Accurate Proteome-wide Missense Variant Effect Prediction with AlphaMissense

By
Nen Bakraniya (4373688)

TABLE OF CONTENTS

01 Introduction and Background

Overview of missense variants, challenges, and AlphaMissense objectives.

02 Methodology and AI Architecture

Details of the AI model, data pipeline, and training methods.

03 Results and Performance Analysis

Evaluation of AlphaMissense against existing tools and its real-world impact.

04 Limitations, Implications, and Summary

Model limitations, future scope, and takeaways from the study.

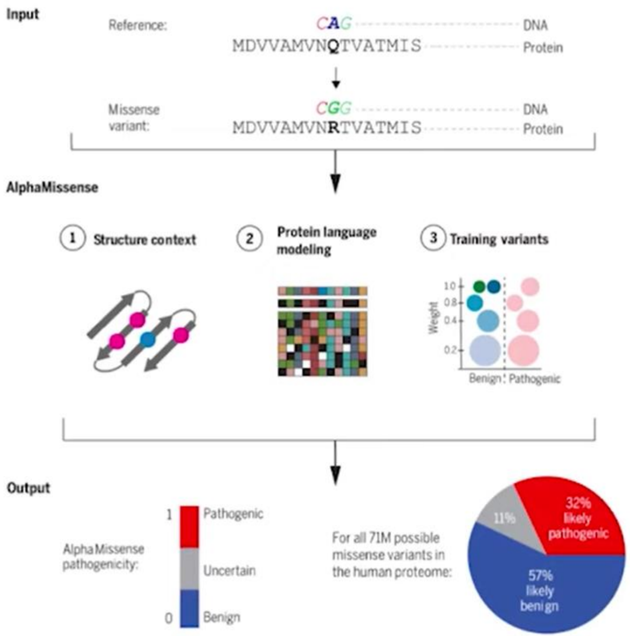
01 Introduction and Background

Overview of missense variants, challenges, and AlphaMissense objectives.

INTRODUCTION

"AlphaMissense: A groundbreaking AI model for accurately predicting the pathogenic effects of genetic missense variants across the entire human proteome, enabling advancements in precision medicine.“

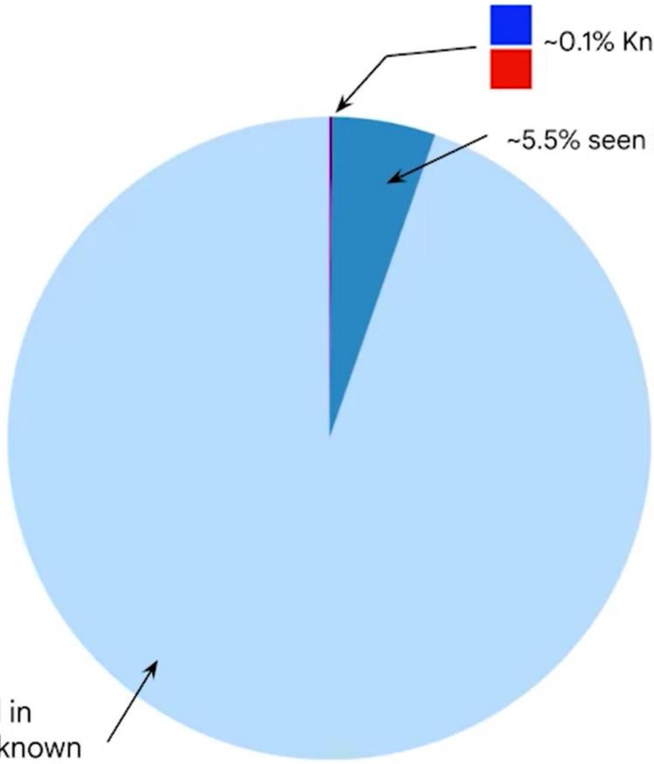
AlphaMissense



- What is the task (and why?)?
- What did we build?
- How do we know it does well at this task?
- Why does it do well?
- What is the output useful for?

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, Žiga Avsec; Science, Sept. 23

Why? A major gap in our knowledge of variant effects



🟦🟥 ~0.1% Known pathogenic or benign (ClinVar)

~5.5% seen in humans but unknown significance

Unobserved in humans (unknown significance)

What is a Missense Variant?

- The **human genome** contains the instructions for building and maintaining the human body, encoded in the form of **DNA**.

- A **missense variant** is a specific type of mutation where a single DNA base is altered, leading to the substitution of one amino acid for another in a protein.

- **Example**:
  - DNA Sequence Before Mutation: GGA ⟶ Codes for Glycine
  - DNA Sequence After Mutation: GTA ⟶ Codes for Valine
  - Result: The protein structure and function may change, which can potentially cause diseases.

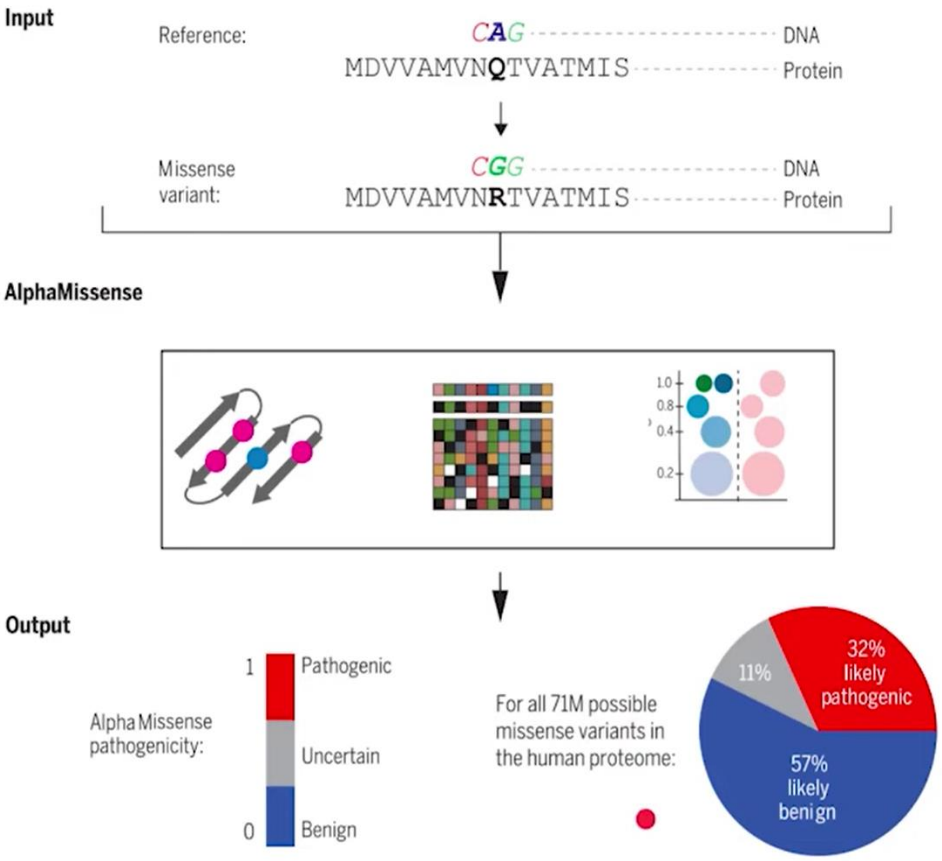
- Missense variants can:
  - Be **benign** (no harm caused).
  - Be **pathogenic** (leading to diseases like cystic fibrosis or sickle cell anemia).

02

Methodology and AI Architecture

Details of the AI model, data pipeline, and training methods.

