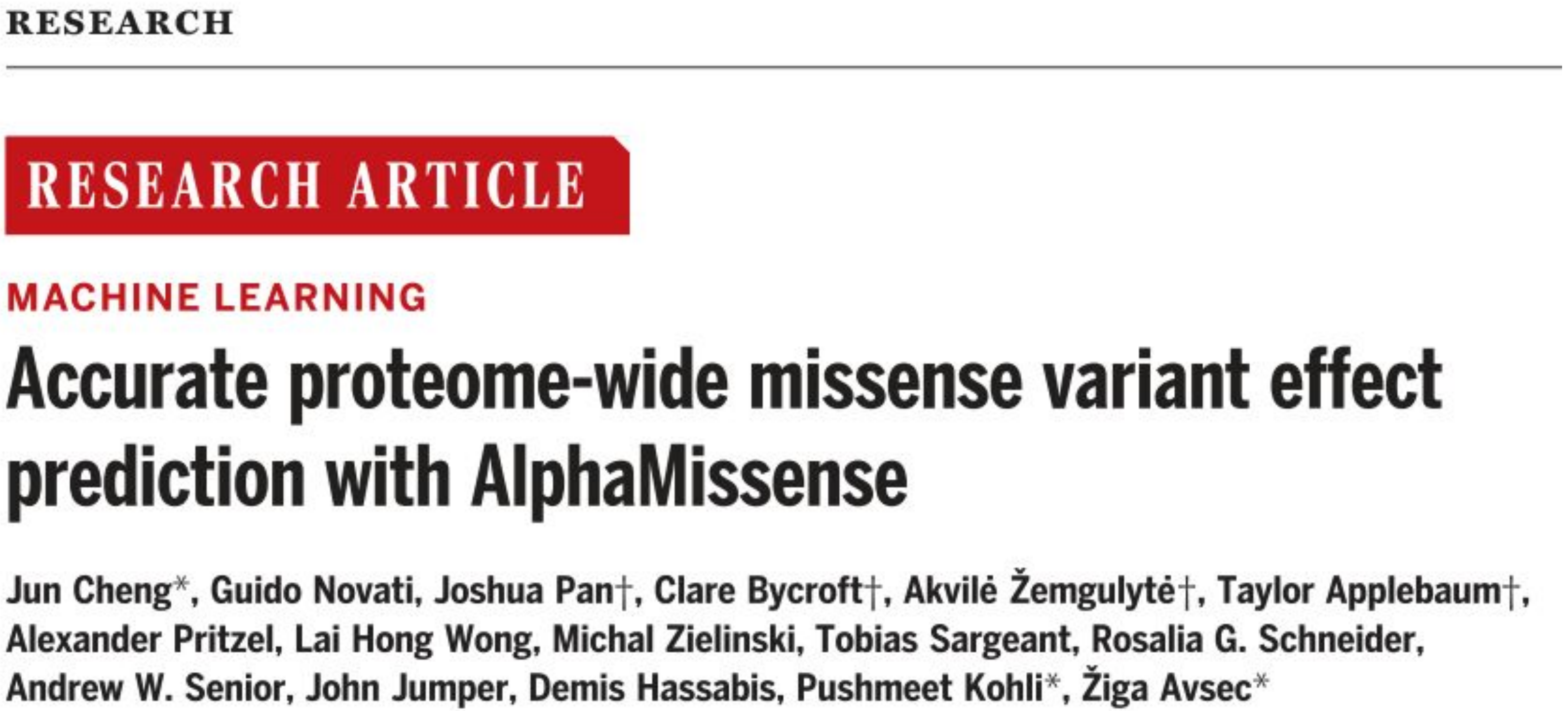


RESEARCH

RESEARCH ARTICLE

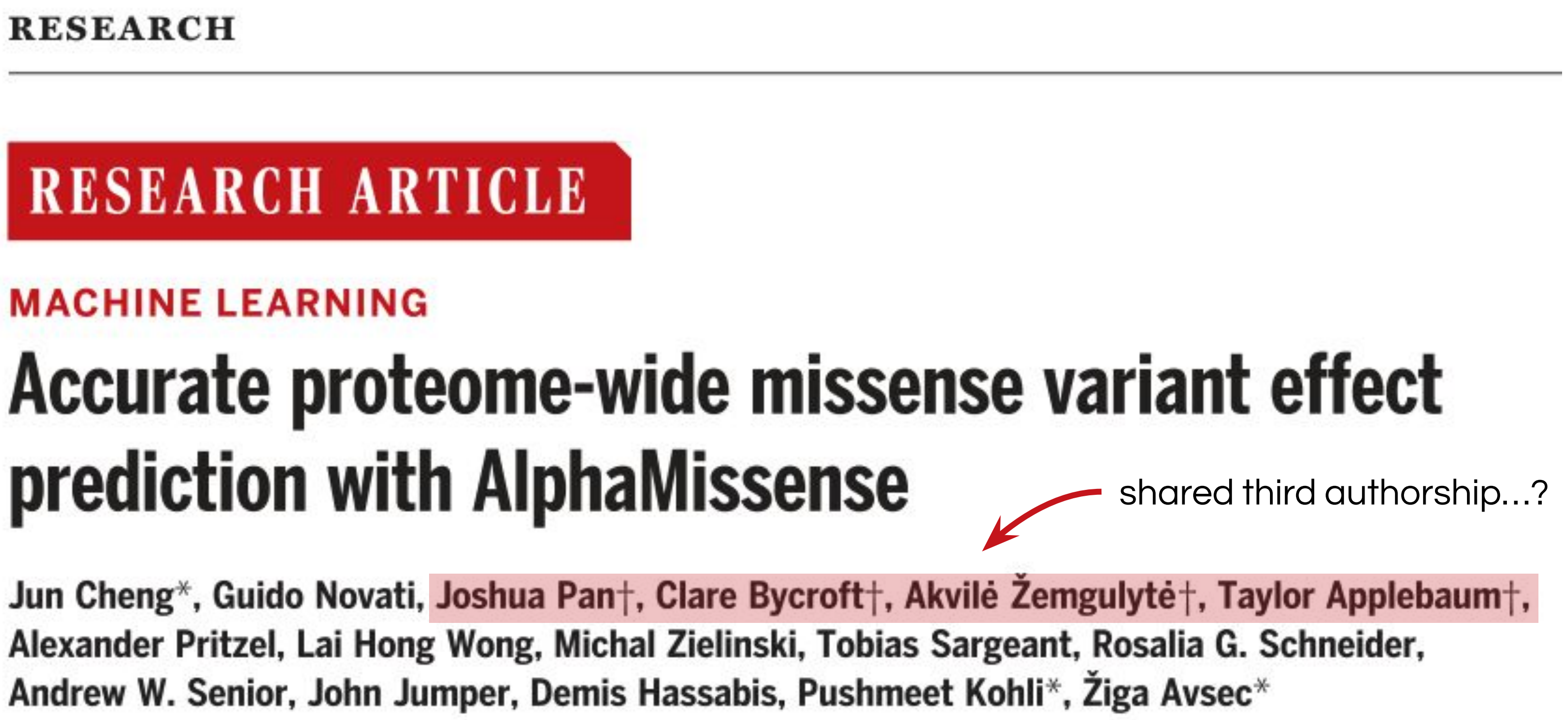
MACHINE LEARNING

Accurate proteome-wide missense variant effect prediction with AlphaMissense

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*

Google DeepMind, London, UK

Science **381**, eadg7492(2023). DOI:10.1126/science.adg7492



RESEARCH

RESEARCH ARTICLE

MACHINE LEARNING

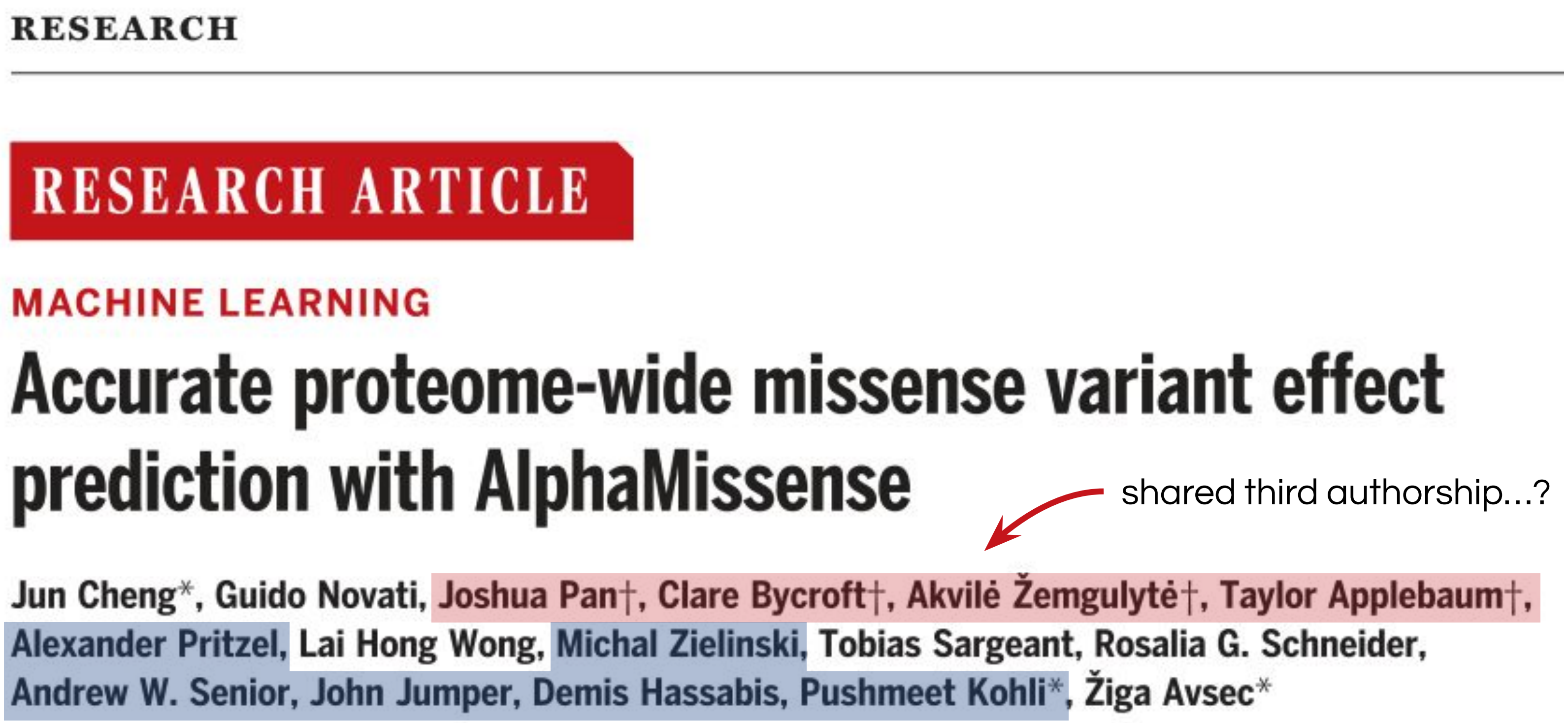
Accurate proteome-wide missense variant effect prediction with AlphaMissense

shared third authorship...?

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[*]

Google DeepMind, London, UK

Science **381**, eadg7492(2023). DOI:10.1126/science.adg7492



RESEARCH

RESEARCH ARTICLE

MACHINE LEARNING

Accurate proteome-wide missense variant effect prediction with AlphaMissense

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[*]

shared third authorship...?

also authored AlphaFold

Google DeepMind, London, UK

Science **381**, eadg7492(2023). DOI:10.1126/science.adg7492

RATIONALE

- more than **4 million** observed **missense variants** (altering AA sequence)
- **only 2% of them categorized** as either benign or pathogenic
- remaining ones are **VUS** (variants of unknown significance)
- experimental categorization and validation is **expensive** (multiplexed assays of variant effect (MAVEs))

Limits:

- diagnosis of rare diseases
- clinical treatments targeting specific genomic functional alterations

… and it would probably be great to **demonstrate the usefulness of AlphaFold** which they already have

PRIOR ADVANCES

1. trained on **human-curated** databases (eg. ClinVar)
    - inherit biases, data leakage between training and test sets
    - *PolyPhen-2, REVEL, VARITY, gMVP*
2. trained on **weak labels** (‘bening’ ~ frequent, ‘pathogenic’ ~ unobserved)
    - many false labels, but no human curation biases
    - *CADD*
3. unsupervised; model **AA distribution** conditioned on **sequence context**
    - ‘pathogenic’ ~ alternate AA not likely to appear naturally
    - *SIFT, EVmutation, GEMME*
4. leveraging **protein structure**
    - *AlphScore, COSMIS*

