Google DeepMind

AlphaMissense predictions for human genetic variation research

gtatatatcatgatatgtatacgcatagatcatgagctca

Webinar objectives

- *Describe the principles behind AlphaMissense and its approach to variant classification*
- *Identify how AlphaMissense utilises protein structure and sequence information for variant assessment*

Talk outline

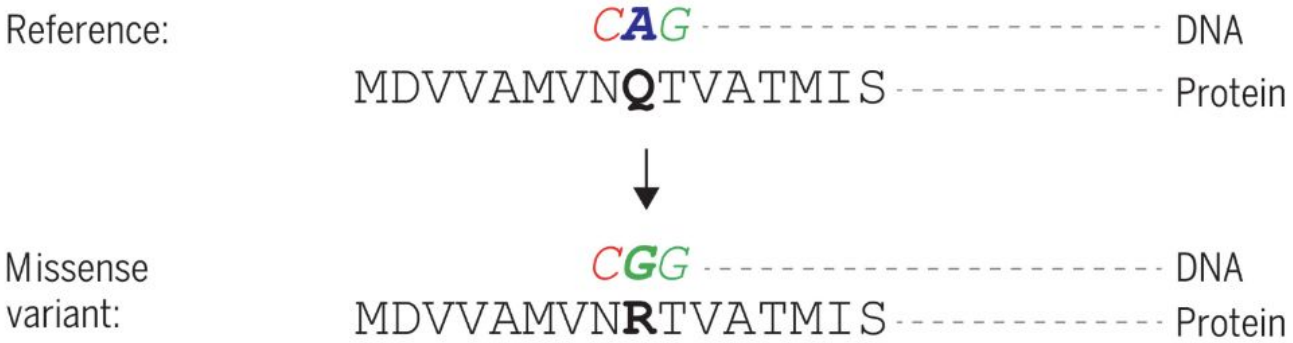
- **Overview of AlphaMissense**
- Design principles
- Evaluating predictions
- Using the scores

Note: Key slides designated with this:

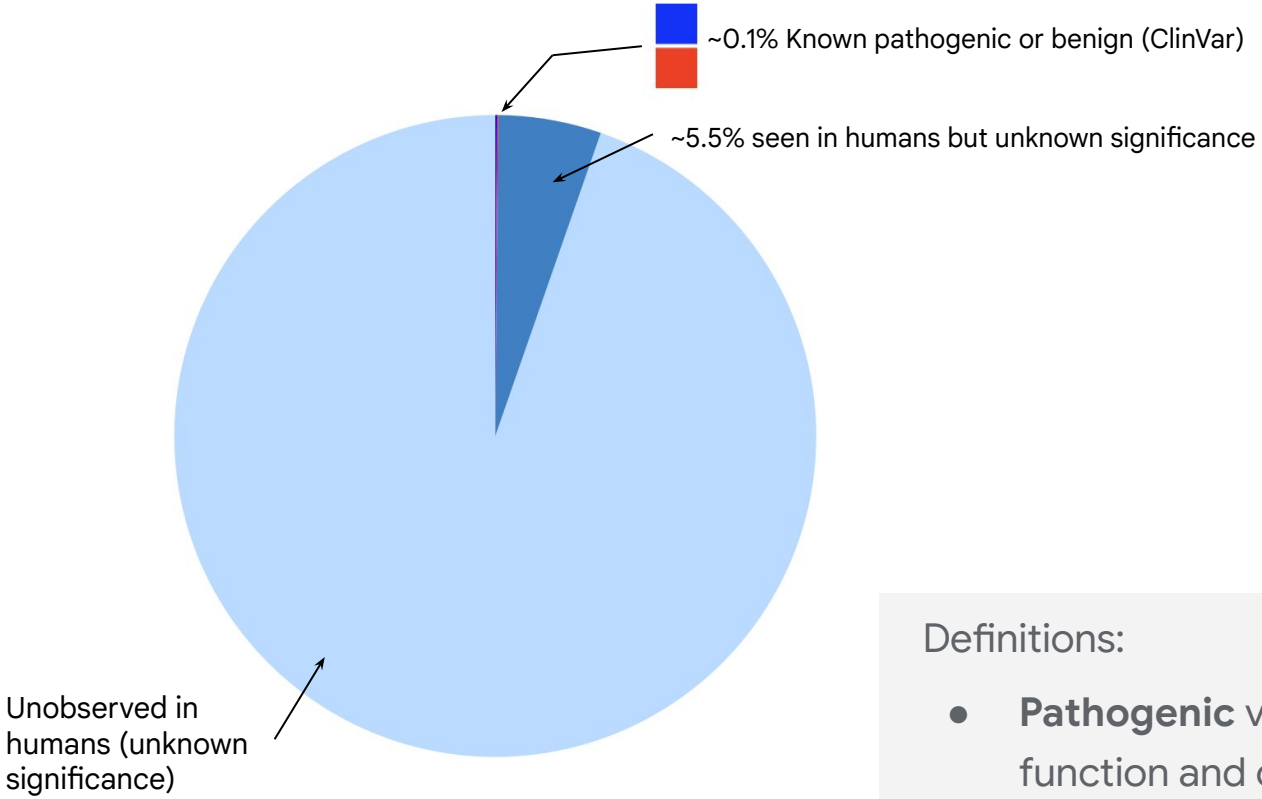
💡 Important 💡

Google DeepMind and science

- Mission: **build AI responsibly to benefit humanity**
- Science group which approaches fundamental science problems
- One such problem is genome variant Interpretation
- Definition: A **missense variant** is a variant affecting a protein, by changing a reference amino acid to an alternate amino acid.



Why missense? A major gap in our knowledge of variant effects

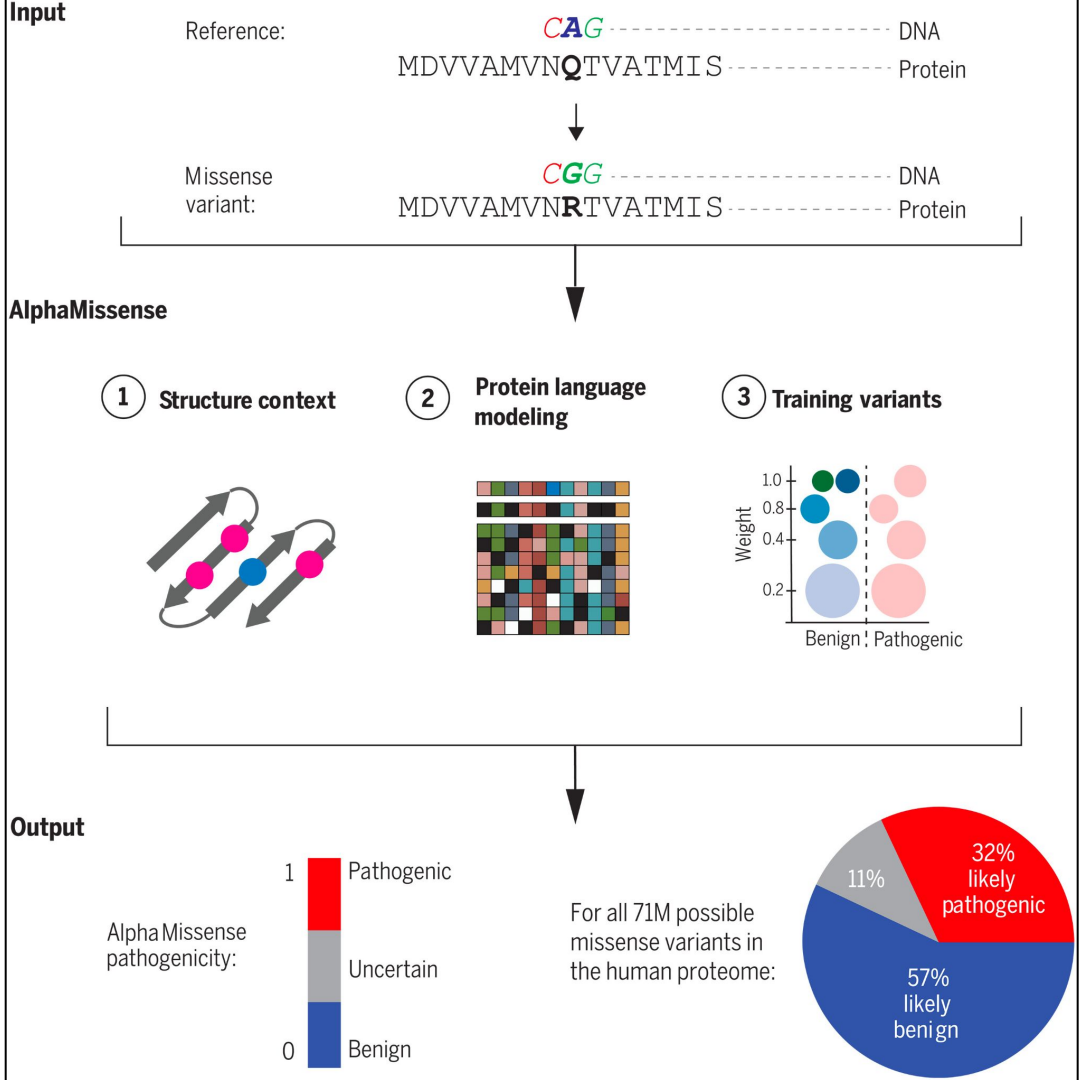


Important

Definitions:

- **Pathogenic** variants: disrupt protein function and often lead to disease
- **Benign** variants: have limited effect