The task



Input

Reference: CAG DNA
MDVVAMVNQTVATMIS Protein

Missense variant: CGG DNA
MDVVAMVNRTVATMIS Protein

AlphaMissense

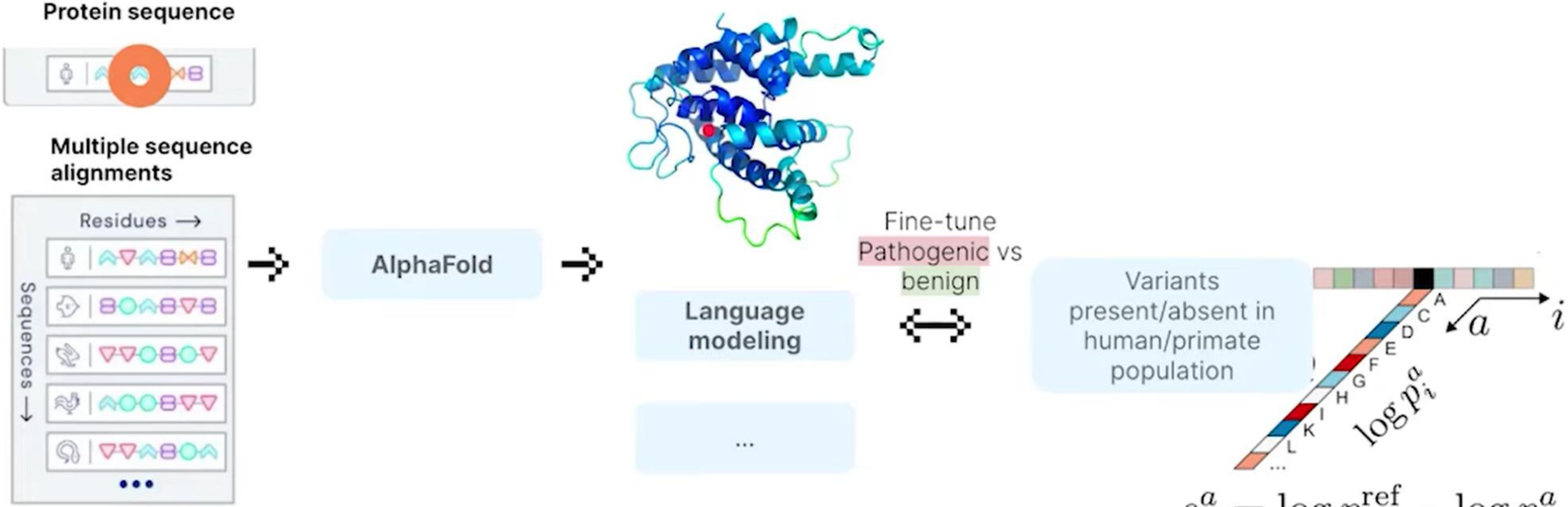
Output

AlphaMissense pathogenicity:
1 Pathogenic
Uncertain
0 Benign

For all 71M possible missense variants in the human proteome:

32% likely pathogenic
11%
57% likely benign

AlphaMissense: Fine-tuning AlphaFold to predict variant pathogenicity



Protein sequence

Multiple sequence alignments

Residues →

Sequences →

AlphaFold

Language modeling

...

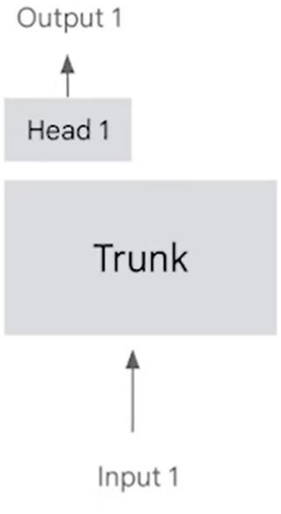
Fine-tune Pathogenic vs benign

Variants present/absent in human/primate population

$$s_i^a = \log p_i^{\mathrm{ref}} - \log p_i^a$$

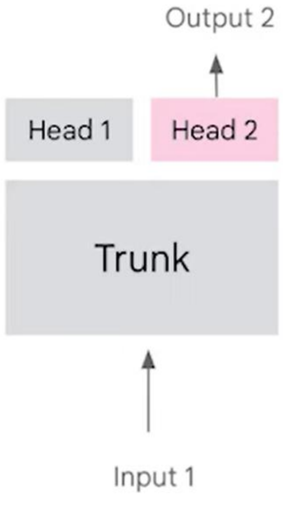
Many options for fine-tuning models on new tasks

New dataset, same head



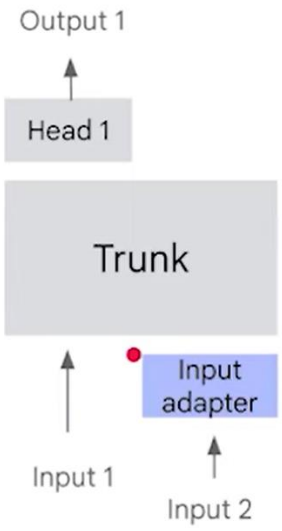
Example: Language model fine-tuning

New dataset, new head



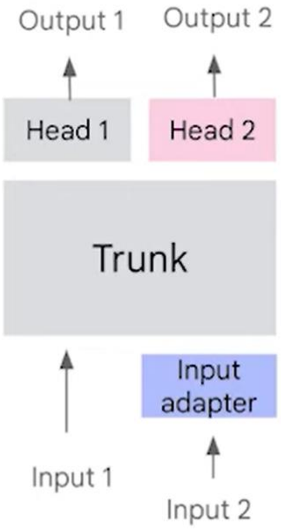
Example: Image segmentation models pre-trained on general images

New dataset, additional input



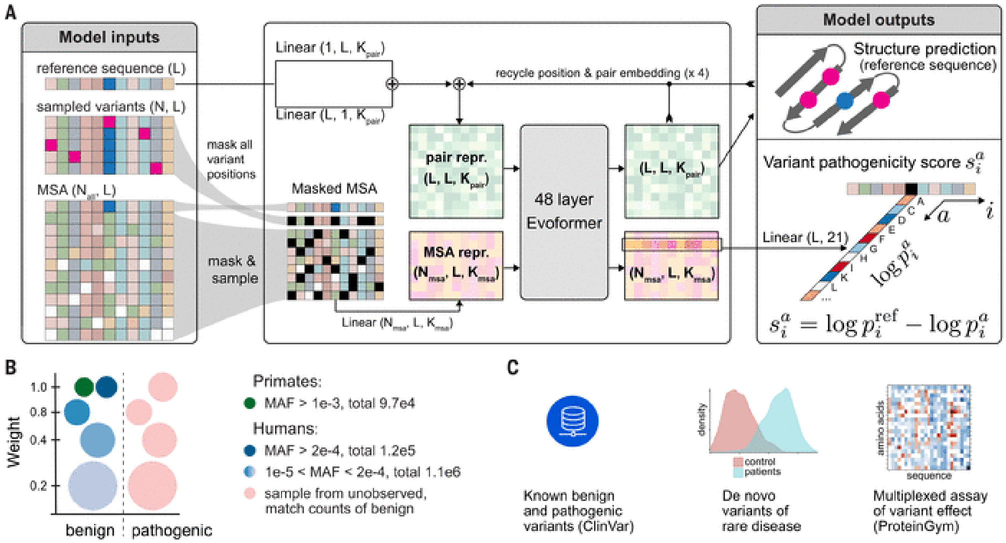
Example: Image-to-text (Flamingo)

Both, keep training on the original objective



Example: AlphaMissense

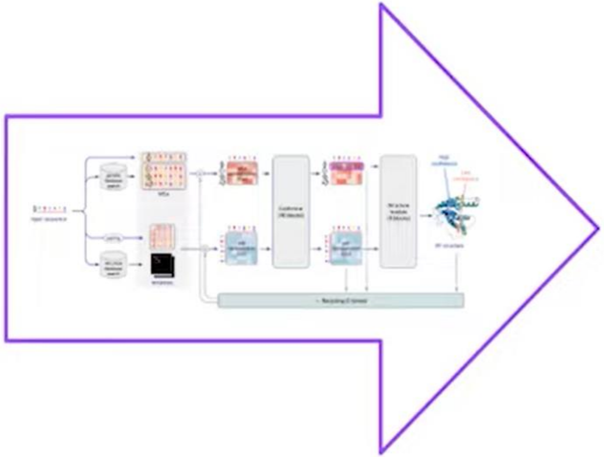
Science | AAAS

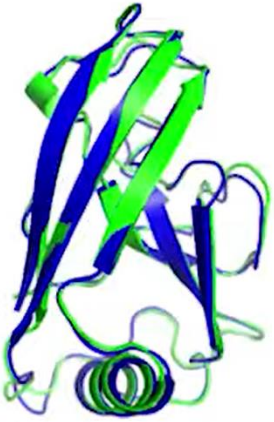

A
Model inputs
reference sequence (L)
sampled variants (N, L)
MSA (N_{all}, L)
mask all variant positions
mask & sample
Linear (1, L, K_{pair})
Linear (L, 1, K_{pair})
Masked MSA
recycle position & pair embedding (x 4)
pair repr. (L, L, K_{pair})
MSA repr. (N_{msa}, L, K_{msa})
48 layer Evoformer
(L, L, K_{pair})
(N_{msa}, L, K_{msa})
Linear (N_{msa}, L, K_{msa})
Model outputs
Structure prediction (reference sequence)
Variant pathogenicity score s_i^a
Linear (L, 21)
A C D E F G H I K L ...
a
i
$\log p_i^a$
$s_i^a = \log p_i^{\mathrm{ref}} - \log p_i^a$
B
Weight
1.0
0.8
0.4
0.2
benign
pathogenic
Primates:
● MAF > 1e-3, total 9.7e4
Humans:
● MAF > 2e-4, total 1.2e5
● 1e-5 < MAF < 2e-4, total 1.1e6
● sample from unobserved, match counts of benign
C
Known benign and pathogenic variants (ClinVar)
density
control
patients
De novo variants of rare disease
amino acids
sequence
Multiplexed assay of variant effect (ProteinGym)