DATA LEAKAGE, BIASES

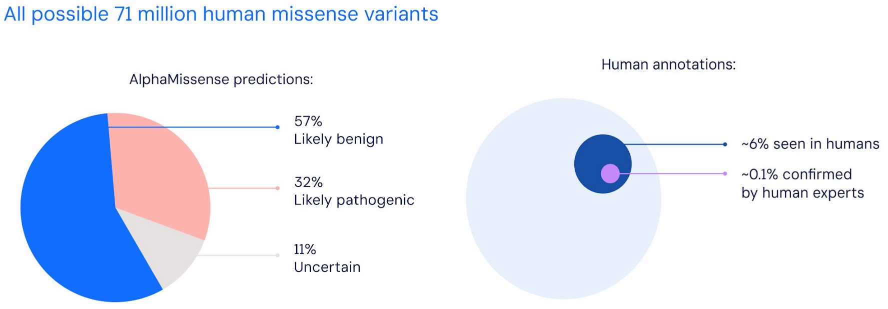
- training variants in the test sets (especially in ensemble models)
- **variants causing the same AA substitution** in both training and test sets
- data leakage from **paralogous genes or homologous protein domains**
- **label circularity** (predictions of some models influence the classification labels of newly curated variants)

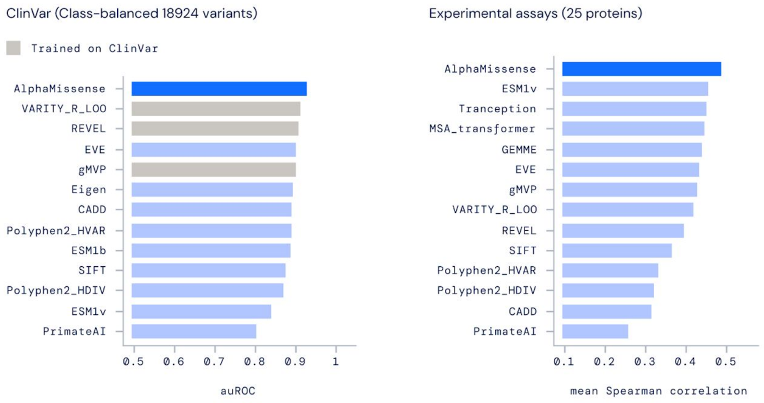
- gene label bias (using the **percentage of pathogenic variants per gene** in ClinVar as a predictor achieves auROC = 0.914 on a ClinVar test set)
- a model might perform better on **well-studied genes** but not on others due to the lack of training data

AlphaMissense

A combination of

1) training on weak labels from population frequency data (no human-curated DBs)
2) unsupervised protein language modeling (AA distributions conditioned on sequence context)
3) structural context by using an AlphaFold-derived system.