AlphaMissense: proteome-wide missense variant effect prediction



Webinar objectives

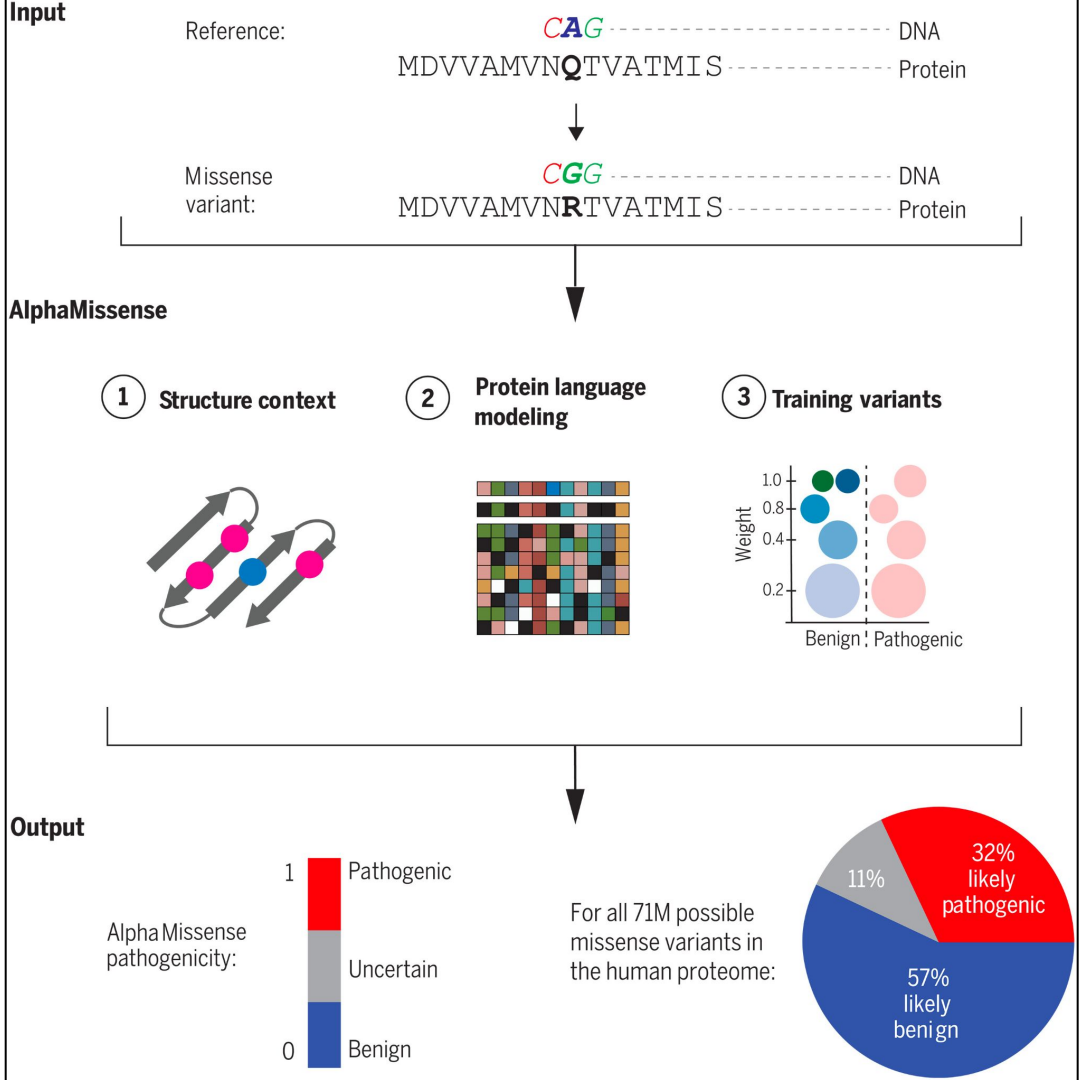
- *Describe the principles behind AlphaMissense and its approach to variant classification*
- *Identify how AlphaMissense utilises protein structure and sequence information for variant assessment*

Talk outline

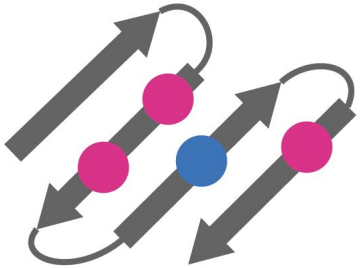
- Overview of AlphaMissense
- **Design principles**
- Evaluating predictions
- Using the scores

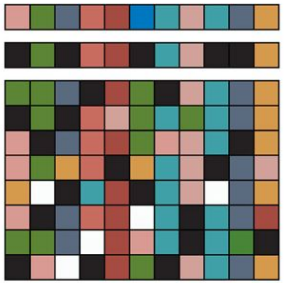
Why is machine learning an appropriate choice for this problem?

- **Clear task**
  - Input: Variant
  - Output: Score from 0-1
- **Clear metrics**
  - Held out ClinVar variants
  - Deep mutational scans
- **Clear impact**
  - Fill in the gaps in an important database
- **Clear inductive biases**

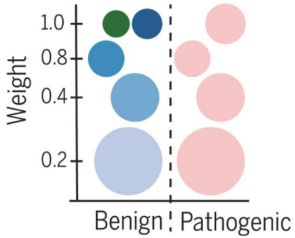


What inductive biases (principles) inform the model?



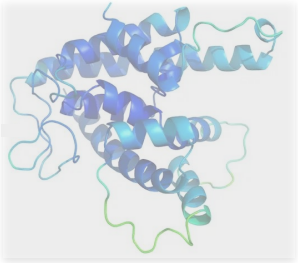


- Protein context matters
  - Sequence
  - Structure
- Protein evolutionary history matters

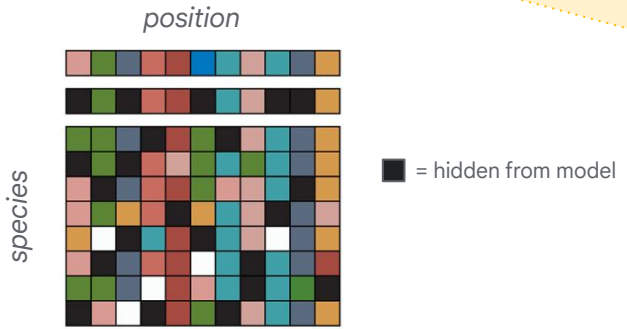


- Observed sequence variation in humans matters

Under the hood: Adapting AlphaFold into AlphaMissense



AlphaFold2 takes protein sequence as input and returns its **structure**. This is the *starting point* for AlphaMissense.



AlphaFold2 has a component that predicts the likelihood of amino acids belonging in *hidden entries* in the multiple sequence alignment of the input protein - thereby learning the statistics of **protein evolutionary history**.

Weak labels

‘Benign’: Variants present in human/primate population

‘Pathogenic’: Variants absent in human population

AlphaMissense fine tunes this likelihood via an additional challenge: given a human protein variant, classify whether it is present or absent in the human population, thereby weakly supervising on **human sequence variation**.

Finally, AlphaMissense is evaluated on ClinVar to assess performance. ***Note***: ClinVar variants are EXCLUDED from the previous step. The model is NOT trained on ClinVar data directly.

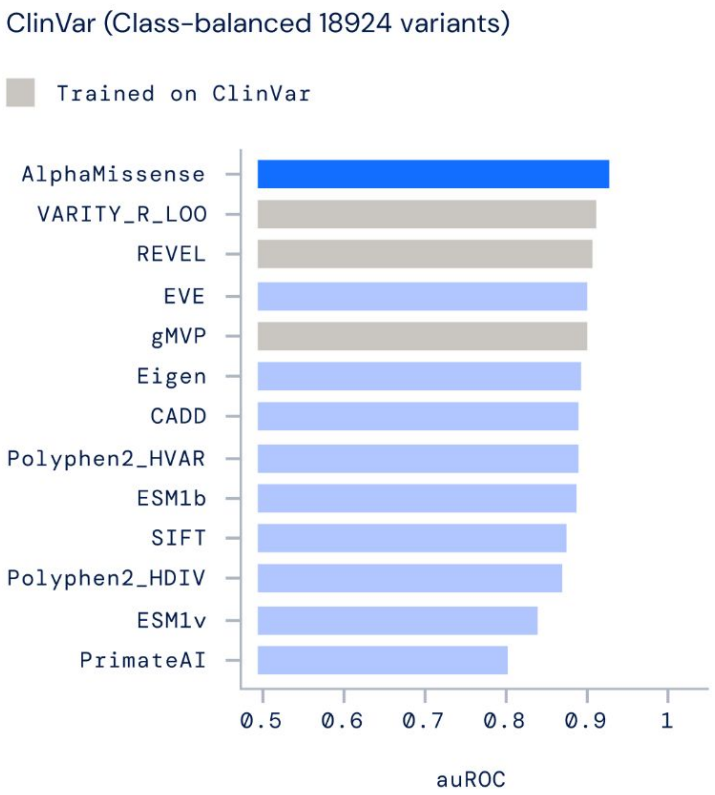
Webinar objectives

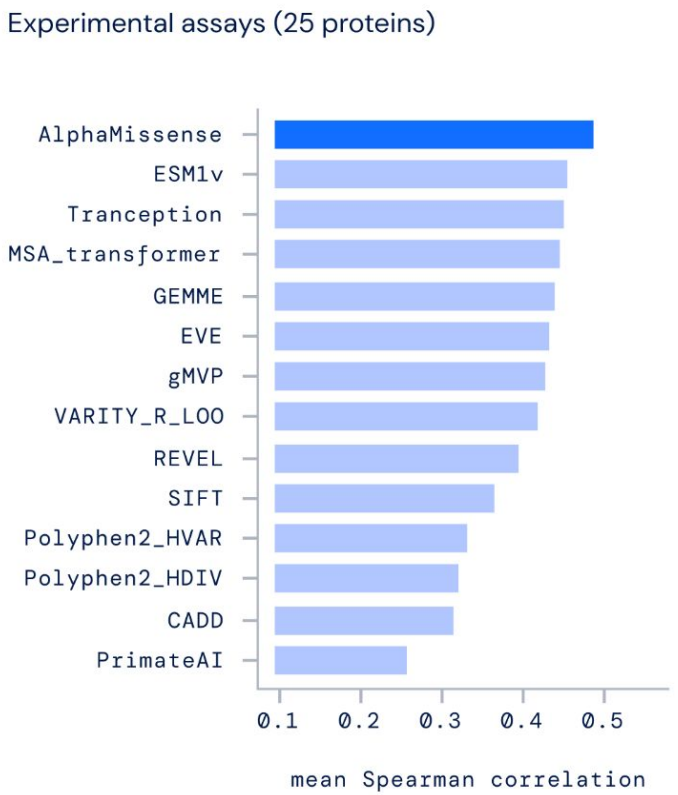
- *Describe the principles behind AlphaMissense and its approach to variant classification*
- *Identify how AlphaMissense utilises protein structure and sequence information for variant assessment*