Accurate proteome-wide missense variant effect prediction with AlphaMissense, Volume: 381, Issue: 6664, DOI: (10.1126/science.adg7492)

AlphaFold

SIFSYITESTGTPSNATYT
YVIERWDPETSGILNPCYG
WPVCYVVTVNHKHTVNGTGG
NPAFQIARIEKLRTLAEVR
DVVLKNRSFPIEGQTTHRG
PSLNSNQECVGLFYQPNSS
GISPRGKLLPGSLCGIAPP
PVHHHHHH



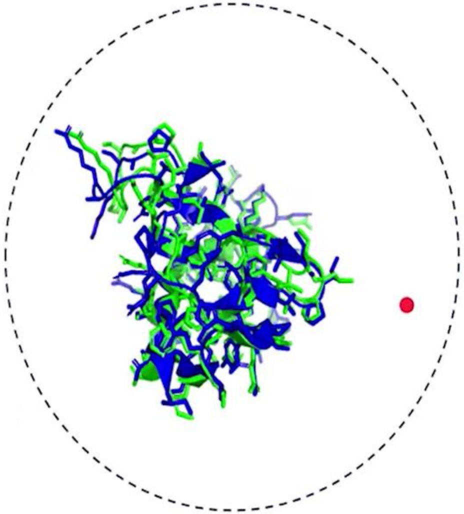


T1049 / 6y4f
93.5 GDT
(adhesin tip)

AlphaFold blind predictions

Ground truth

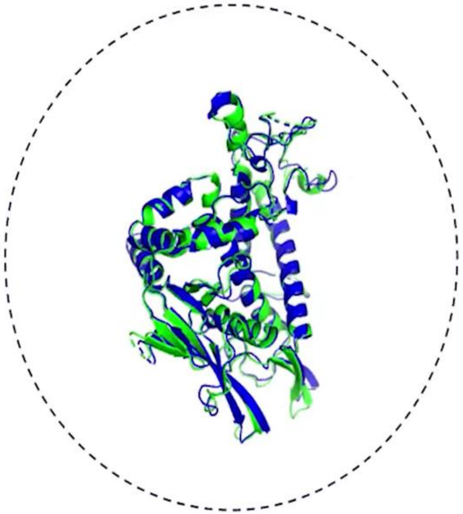
Prediction



T1064 / 7jtl

87.0 GDT

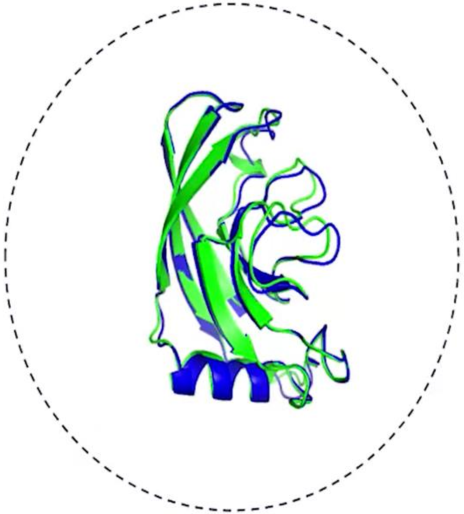
(ORF8 from SARS-CoV-2)



T1037 / 6vr4

90.7 GDT

(RNA polymerase domain)

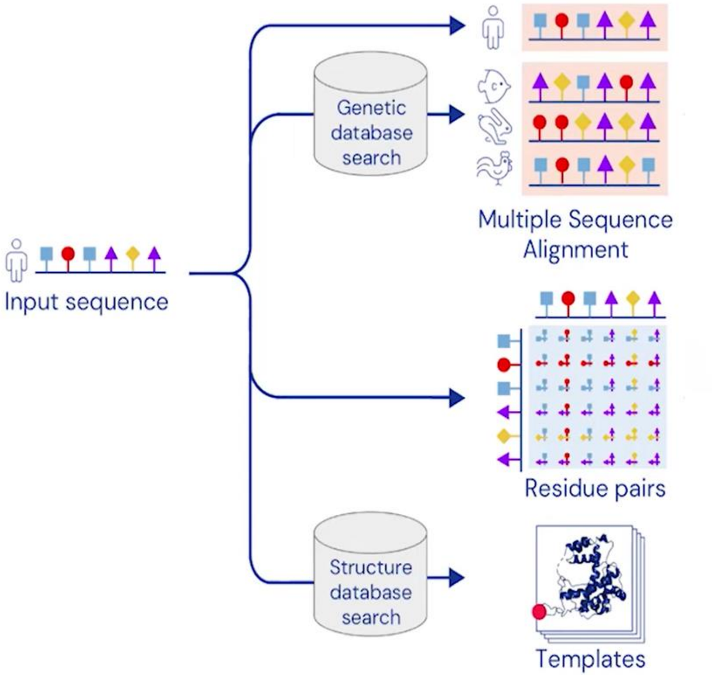


T1049 / 6y4f

93.3 GDT

(adhesin tip)

AlphaFold overview - inputs



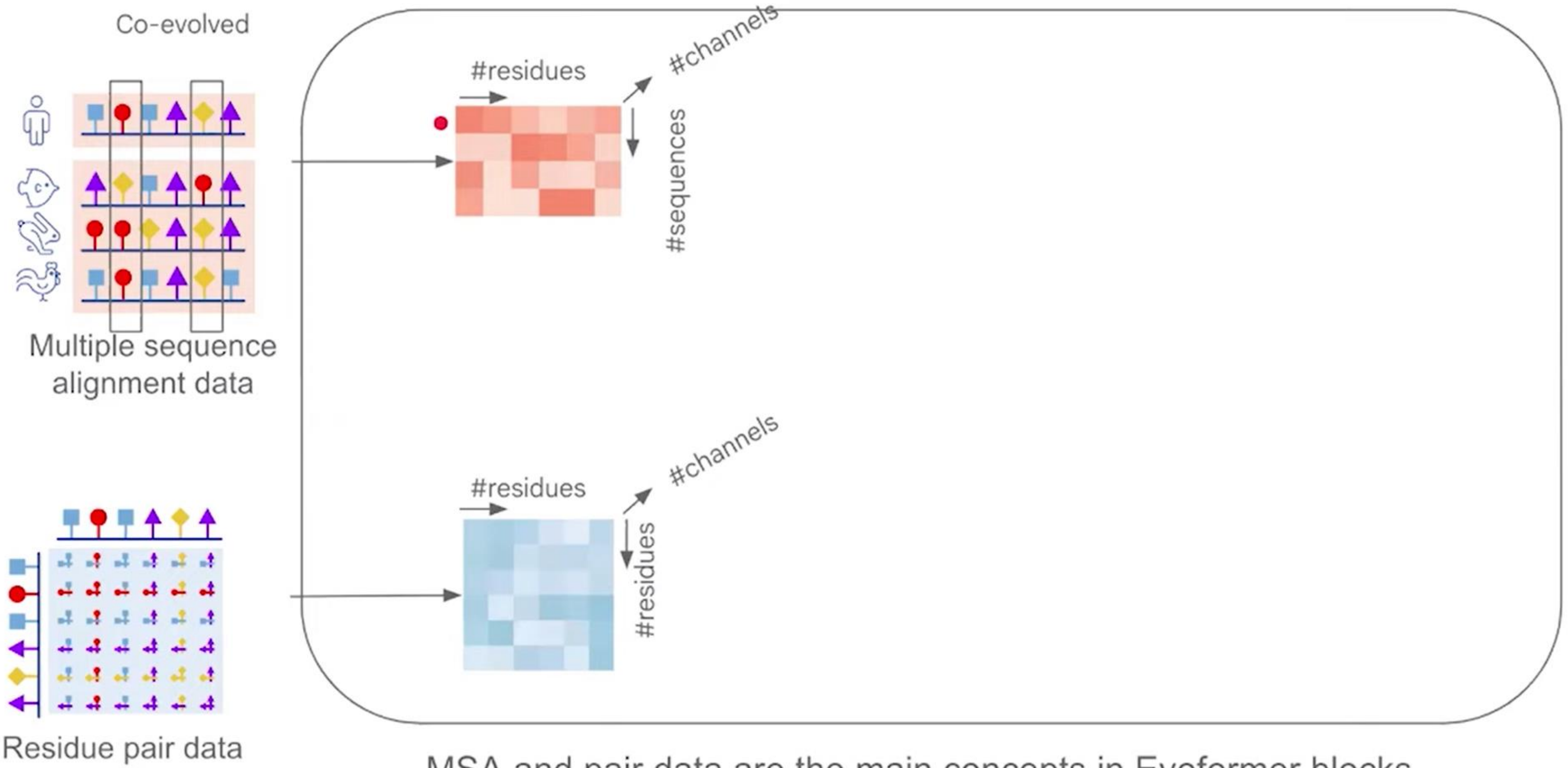
- Amino acid sequence
- MSA of related sequences
- Template structures from PDB