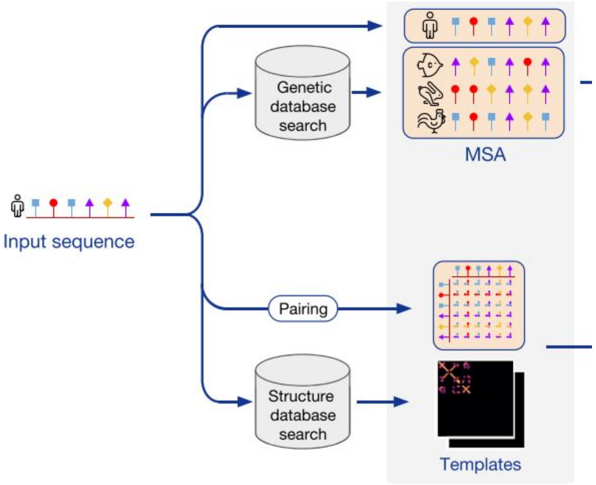


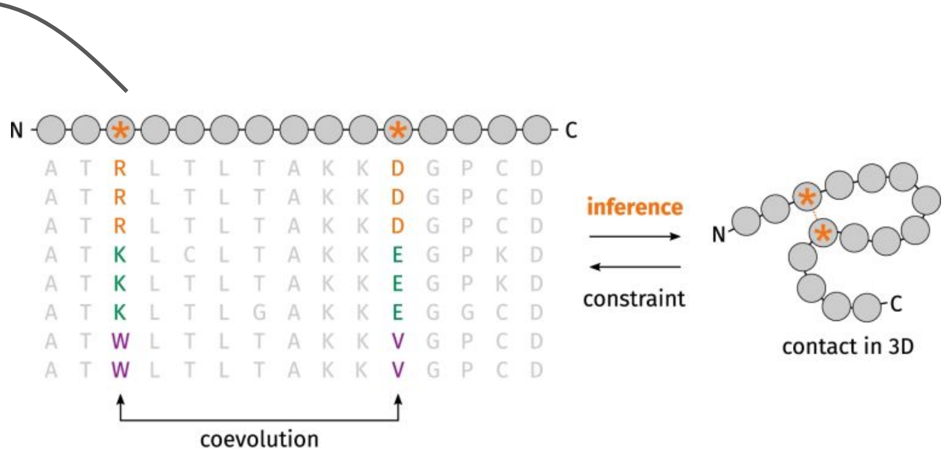
DeepMind blog

ARCHITECTURE

AlphaFold

preprocessing





```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

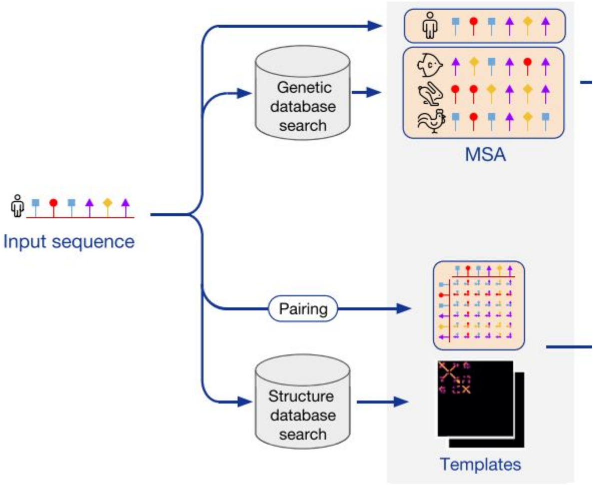
- querying various DBs for similar sequences → **MSA** (evolutionary constraints)
- finding existing structural templates → **pair representation**

Oxford Protein Informatics Group

ARCHITECTURE

AlphaFold

preprocessing



residues in the MSA are **randomly masked or mutated**

→ **original ones are predicted within the model**

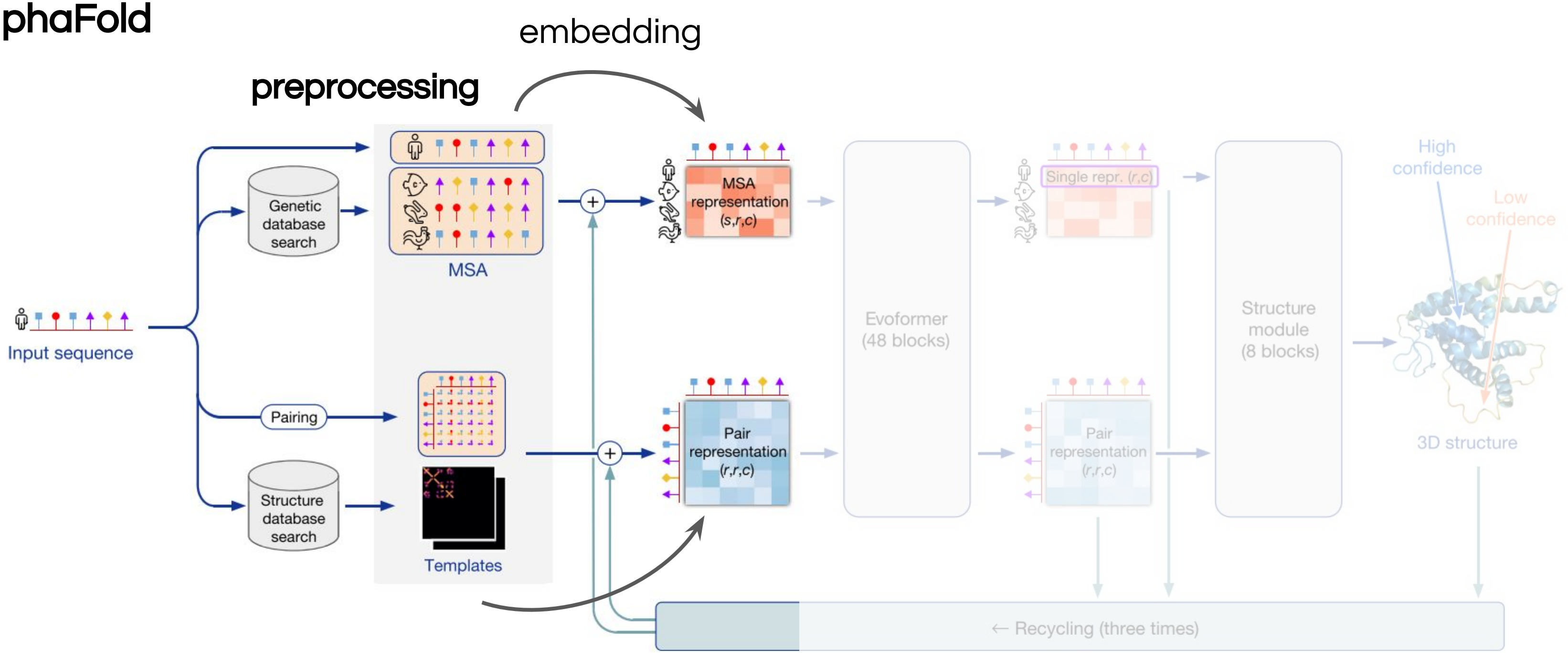
→ serves as an **"auxiliary loss"**

(added to the global loss function alongside losses measuring structural accuracy)

→ encourages the network to learn to interpret phylogenetic and covariation relationships

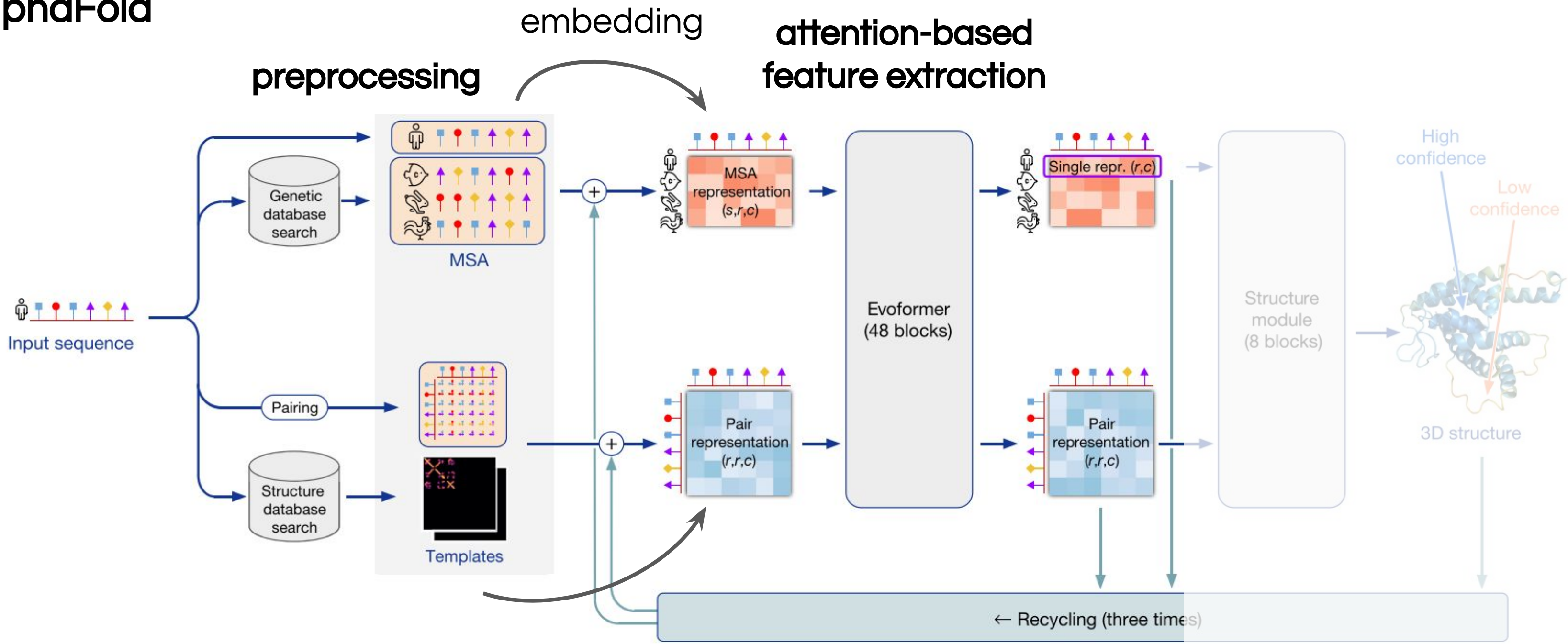
ARCHITECTURE

AlphaFold



ARCHITECTURE

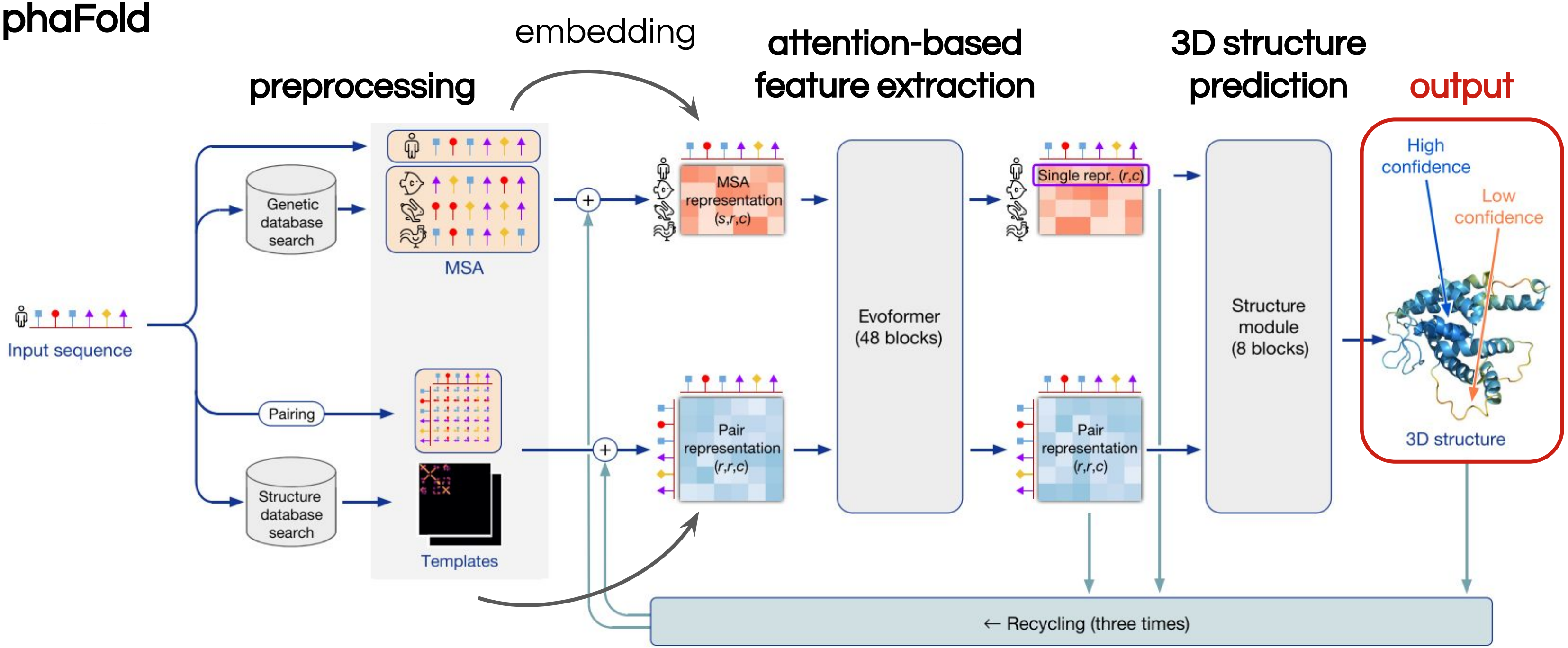
AlphaFold


embedding
preprocessing
attention-based feature extraction
