRD
  - Scientific Literature
    - E.g., PubMed, PubMed Central
  - Web Data
    - E.g., COMETA (from Reddit), WebText

A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics, 2025

[Submitted on 10 Dec 2025]

Stanford Sleep Bench: Evaluating Polysomnography Pre-training Methods for Sleep Foundation Models

Magnus Ruud Kjaer, Rahul Thapa, Gauri Ganjoo, Hyatt Moore IV, Poul Joergen Jennum, Brandon M. Westover, James Zou, Emmanuel Mignot, Bryan He, Andreas Brink-Kjaer

Polysomnography (PSG), the gold standard test for sleep analysis, generates vast amounts of multimodal clinical data, presenting an opportunity to leverage self-supervised representation learning (SSRL) for pre-training foundation models to enhance sleep analysis. However, progress in sleep foundation models is hindered by two key limitations: (1) the lack of a shared dataset and benchmark with diverse tasks for training and evaluation, and (2) the absence of a systematic evaluation of SSRL approaches across sleep-related tasks. To address these gaps, we introduce Stanford Sleep Bench, a large-scale PSG dataset comprising 17,467 recordings totaling over 163,000 hours from a major sleep clinic, including 13 clinical disease prediction tasks alongside canonical sleep-related tasks such as sleep staging, apnea diagnosis, and age estimation. We systematically evaluate SSRL pre-training methods on Stanford Sleep Bench, assessing downstream performance across four tasks: sleep staging, apnea diagnosis, age estimation, and disease and mortality prediction. Our results show that multiple pretraining methods achieve comparable performance for sleep staging, apnea diagnosis, and age estimation. However, for mortality and