Good predictions possible without a template, based on MSA or vice versa

Model isn’t forced to be similar to template

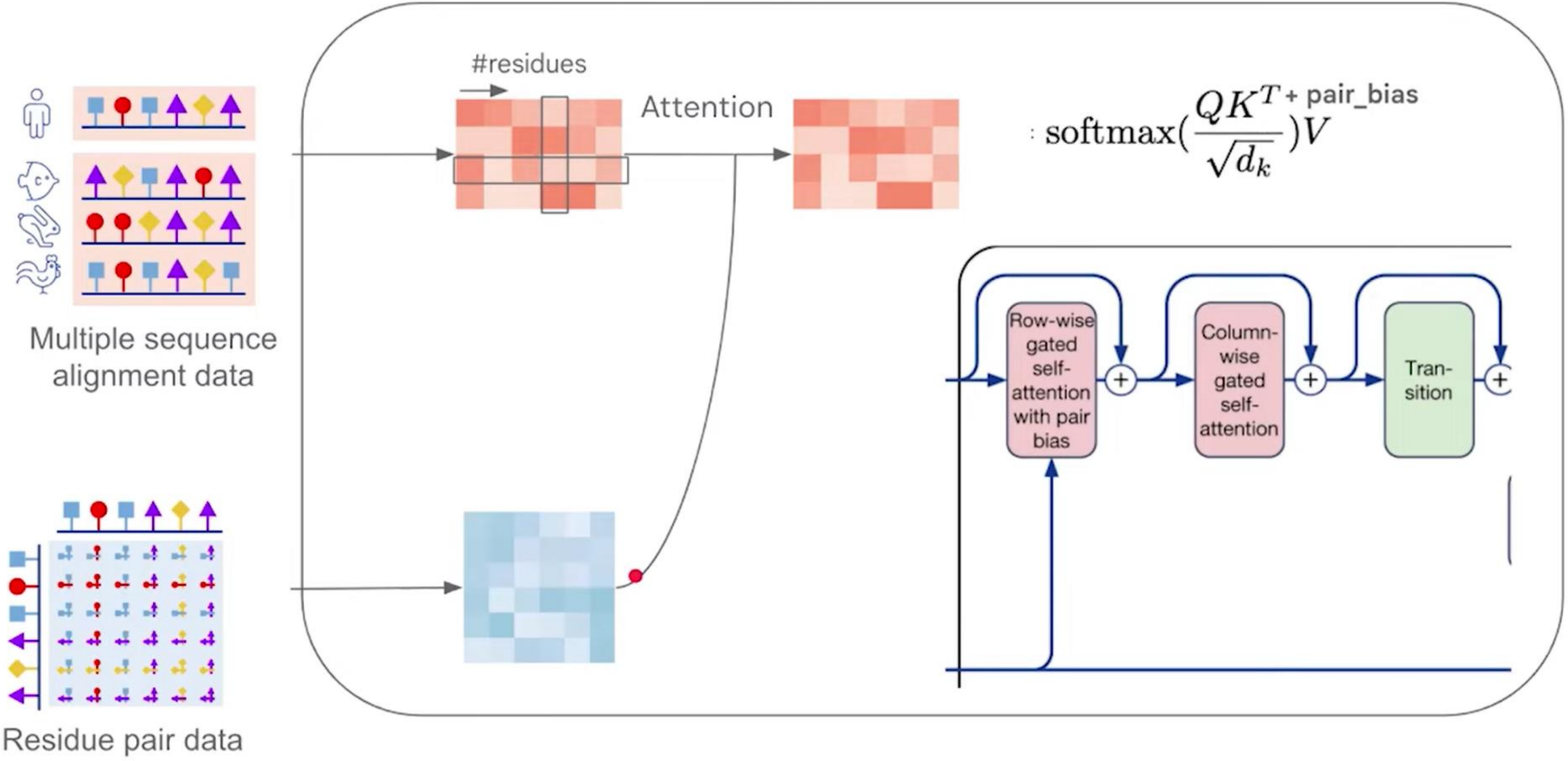
Generally good predictions on designed proteins without MSA and Templates

Evoformer



MSA and pair data are the main concepts in Evoformer blocks

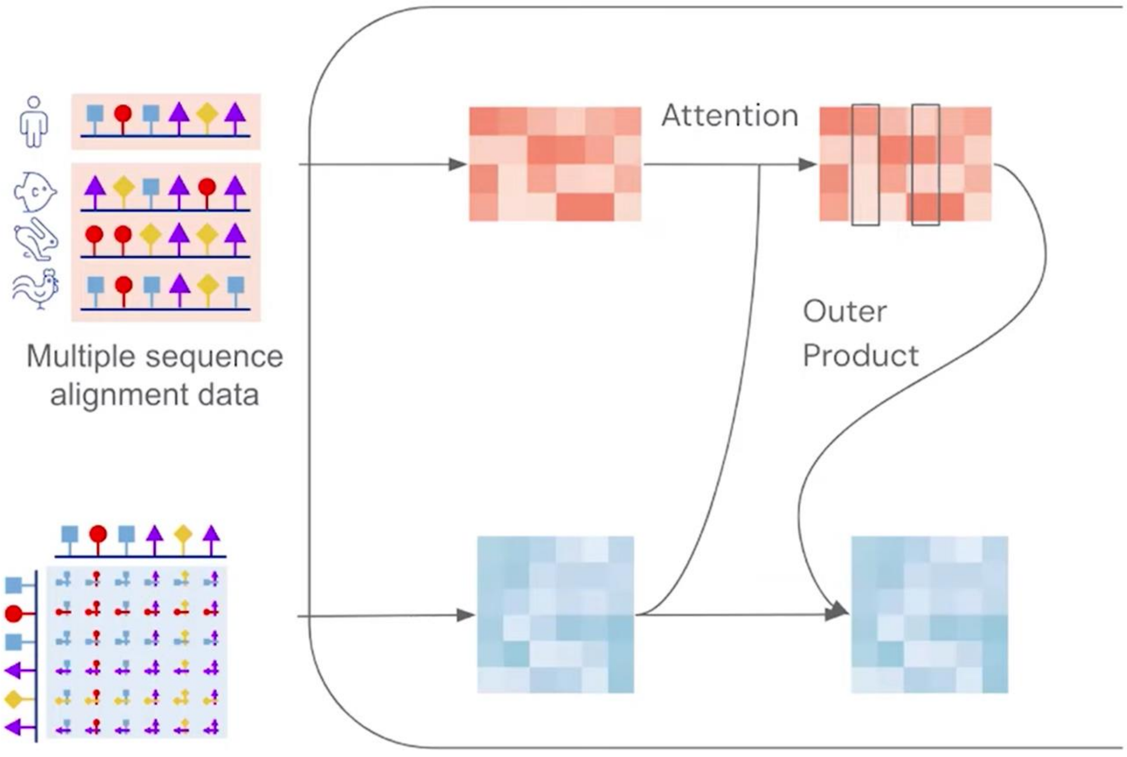
Evoformer


#residues
Attention
Multiple sequence alignment data
Residue pair data
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Transition
+

$$\mathrm{softmax}\left(\frac{QK^T + \text{pair\_bias}}{\sqrt{d_k}}\right)V$$

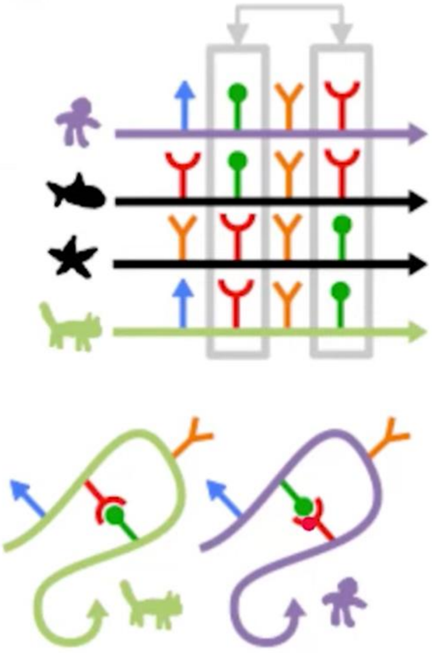
Attention is augmented by the network's belief about residue pairs

