RESEARCH ARTICLE SUMMARY

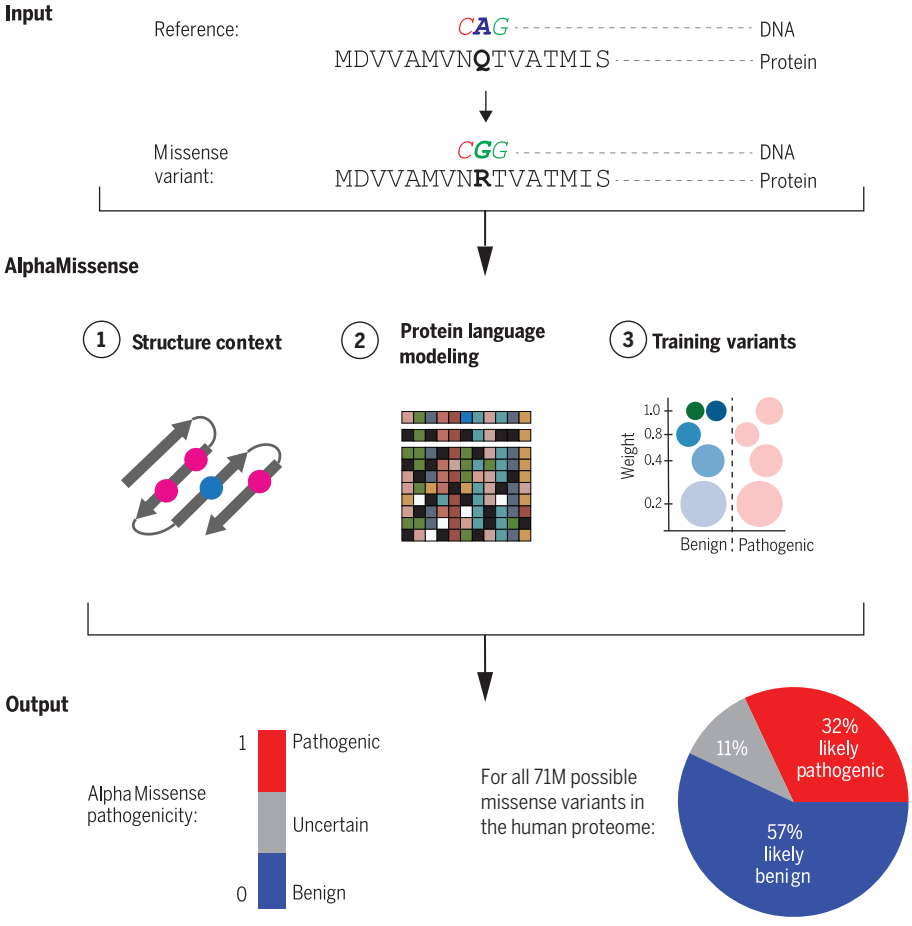
MACHINE LEARNING

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*

INTRODUCTION: Genome sequencing has revealed extensive genetic variation in human populations. Missense variants are genetic variants that alter the amino acid sequence of proteins. Pathogenic missense variants disrupt protein function and reduce organismal fitness, while benign missense variants have limited effect.

RATIONALE: Classifying these variants is an important ongoing challenge in human genetics. Of more than 4 million observed missense variants, only an estimated 2% have been clinically classified as pathogenic or benign, while the vast majority of them are of unknown clinical significance. This limits the diagnosis of rare diseases, as well as the development or application of clinical treatments that target the underlying genetic cause. Machine learning approaches could close the variant interpretation gap by exploiting patterns in biological data to predict the pathogenicity of unannotated variants. Specifically, AlphaFold, which accurately predicts protein structure from protein sequence, may be used as a foundation to predict the pathogenicity of variants on proteins.



AlphaMissense pathogenicity prediction. AlphaMissense takes as input a missense variant and predicts its pathogenicity. We fine-tuned AlphaFold on human and primate variant population frequency data and calibrated the confidence on known disease variants. AlphaMissense predicts the probability of a missense variant being pathogenic and classifies it as either likely benign, likely pathogenic, or uncertain. We provide predictions for all possible human missense variants as a resource for the community.

RESULTS: We developed AlphaMissense to leverage advances on multiple fronts: (i) unsupervised protein language modeling to learn amino acid distributions conditioned on sequence context; (ii) incorporating structural context by using an AlphaFold-derived system; and (iii) fine-tuning on weak labels from population frequency data, thereby avoiding bias from human-curated annotations. AlphaMissense achieves state-of-the-art missense pathogenicity predictions in clinical annotation, de novo disease variants, and experimental assay benchmarks without explicitly training on such data. As a resource to the community, we provide a database of predictions for all possible single amino acid substitutions in the human proteome. We classify 32% of all missense variants as likely pathogenic and 57% as likely benign using a cutoff yielding 90% precision on the ClinVar dataset, thereby providing a confident prediction for most human missense variants.

We show how this resource can be used to accelerate research in multiple fields. Molecular biologists could use the database as a starting point for designing and interpreting experiments that probe saturating amino acid substitutions across the human proteome. Human geneticists could combine gene-level AlphaMissense predictions with population cohort–based approaches to quantify the functional significance of genes, especially for shorter human genes where cohort-based approaches lack statistical power. Finally, clinicians could benefit from the boost in coverage of confidently classified pathogenic variants when prioritizing de novo variants for rare disease diagnostics, and AlphaMissense predictions could inform studies of complex trait genetics that use annotations of rare, likely deleterious variants.

CONCLUSION: AlphaMissense predictions may illuminate the molecular effects of variants on protein function, contribute to the identification of pathogenic missense mutations and previously unknown disease-causing genes, and increase the diagnostic yield of rare genetic diseases. AlphaMissense will also foster further development of specialized protein variant effect predictors from structure prediction models. ■

Google DeepMind, London, UK.
*Corresponding author. Email: jucheng@google.com (J.C.); pushmeet@google.com (P.K.); avsec@google.com (Ž.A.)
†These authors contributed equally to this work.
Cite this article as J. Cheng *et al.*, *Science* **381**, eadg7492 (2023). DOI: 10.1126/science.adg7492

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adg7492

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*

The vast majority of missense variants observed in the human genome are of unknown clinical significance. We present AlphaMissense, an adaptation of AlphaFold fine-tuned on human and primate variant population frequency databases to predict missense variant pathogenicity. By combining structural context and evolutionary conservation, our model achieves state-of-the-art results across a wide range of genetic and experimental benchmarks, all without explicitly training on such data. The average pathogenicity score of genes is also predictive for their cell essentiality, capable of identifying short essential genes that existing statistical approaches are underpowered to detect. As a resource to the community, we provide a database of predictions for all possible human single amino acid substitutions and classify 89% of missense variants as either likely benign or likely pathogenic.

Genome sequencing has revealed extensive genetic variation in human populations (*1*–*3*). Missense variants are genetic variants that alter the amino acid sequence of proteins. Pathogenic missense variants severely disrupt protein function and reduce organismal fitness, whereas benign missense variants have limited effects. Of the more than 4 million observed missense variants, only an estimated 2% have been clinically classified as pathogenic or benign. Classifying the remaining variants of unknown significance is an important ongoing challenge in human genetics (*3*). Lack of accurate missense variant functional predictions limits the diagnostic rate of rare diseases, as well as the development or application of clinical treatments that target the underlying genetic cause. Although multiplexed assays of variant effect (MAVEs) systematically measure protein variant effects (*4*) and can accurately predict the clinical outcomes of variants (*5*), a proteome-wide survey of variant pathogenicity remains incomplete because of the cost and labor required for MAVE experiments (*6*).

Machine learning approaches could close this variant interpretation gap by exploiting patterns in biological data to predict the pathogenicity of unannotated variants. Machine learning methods follow four broad strategies. The first class of methods train directly on human-curated variant databases (*7*–*10*), thereby leveraging prior knowledge to inform the status of unannotated variants. Such strategies will inherit biases from the human curators and previous in silico predictors, and they are prone to leaking data between training and test splits (*11*).

To overcome such circularity, the second class of methods train with weak labels that do not depend on human classification (*12*, *13*). In the training data, "benign" variants are defined as variants frequently observed in human or other primate species. The "pathogenic" class is approximated with hypothetical variants unobserved in the human population. Such an approach represents a promising direction to mitigate potential human curation biases. However, because the training data contain many false labels, such models require evaluation on more-reliable labels to assess their true performance.

A third class of methods avoid training on variant annotations directly and instead use unsupervised approaches to model the distribution of amino acids at a given sequence position conditioned on an amino acid sequence context (*14*–*16*). Recently, deep learning models that learn high-order dependencies between amino acids from protein sequences, such as autoencoders or language models, have achieved strong performance (*17*–*19*). In such models, pathogenicity is interpreted as the difference in predicted log-likelihood between reference and alternate sequences. Although such models effectively capture the distribution of naturally evolved sequences, they lack the state-of-the-art understanding of protein structure achieved by AlphaFold (AF) (*20*, *21*).

A fourth strategy is to exploit protein structure to reason about pathogenicity, as the structural context of an altered amino acid provides crucial information to interpret its effects on the protein. Initial explorations with predicted protein structures showed promise (*22*, *23*), and estimates of genetic evolutionary constraint have been aided by predicted protein structures (*24*). Although this strategy has improved genetic constraint quantification, using this approach for pathogenicity prediction directly has shown only moderate performance on ClinVar variants (*24*), likely because of low genetic diversity observed in current human sequence databases.