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)

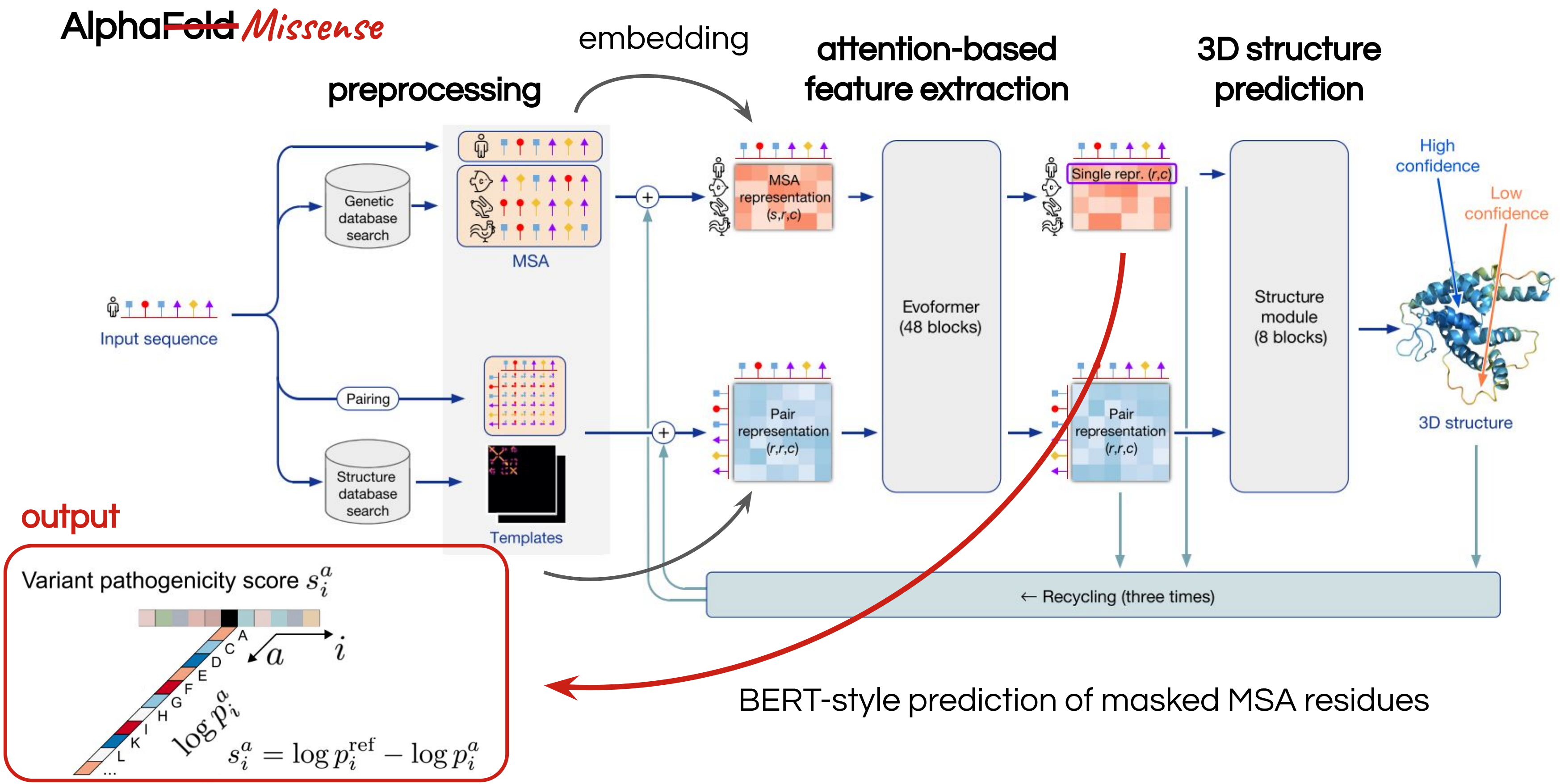
both the MSA and the pair representation are **iteratively refined while exchanging information** through two communicating **transformers** (“two-tower architecture”)

ARCHITECTURE

AlphaFold



ARCHITECTURE



Alpha~~Fold~~ Missense

preprocessing

embedding

attention-based feature extraction

3D structure prediction

output

Variant pathogenicity score s_i^a

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

BERT-style prediction of masked MSA residues

ARCHITECTURE

Alpha~~Fold~~ Missense

1. Pre-training

- training a traditional AlphaFold model with **increased weight of the MSA-prediction loss**

2. Fine-tuning

- human & primate proteins
- binary classification of benign vs. pathogenic variants
- stop training once it starts to overfit on the validation set (ClinVar)

3. Calibration

- linear logistic rescaling to convert raw pathogenicity score to label probability + **thresholding**

The model does not:

- predict mutated protein structure