Evoformer



Multiple sequence alignment data

Residue pair data

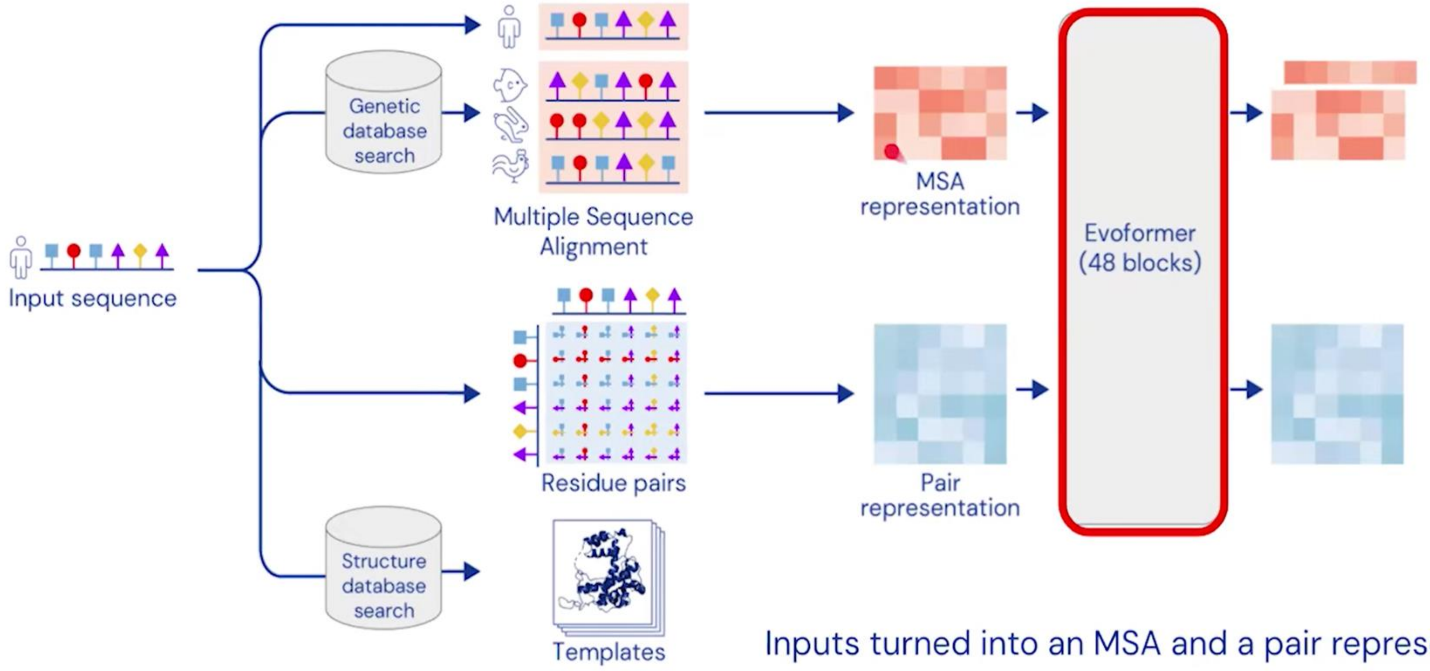


Co-evolution: residues in contact must mutate together.

Coevolution cartoon by Sergey Ovchinnikov (https://jgi.doe.gov/seeking-structure-metagenome-sequences/cartoon-coevolution-sergey-o/)

Outer product allows generalized correlation similar to co-evolution

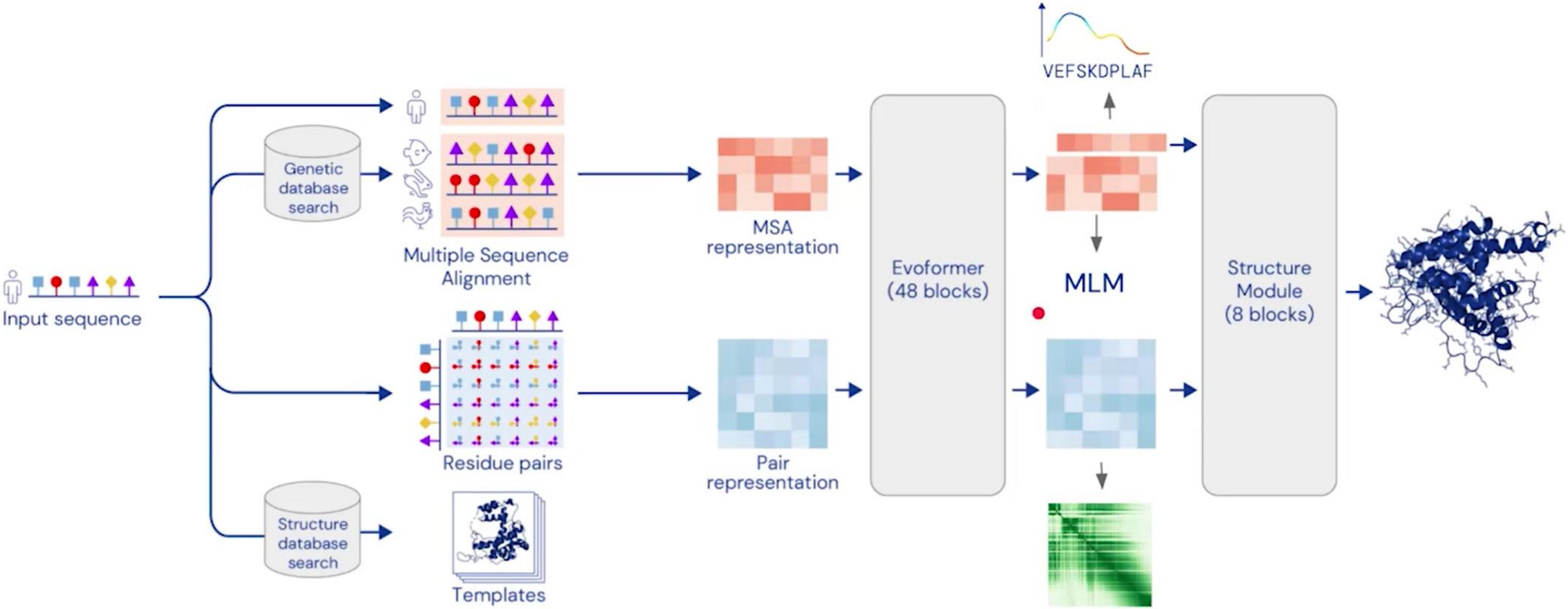
AlphaFold overview - processing



Inputs turned into an MSA and a pair representation

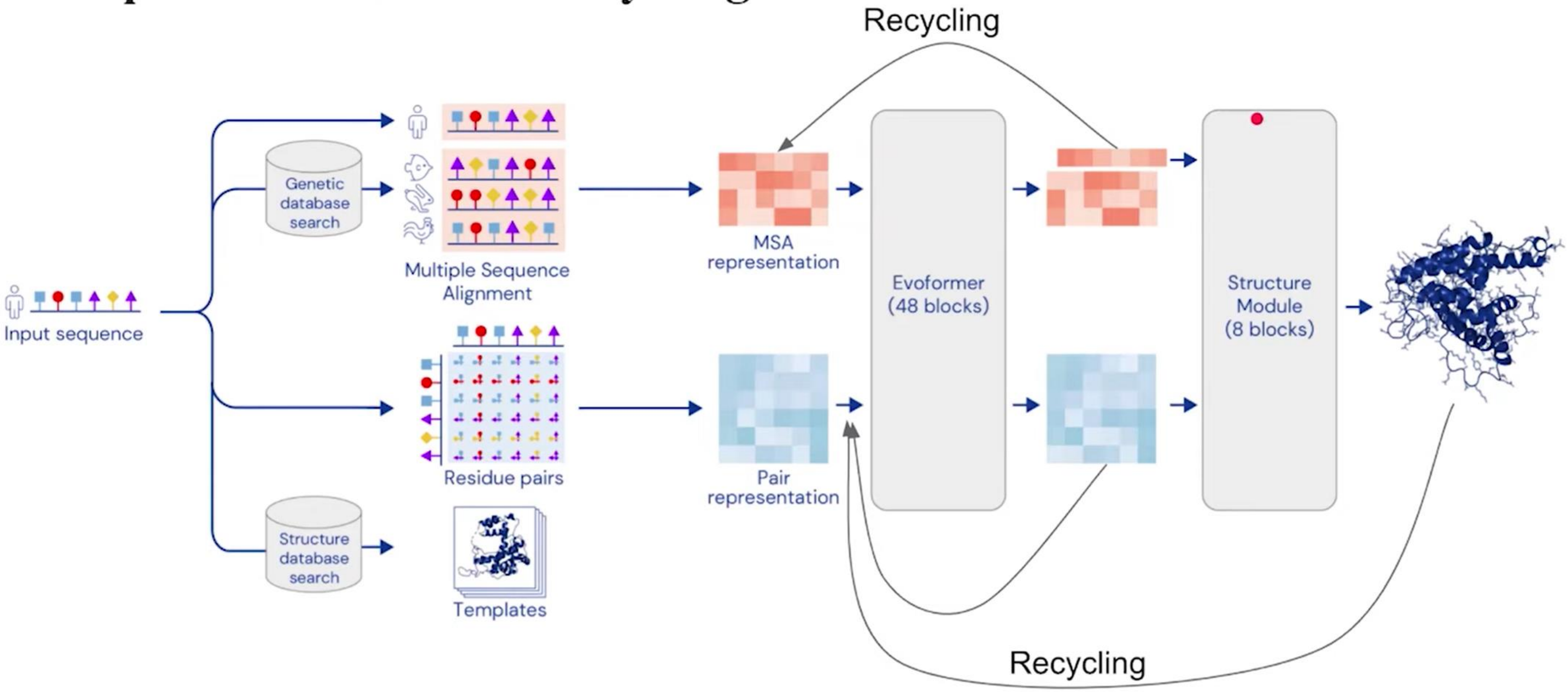
Evoformer: repeatedly updates these to build up information about the relationship between residues