AF has recently shown that highly accurate protein structures can be predicted at scale using protein sequences as input (*21*, *25*). Such protein structure models may act as foundations for understanding other aspects of protein biology, such as variant pathogenicity. Although AF is largely insensitive to input sequence variation and cannot accurately predict structural changes upon point mutation (*26*), we hypothesized that AF's intrinsic understanding of multiple sequence alignments (MSAs) and protein structure provides a valuable starting point for models directly predicting the pathogenicity of missense variants.

Here, we present AlphaMissense, which combines the following elements of existing strategies: (i) training on weak labels from population frequency data, avoiding circularity by not using human annotations; (ii) incorporating an unsupervised protein language modeling task to learn amino acid distributions conditioned on sequence context; and (iii) incorporating structural context by using an AF-derived system. We achieve state-of-the-art predictions in clinical annotation, de novo disease variants, and experimental MAVE benchmarks, without explicitly training our model on such data. We predict and characterize the pathogenicity of all single amino acid substitutions in the human proteome and make these predictions available to the community.

AlphaMissense: Fine-tuning AlphaFold for variant effect prediction

AlphaMissense takes as input an amino acid sequence and predicts the pathogenicity of all possible single amino acid changes at a given position in the sequence. AlphaMissense leverages two key capabilities of AF: its highly accurate model of protein structure and its capacity to learn evolutionary constraints from related sequences (*21*). Accordingly, the implementation of AlphaMissense closely follows that of AF, with minor architectural differences (Fig. 1 and fig. S1; and see methods in the supplementary materials). Notably, AlphaMissense does not predict the structural changes of the mutated amino acid sequences but instead predicts pathogenicity as scalar values.

AlphaMissense is trained in two stages. In the first stage, the network is trained like AF to perform single-chain structure prediction (AF pretraining) along with protein language modeling by predicting the identity of the amino acids masked at random positions in

Google DeepMind, London, UK.
*Corresponding author. Email: jucheng@google.com (J.C.); pushmeet@google.com (P.K.); avsec@google.com (Ž.A.)
†These authors contributed equally to this work.

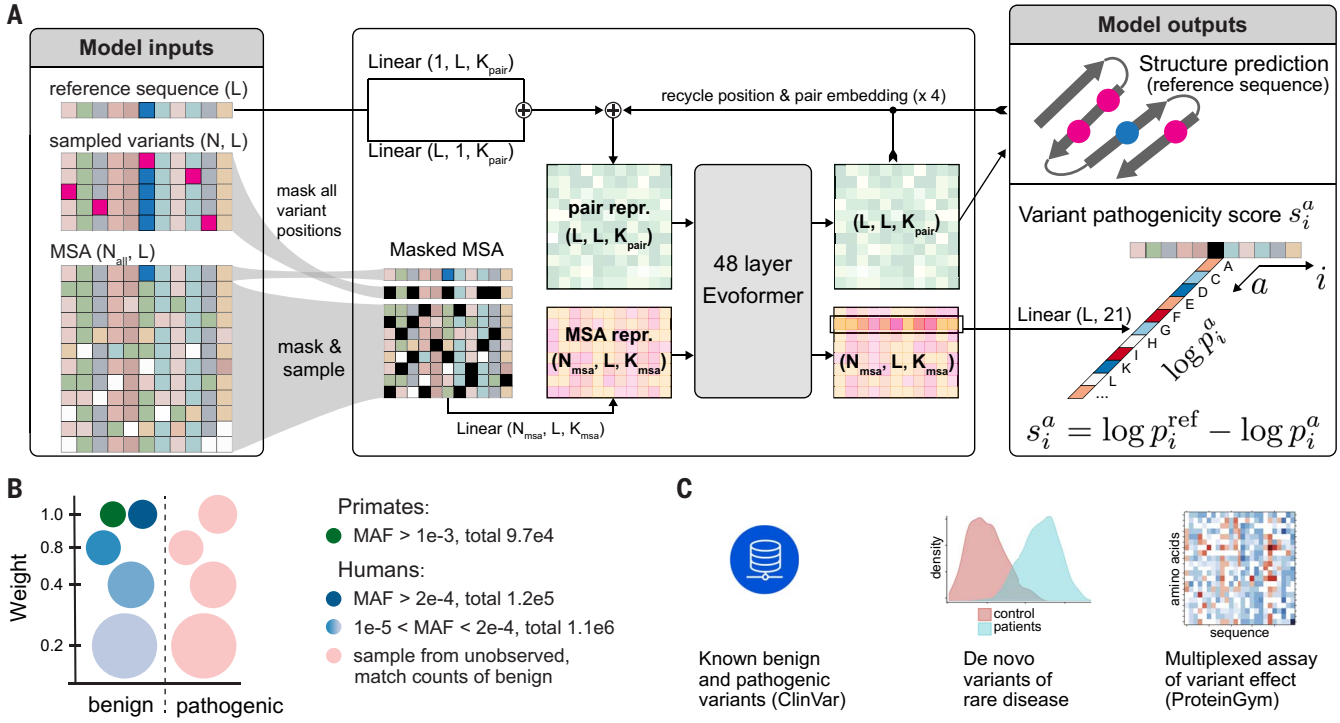


Fig. 1. Overview of AlphaMissense. (**A**) AlphaMissense architecture. The model inputs consist of the reference protein sequence [cropped to length (L) = 256 residues], a set of variants sampled from the training set for the same sequence (up to N = 50 variants), and multiple sequence alignments (MSAs, up to N_{all} = 2048). Inference is performed for one variant at a time (N = 1). The reference sequence is repeated in the second row of the MSA with all sampled variant positions masked (see methods). As in AlphaFold, the model constructs the pair representation (i.e., encodes information about two-way interactions between residues) from the reference sequence (embedding size K_{pair}), and the MSA representation from the masked MSA (embedding size K_{msa}). The MSA and pair representations are processed by a stack of Evoformer layers with recycling. Finally, the model predicts the structure of the reference sequence and the pathogenicity score (s_i^a) for the variant, which is derived from the masked residue prediction head as the log-likelihood difference between residue a relative to the reference residue at position i (see methods). (**B**) The pathogenicity score is fine-tuned as a binary classification of variants as benign (observed or frequent missense variants in human or primate populations) or pathogenic (unobserved human missense variants). We split the benign variants into clusters by their minor allele frequency (MAF) and introduce weights in the loss function that reduce the contribution of rare variants. For each observed variant in the benign set, we sample a missense variant from the pathogenic set and assign it the same loss weight as for the benign variant (see methods). (**C**) We evaluated AlphaMissense on a diverse set of benchmark datasets, including annotated missense variants in ClinVar (*30*), de novo disease variants (*54*), and MAVE data collected in ProteinGym (*19*).

the MSA. We introduced a few minor architecture modifications to AF and increased the loss weight toward the protein language modeling while still achieving structure prediction performance comparable to that of AF (see methods). After pretraining, the masked language modeling head can already be used for variant effect prediction by computing the log-likelihood ratio between the reference and alternative amino acid probabilities, as done in MSA Transformer (*27*) and Evolutionary Scale Modeling [ESM (*28*)].

In the second stage (Fig. 1A), the model is fine-tuned on human proteins with an additional variant pathogenicity classification objective defined for a variant sequence presented in the second row of the MSA (Fig. 1A). For the training set, we assign benign labels to variants frequently observed in the human and primate populations, and pathogenic labels to variants absent from human and primate populations, as is done in PrimateAI (*12*) (Fig. 1B; see methods). We stop training the model once it starts to overfit on the validation set (2526 ClinVar variants with an equal number of pathogenic and benign variants per gene; see methods).

Our training set is inherently noisy, because many unobserved variants are potentially benign, but it offers enough learning signal to improve the variant pathogenicity score compared with pretraining alone. To increase the quality and size of the training set, we employ self-distillation by using preliminary AlphaMissense models to filter out unobserved variants predicted to be likely benign. The fine-tuning stage is then repeated using this filtered training set (see methods). Further innovations include a custom classification loss function, sampling multiple variants during training, improving the matched sampling of variants, and weight decay during fine-tuning toward the pretrained parameter values (see methods and the ablation studies section below).

Improved pathogenicity classification across multiple clinical benchmarks

Clinical databases collect missense variants that cause human disease. These databases can be used to benchmark pathogenicity prediction models, but such data contain human biases and may misrepresent the true distribution of clinically relevant variants (see supplementary note in the supplementary materials). Models trained on these databases (ClinVar, for example) inherit these biases and often fail to generalize to other benchmarks (*11*, *29*). We avoid training directly on clinically curated labels to mitigate such issues and enable faithful evaluation on diverse benchmarks, including the held-out test set of annotated missense variants in ClinVar (*30*), de novo variants from patients with rare developmental

disorders and controls (*12*), MAVE benchmarks in ProteinGym (*19*), and additional MAVE benchmarks curated in this study (Fig. 1C; see methods).

We first evaluated our model on ClinVar missense variants. After balancing the number of pathogenic and benign variants per gene, AlphaMissense achieves an area under the receiver operator curve (auROC) of 0.940 on 18,924 ClinVar test variants, compared with an auROC of 0.911 achieved by the Evolutionary model of Variant Effect (EVE; P = 0.001, bootstrap), the next best model that did not train directly on ClinVar (*17*) (Fig. 2A). AlphaMissense also outperforms models trained directly on ClinVar, despite these models exhibiting data leakage and label circularity (Fig. 2A; see supplementary note) (*11*, *17*, *29*). Furthermore, we observe that AlphaMissense is capable of distinguishing pathogenic from benign ClinVar variants within regions of high evolutionary constraint (*31*), and it outperforms the best competing models on this task (ESM1b, P = 0.001, bootstrap) (fig. S2A). This result suggests that the model is not merely relying on identifying constrained domains but is capturing differences in the effect of individual variants within those domains. Our model performance is consistent across different AlphaFold confidence levels (fig. S2B). However, we note reduced performance on variants from residues predicted to be disordered (fig. S2C).

