nature communications

Article

https://doi.org/10.1038/s41467-025-59937-2

A probabilistic graphical model for estimating selection coefficients of nonsynonymous variants from human population sequence data

Received: 4 April 2024

Accepted: 6 May 2025

Published online: 20 May 2025

Check for updates

Yige Zhao[1,2], Tian Lan[1], Guojie Zhong[1,2], Jake Hagen[1,3], Hongbing Pan[4], Wendy K. Chung[3] & Yufeng Shen[1,4,5]

Accurately predicting the effect of missense variants is important in discovering disease risk genes and clinical genetic diagnostics. Commonly used computational methods predict pathogenicity, which does not capture the quantitative impact on fitness in humans. We develop a method, MisFit, to estimate missense fitness effect using a graphical model. MisFit jointly models the effect at a molecular level (d) and a population level (selection coefficient, s), assuming that in the same gene, missense variants with similar d have similar s. We train it by maximizing probability of observed allele counts in 236,017 individuals of European ancestry. We show that s is informative in predicting allele frequency across ancestries and consistent with the fraction of de novo mutations in sites under strong selection. Further, s outperforms previous methods in prioritizing de novo missense variants in individuals with neurodevelopmental disorders. In conclusion, MisFit accurately predicts s and yields new insights from genomic data.

Missense variants, which cause single amino acid changes in proteins, are the most common type of variant in protein-coding regions. They are major contributors to genetic risk of developmental disorders[1–3], cancer, and other diseases. However, as missense variants have a potentially broad range of functional impact but generally a low chance of recurrence, most missense variants identified in cohorts or clinical sequencing have uncertain effect[4–9]. Deep mutational scanning (DMS) assays can assist with interpretation of missense variants[10–31], but there is limited scalability as different proteins have different and multifaceted functions that require different functional assays. Therefore, computationally predicting the effect of missense variants is important to support the scale required for novel disease gene discovery and interpretation.

Although many methods have been developed to predict variant effects, there is a long-standing ambiguity of the concepts used to describe variant effect. We adopt a set of definitions[32] to explain the related causes and consequences specifically for different aspects of missense variant effect (Supplementary Fig. 1). At the molecular level, we define the effect (d) as change of abundance, localization, or function of a protein. At organism level, a damaging variant (with larger d) is defined as pathogenic if it increases the risk of diseases or conditions. Pathogenic variants are often the focus in human genetic studies and clinical testing. Databases like ClinVar[9] and HGMD[33] have curated pathogenic variants, which are used as the training labels in supervised methods, such as CADD[34], REVEL[35], M-CAP[36], gMVP[37], VEST[38], MetaSVM[39], MVP[40] and MPC[41]. Although these methods have proven helpful, they usually suffer from inconsistent performance across genes, since most of the curated pathogenic variants are from only a few thousand genes that are well-established as

[1]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA. [2]The Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University, New York, NY, USA. [3]Department of Pediatrics, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA. [4]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA. [5]JP Sulzberger Columbia Genome Center, Columbia University, New York, NY, USA. e-mail: ys2411@cumc.columbia.edu

disease-associated. We suggest that *predicted pathogenicity* is an uncertain aggregation of variant functional effect, gene risk, and even disease properties. Our knowledge of gene to disease association is incomplete and in fact, identification of new associations is a primary goal of predicting variant effect in genetic studies. Therefore, we seek other metrics for describing missense effects in prediction that can be quantified without knowing gene-disease associations.

One such metric is selection coefficient (s) which quantifies the fitness effect of variants in a population. A pathogenic variant is usually subject to negative selection in human populations. Although s of a variant depends on the penetrance of the variant to various conditions and the total fitness effect of the conditions, the consequence of s, especially of heterozygotes, can be observed in allele frequencies in human populations[42]. It is therefore theoretically feasible to estimate s without knowing any traits with which the variant is associated. Biobank-scale genomic sequencing efforts[4–8] have generated a large number of human population genome sequences that enable estimation on heterozygous selection coefficient of gene-aggregated protein truncating variants (PTVs)[43–46]. However, estimation for missense variants is much more challenging because we cannot reasonably assume all or most of missense variants in one gene have the same selection coefficient. Existing prediction of selection for individual variants[47] does not directly utilize protein context, and is still based on a very small sample size.