AlphaFold overview - outputs



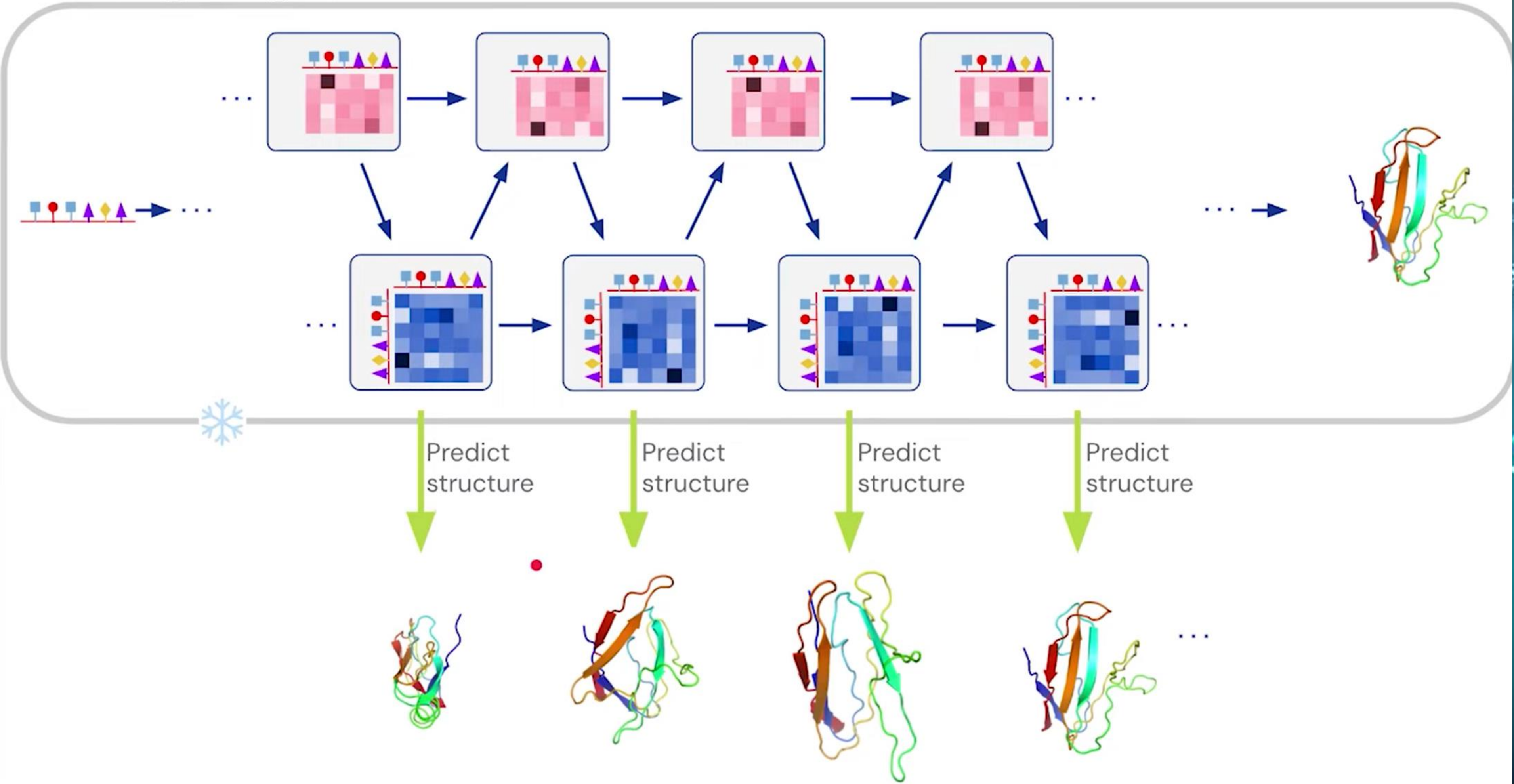
Structure Module: predicts a rotation and translation to place each residue. Small networks predict side chain orientations, and confidence metrics

AlphaFold overview - recycling



Recycling: Run Model multiple times feeding the previous output back in

Interrogating the Network



Importance of Predicting Variant Effects

- Proteins are **molecular machines** essential for every biological process. A single amino acid substitution can disrupt a protein's function, leading to diseases.

- Predicting the effects of missense variants helps:
  - **Diagnose genetic disorders**: For instance, understanding a variant's pathogenicity can identify a disease-causing mutation in a patient.
  - **Personalized Medicine**: Enables tailored treatments based on an individual's genetic makeup.
  - **Drug Development**: Identifies targets for new therapies.

- **Example in Medicine**:
  - In cancer, certain missense variants in genes like **BRCA1** increase cancer risk. Predicting these variants' effects allows early intervention and targeted treatments.

Key components

MSA Embeddings

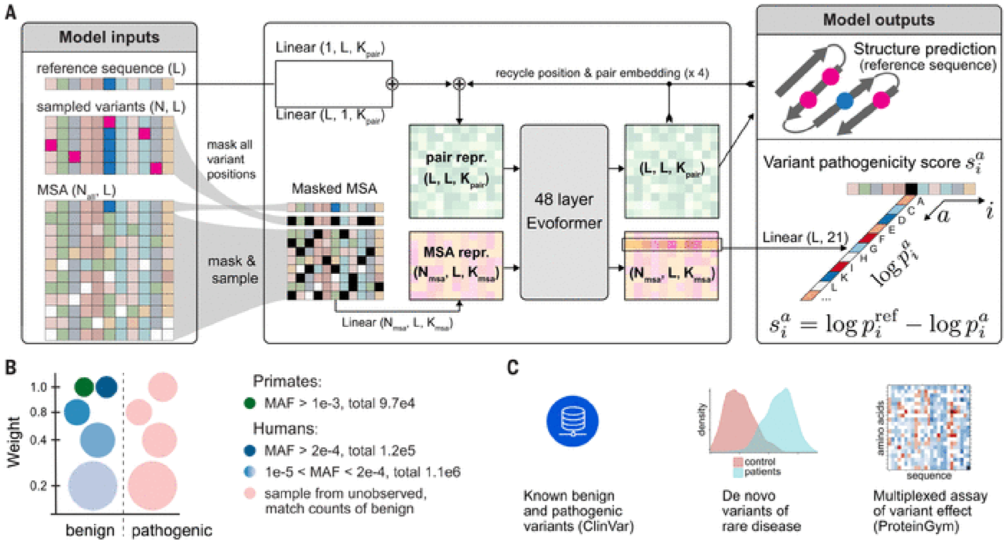
Multi-Sequence Alignments (MSA) represent evolutionary relationships among proteins.

Recycling Process

Iteratively updates embeddings to improve predictions.

Variant Masking

Specific masking techniques ensure the model focuses on variant pathogenicity prediction.