- consider interactions with other proteins
- support indels

TRAINING SET

Bening variants

derived from **observed variants** in human and primate species (gnomAD, etc.)

Pathogenic variants

with very sophisticated methods

sampled from the remaining 65,314,044 variants, out of all the possible missense variants, that were **not observed** in any primate or human population

→ **“weak labels”**

+ Data self-distillation

pathogenic variant sampling **refined with initial predictions** of pathogenicity (variants predicted as bening are not included in further training)

EVALUATION SETS

ClinVar variants

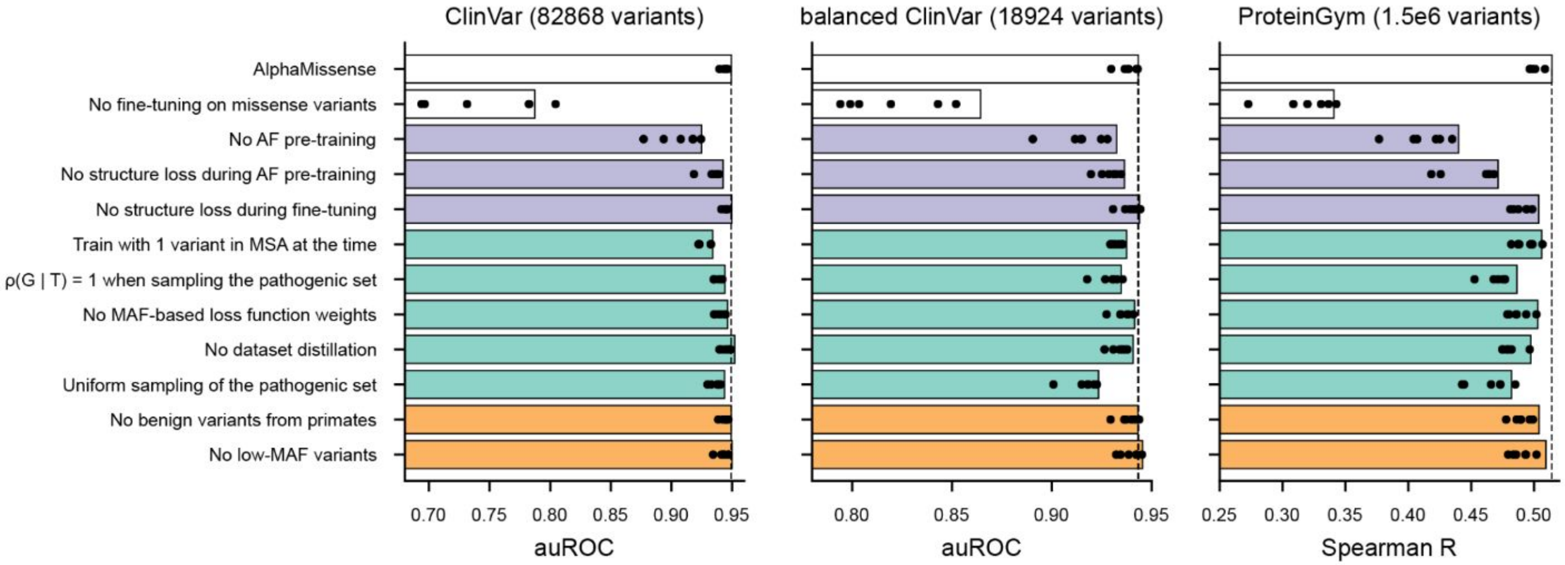
- labels “bening”, “likely bening”, “likely pathogenic”, “pathogenic”
- *Variants with the term* ***“splice”*** *in their description* ***were removed*** *(29 variants), since our missense predictor is unlikely to be able to predict variants affecting splicing.*
- subset: randomly selected 300 proteins → maximum possible equal number of positive and negative variants → (balanced) **validation set** (2,526)