Clinical assessment of variants often focuses on specific disease-associated genes, and discriminating between benign and pathogenic variants within such genes is an important, clinically relevant task for predictive models. To understand AlphaMissense model performance on this task, we analyzed the 612 genes with at least five pathogenic and five benign variants in the ClinVar test set. For these genes, we calculated the gene-level auROC, which captures the model's performance at classifying variants within an individual gene. When evaluated in this way, AlphaMissense outperforms the next best method that did not train directly on ClinVar, EVE (*17*), with average gene-level auROC of 0.950 versus 0.921 (P = 0.001, bootstrap) (Fig. 2B).

We further assessed the performance of AlphaMissense on two important sets of proteins. The first set comprises proteins encoded by the clinically actionable genes prioritized by the American College of Medical Genetics (ACMG) (*32*), which has recommended that clinical exome and genome sequencing of these genes be returned as secondary findings in the clinic because of their clear disease phenotypes and highly penetrant mutations. For the 34 ACMG genes with sufficient ClinVar labels and scores from both methods, 26 genes (77%) see improvements using AlphaMissense pathogenicity predictions over EVE (fig. S3A). The second set are proteins prioritized for future MAVE studies by the community on the basis of clinical relevance and experimental tractability (*33*). For the 20 genes with sufficient ClinVar labels and scores from both methods, improvements were seen relative to EVE for 16 genes (80%) using AlphaMissense pathogenicity predictions (fig. S3B).

Finally, we evaluated AlphaMissense on the Deciphering Developmental Disorders (DDD) benchmark, where AlphaMissense achieves an auROC of 0.809, on par with PrimateAI (auROC = 0.797, P = 0.31, bootstrap) (*12*) (Fig. 2C). We also evaluated our model on classifying cancer hotspots, where AlphaMissense achieves an auROC of 0.907 compared with 0.885 for the next-best model, VARITY (P = 0.001, bootstrap) (*9*) (fig. S2D). Overall, AlphaMissense achieves state-of-the-art performance across all curated clinical benchmarks, whereas no other previously reported model consistently ranks highly across these benchmarks.

Calibrated AlphaMissense predictions expand the number of confidently classified variants relative to other methods

Having established state-of-the-art performance of AlphaMissense on clinical benchmarks, we next generated and analyzed proteome-wide predictions. We used AlphaMissense to predict the pathogenicity of all 216 million possible single amino acid changes across the 19,233 canonical human proteins, resulting in 71 million missense variant predictions saturating the human proteome (see methods).

Practical use of predicted scores requires careful calibration against the gold-standard set of clinically curated pathogenic and benign variants. We used the balanced validation set with 2526 variants from ClinVar (see methods) to calibrate our predictions using a univariate logistic regression model. This approach yields calibrated scores, as shown on the ClinVar test set (Fig. 2D; see methods). Calibrated AlphaMissense scores (ranging between 0 and 1) can be interpreted as the approximate probability of a variant being clinically pathogenic. We note that as the majority of predictions are close to 0 or 1, the calibrations for scores between 0.2 and 0.8 are likely less accurate.

Next, we used our calibrated prediction scores to classify variants into three discrete categories similar to ACMG terminology (*32*, *34*): likely pathogenic, likely benign, and ambiguous [cutoffs were chosen such that variants classified as likely pathogenic or likely benign have 90% expected precision estimated from ClinVar for both classes, as done in (*17*)] (fig. S4A). Owing to higher predictive performance, the fraction of ClinVar test variants that we can confidently classify with 90% precision is increased by 25.8 percentage points (from 67.1% to 92.9%) compared with the recent well-performing unsupervised model EVE. (*17*) (Fig. 2E and fig. S4B). This approach offers a major expansion in the number of variants with confident predictions in a proteome-wide context.

Overall properties and examples of AlphaMissense predictions

To understand the overall properties of the predictions, we compared them against the effective number of sequence alignments (N_{eff} score), genetic constraint, and predicted protein disorder (fig. S4, C to F). Residues with a low effective number of aligned sequences and hence lower conservation levels tend to have lower predicted pathogenicity (fig. S4C). This relationship is less pronounced when looking at aggregated protein-level results (fig. S4D), suggesting that AlphaMissense captures domain conservation within a protein, rather than overall protein-level evolutionary conservation. Similarly, variants located in evolutionarily constrained genes are systematically predicted as more pathogenic compared with those in unconstrained genes (fig. S4E). Variants located in structured regions, which may alter protein stability (*35*, *36*), are associated with higher pathogenicity scores than variants located in disordered regions (fig. S4F; protein disorder is predicted with AlphaFold). This is consistent with recent observations that known disease-causing variants are more likely to reside in thermally stable proteins (*37*).

To further understand properties of amino acid substitutions learned by AlphaMissense, we computed the mean predicted pathogenicity per amino acid substitution across all human proteins (fig. S4G). As expected, mutations in aromatic amino acids or cysteine are more likely to be pathogenic given their role in maintaining protein structure. The predicted substitution scores are asymmetric, as previously reported (*38*), and correlate with the BLOSUM62 (*39*) substitution matrix overall [correlation coefficient (r) = −0.61; fig. S4H] and per reference amino acid (fig. S4I). Together, these results suggest that the model is using both the structural information and evolutionary information present in the MSA to make predictions consistent with known biology.

We visualized pathogenicity predictions alongside ClinVar labels (Fig. 2, F and G, left panels) and AF predicted protein structures (Fig. 2, F and G, right panels). General trends can be observed for these specific proteins. For instance, structurally disordered regions are aligned with benign predictions and benign clinical annotations, consistent with the proteome-wide results (fig. S4F). In particular cases, the pathogenicity predictions make sense in light of the protein function. For example, we predict the transmembrane domain of ACVRL1 (amino acids 119 to 141) to be more tolerant to mutation than either of the globular domains, which represent enzymatic or protein-protein interaction sites (Fig. 2F).

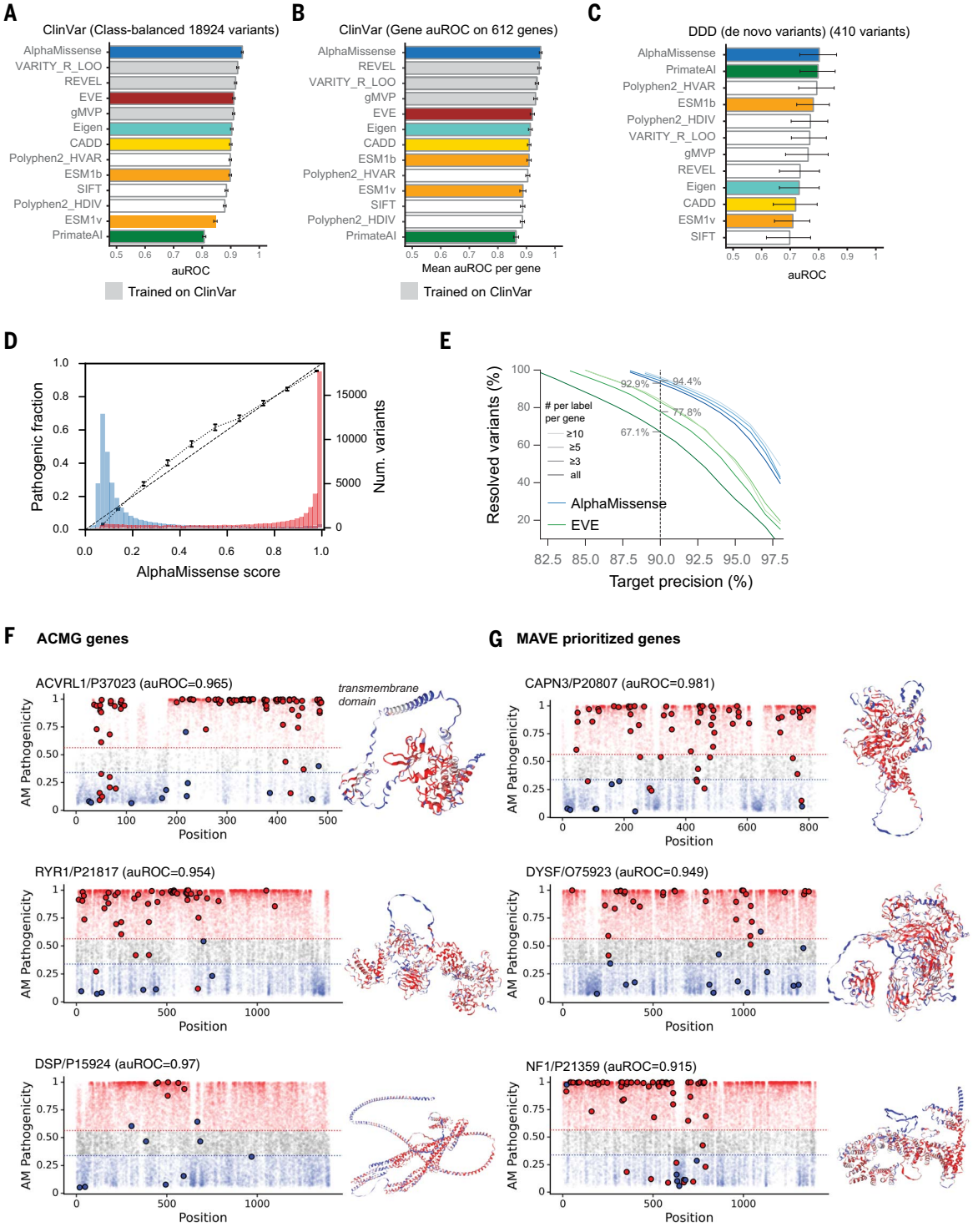


Fig. 2. Performance of AlphaMissense on clinically curated classification benchmarks. Benchmarks are evaluated by area under the receiver operator curve (auROC). Error bars show the 95% confidence interval of 1000 bootstrap resamples (see methods). A few manually chosen methods are colored to illustrate the relative position on different benchmarks. (**A**) Performance on classification of ClinVar variants (9462 pathogenic and 9462 benign variants from 999 proteins) balancing the number of positive and negative variants per gene. Methods shown in gray were trained directly on ClinVar. Some of their training variants are contained in this test set, so their performances are likely overestimated. Error bars show the 95% confidence interval of 1000 bootstrap resamples (see methods). (**B**) Average per-gene auROC on the ClinVar test set. A total of 612 proteins with at least five benign and five pathogenic ClinVar test variants are considered. (**C**) Comparison of AlphaMissense and other predictors on distinguishing de novo variants from DDD cohort patients and healthy controls (*12*). A total of 353 patient variants and 57 control variants from 215 DDD related genes are considered. We excluded EVE because of its low coverage of variants in this dataset (227/410 variants). (**D**) The AlphaMissense scores were calibrated on the class-balanced ClinVar validation set (see methods). The figure shows the calibration curve, which plots the average score against the fraction of pathogenic variants per bin, computed on