Talk outline

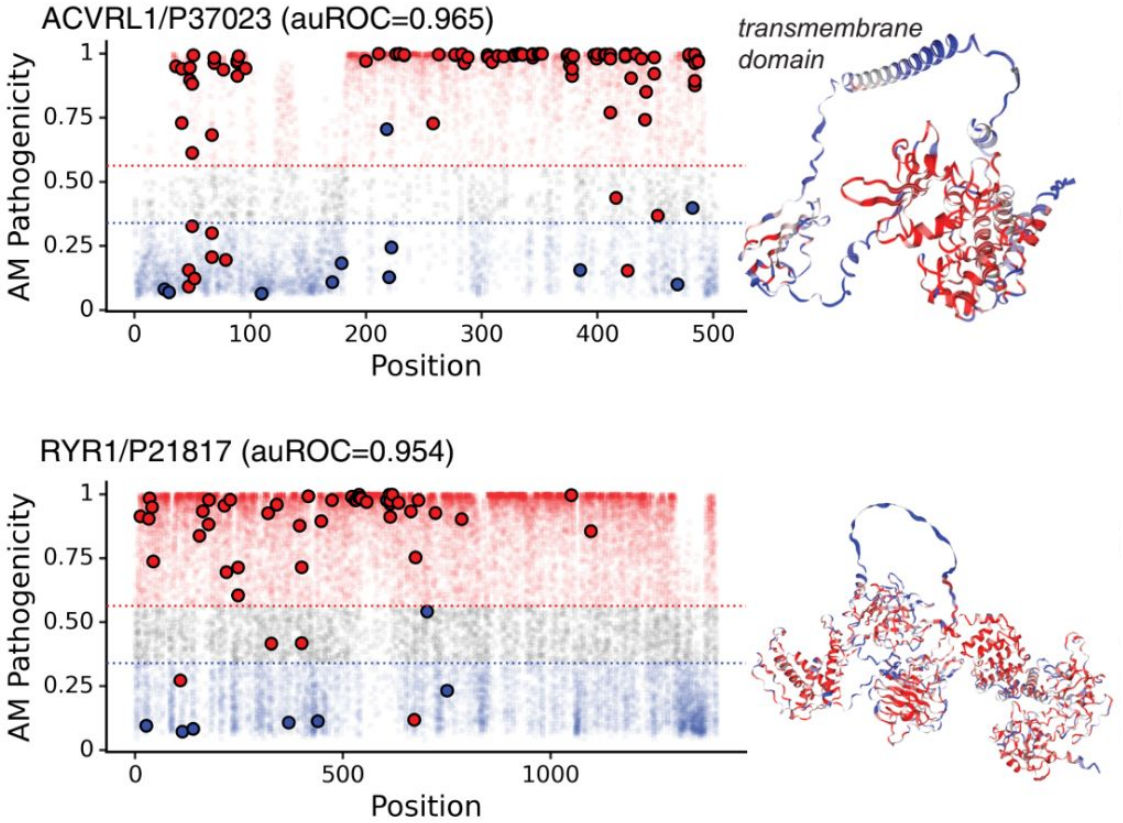
- Overview of AlphaMissense
- Design principles
- **Evaluating predictions**
- Using the scores

Good performance across a diverse set of benchmarks





Example: AlphaMissense predictions for clinically actionable genes



All about the 0 to 1 score

What does a score between 0-1 mean?

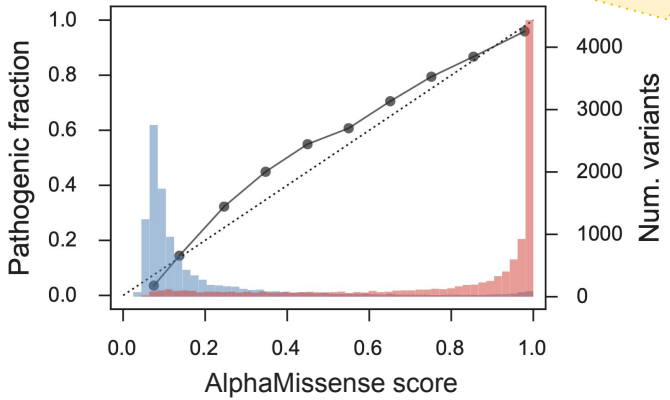
- Score of 0.8 implies: for 10 variants that score in that range, you will expect 8 of them will say “pathogenic/likely-pathogenic” in ClinVar, and 2 of them to say something else (uncertain or likely benign)

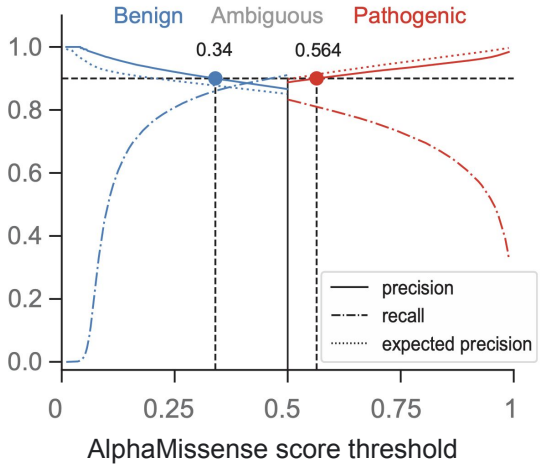
How did they get to be in that scale?

- Model output was transformed to [0,1] using a logistic regression classifier on a set of ClinVar validation variants

How were class thresholds drawn?

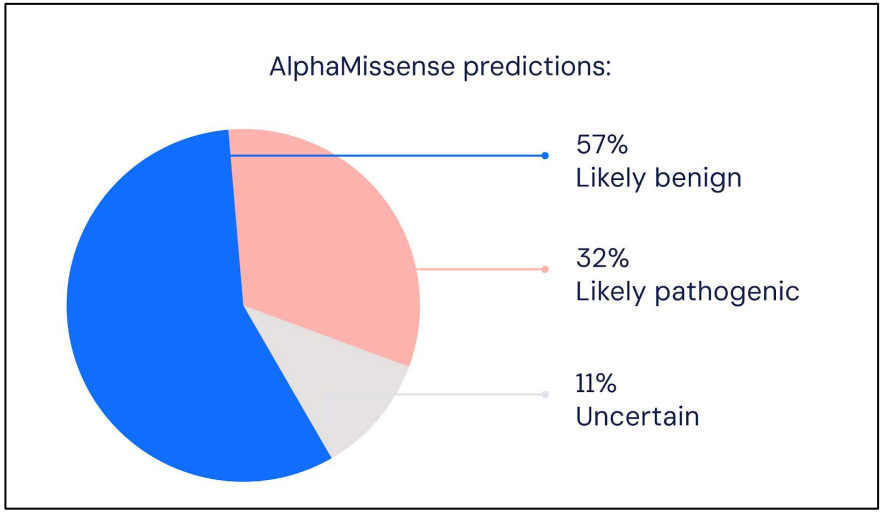
- Likely benign [0-0.34]: The set of predictions in this range obtain 90% negative-class precision (benign)
- Likely pathogenic [0.564-1]: Same, but for 90% positive-class precision (pathogenic).
- Class labels chosen to match ClinVar naming conventions - however, these should not be interpreted as equivalent to ACMG labels





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. Fractions are based on thresholds that yield 90% precision on ClinVar variants



Webinar objectives

- *Describe the principles behind AlphaMissense and its approach to variant classification*
- *Identify how AlphaMissense utilises protein structure and sequence information for variant assessment*

Talk outline

- Overview of AlphaMissense
- Design principles
- Evaluating predictions
- **Using the scores**

How should I reason about the scores?

If you are a clinical researcher:

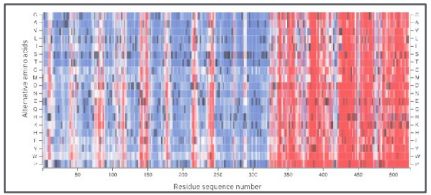
- Scores for genetic variants are available via familiar portals such as VEP, DECIPHER
- There is a lot of thinking and guidelines around use of informatic evidence in clinical settings that should inform use of AlphaMissense (and other predictive approaches)
    - AlphaMissense scores are a research tool
    - Should be used alongside other sources of evidence when drawing scientific conclusions
    - Not a substitute for professional medical advice, diagnosis or treatment
- Recommend benchmarking for your specific research use-case

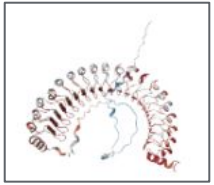
How should I reason about the scores?

If you are a molecular biologist:

Recommend three levels of granularity with which to approach the scores:

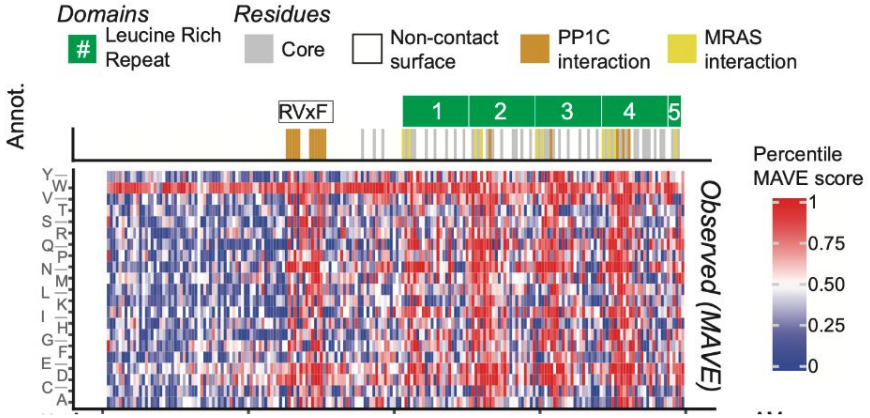
- Individual protein - all variants
  - Residue position x 20 possible amino acids (1 ref, 19 alt)
  - Heatmap visualization recommended
- Individual protein - position averages
  - Single value for each position in your protein
  - Use in conjunction with protein structure (measured or predicted) to aid in interpretation
- Protein-level averages