- remaining: (non-balanced) **test set**

Additional **test sets**

- cancer hotspot mutations
- de novo variants from rare disease patients
- ProteinGym + additional MAVE datasets

ABLATION RESULTS

Does it *really* need to be this complicated?



no structural context

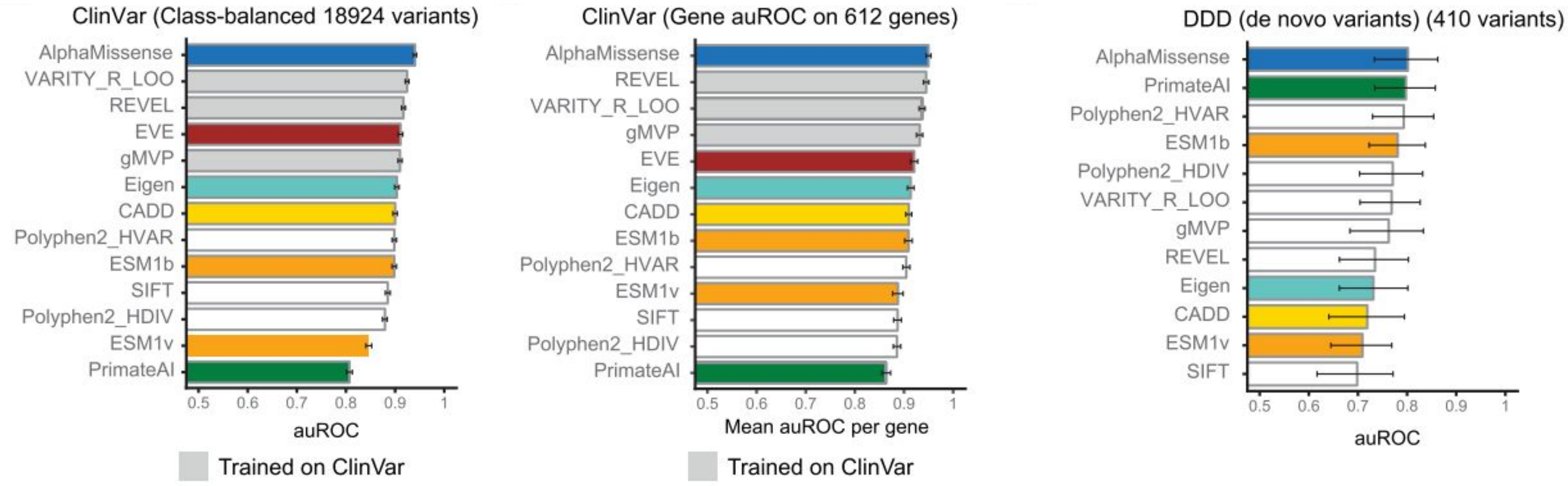
modified variant sampling

excluded data sources

... questionable.

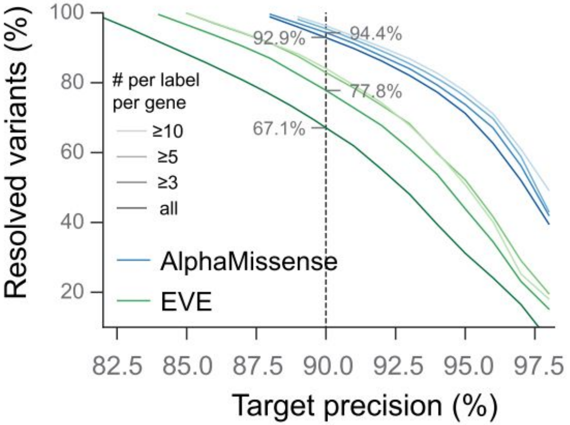
RESULTS

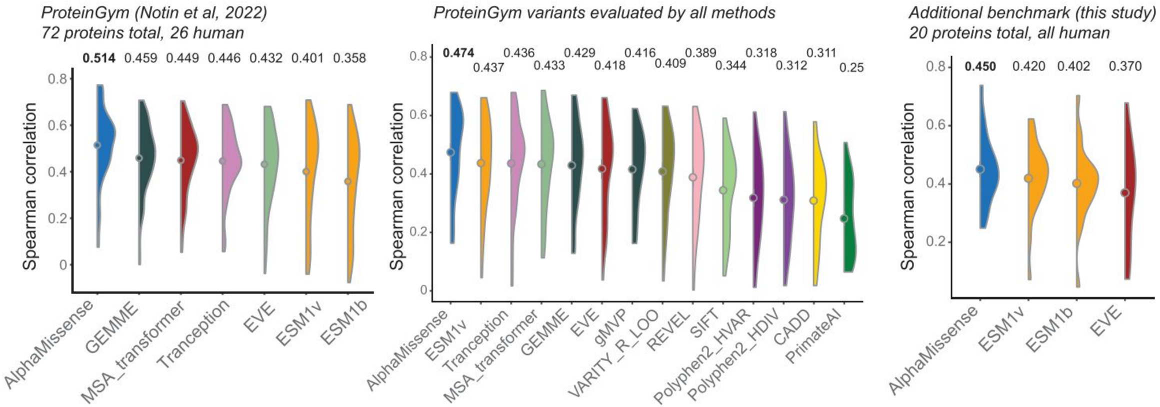
- **consistently high performance** across clinical benchmarks even when compared to methods trained on ClinVar



RESULTS

- **increase the number of confidently classified variants** compared to other methods
- **more consistent with MAVE results** than other methods

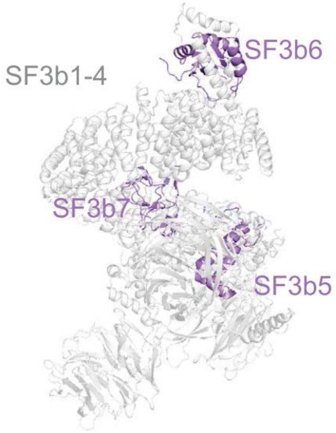




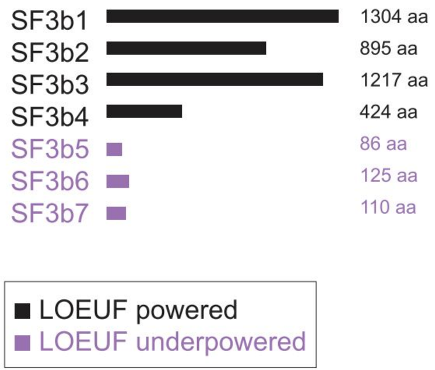
RESULTS

- better predictions for **gene essentiality** than LOEUF

Human SF3b complex
(7 subunits, PDB: 5Z56)

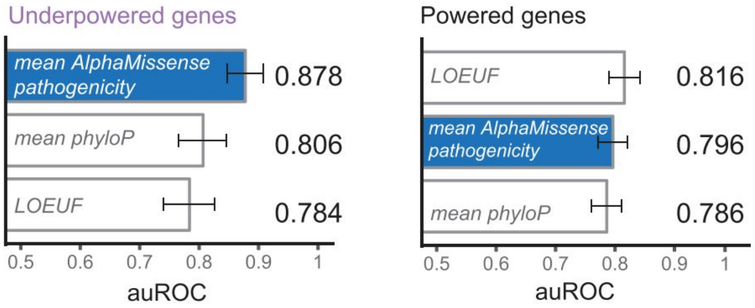


Protein sequence length