the ClinVar test set (82,872 variants). The error bars represent 95% confidence intervals computed from 1000 bootstrap resamples. The histograms show the distribution of scores among pathogenic (red) and benign (blue) variants. (**E**) Fraction of resolved (unambiguous) missense variants at different levels of target precision. Precision is defined as the fraction of true predictions in both pathogenic and benign class prediction. The resolved fractions are computed with ClinVar test set variants from proteins scored by EVE (dark lines, all) and then filtered to proteins with at least 3, 5, or 10 variants for each label (lighter lines). (**F**) Example proteins chosen from ACMG clinically actionable genes (*32*). Protein names are written as "[HUGO symbol]/[Uniprot accession ID]." (Left panels) Missense variants, represented as points, are plotted against AlphaMissense (AM) pathogenicity scores (*y* axis) and amino acid positions (*x* axis). Variants predicted as likely pathogenic are shown in red, variants predicted as likely benign are shown in blue, and ambiguous variants are shown in gray. If a variant contains a clinical label in ClinVar, it is plotted as a solid circle. For proteins longer than 1400 amino acids, the first 1400 are shown. (Right panels) The protein structure prediction from AlphaFold is shown for the selected region. Each residue in the predicted structure is colored according to the average AlphaMissense pathogenicity score of that residue (out of 19 possible amino acid changes per residue). See also fig. S3A. (**G**) The same as (F), but for examples chosen from genes prioritized by the MAVE community for further study (*33*). See also fig. S3B.

AlphaMissense achieves state-of-the-art agreement with multiplexed assays of variant effect

MAVE experiments generate "proactive" maps of variant effects (*40*) by expressing protein variants in cells and measuring activity using growth or fluorescence readouts. Because MAVE experiments densely cover (and often saturate) the protein of interest, they provide valuable information on protein regions otherwise missed by sparse clinical curations, although the direct clinical utility of MAVE data depends on the assay readout and experimental quality (*41*).

To assess the agreement between AlphaMissense and MAVE studies, we benchmarked predictions against two sources of MAVE data: 1.5 million variants from 72 proteins collected in ProteinGym (*19*) and an additional benchmark set consisting of 20 recently published human proteins not contained in ProteinGym (see methods). Relative to other methods, AlphaMissense agrees with MAVE data the most strongly (mean Spearman correlation on ProteinGym: 0.514; on the additional MAVE benchmark: 0.450; Fig. 3, A to C). When restricting to only those amino acid variants from 25 human proteins that are scored by all methods, AlphaMissense remains the highest scoring in ProteinGym out of the 13 methods (mean Spearman correlation: 0.474; Fig. 3B). AlphaMissense improves predictions for most proteins within both benchmarks compared with the next-best model [62/72 relative to the Global Epistatic Model for Predicting Mutational Effects (GEMME) (*16*) in ProteinGym, 60/72 relative to EVE in ProteinGym, and 13/20 relative to ESM1v in the additional MAVE benchmark] (fig. S5, A and B).

We compared the observed MAVE data and available model predictions against the experimentally resolved protein structures and domain annotations for disease-relevant proteins. The SHOC2 protein forms a complex with MRAS and PP1C to activate the Ras-MAPK (mitogen-activated protein kinase) signaling pathway in cancer (*42*). AlphaMissense pathogenicity correlates with MAVE data that measure the impact of SHOC2 variants on Ras-activated cancer cell fitness (*43*) (Spearman correlation: 0.47), outperforming ESM1v, ESM1b, and EVE (Spearman correlation: 0.41, 0.40, and 0.32, respectively; fig. S5B).

We investigated whether AlphaMissense better captures pathogenicity driven by specific domains within SHOC2, which would be reflected by the average pathogenicity at each amino acid position. AlphaMissense per-position average pathogenicity agrees strongly with the MAVE per-position average (positional Spearman correlation: 0.64), outperforming ESM1b, ESM1v, and EVE (positional Spearman correlation: 0.56, 0.55, and 0.48, respectively; fig. S5C). Of the first 80 amino acids of SHOC2, positions 63 to 74 were pathogenic according to the MAVE assay (Fig. 3D). This region was structurally shown to bind PP1C through an RVxF motif (*43*) (Fig. 3E). AlphaMissense is the only model to correctly predict pathogenic effects of mutations in this functionally important region (Fig. 3D and fig. S5D). Additionally, after the 80th position, both the MAVE data and AlphaMissense predictions peak in pathogenicity approximately every 23 amino acids (Fig. 3D), corresponding to the 20 leucine-rich repeat domains that contact MRAS and PP1C approximately every 23 amino acids (Fig. 3, D and E). Overall, residues directly contacting either MRAS or PP1C score as highly pathogenic (median AlphaMissense pathogenicity: 0.98 and 0.96, respectively), nearly as highly as core hydrophobic residues (median AlphaMissense pathogenicity: 0.99) and higher than surface residues that do not form protein-protein contacts (median AlphaMissense pathogenicity: 0.51; fig. S5E).

Next, we sought to determine whether the average substitution effect of each of the 20 possible amino acids, driven by their chemical properties, is better reflected in our model. For SHOC2, AlphaMissense agrees most strongly with the measured per–amino acid average substitution effect compared with other models (fig. S5C). Overall, when calculated this way across all proteins in ProteinGym and the additional MAVE benchmark, AlphaMissense displays the highest average performance across both the amino acid substitution and the positional metrics, suggesting that improvements in domain-level pathogenicity prediction and amino acid properties both underlie model performance (mean positional Spearman correlation on ProteinGym: 0.54; mean substitution Spearman correlation on ProteinGym: 0.545; fig. S5F).

Another example protein is the human glucose sensor GCK. Variants that decrease GCK activity can cause maturity-onset diabetes of the young (MODY) (*44*). AlphaMissense pathogenicity correlates with MAVE data measuring the fitness of auxotrophic yeast strains expressing human GCK variants in the presence of glucose (*45*) (Spearman correlation: 0.53), outperforming ESM1v, EVE, and ESM1b (Spearman correlation: 0.49, 0.48, and 0.45, respectively; fig. S5B). GCK primarily functions to catalyze glucose; the catalytic residue Asp^{205} (D205) is the highest-ranked residue by average AlphaMissense pathogenicity (0.999), and other residues in direct contact with the ligand were similarly pathogenic (Fig. 3F). AlphaMissense pathogenicity is associated with decreased fasting glucose in patients harboring missense variants in GCK (Spearman correlation: −0.49) (*45*). AlphaMissense pathogenicity exhibits a log-linear relationship with in vitro GCK activity measurements for 36 clinical variants (Spearman correlation: −0.65; Fig. 3G), falling short of experimental accuracy, as estimated by the MAVE data (Spearman correlation: 0.75), but closer than other prediction methods (Spearman correlation for ESM1v: 0.61; ESM1b: 0.50; EVE: −0.50; fig. S5G). Highly pathogenic variants according to AlphaMissense exhibit orders of magnitude lower GCK activity, consistent with the fact that most of the clinically confirmed pathogenic GCK variants are MODY variants with decreased activity. On the other hand, a small number of hyperactive pathogenic variants [clustered near the allosteric site, e.g., Thr^{65}→Ile (T65I)] (Fig. 3F and fig. S5H) can cause hyperinsulinemic hypoglycemia (*44*). AlphaMissense more often classifies these as ambiguous or benign (Fig. 3G).