Is AlphaMissense informative of function? Example SHOC2

Experiment (cancer cell fitness) Kwon et al, 2022, Nature



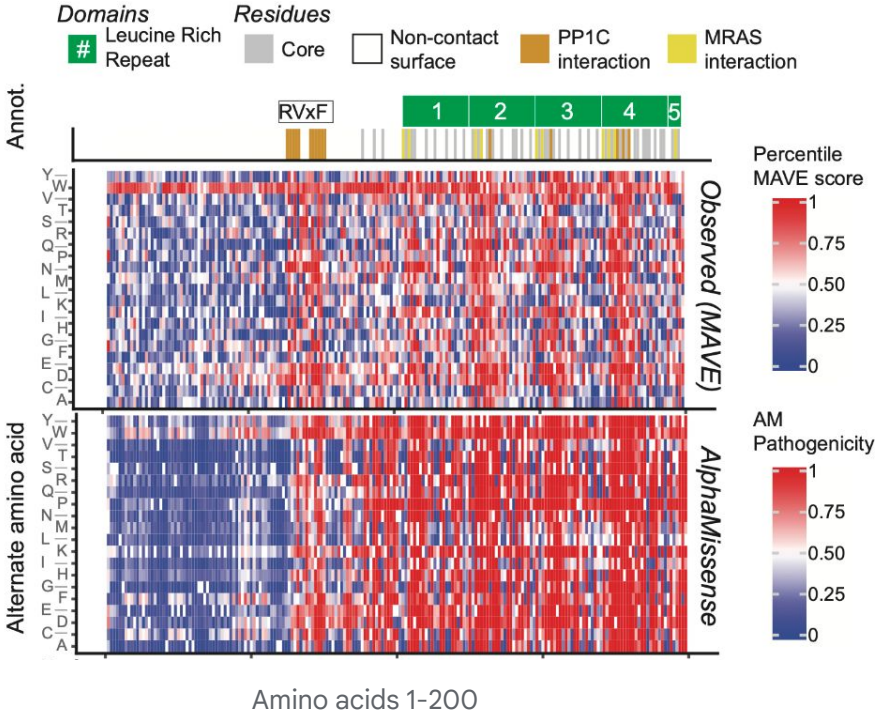
Amino acids 1-200

Is AlphaMissense informative of function? Example SHOC2

Experiment (cancer cell fitness) Kwon et al, 2022, Nature

AlphaMissense

Spearman R = 0.47



Is AlphaMissense informative of function? Example SHOC2

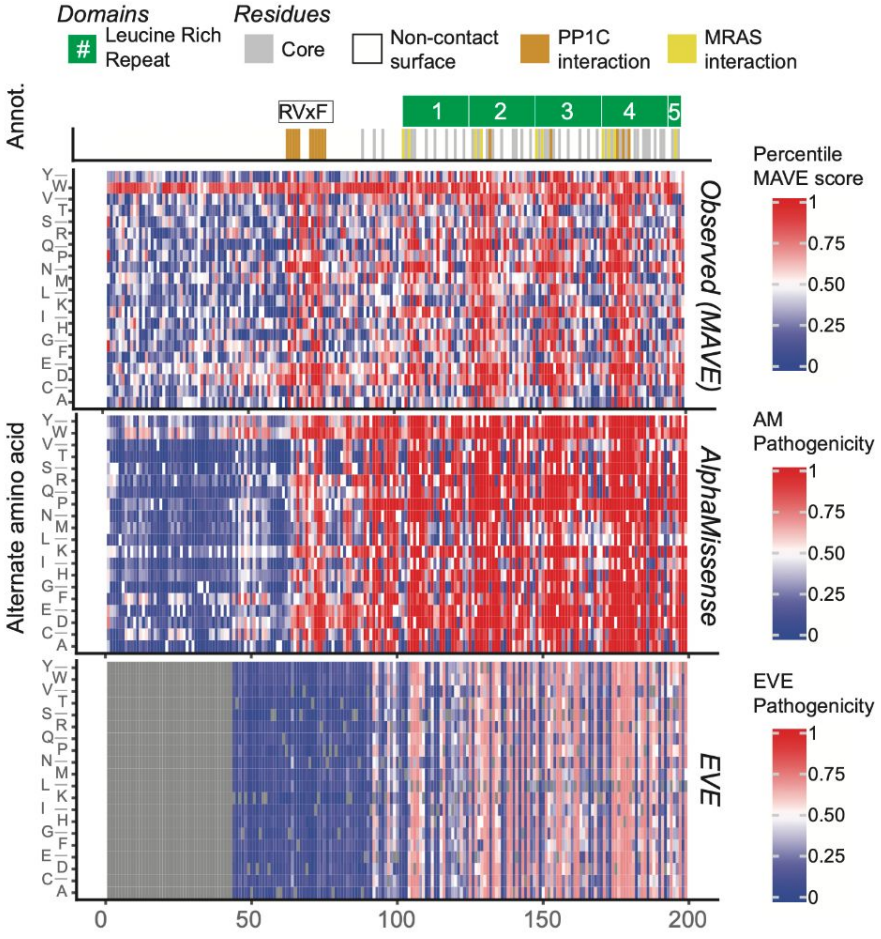
Experiment (cancer cell fitness) Kwon et al, 2022, Nature

AlphaMissense

Spearman R = 0.47

EVE

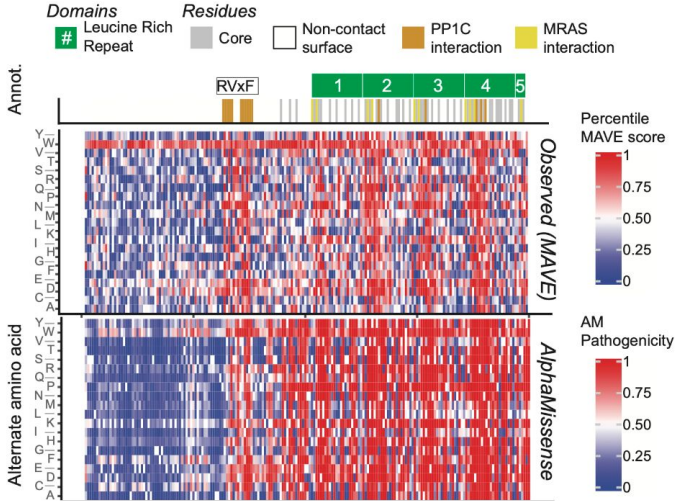
Spearman R = 0.32

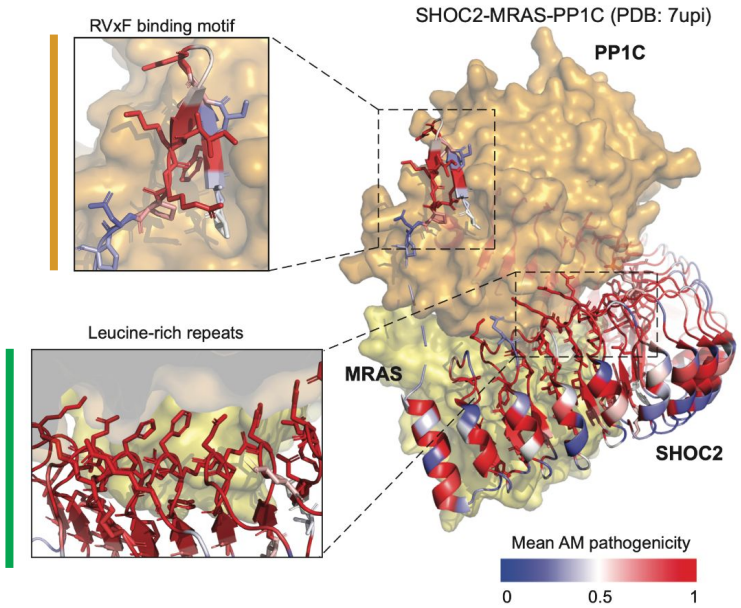


Is AlphaMissense informative of function? Example SHOC2

Experiment (cancer cell fitness)

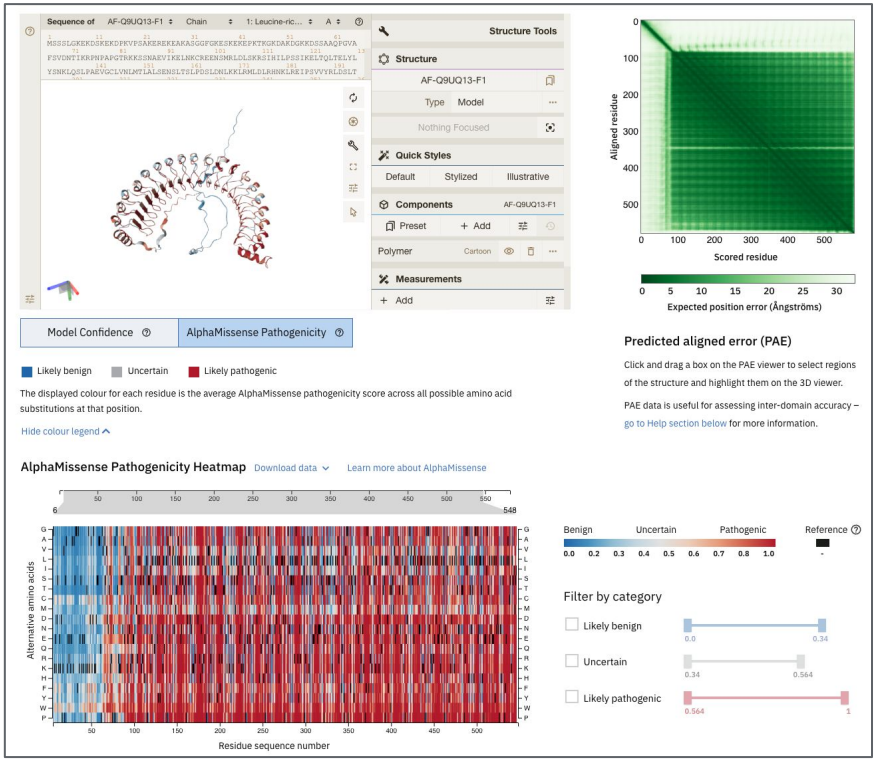
AlphaMissense
Spearman R = 0.47

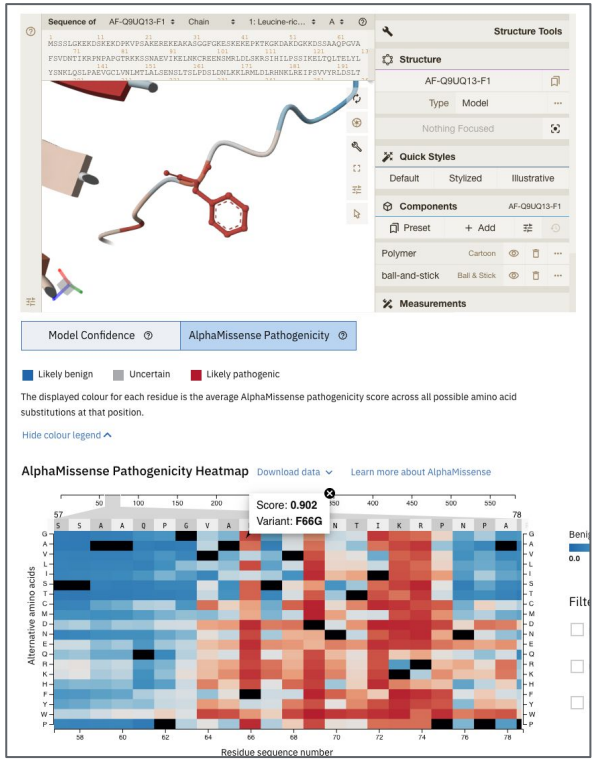




- Regions with high AlphaMissense predictions reflect recently discovered functionally-significant domains.