A
Model inputs
reference sequence (L)
sampled variants (N, L)
MSA (N_all, L)
mask all variant positions
mask & sample
Linear (1, L, K_pair)
Linear (L, 1, K_pair)
Masked MSA
recycle position & pair embedding (x 4)
pair repr. (L, L, K_pair)
MSA repr. (N_msa, L, K_msa)
48 layer Evoformer
(L, L, K_pair)
(N_msa, L, K_msa)
Linear (N_msa, L, K_msa)
Model outputs
Structure prediction (reference sequence)
Variant pathogenicity score s_i^a
Linear (L, 21)
a
i
$\log p_i^a$
A C D E F G H I K L ...
$s_i^a = \log p_i^{\text{ref}} - \log p_i^a$
B
Weight
1.0
0.8
0.4
0.2
benign
pathogenic
Primates:
MAF > 1e-3, total 9.7e4
Humans:
MAF > 2e-4, total 1.2e5
1e-5 < MAF < 2e-4, total 1.1e6
sample from unobserved, match counts of benign
C
Known benign and pathogenic variants (ClinVar)
density
control
patients
De novo variants of rare disease
amino acids
sequence
Multiplexed assay of variant effect (ProteinGym)

Accurate proteome-wide missense variant effect prediction with AlphaMissense, Volume: 381, Issue: 6664, DOI: (10.1126/science.adg7492)

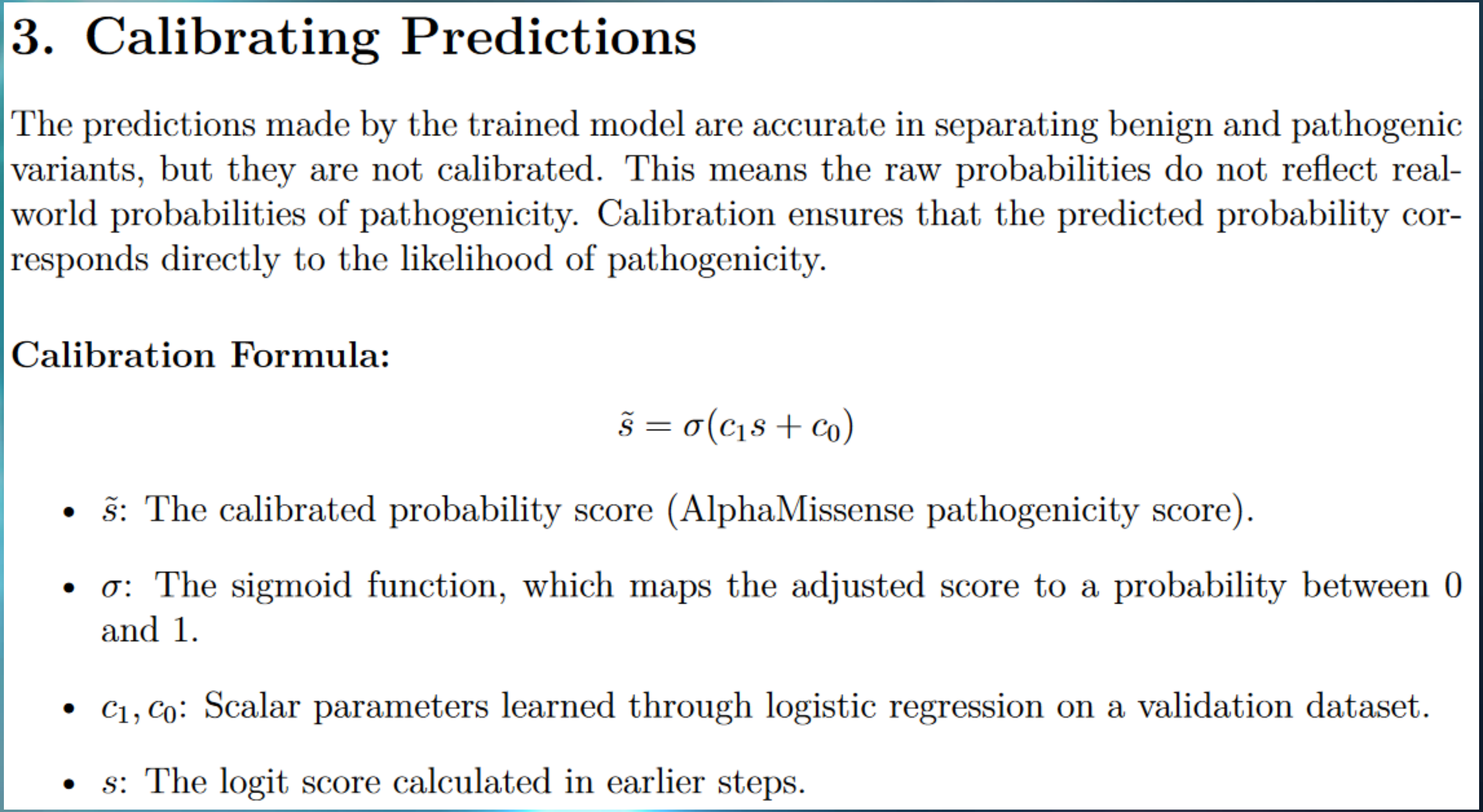
Logit Score Calculation: Understanding Variant Impact

The first step in variant pathogenicity prediction is calculating the logit score (s_i^a) for each variant. This score quantifies how much the substitution of one amino acid (variant a) at a specific position i in a protein sequence differs from the original reference amino acid.

$$s_i^a = \log p_i^{\text{ref}} - \log p_i^a$$

What this means:

- p_i^{ref}: The probability of the reference amino acid being present at position i.
- p_i^a: The probability of the alternative (variant) amino acid being present at position i.
- The formula computes the logarithmic difference between these probabilities.



3. Calibrating Predictions

The predictions made by the trained model are accurate in separating benign and pathogenic variants, but they are not calibrated. This means the raw probabilities do not reflect real-world probabilities of pathogenicity. Calibration ensures that the predicted probability corresponds directly to the likelihood of pathogenicity.

Calibration Formula:

$$\tilde{s} = \sigma(c_1 s + c_0)$$

- $\tilde{s}$: The calibrated probability score (AlphaMissense pathogenicity score).
- σ: The sigmoid function, which maps the adjusted score to a probability between 0 and 1.
- c_1, c_0: Scalar parameters learned through logistic regression on a validation dataset.
- s: The logit score calculated in earlier steps.

Summary of Methodology

AlphaMissense achieves accuracy and scalability by leveraging spatially cropped sequences, evolutionary insights through multiple sequence alignments, and advanced fine-tuning techniques to predict the pathogenicity of missense variants across the entire human proteome with high reliability.

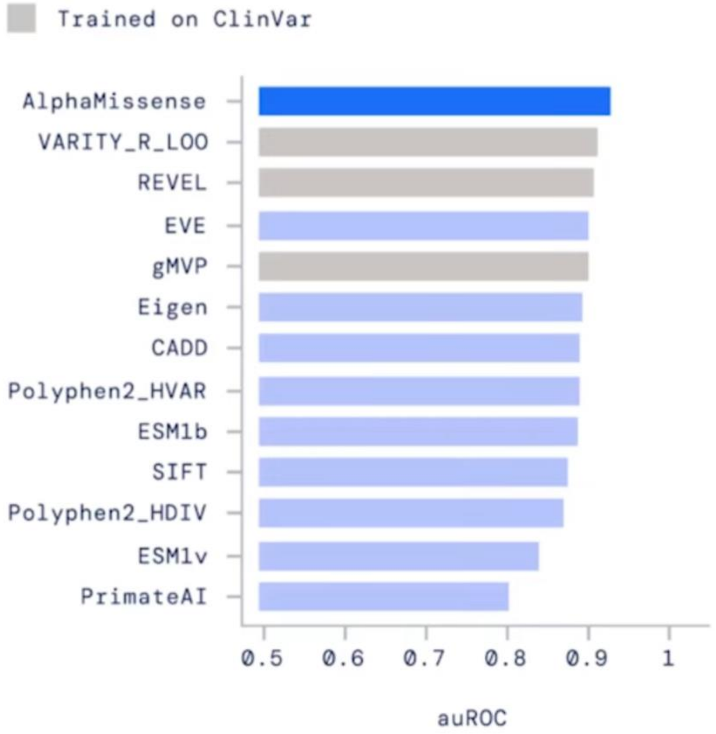
03

Results and Performance Analysis

Evaluation of AlphaMissense against existing tools and its real-world impact.