Ablating components of AlphaMissense reveals key drivers of performance

Given the improved performance of AlphaMissense on different benchmarks, we next investigated which components are necessary for its high performance on ClinVar and ProteinGym test sets by systematically removing components of the model in an ablation study. We focused on three types of components:

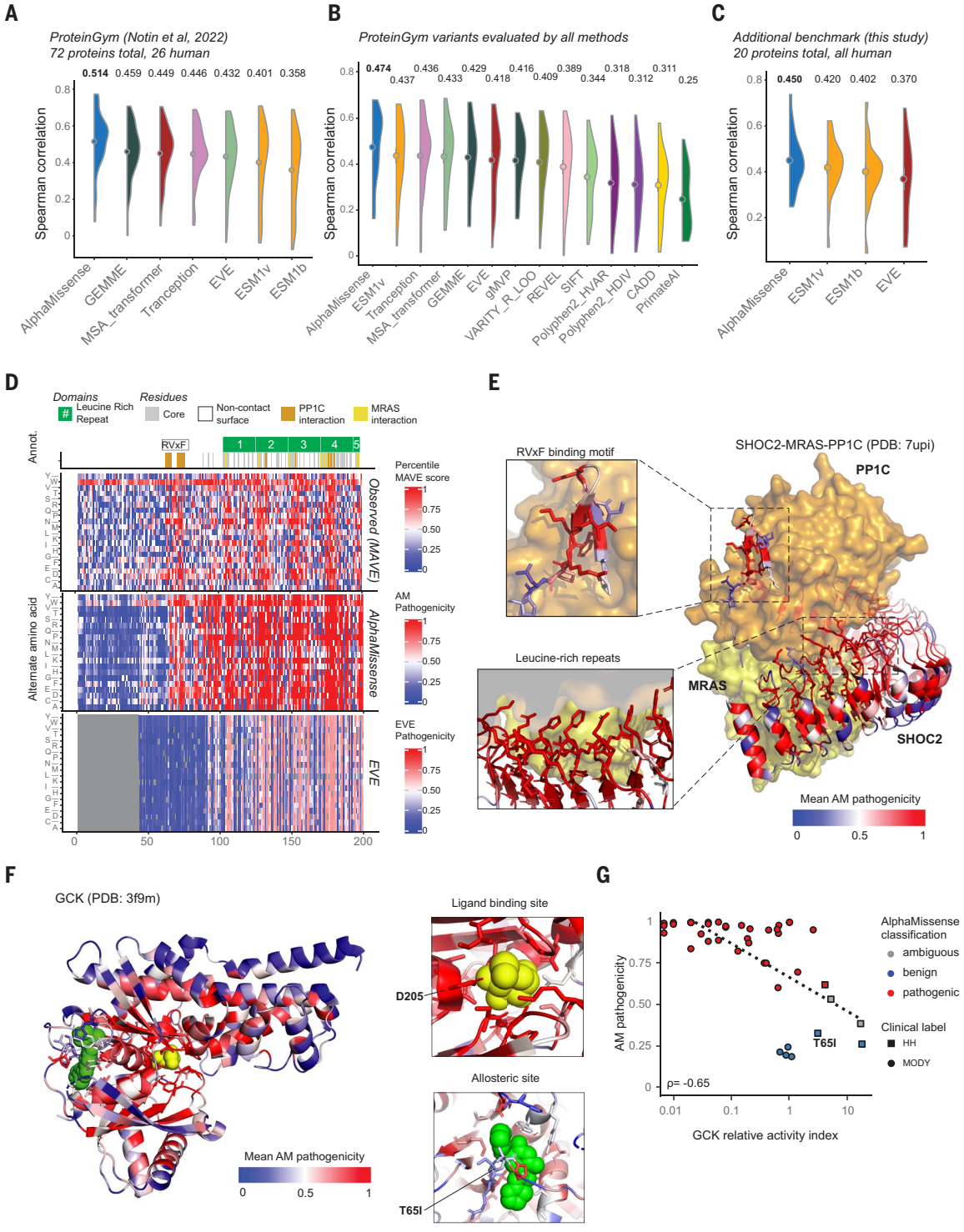


Fig. 3. AlphaMissense achieves state-of-the-art agreement with multiplexed assays of variant effect. (**A**) Performance on MAVE benchmarks. ProteinGym (*19*) is a collection of 72 proteins with MAVE data. The distribution of per-protein Spearman correlations between predictions and ProteinGym MAVE data for each model is shown, with mean value shown as a dot (and numerically above the violin plot). (**B**) Performance comparison on a subset of ProteinGym variants (608,175 variants, 25 human proteins) that were scored by all methods. Dots represent mean Spearman correlation across proteins per method, which are also represented numerically above each violin plot. (**C**) We curated an additional benchmark dataset of 20 human proteins not included in ProteinGym. The distribution of per-protein Spearman correlations between predictions and additional MAVE data is shown. (**D**) Heatmaps of observed and predicted effects of amino acid substitutions on the first 200 amino acids of SHOC2. (Top heatmap) Observed pathogenicity as measured by a MAVE assay of cell growth in cancer cells dependent on SHOC2 (*43*). Scores are percentile normalized measurements from the experimental assay. Variants with scores closer to zero (blue) retain SHOC2 function, whereas scores closer to one (red) lose SHOC2 function. (Middle and bottom heatmaps) AlphaMissense (AM) and EVE pathogenicity scores, respectively. Both scores range from zero to one, with higher scores corresponding to increased pathogenicity. Variants with no

prediction are colored gray (see EVE heatmap). Domain-level annotations (Annot.), including RVxF and leucine-rich repeat (LRR) regions, are shown above the heatmaps. Residue-level annotations are also shown [as calculated in (*43*) from Protein Data Bank (PDB) ID 7UPI], representing surface, core, and protein-protein interaction residues. (**E**) Experimentally derived structure of SHOC2 (blue and red) in complex with MRAS (yellow) and PP1C (gold) [PDB ID 7UPI (*43*)]. The mean AlphaMissense pathogenicity score per position is shown in the SHOC2 structure, with blue corresponding to benign and red corresponding to pathogenic. (Insets) Close-ups of the RVxF binding region of SHOC2 contacting PP1C, and the LRR region contacting MRAS and PP1C. (**F**) Experimentally derived structure of GCK (blue and red) [PDB ID 3F9M (*55*)]. The mean AlphaMissense pathogenicity score per position is shown in the GCK structure, with blue corresponding to benign and red corresponding to pathogenic. The active site ligand (yellow) and an allosteric inhibitor (green) are also shown. (Insets) Close-ups of residues that contact the ligand (such as D205, the catalytic site) and the residues that bind the allosteric inhibitor (such as T65I). (**G**) Comparison of relative activity index for glucokinase mutants (*56*) against AlphaMissense pathogenicity. On the log *x* axis, a score of one indicates in vitro activity equivalent to wild type, a score lower than one indicates less activity, and a score above one indicates hyperactivity. Spearman correlation is shown in the lower left of the panel. Each dot represents a different protein variant, colored according to AlphaMissense classification thresholds. The shape indicates the clinical label (*45*). The dashed line shows the linear fit between in vitro measurement and AlphaMissense pathogenicity. T65I, which causes hyperinsulinemic hypoglycemia (HH), is labeled. MODY, maturity-onset diabetes of the young.

structure prediction, variant sampling, and training data.

We found that both AF pretraining and fine-tuning stages are essential for good performance ("No AF pretraining" and "No fine-tuning on missense variants" in fig. S6). Furthermore, we found that pretraining with masked MSA alone without structure prediction is not sufficient for good performance ("No structure loss during AF pretraining"), suggesting that both structure prediction and the protein language modeling across a large corpus of samples contributed to the overall performance (fig. S6). Sampling variants to account for gene bias in the training set and sampling multiple variants with the training sequence crop are both important to reduce gene-level bias and regularize the model (fig. S6). Variant self-distillation helped on the ProteinGym task but not on the ClinVar task. Similarly, we found that additional training variants from primates or the extremely low minor allele frequency (MAF) variants from humans are only mildly helpful on the ProteinGym task and not on ClinVar (fig. S6). Overall, these results emphasize the importance of both training stages: pretraining on a large database of structures and fine-tuning directly for the target application.

Gene-level AlphaMissense pathogenicity predicts cell essentiality

An important endeavor in human genetics is quantifying the functional significance of a protein in human survival or fitness over evolutionary time. A common approach is to measure, among a healthy population cohort, depletion in the observed number of variants that likely ablate, or severely disrupt, the function of a protein compared with the expectation under neutral selection (*1*, *3*, *31*). However, the reliability of such estimates depends on the expected number of such variants in a gene, which in turn depends on the coding sequence length (*3*). As noted by the authors of one such approach, LOEUF (loss-of-function observed/expected upper bound fraction) (*3*), many genes are too small for the metric to be a reliable measure in current sample sizes (22% of protein coding genes; Fig. 4A). Given the observation that the average AlphaMissense pathogenicity of all possible missense variants within a gene is correlated with LOEUF (Spearman correlation: −0.48, $P < 2.2 \times 10^{-16}$; fig. S4E), we investigated whether AlphaMissense is capable of predicting genes known to be sensitive to functionality-altering perturbations in humans, particularly among ~4000 genes that would otherwise be underpowered in population cohort–based approaches.

Overall, we find that a gene's average AlphaMissense pathogenicity shares similar properties with LOEUF across a broad range of biological measures of intolerance to perturbation in humans, such as depletion in observed large structural deletions, and an enrichment of genes known to cause severe developmental disorders among more-pathogenic genes (fig. S7; see supplementary note). Furthermore, most of the properties of genes in the most-pathogenic decile of AlphaMissense predictions remain consistent among genes underpowered for LOEUF, supporting the generalizability of the scores to an additional 4252 small genes (fig. S7; see supplementary note). Genes experimentally identified as essential to cell survival across a variety of human cell lines (*46*) showed a strong enrichment among the most-pathogenic decile of AlphaMissense. The enrichment is both stronger than LOEUF (3.8-fold versus 2.3-fold enrichment) for the most-pathogenic decile and remains significant among smaller genes [5.9-fold, hypergeometric *P* value (P_{hyper}) = 5.6×10^{-46}; Fig. 4C]. AlphaMissense outperforms LOEUF and PhyloP, a conservation-based measure (*47*), at distinguishing experimentally determined cell-essential from nonessential genes (*48*) in the context of smaller genes (Fig. 4B). AlphaMissense achieves an auROC of 0.88 versus 0.81 for LOEUF (P = 0.001, bootstrap), while maintaining performance among the rest of the proteome (auROC of 0.80 versus 0.82, P = 0.092). The advantage of AlphaMissense in this context is exemplified by the spliceosome protein complex SF3b, which involves a gene with one of the highest average AlphaMissense pathogenicity scores, *PHF5A/SF3B7* (Fig. 4, D to F and data S1). All seven primary protein components (*49*) are experimentally classified as cell-essential (*48*). Four of these are sufficiently large genes for LOEUF to reliably predict their functional importance, and they are all strongly depleted for observed predicted loss of function (pLoF) variants (in the lowest decile of LOEUF) (Fig. 4, E and F). The other three are small genes (maximum: 125 amino acids), such that LOEUF is too underpowered to be informative. AlphaMissense predicts all three of these subunits to be more pathogenic than 96% of all human protein coding genes (Fig. 4F).