Try it yourself using AlphaFold Database!





https://alphafold.ebi.ac.uk/entry/Q9UQ13

How should I reason about the scores?

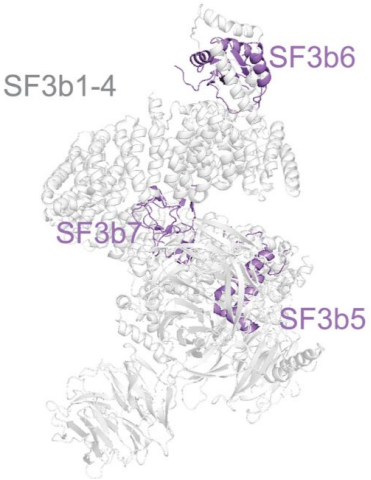
If you are a molecular biologist:

There are three levels of granularity with which to approach the scores:

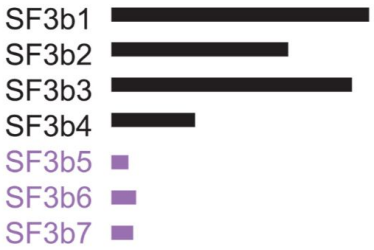
- Individual protein - all variants
  - Residue position x 20 possible amino acids (1 ref, 19 alt)
  - Heatmap visualization recommended
- Individual protein - position averages
  - Single value for each position in your protein
  - Use in conjunction with protein structure (measured or predicted) to aid in interpretation
- **Protein-level averages**
  - Single number for your protein of interest
  - Good for estimating the protein intolerance to loss-of-function variation, LOEUF-style

Predicting gene essentiality by averaging AM scores

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



Protein sequence length



		Cell essential	mean AM decile	Expected pLoF	LOEUF decile
SF3b1	1304 aa	Yes	0%	72.1	0%
SF3b2	895 aa	Yes	10%	53.2	0%
SF3b3	1217 aa	Yes	0%	66.4	0%
SF3b4	424 aa	Yes	0%	14.4	0%
SF3b5	86 aa	Yes	0%	2.7	50%
SF3b6	125 aa	Yes	0%	7.0	30%
SF3b7	110 aa	Yes	0%	6.6	20%

■ LOEUF powered
■ LOEUF underpowered

Limitations!

Shares some of the same limitations of AlphaFold2

- Certain length restrictions
- Certain non-canonical residues
- Lack of context - PPI, DNA, RNA, etc.

The 0 to 1 score does not inform on the following:

Clinical:

- Penetrance
- Haploinsufficiency
- Specific diseases or phenotypes
- Epistasis (combinations of variants)

Molecular biology:

- Biochemical details (stability vs. function)
- Not directly equivalent to a mutational experiment (which are tied to specific biochemical readout), though the information can be complementary
- Does not return the structures of proteins containing the alternative amino acid

Other notes:

- Gain of function variants are challenging
- Thresholds might need to be tweaked for particular use cases

Not a software: main artifact is the database

Webinar objectives

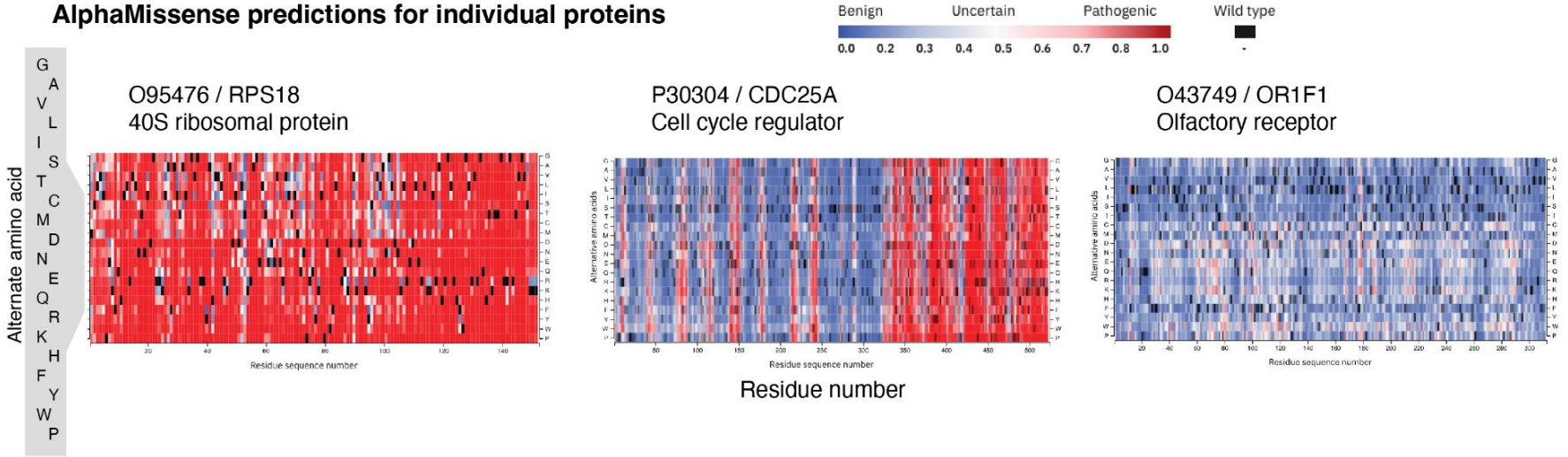
- *Describe the principles behind AlphaMissense and its approach to variant classification*
- *Identify how AlphaMissense utilises protein structure and sequence information for variant assessment*

Talk outline

- **Bonus: Field notes**

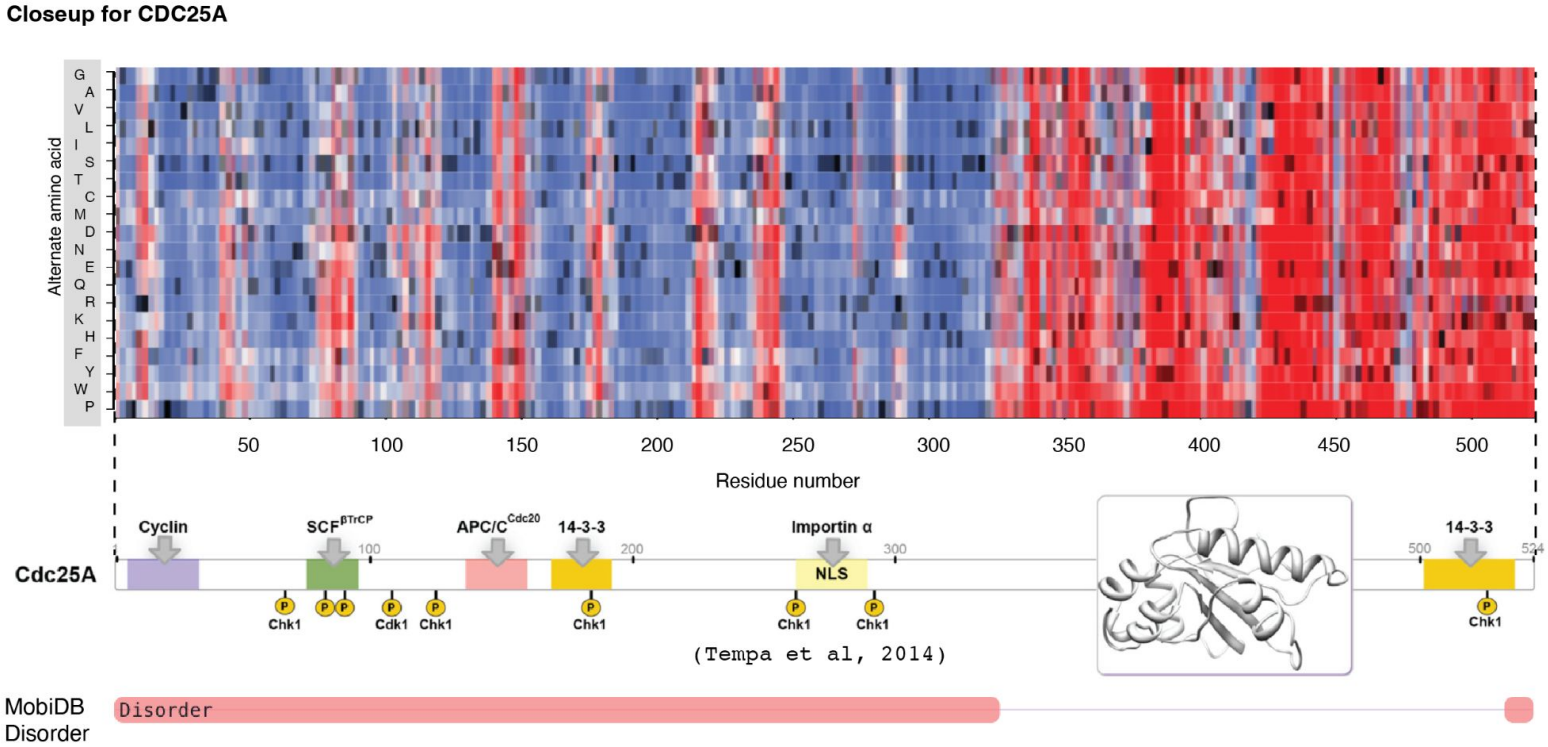
Field notes

(***Note***: These are anecdotes, not from the peer reviewed paper.)