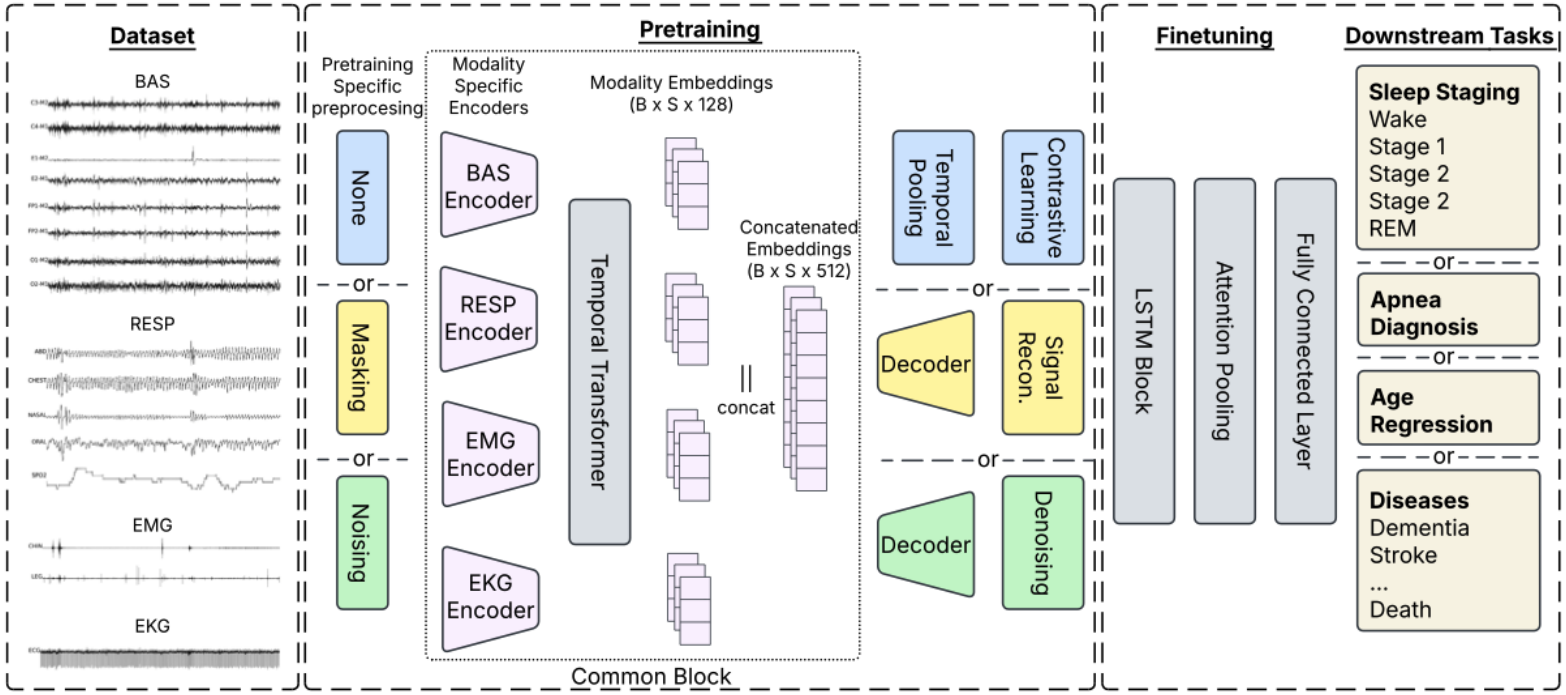


Fig. 1 Overview of Stanford Sleep Bench and the training pipeline for multiple SSRL methods. Stanford Sleep Bench includes four signal modalities: brain activity (BAS), respiration (RESP), electrocardiogram (EKG), and electromyography (EMG). Each modality is encoded independently using a CNN-based encoder, followed by a temporal transformer. SSRL methods—contrastive learning, signal reconstruction, and denoising—are trained separately, with no shared parameters. After pretraining, an LSTM-based prediction head with attention pooling is added for downstream tasks. Note that while all SSRL methods are shown, they are trained as separate models.

nature health

Perspective

https://doi.org/10.1038/s44360-025-00024-7

Large language models in global health

Received: 23 June 2025

Accepted: 11 November 2025

Published online: 15 January 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

Large language models (LLMs) are emerging as powerful tools in healthcare, with a growing role in global health, particularly in low- and middle-income countries (LMICs). This Perspective examines the current progress, challenges and prospects of LLMs in addressing health system disparities and supporting the achievement of the Sustainable Development Goals (SDGs). While high-income countries dominate the development and deployment of LLMs, LMICs face substantial barriers. These include limited digital infrastructure, a scarcity of locally relevant data, regulatory gaps, under-representation of local languages and dialects, and challenges related to privacy and data security. The limited availability of local expertise, capacity building programmes and sustained technical support remains a key barrier to scaling LLMs in LMICs. Nonetheless, case studies highlight how mobile-based LLM applications, hybrid artificial intelligence systems and open-weight models like DeepSeek are enhancing access to care

Clinical diagnosis	Patient education	Patient education
Malaria infection detection	Maternal and baby health	Diabetes management
Transformer-based model	**LLM-based chatbot**	**LLM-based system**
Smartphone app to automate detection of *Plasmodium* spp. in blood smear samples using a transformer model	SMS or WhatsApp-based chatbot to triage health enquiries to urgent and non-urgent, and generate responses in natural language	Integrated image-based deep learning model with LLM to augment primary care physicians (PCPs) in diabetes care
Better scalability	**Operational efficiency**	**Improved outcomes**
Greater scalability and accessibility over conventional vision models on microscopes	Reduction in urgent health enquiries that are missed by phone operators; improved productivity	Improved diabetes control and medication adherence; support untrained PCPs to manage better

Fig. 2 | Illustrative examples of LLM- or transformer-based applications deployed across different domains of global health. These case studies or pilot projects leverage existing telecommunications devices, such as smartphones

Table 1 | Examples of LLM applications and how they can potentia

	SDG 3: ens
Specific targets	**Supporting**
3.1. By 2030, reduce the global maternal mortality ratio to less than 70 per 100,000 live births	LLM-based a support[107]; c patient educ abnormal ca
3.2. By 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1,000 live births	LLMs to iden children pre neonatal car systems to c
3.3. By 2030, end the epidemics of AIDS, tuberculosis, malaria and neglected tropical diseases and combat hepatitis, waterborne diseases and other communicable diseases	LLMs in ansv during medi tuberculosis mitigate HIV
3.4. By 2030, reduce by one third premature mortality from noncommunicable diseases through prevention and treatment and promote mental health and well-being	LLM-based c intervention clinical note for patients
3.5. Strengthen the prevention and treatment of substance abuse, including narcotic drug abuse and harmful use of alcohol	LLMs to prec disorder fror alcohol use