Together, these observations indicate that methodology that combines both AlphaMissense predictions and population cohort–based approaches could be beneficial for quantifying functional significance, especially for the large subset of short human genes where population cohort–based approaches lack statistical power.

AlphaMissense predictions as a community resource

We have released four resources for the research community. The first is a dataset of 71 million missense variant predictions saturating the human proteome. Each missense variant is defined by the single nucleotide change resulting in a changed amino acid (Fig. 5A). Out of the 71 million missense variants, 32% (22.8 million) are classified as likely pathogenic and 57% (40.9 million) as likely benign, using score cutoffs achieving 90% precision on the ClinVar dataset (fig. S4A). We note that choice of the cutoff can be adjusted by users to better match different use cases or accuracy trade-offs, or to achieve the desired precision on a different labeled dataset. The second resource is gene-level AlphaMissense pathogenicity predictions, defined as the average pathogenicity over all possible missense variants in a gene. The third is the expanded dataset of all 216 million possible single amino acid substitutions across the 19,233 canonical human proteins. Finally, we provide predictions for all possible missense variants and amino acid substitutions across 60,000 alternative transcript isoforms for future research and evaluation of isoform-specific

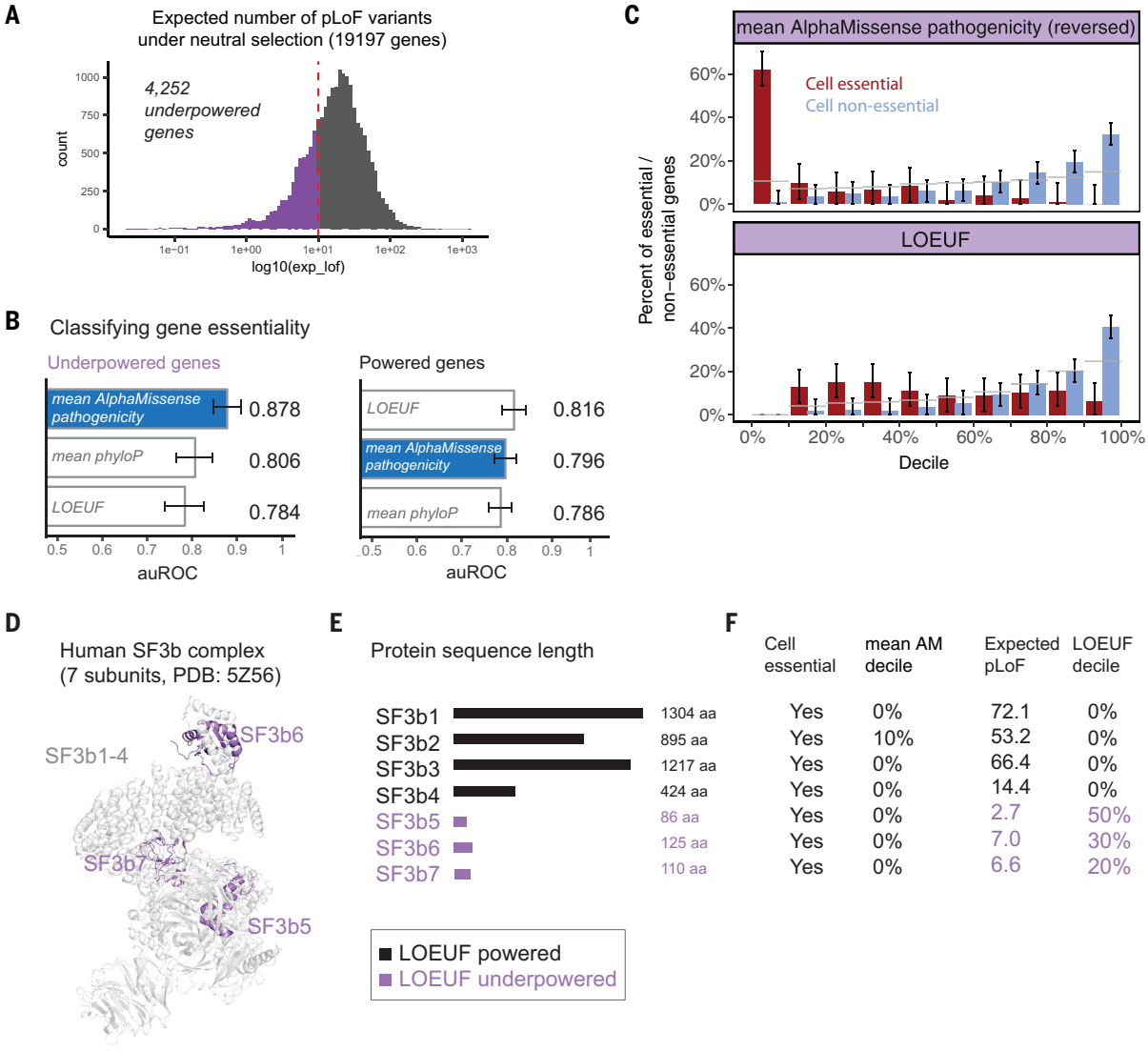


Fig. 4. AlphaMissense predicts cell essentiality without constraints on sequence length. (**A**) Distribution of the expected number of pLoF variants per gene under neutral selection within a cohort of 125,000 individuals, as estimated in (*3*). Briefly, pLoF variants are a class of genetic variants that introduce a major change to the protein coding sequence, such as a premature stop codon, which likely results in loss of protein function [see (*3*) for a full definition]. We refer to underpowered genes (expected pLoF ≤ 10) as those with insufficient statistical power, as determined in (*3*), to be classified among the most constrained genes by LOEUF. (**B**) Performance (auROC) of gene-level scores at classifying cell essentiality among genes underpowered (left) and well-powered (right) for LOEUF. The positive and negative examples for classification are 1247 cell essential genes and 728 cell nonessential genes, respectively, queried from DepMap (*48*). Within the LOEUF underpowered genes, there are 190 positive and 290 negative examples. Conversely, within the LOEUF powered genes, there are 1084 positive and 438 negative examples (see methods). (**C**) Distribution of experimentally determined cell essential and cell nonessential genes (*46*) across the deciles of mean AlphaMissense pathogenicity and LOEUF among genes underpowered for LOEUF (see methods). To be consistent with LOEUF, where low values indicate high gene constraint, AlphaMissense deciles are defined such that low deciles correspond to higher pathogenicity. Error bars show 95% confidence intervals of multinomial proportions. Horizontal gray lines show the percentage of all underpowered genes in each decile bin, which represents the expected percent if there is no enrichment or depletion of cell essential or nonessential genes. (**D**) Experimentally determined structure of the SF3b protein complex (PDB ID 5Z56) that is a crucial component of the U2 small nuclear ribonucleoprotein (*49*). The locations of the small protein subunits underpowered for LOEUF are highlighted in purple. (**E**) Lengths of each protein in SF3b (canonical UniProt isoforms). aa, amino acids. (**F**) Additional characteristics of SF3b proteins listed in (E). "Cell essential" means that it is in the list of "common essential" genes as determined by DepMap (*48*). "Expected pLoF" is the number of expected pLoF variants under neutral selection within the cohort from which LOEUF is derived [as in (A)]. For "Mean AM decile" and "LOEUF decile," 0% indicates the most pathogenic or constrained decile, respectively. For further information, see data S1.

effects. These resources benefit from the expanded coverage of confident predictions and have value in several contexts.

The predictions of all possible missense variants could assist clinicians in prioritizing variants for rare disease diagnostics, as they offer an important increase in the coverage of confidently classified missense variants (which would otherwise remain variants of unknown significance) without being biased toward the existing human curation process or well-studied genes. Out of 69.5 million variants unobserved in gnomAD, we were able to make a confident prediction for 61.7 million (88.8%) missense variants by classifying them as likely

A

CHROM	POS	REF	ALT	genome	uniprot_id	transcript_id	protein_variant	am_pathogenicity	am_class
chr7	44149825	T	C	hg38	P35557	ENST00000403799	D205G	0.999516	likely_pathogenic
...	...	...	...	...	...	...	...	...	...