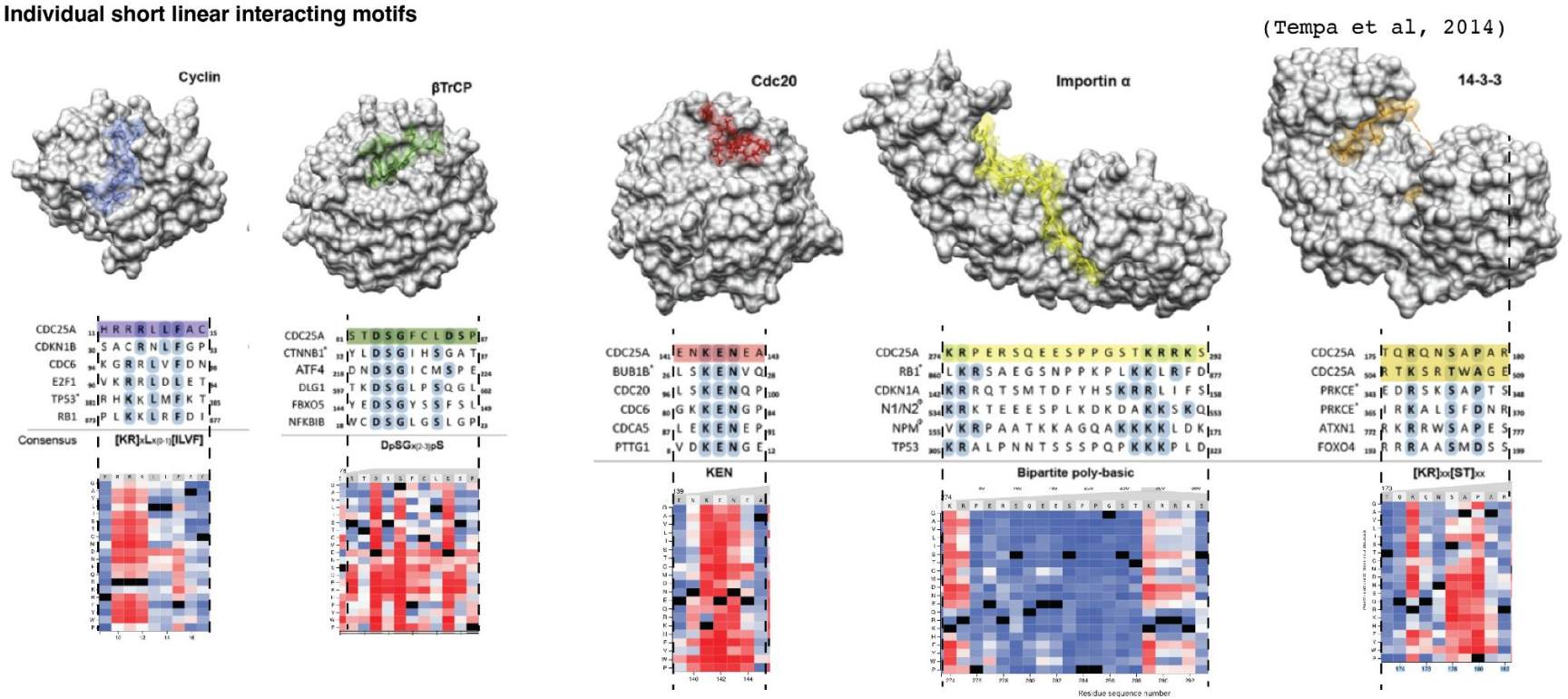
Field notes

(***Note***: These are anecdotes, not from the peer reviewed paper.)



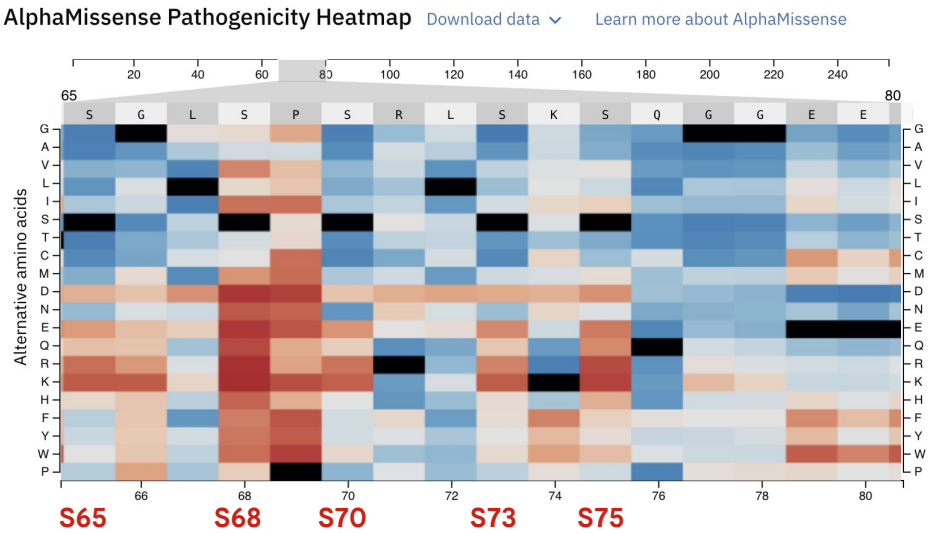
Field notes - interaction motifs

(***Note***: These are anecdotes, not from the peer reviewed paper.)

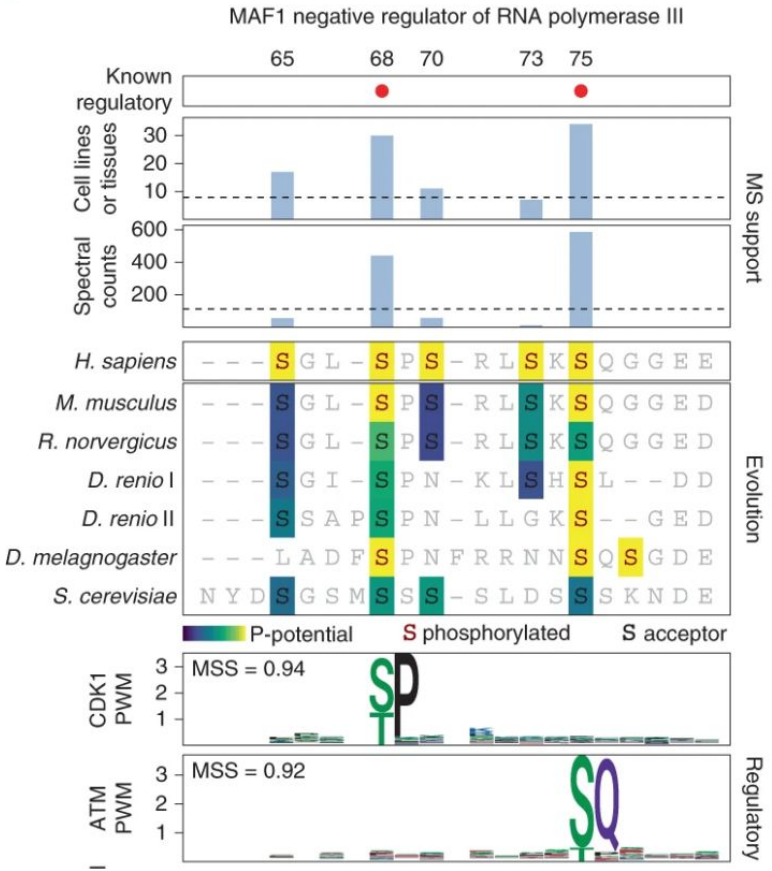


Field notes - phosphorylation sites

(***Note***: These are anecdotes, not from the peer reviewed paper.)



Ochoa et al, 2019



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.
- Increased the number of confidently classified variants (using ClinVar to estimate precision) proteome-wide
- Widely available predictions (Stay on the webinar to learn how to access!)

Citation

Cheng *et al.*, Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science***381**,eadg7492(2023). DOI:10.1126/science.adg7492

Acknowledgements

AlphaMissense

- Jun Cheng
- Guido Novati
- 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

Contact us:
alphamissense@google.com

Thank you.

Integration of AlphaMissense data into EMBL-EBI resources

7th June 2024

Paulyna Magaña
Bioinformatician

EMBL-EBI

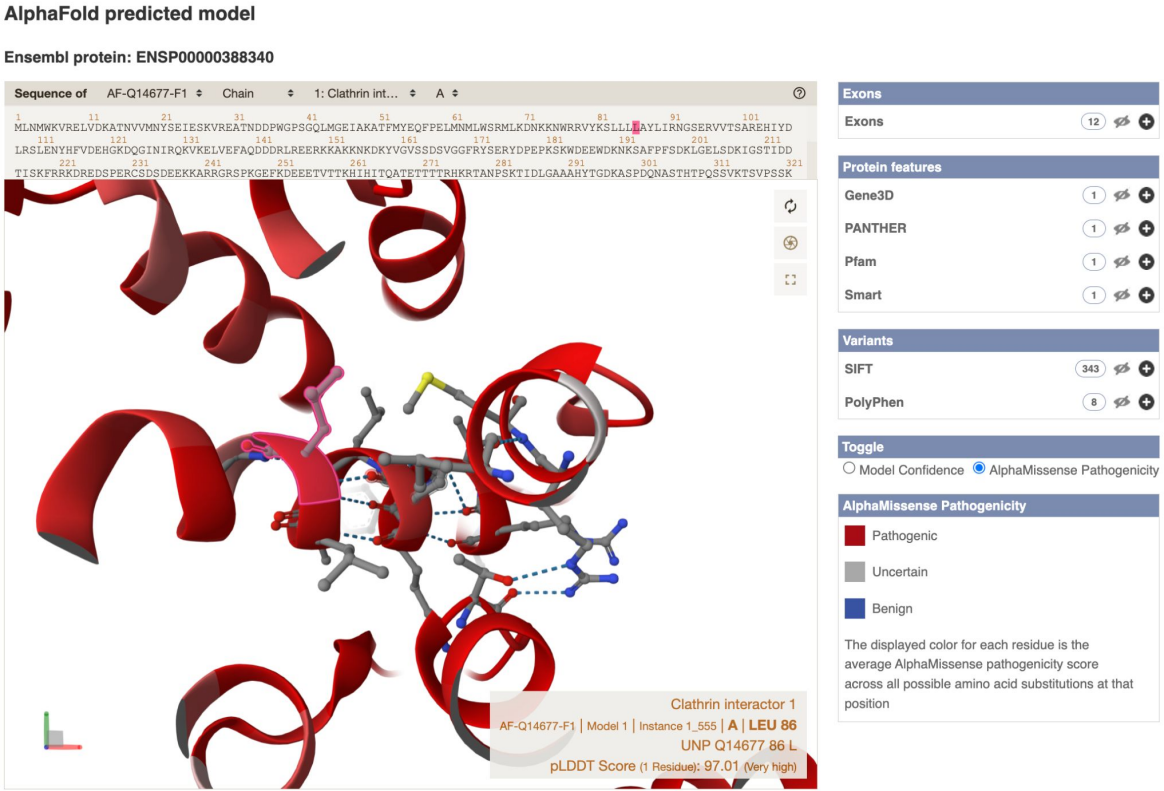
AlphaMissense in Ensembl

- The Ensembl genome browser (www.ensembl.org) supports research in comparative genomics, evolution, sequence variation and transcriptional regulation.
- AlphaMissense scores are integrated into the Ensembl Variant Effect Predictor (VEP) tool enabling annotation of variants via its web interface, REST API and the command line.