	Cell essential	mean AM decile	Expected pLoF	LOEUF decile
SF3b1	Yes	0%	72.1	0%
SF3b2	Yes	10%	53.2	0%
SF3b3	Yes	0%	66.4	0%
SF3b4	Yes	0%	14.4	0%
SF3b5	Yes	0%	2.7	50%
SF3b6	Yes	0%	7.0	30%
SF3b7	Yes	0%	6.6	20%

Classifying gene essentiality



COMMUNITY RESOURCE

https://console.cloud.google.com/storage/browser/dm_alphamissense

1. **71 million missense variant predictions** saturating the human proteome

`AlphaMissense_hg19.tsv.gz, AlphaMissense_hg38.tsv.gz`

2. **gene-level AlphaMissense pathogenicity predictions,** defined as the average pathogenicity over all possible missense variants in a gene

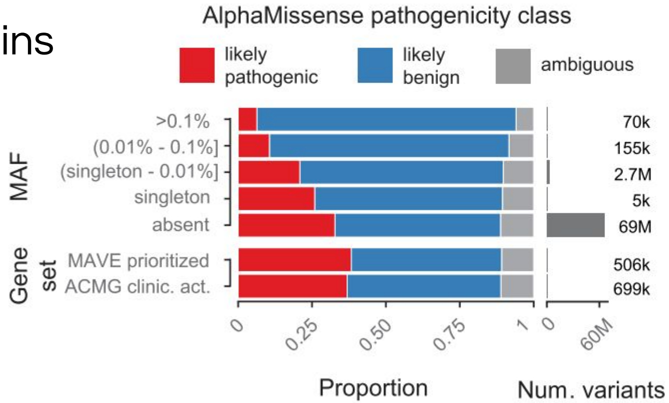
`AlphaMissense_gene_hg19.tsv.gz, AlphaMissense_gene_hg38.tsv.gz`

3. expanded dataset of **all 216 million possible single amino acid substitutions** across the 19,233 canonical human proteins

`AlphaMissense_aa_substitutions.tsv.gz`

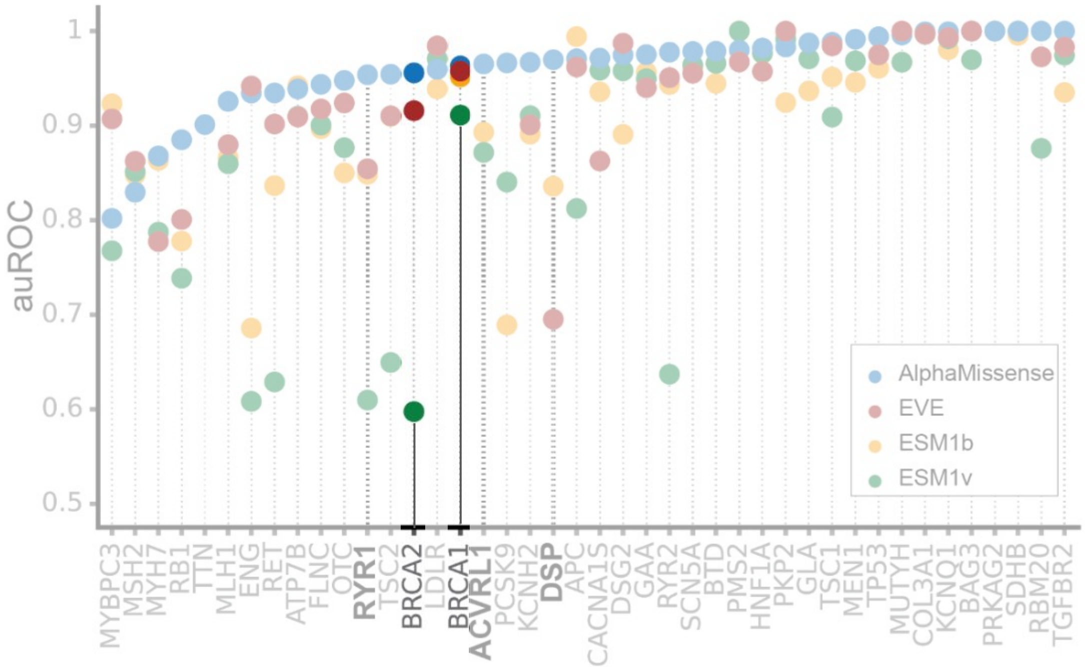
4. predictions for all possible missense variants and amino acid substitutions **across 60,000 alternative transcript isoforms**