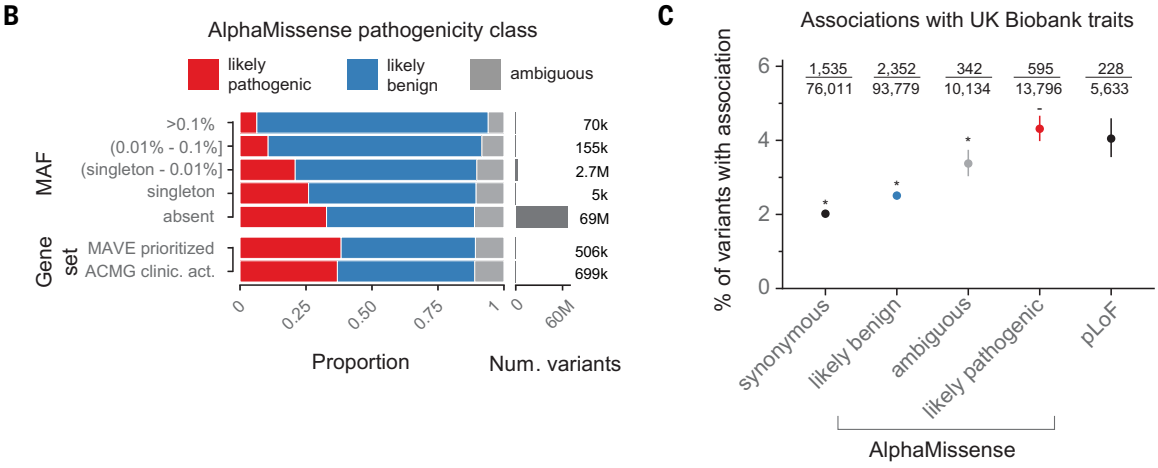


Fig. 5. AlphaMissense predictions as a community resource. (**A**) Example row from the AlphaMissense (AM) proteome-wide pathogenicity prediction dataset. (**B**) Pathogenicity class proportions for different variant MAF ranges in gnomAD (top) and two different gene sets (bottom): prioritized genes for MAVE studies in (*33*) and clinically actionable genes prioritized by ACMG (*32*). (**C**) The proportion of rare variants (MAF < 0.01) that have a statistical association ($P < 1 \times 10^{-5}$) with at least one of ~4000 UK Biobank traits (see methods). Counts above each variant set are the number of variants with an association over the total number in that set. An asterisk indicates that the proportion is significantly different from the pLoF set (Fisher exact test, $P < 0.05$), and a minus sign indicates that there is no evidence of a difference.

benign (38.9 million, 56.0%) or likely pathogenic (22.8 million, 32.8%) (Fig. 5B). This coverage and predictive performance on ClinVar remain high among clinically actionable genes prioritized by ACMG (*32*) (88.9% resolved, average auROC of 0.959; fig. S3A). The fraction of predicted pathogenic variants decreases with increasing allele frequency, as expected by purifying selection (Fig. 5B). The MAVE and ACMG prioritized proteins have a higher proportion of predicted pathogenic variants than variants absent from gnomAD proteome-wide (38.4% and 36.8%, respectively, versus 32.8%), in line with the high evolutionary constraint and functional importance of these two gene sets (Fig. 5B and fig. S3, A and B).

This resource could also inform studies of complex trait genetics (*50*, *51*). We compared the proportion of rare variants (MAF < 0.01) that are statistically associated with any of 4000 traits in the UK Biobank (*52*) for different classes of variation (see methods). We found that missense variants predicted as likely pathogenic by AlphaMissense contained twice as many trait associations compared with synonymous variants (Fig. 5C and fig. S8A), and that the rate of associations among predicted likely pathogenic variants is statistically indistinguishable ($P = 0.43$, Fisher exact test) from pLoF variants (Fig. 5C). In contrast, the rates among both ambiguous and likely benign variant sets are significantly lower ($P <$ 0.05, Fisher exact test), with likely benign variants having the most similar rate to synonymous variants (Fig. 5C). By combining variants from both AlphaMissense pathogenic and pLoF categories, we increase the number of candidate deleterious rare variants by 3.2-fold, translating to ~7000 additional genes that would be testable in gene-level association analyses in large-scale cohorts such as UK Biobank (*2*) (cumulative allele count > 50 in UK Biobank; fig. S8B). As such, our annotation of missense variants could be a powerful tool for discovery of previously unknown genes underlying complex traits (*50*, *51*).

The predictions of all possible amino acid substitutions are intended for studying the full range of single-residue perturbations. For example, predictions could be used as a starting point for designing and interpreting experiments that probe saturating amino acid substitutions across the human proteome, as performed by the MAVE community. Such scores can be used alongside the AlphaFold Structure Database (*21*, *53*) to assess the predicted pathogenicity in the context of predicted protein structures for every single human protein. Together, AlphaMissense predictions have the potential to accelerate our understanding of the molecular effects of variants on protein function, contribute to the discovery of disease-causing genes, and increase the diagnostic yield of rare genetic diseases.

Materials and methods summary

Full details of the methods are described in the supplementary materials and are summarized here. The model architecture is similar to that of AlphaFold (*21*), with minor modifications. AlphaMissense was trained in two stages: structure pretraining and variant fine-tuning. The pretraining stage is the same as described in AlphaFold, except with higher weights on the masked MSA reconstruction loss. During fine-tuning, the model is optimized to predict both variant pathogenicity and structure of the reference sequence. The benign training variants are derived from observed variants in human and primate species following the PrimateAI approach (*12*). Pathogenic training variants are sampled from unobserved variants with sampling weights depending on the trinucleotide context and the gene. A small subset of ClinVar (1263 pathogenic and 1263 benign) variants are used as the evaluation set for model selection and hyperparameter optimization. The variant effect prediction score is defined as the log-likelihood difference between the reference amino acid and the alternative amino acid. The final model predictions are the average of six models: three independently trained models (with minor hyperparameter differences) each run twice, once with diversity filtering on the MSA and once without. Raw model prediction scores are calibrated with the ClinVar evaluation set

to represent approximate probabilities (we refer to the calibrated scores as AlphaMissense pathogenicity). Finally, we define threshold score values to interpret a variant as "likely pathogenic," "ambiguous," or "likely benign." These values are derived such that the labels are assigned with 90% precision on ClinVar variants, following the approach of EVE (17). The model performance was compared with previous computational methods using multiple evaluation datasets: ClinVar (*30*), de novo variants from the DDD cohort (*54*), cancer hotspot mutations (*10*), MAVE data of 72 proteins from ProteinGym (https://www.proteingym.org/) and additional MAVE data of 20 proteins collected from the literature. Transcript-level mean AlphaMissense pathogenicity is calculated by averaging the pathogenicity scores of all possible single-nucleotide missense variants per transcript. For methods associated with the analysis of properties of the model outputs beyond the primary evaluation metrics (e.g., relationship with allele frequencies and cell essentiality) we refer readers to the supplementary materials.