Location	Allele	Consequence	Symbol	Protein position	Amino acids	SIFT	PolyPhen	Pubmed	AlphaMissense classification	AlphaMissense pathogenicity score
7:117587806-117587806	A	missense_variant	CFTR	551	G/D	0.01	0.999	73 PubMed IDs	likely_pathogenic	0.9897
7:117587806-117587806	T	missense_variant	CFTR	551	G/V	0	1	73 PubMed IDs	likely_pathogenic	0.9593
10:62813413-62813413	A	missense_variant	EGR2	409	R/W	0	0.996	20301384 9537424 12525712	likely_pathogenic	0.9952
10:62813413-62813413	A	missense_variant	EGR2	409	R/W	0	0.996	20301384 9537424 12525712	likely_pathogenic	0.9952
14:73219188-73219188	G	missense_variant	PSEN1	435	L/V	0	-	-	likely_pathogenic	0.9848
14:73219188-73219188	T	missense_variant	PSEN1	435	L/F	0	-	-	likely_pathogenic	0.9978

AlphaMissense in Ensembl

- Average AlphaMissense pathogenicity scores for each amino acid can be visualised on the AlphaFold predicted 3D protein structure, available from the associated Ensembl transcript page.
- The interactive view allows switching between variants, domains, exons and AlphaMissense results to support interpretation of different regions of the protein and their sensitivity to change.



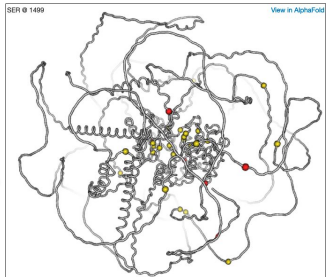
Ensembl

EMBL-EBI

AlphaMissense in DECIPHER

DECIPHER enables the sharing and interpretation of phenotype-linked genomic variants.

It aggregates the latest information about variants and makes it available via easy-to-use tools and visualisations.



EMBL-EBI

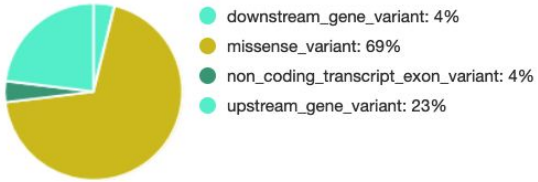
DECIPHER

Gene | Browser | Protein | Annotation | Matching patient variants 107 | Matching DDD research variants 0

Consequence prediction (VEP) | Allele frequency | Functional | ClinVar | Disease cohorts

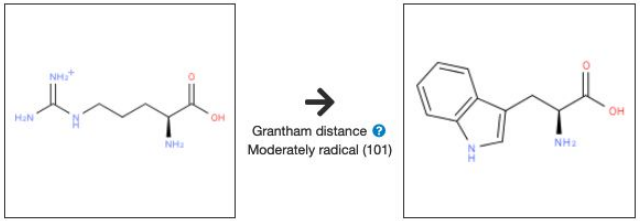
Ensembl Variant Effect Predictor (VEP)

Consequences



Amino acid substitution

Images generated by ChEBI



Arginine (R)
Positive, polar, hydrophilic

Tryptophan (W)
Uncharged, non-polar, hydrophobic

Annotations for CHD4: 1 to 26 of 26

Filter...

Transcript	Protein change	VEP Consequence	Other annotations
ENST00000544040.7 - NM_001273.5 MANE Select · Selected transcript ENST00000544040.7:c.1933C>T	R/W at position 645 of 1912 ENSP00000440542.2:p.Arg645Trp	missense_variant	Sift: Deleterious low confidence (0) PolyPhen: Probably damaging (0.994) CADD: 32 REVEL: 0.838 AlphaMissense: Likely Pathogenic (0.9183) SpliceAI: ≤ 0.2 phyloP: 2.765
ENST00000645022.1 NM_001297553.2 ENST00000645022.1:c.1912C>T	R/W at position 638 of 1905 ENSP00000496163.1:p.Arg638Trp	missense_variant	Sift: Deleterious low confidence (0) PolyPhen: Probably damaging (0.967) CADD: 32 REVEL: 0.838 SpliceAI: ≤ 0.2 phyloP: 2.765

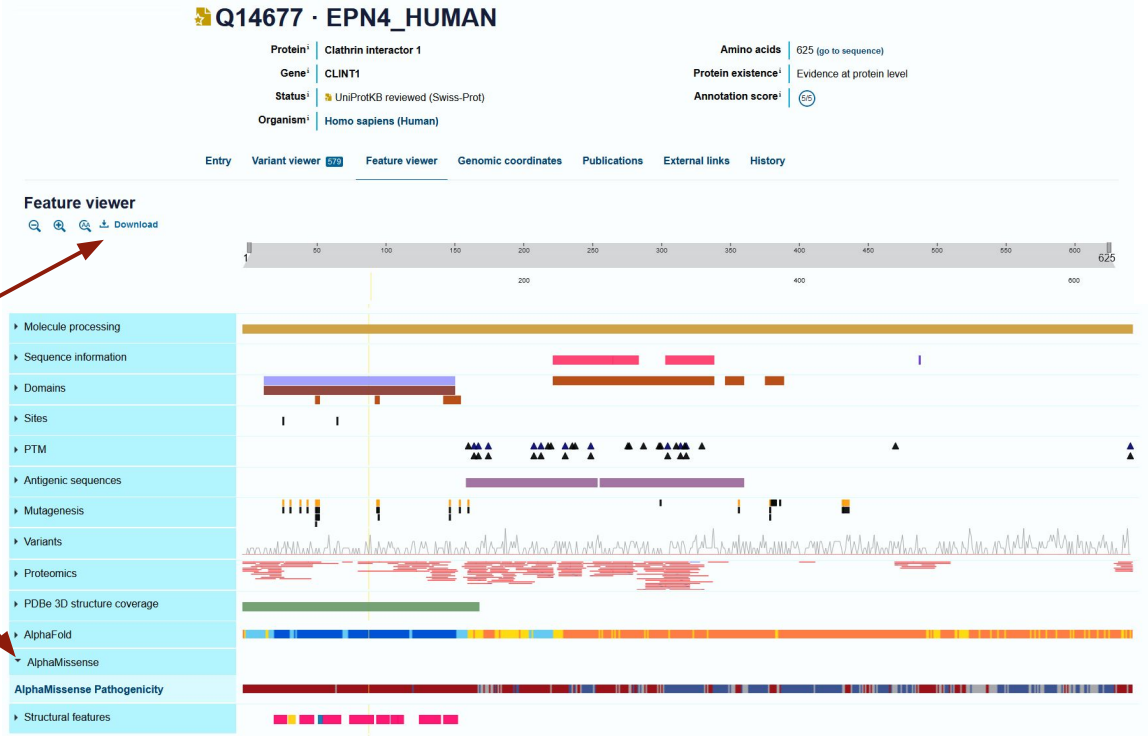
AlphaMissense scores and classifications of likely pathogenicity are displayed on DECIPHER variant pages.

They are only for the transcript which matches the UniProt protein used in score calculation.

EMBL-EBI

AlphaMissense in UniProt

- UniProt is a comprehensive, high-quality and freely accessible resource of protein sequence and functional information for a large variety of species.
- AM annotations can be downloaded by opening the Download panel.
- AlphaMissense scores are integrated in the feature viewer, allowing the visualization of overlaps of AM scores in context with protein features such as domains, motifs, PTMs or variants *(heatmap view coming soon)*.



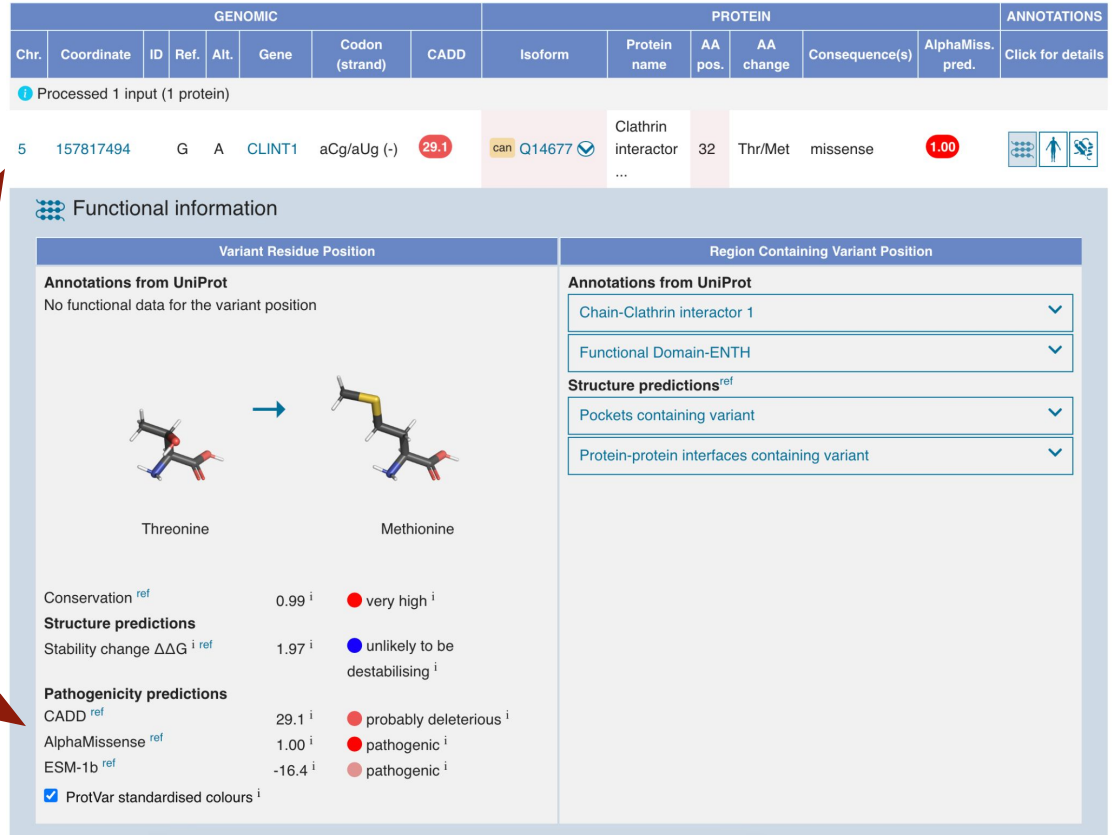
https://www.uniprot.org/uniprotkb/Q14677/feature-viewer

UniProt

EMBL-EBI

AlphaMissense in ProtVar

- ProtVar is a resource to investigate SNV missense variation in humans by presenting annotations which may be relevant to interpretation. Developed in the Protein Function Development group.
- AM predictions can be viewed for 1 or multiple variants on any given list of human proteins.
- AlphaMissense scores are shown alongside other pathogenicity prediction scores and other functional annotations from UniProt.