```
AlphaMissense_isoforms_hg38.tsv.gz,
AlphaMissense_isoforms_aa_substitutions.tsv.gz
```



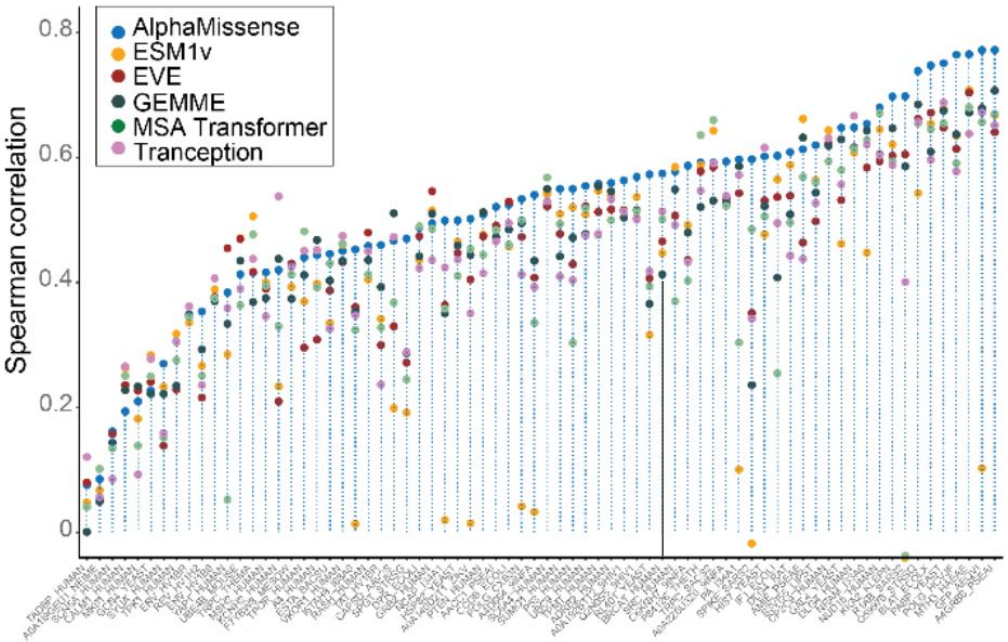
BRCA1, BRCA2

ClinVar test variants per gene



BRCA1, BRCA2

ProteinGym benchmark by protein


● AlphaMissense
● ESM1v
● EVE
● GEMME
● MSA Transformer
● Tranception
Spearman correlation
TADBP_HUMAN
A0A1I9GEU1_NEIME
CALM1_HUMAN
SYUA_HUMAN
TPK1_HUMAN
SUMO1_HUMAN
SPIKE_SARS2
GFP_AEQVI
A4GRB6_PSEAI