REFERENCES AND NOTES

1. M. Lek *et al*., Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016). doi: 10.1038/nature19057; pmid: 27535533
2. C. Bycroft *et al*., The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018). doi: 10.1038/s41586-018-0579-z; pmid: 30305743
3. K. J. Karczewski *et al*., The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi: 10.1038/s41586-020-2308-7; pmid: 32461654
4. D. M. Fowler, S. Fields, Deep mutational scanning: A new style of protein science. *Nat. Methods* **11**, 801–807 (2014). doi: 10.1038/nmeth.3027; pmid: 25075907
5. G. M. Findlay *et al*., Accurate classification of *BRCA1* variants with saturation genome editing. *Nature* **562**, 217–222 (2018). doi: 10.1038/s41586-018-0461-z; pmid: 30209399
6. AVE Alliance Founding Members, The Atlas of Variant Effects (AVE) Alliance: understanding genetic variation at nucleotide resolution, version 4a, Zenodo (2021); https://doi.org/10.5281/zenodo.7508716.
7. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7.20.1–7.20.41 (2013). doi: 10.1002/0471142905.hg0720s76; pmid: 23315928
8. N. M. Ioannidis *et al*., REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016). doi: 10.1016/j.ajhg.2016.08.016; pmid: 27666373
9. Y. Wu *et al*., Improved pathogenicity prediction for rare human missense variants. *Am. J. Hum. Genet.* **108**, 1891–1906 (2021). doi: 10.1016/j.ajhg.2021.08.012; pmid: 34551312
10. H. Zhang, M. S. Xu, X. Fan, W. K. Chung, Y. Shen, Predicting functional effect of missense variants using graph attention neural networks. *Nat. Mach. Intell.* **4**, 1017–1028 (2022). doi: 10.1038/s42256-022-00561-w; pmid: 37484202
11. D. G. Grimm *et al*., The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015). doi: 10.1002/humu.22768; pmid: 25684150
12. L. Sundaram *et al*., Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018). doi: 10.1038/s41588-018-0167-z; pmid: 30038395
13. M. Kircher *et al*., A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014). doi: 10.1038/ng.2892; pmid: 24487276
14. P. C. Ng, S. Henikoff, SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003). doi: 10.1093/nar/gkg509; pmid: 12824425
15. T. A. Hopf *et al*., Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017). doi: 10.1038/nbt.3769; pmid: 28092658
16. E. Laine, Y. Karami, A. Carbone, GEMME: A simple and fast global epistatic model predicting mutational effects. *Mol. Biol. Evol.* **36**, 2604–2619 (2019). doi: 10.1093/molbev/msz179; pmid: 31406981
17. J. Frazer *et al*., Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021). doi: 10.1038/s41586-021-04043-8; pmid: 34707284
18. J. Meier *et al*., Language models enable zero-shot prediction of the effects of mutations on protein function. bioRxiv 2021.07.09.450648 [Preprint] (2021); https://doi.org/10.1101/2021.07.09.450648.
19. P. Notin *et al*., Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. *Proc. Mach. Learn. Res.* **162**, 16990–17017 (2022).
20. Z. Lin *et al*., Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). doi: 10.1126/science.ade2574; pmid: 36927031
21. J. Jumper *et al*., Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). doi: 10.1038/s41586-021-03819-2; pmid: 34265844
22. S. Ittisoponpisan *et al*., Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated? *J. Mol. Biol.* **431**, 2197–2212 (2019). doi: 10.1016/j.jmb.2019.04.009; pmid: 30995449
23. A. Schmidt *et al*., Predicting the pathogenicity of missense variants using features derived from AlphaFold2. *Bioinformatics* **39**, btad280 (2023). doi: 10.1093/bioinformatics/btad280; pmid: 37084271
24. B. Li, D. M. Roden, J. A. Capra, The 3D mutational constraint on amino acid sites in the human proteome. *Nat. Commun.* **13**, 3273 (2022). doi: 10.1038/s41467-022-30936-x; pmid: 35672414
25. K. Tunyasuvunakool *et al*., Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021). doi: 10.1038/s41586-021-03828-1; pmid: 34293799
26. G. R. Buel, K. J. Walters, Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* **29**, 1–2 (2022). doi: 10.1038/s41594-021-00714-2; pmid: 35046575
27. R. M. Rao *et al*., MSA Transformer. *Proc. Mach. Learn. Res.* **139**, 8844–8856 (2021).
28. A. Rives *et al*., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021). doi: 10.1073/pnas.2016239118; pmid: 33876751
29. B. J. Livesey, J. A. Marsh, Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380 (2020). doi: 10.15252/msb.20199380; pmid: 32627955
30. M. J. Landrum *et al*., ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018). doi: 10.1093/nar/gkx1153; pmid: 29165669
31. J. M. Havrilla, B. S. Pedersen, R. M. Layer, A. R. Quinlan, A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019). doi: 10.1038/s41588-018-0294-6; pmid: 30531870
32. D. T. Miller *et al*., ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **24**, 1407–1414 (2022). doi: 10.1016/j.gim.2022.04.006; pmid: 35802134
33. D. Kuang *et al*., Prioritizing genes for systematic variant effect mapping. *Bioinformatics* **36**, 5448–5455 (2021). doi: 10.1093/bioinformatics/btaa1008; pmid: 33300982
34. S. Richards *et al*., Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015). doi: 10.1038/gim.2015.30; pmid: 25741868
35. M. H. Høie, M. Cagiada, A. H. Beck Frederiksen, A. Stein, K. Lindorff-Larsen, Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep.* **38**, 110207 (2022). doi: 10.1016/j.celrep.2021.110207; pmid: 35021073
36. K. A. Matreyek *et al*., Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018). doi: 10.1038/s41588-018-0122-z; pmid: 29785012
37. A. Laddach, J. C. F. Ng, F. Fraternali, Pathogenic missense protein variants affect different functional pathways and proteomic features than healthy population variants. *PLOS Biol.* **19**, e3001207 (2021). doi: 10.1371/journal.pbio.3001207; pmid: 33909605
38. D. Munro, M. Singh, DeMaSk: A deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics* **36**, 5322–5329 (2021). doi: 10.1093/bioinformatics/btaa1030; pmid: 33325500
39. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915–10919 (1992). doi: 10.1073/pnas.89.22.10915; pmid: 1438297
40. S. Fayer *et al*., Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in *BRCA1*, *TP53*, and *PTEN*. *Am. J. Hum. Genet.* **108**, 2248–2258 (2021). doi: 10.1016/j.ajhg.2021.11.001; pmid: 34793697
41. B. J. Livesey, J. A. Marsh, Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol. Syst. Biol.* **19**, e11474 (2023). doi: 10.15252/msb.202211474; pmid: 37310135
42. J. J. Kwon, W. C. Hahn, A leucine-rich repeat protein provides a SHOC2 the RAS circuit: a structure-function perspective. *Mol. Cell. Biol.* **41**, e00627-20 (2021). doi: 10.1128/MCB.00627-20; pmid: 33526449
43. J. J. Kwon *et al*., Structure-function analysis of the SHOC2-MRAS-PP1C holophosphatase complex. *Nature* **609**, 408–415 (2022). doi: 10.1038/s41586-022-04928-2; pmid: 35831509
44. S. M. Sternisha, B. G. Miller, Molecular and cellular regulation of human glucokinase. *Arch. Biochem. Biophys.* **663**, 199–213 (2019). doi: 10.1016/j.abb.2019.01.011; pmid: 30641049
45. S. Gersing *et al*., A comprehensive map of human glucokinase variant activity. *Genome Biol.* **24**, 97 (2023). doi: 10.1186/s13059-023-02935-8; pmid: 37101203
46. T. Hart *et al*., Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3* **7**, 2719–2727 (2017). doi: 10.1534/g3.117.041277; pmid: 28655737
47. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010). doi: 10.1101/gr.097857.109; pmid: 19858363
48. J. M. Dempster *et al*., Extracting biological insights from the Project Achilles genome-scale CRISPR screens in cancer cell lines. bioRxiv 720243 [Preprint] (2019); https://doi.org/10.1101/720243.
49. C. Sun, The SF3b complex: Splicing and beyond. *Cell. Mol. Life Sci.* **77**, 3583–3595 (2020). doi: 10.1007/s00018-020-03493-z; pmid: 32140746
50. L. Bomba, K. Walter, N. Soranzo, The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017). doi: 10.1186/s13059-017-1212-4; pmid: 28449691
51. S. Lee, G. R. Abecasis, M. Boehnke, X. Lin, Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014). doi: 10.1016/j.ajhg.2014.06.009; pmid: 24995866
52. K. J. Karczewski *et al*., Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**, 100168 (2022). doi: 10.1016/j.xgen.2022.100168; pmid: 36778668
53. M. Varadi *et al*., AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022). doi: 10.1093/nar/gkab1061; pmid: 34791371
54. Deciphering Developmental Disorders Study, Prevalence and architecture of *de novo* mutations in developmental disorders. *Nature* **542**, 433–438 (2017). doi: 10.1038/nature21062; pmid: 28135719
55. P. Petit *et al*., The active conformation of human glucokinase is not altered by allosteric activators. *Acta Crystallogr. D* **67**, 929–935 (2011). doi: 10.1107/S0907444911036729; pmid: 22101819
56. A. L. Gloyn *et al*., in *Glucokinase and Glycemic Disease: From Basics to Novel Therapeutics*, F. M. Matschinsky,

M. A. Magnuson, Eds., vol. 16 of *Frontiers in Diabetes* (S. Karger AG, 2004), pp. 92–109.
57. J. Cheng *et al*., Source code for AlphaMissense, version 1.0.0, Zenodo (2023); https://doi.org/10.5281/zenodo.8208697.
58. J. Cheng *et al*., Predictions of AlphaMissense, version 1.0.0, Zenodo (2023); https://doi.org/10.5281/zenodo.8208688.

ACKNOWLEDGMENTS

We thank K. Tunyasuvunakool, R. Fergus, and E. Papa for their insights and manuscript reviews; D. La for feedback on structural representations and analyses; Z. Wu and S.-J. Dunn for their contributions at the early stage of the project; the Research Platform colleagues for their continuous support; and other colleagues at DeepMind and Google for their encouragement and support. This research has been conducted using summary statistics generated from the UK Biobank resource (under applications 26041 and 48511), accessed at https://app.genebass.org/ (*52*). **Funding:** All research in this study was funded by DeepMind and Alphabet. There was no external funding. **Author contributions:** J.C. and Ž.A. conceptualized the study with input from J.J., A.W.S., P.K., and D.H.; J.C. and Ž.A. managed and supervised the project; J.C. and G.N. developed the model with input from A.P., Ž.A., J.J., and A.W.S.; J.C., G.N., Ž.A., A.Ž., and R.G.S. developed the data pipeline; J.C. and G.N. performed modeling experiments with help from Ž.A.; J.C., G.N., J.P., C.B., T.A., and Ž.A. analyzed data, prepared figures, and wrote the manuscript; J.C. and T.A. developed software infrastructure for model inference; T.A. developed software for data ingestion (complex traits, competitor methods) and generated genome-to-proteome maps with support from A.Ž.; J.P. analyzed multiplexed assays of variant effect (MAVEs) data with help from J.C.; J.P. curated and analyzed structural data; C.B. and T.A. conceived of and executed analysis of complex traits; C.B. conceived of and executed analysis of gene constraint and cell essentiality with help from J.P.; T.A. performed inference for ESM1v with help from A.Ž.; A.Ž. helped with software infrastructure and generated structure visualization utilities; L.H.W. reviewed the DMS literature, annotated ProteinGym data, and managed and coordinated project planning and execution; M.Z. reviewed code, provided feedback on the methodology, and contributed to the proteome-wide analysis; T.S. contributed to the training data preparation and software infrastructure; D.H. and P.K. contributed to management and supervision; J.C., G.N., J.P., C.B., T.A., A.Ž., A.P., L.H.W., M.Z., T.S., A.W.S., and Ž.A. edited the manuscript. All authors reviewed the manuscript. **Competing interests:** This work was done in the course of employment at DeepMind, with no other competing financial interests. J.C., G.N., and Ž.A. have filed provisional patent applications relating to machine learning for predicting missense variant effects (US Provisional Patent nos. 63/415,117 and 63/479,653). **Data and materials availability:** The source code of AlphaMissense is available at Zenodo (*57*) and https://github.com/deepmind/alphamissense. Predictions for all human missense variants and amino acid substitutions are available at Zenodo (*58*) or https://console.cloud.google.com/storage/browser/dm_alphamissense. Researchers interested in predictions not yet provided, and for noncommercial use, can send an expression of interest to alphamissense@google.com. As part of our commitment to releasing our research breakthroughs safely and responsibly, we will not be sharing model weights, to prevent use in potentially unsafe applications.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adg7492
Supplementary Note
Materials and Methods
Figs. S1 to S10
Tables S1 to S6
References (*59*–*100*)
MDAR Reproducibility Checklist
Data S1 to S9

Submitted 19 January 2023; accepted 23 August 2023
Published online 19 September 2023
10.1126/science.adg7492