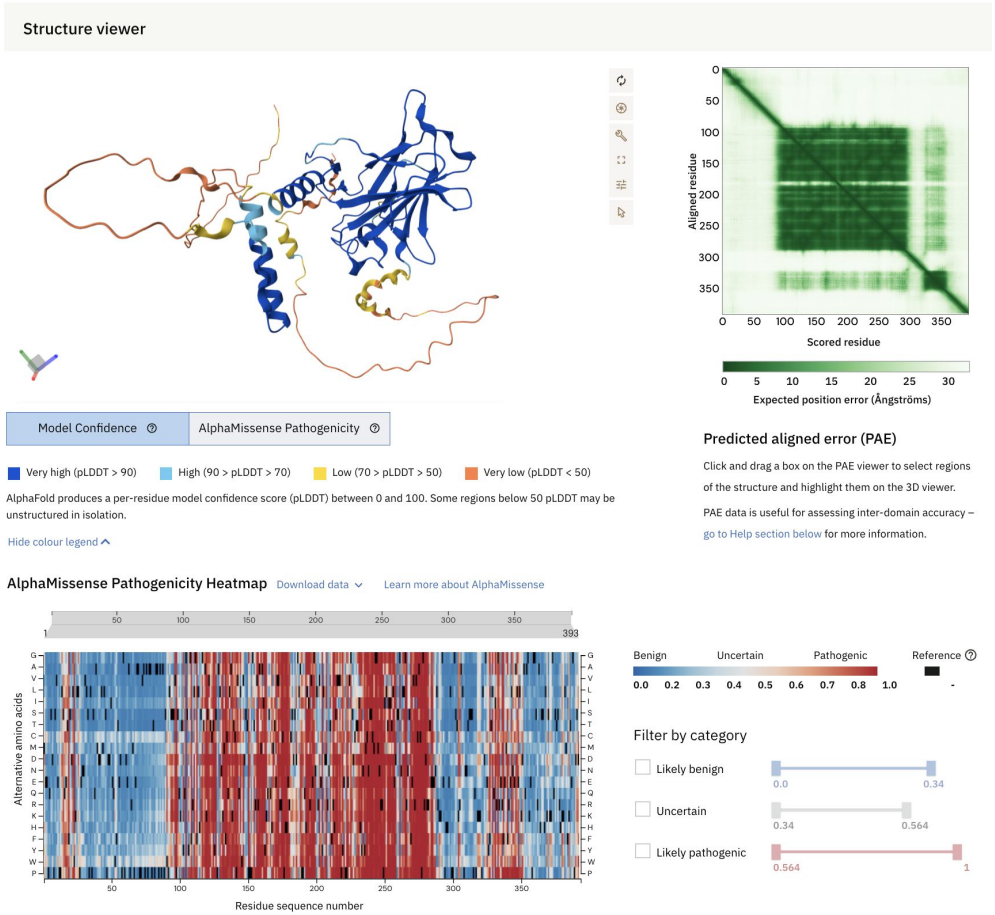


ProtVar

Open Targets

EMBL-EBI

AlphaFold Protein Structure Database



Amino acid-level confidence measure, pLDDT, a local accuracy metric.

Measures the confidence in the relative position of two amino acids in Ångströms.

Integration of AlphaMissense data

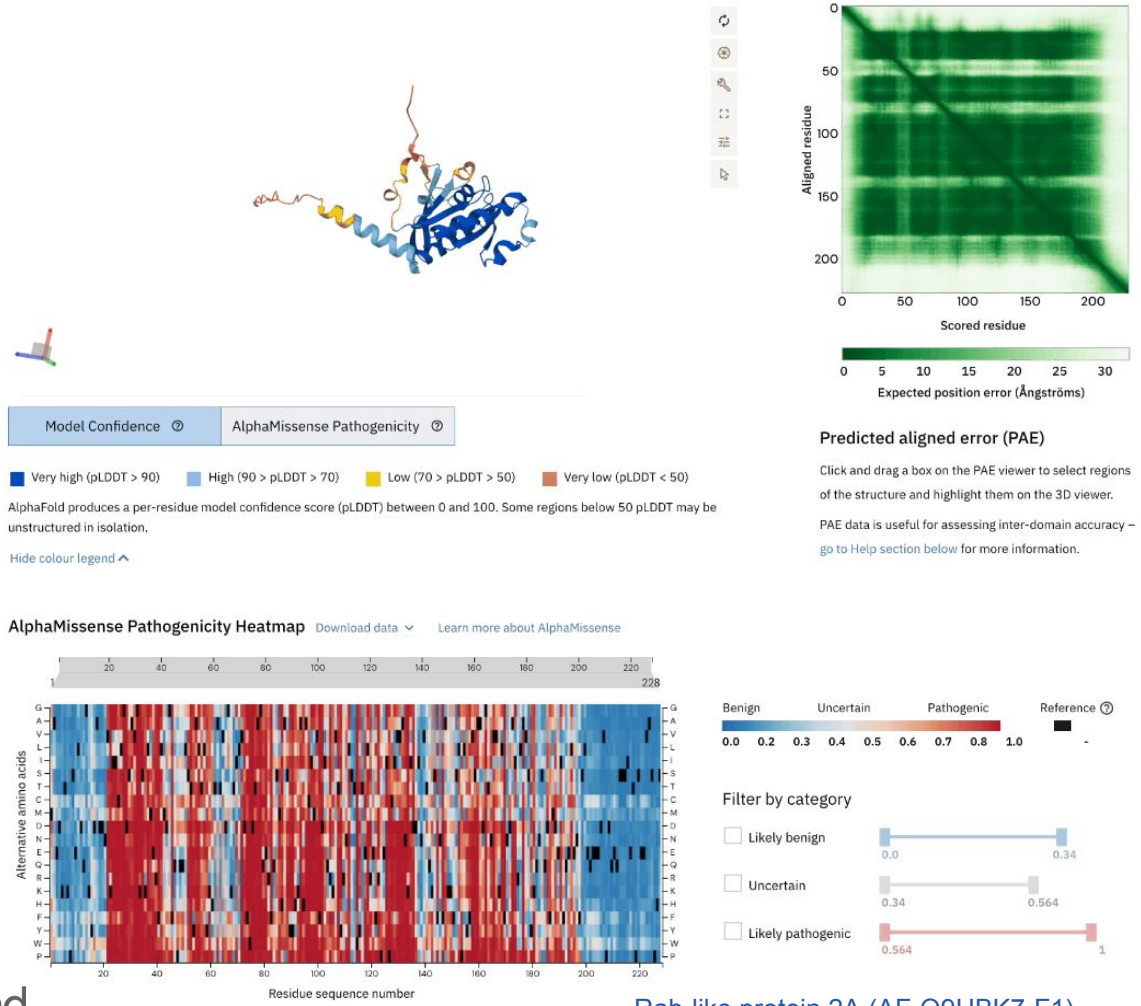
[Cellular tumor antigen p53 (AF-P04637-F1)]

Google DeepMind

EMBL-EBI

You can toggle between pLDDT and average AlphaMissense pathogenicity

Interactive heatmap
Zoom in and select residue of interest



Filter heatmap view by category and adjust range

Google DeepMind

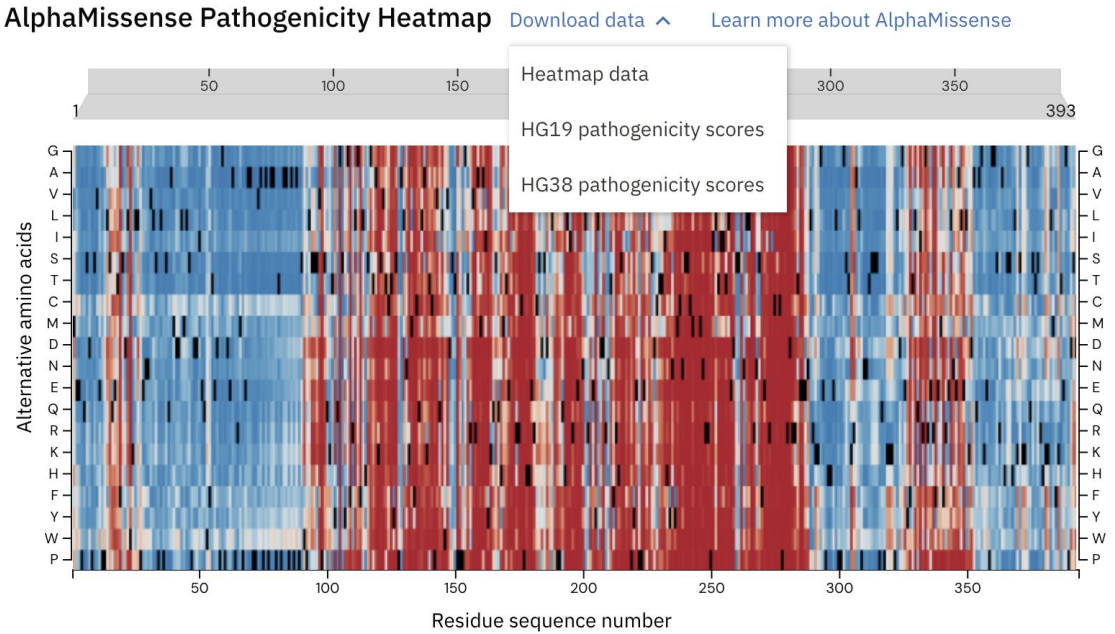
[Rab-like protein 2A (AF-Q9UBK7-F1)]

EMBL-EBI

Download AlphaMissense data

- **Heatmap data:** Contains the scores estimating the likelihood of pathogenicity and classifications for each possible amino acid substitution within the protein. **Used to visualise the heatmap you see on the entry pages.**

- **Pathogenicity scores** (HG19 and HG38). These files include AlphaMissense scores for all possible missense **single nucleotide variants across the human reference genome.** Each file corresponds to a specific genome assembly with information on specific genome position.



Google DeepMind

EMBL-EBI

Download AlphaMissense data via API

Response example for P04637:
https://alphafold.ebi.ac.uk/api/prediction/P04637?key=AIzaSyCeurAJz7ZGjPQUtEaerUkBZ3TaBkXrY94

Google DeepMind

EMBL-EBI

Thank you!

Integration of AlphaMissense data into EMBL-EBI resources

7th June 2024

EMBL-EBI