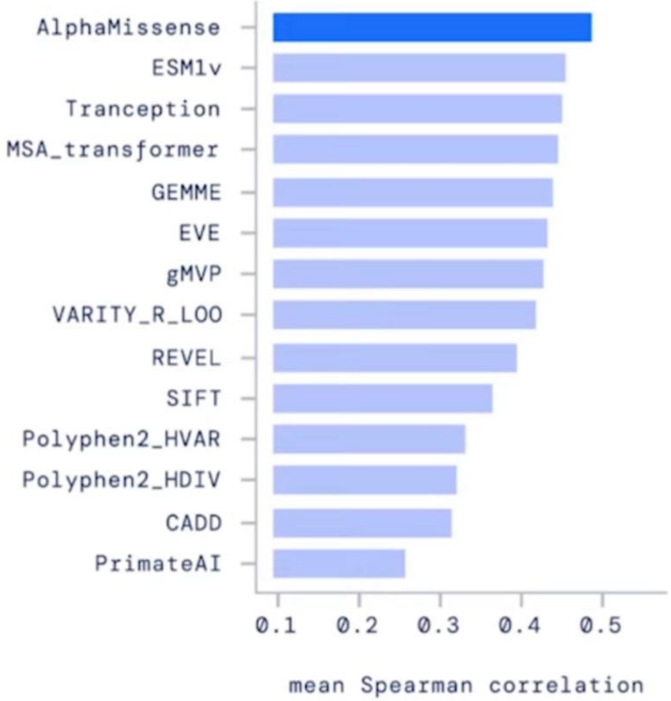
Most accurate predictions across a diverse set of benchmarks

ClinVar (Class-balanced 18924 variants)



Experimental assays (25 proteins)

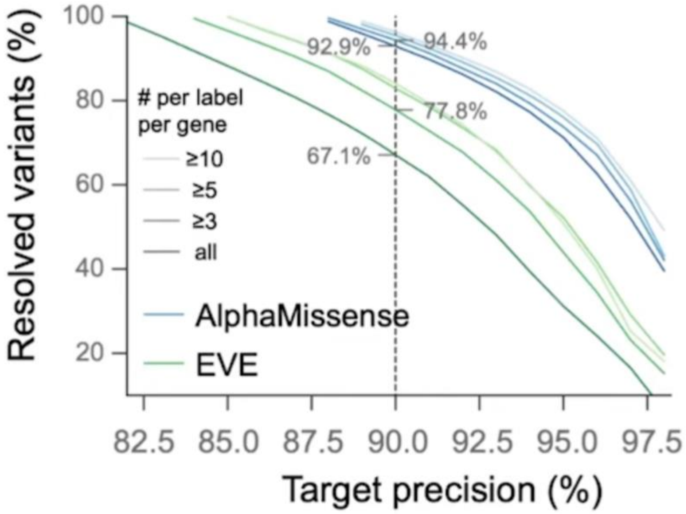
DMS / MAVE, human proteins



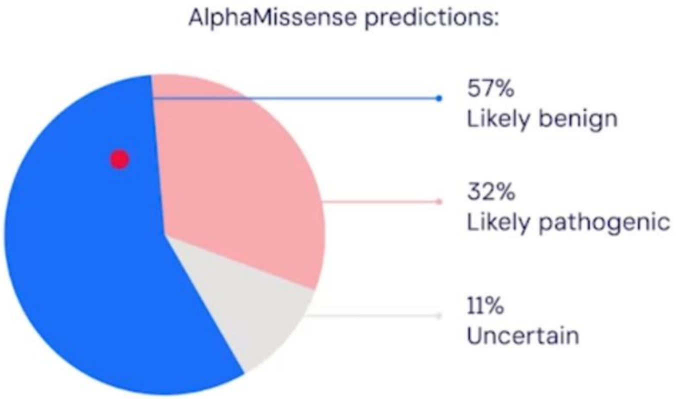
Proteome-wide predictions with more confidently classified variants

- **600M predictions** all AA substitutions and missense variants in 20k canonical gene isoforms and 60k alternative isoforms
- **Higher coverage %** of confident predictions due to better performance (**67% -> 92% of ClinVar**)
- Accessible as **VEP plugin**

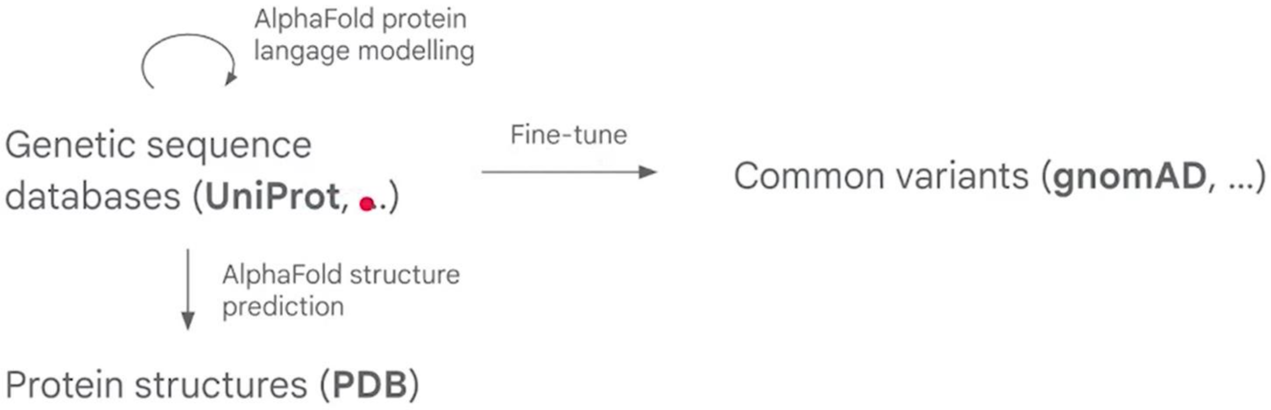
ClinVar variants



All possible 71 million human missense variants



Utilizing different sources of information



AlphaFold protein
langage modelling

Genetic sequence databases (**UniProt**, ...)

Fine-tune

Common variants (**gnomAD**, ...)

AlphaFold structure
prediction

Protein structures (**PDB**)

Training

Evaluation

Clinically relevant variants (**ClinVar**)

Multiplexed assays of variant effect (**ProteinGym**, MaveDB)

Benchmark Comparisons: AlphaMissense vs. Other Tools

- **Comparison Tools**:
    - **PolyPhen-2**, **SIFT**, **PrimateAI**, and other state-of-the-art models for variant pathogenicity prediction.

- **Key Metrics**:
    - **Area Under the Receiver Operating Characteristic (auROC):**
    - AlphaMissense: **0.947**
    - PolyPhen-2: **0.856**
    - SIFT: **0.857**
    - PrimateAI: **0.936** (Higher values indicate better prediction accuracy.)

04

Limitations, Implications, and summary

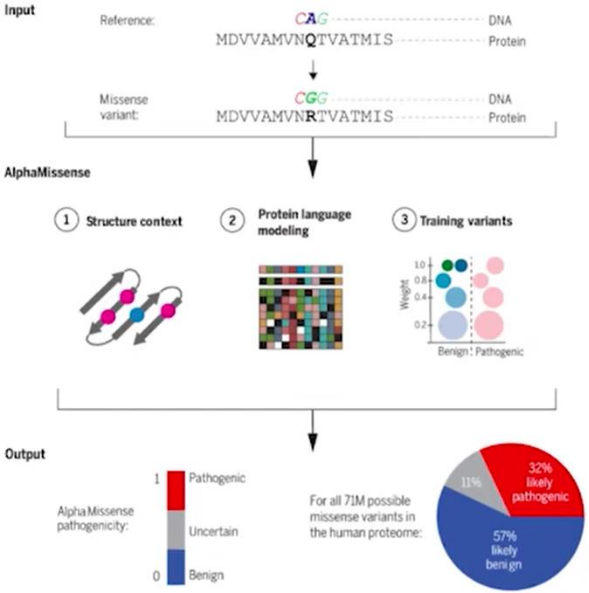
Model limitations, future scope, and takeaways from the study.

Model Limitations of AlphaMissense

- **Structural Impact Prediction**:
  - AlphaMissense focuses on pathogenicity at the sequence level but struggles to predict how variants impact **protein 3D structures** or dynamics.
  - Example: It cannot fully model structural changes in key genes like **TP53**, where subtle folding differences may alter functionality.
- **Multivariant Interactions**:
  - The model evaluates variants individually and does not account for **epistatic effects**, where multiple variants interact to influence a protein's behavior.
  - This limits its application to diseases caused by complex variant combinations.
- **Data Bias and Generalization**:
  - Despite improvements, the model still relies heavily on **annotated datasets**, which may introduce biases in underrepresented gene regions or populations.

Summary

- Fine-tuned AlphaFold to predict pathogenicity of missense variants, without using clinically-ascertained variants in training.
- Outperforms state-of-the-art on multiple diverse benchmarks.
- Highlighted (and accounted for) biases in gold-standard evaluation data set (ClinVar)
- Increased the number of confidently classified variants (using ClinVar to estimate precision) proteome-wide
- High performance holds amongst clinically-actionable genes, and aligns with known functional regions in some cases.
- Average AlphaMissense pathogenicity predicts cell essential genes, outperforming other computational approaches (e.g. LOEUF) for smaller genes.



Thank you