Here we describe a new method, MisFit, to jointly predict molecular effect and human fitness effect of missense variants through a probabilistic graphical model. We aimed to estimate selection coefficient for variants under moderate to strong negative selection. In the model, the molecular effect depends on amino acid change in the protein context, and heterozygous selection coefficient depends on molecular effect of the variant and gene-level importance in selection in human populations. We trained the model using population genome data without pathogenicity labels and evaluated it using deep mutational scan readout data and de novo and inherited variants in developmental disorders.

Results

Using Poisson-Inverse-Gaussian distribution to model allele counts in human populations

The distribution of allele counts (m) in population sequencing samples is determined by heterozygous selection coefficient (s), mutation rate (ν) and number of chromosomes (n). To infer s, we first need to model the probability of observed allele counts $p(m|s;\nu,n)$. Allele frequency q of a variant at equilibrium state equals $q=\frac{\nu}{s}$, and therefore the allele count m follows a Poisson distribution with an expectation $\frac{n\nu}{s}$. When taking genetic drift into account, the distribution has strong overdispersion. Nei's model[48] describes allele count as a Negative Binomial distribution with an additional parameter, effective population size N_e. However, as there was exponential growth in recent generations[49,50], N_e is not a constant, and there is no closed form to describe $p(m|s;\nu,n)$. Here, we used a long-tailed distribution, Inverse-Gaussian (IG) distribution, to approximate the distribution of q, which results in a Poisson-Inverse-Gaussian (PIG) distribution of m. The parameters associated with the PIG distribution are functions of s,ν,n, which are optimized prior to MisFit training steps by simulated allele frequencies given dense grids of s,ν and European effective population size history (Methods). We are mainly interested in those rare variants with relatively large s; therefore, we chose to optimize the distribution for the recent exponential population growth. In this way, we were able to easily obtain a tractable gradient to s with a time complexity independent to n.

To investigate how to approximate allele frequency distribution in a finite and expanding population, we performed a simulation based on a demographic history model of European population[49]. Given ν and s, we sampled each generation by a Wright-Fisher process (Methods). We set the final effective population size to 1.5 million, as it best fits the distribution of observed sample allele counts of rare synonymous C-to-T variants in methylated CpG sites with high roulette mutation rate ($\nu > 10^{-7}$ per generation) (Supplementary Fig. 2). This final population size is smaller than recent work[45,46] (5 million), which is optimized for all variants with gnomAD mutation rate (with an lower average $\nu \sim 6\times10^{-9}$).

We fitted the PIG model parameters (Supplementary Fig. 3) based on simulated variants under different settings of $\nu \in [10^{-9}, 3\times10^{-7}]$, $s \in [10^{-6}, 1]$. When s is small, random drift makes the distribution of allele counts more resemble a Negative Binomial distribution with small N_e. When s is large, the distribution is closer to a Poisson distribution, as these variants are likely to emerge recently when the effective population size is large (Fig. 1a, b). The PIG model fits the simulated results better than other simple distribution models in all ranges (Supplementary Fig. 4).

Feasibility of estimating selection coefficient for a group of variants

Given the generally low mutation rate at 10^{-8}, the highest probability usually lies at 0 count regardless of s (Supplementary Fig. 4), so it is nearly impossible to precisely estimate s for individual single nucleotide variants only using allele counts. We therefore investigated the feasibility of estimating s for a group of variants with similar s. We aggregated certain numbers of sites simulated from the same s as a group (Fig. 1c). We investigated whether the maximum-likelihood-estimation (MLE) for the whole group based on the PIG model is consistent with the simulation condition.

For deleterious variants of $s > 0.01$ with high mutation rate, the accuracy is high. More variants to aggregate and higher mutation rate always helps with better estimation. The PIG model does not provide good performance for $s < 10^{-4}$, because randomly including or excluding a common variant in the group can significantly change the joint likelihood. Notably, increasing sample size in a single population only helps with variants under strong selection ($s > 0.01$) (Fig. 1c, Supplementary Fig. 5). The over-dispersion of allele counts for milder variants mostly comes from the uncertainty of allele frequency (the long-tailed distribution of q) due to genetic drift, rather than from sampling (the Poisson distribution given nq). Adding samples from another population improves accuracy more than from the same population. Based on the results, we implicitly group missense variants by the degree of damage (d) in the same gene in the MisFit model.

MisFit model structure and training process

We describe the architecture of MisFit in Fig. 2. The degree of damage (d) for each single amino acid substitution depends on the protein sequence and structure context. Here we used the embeddings (x) from the last layer in the masked protein language model, ESM-2 (650 M)[51] to capture the protein sequence and structure context. We added additional transformer blocks and fully-connected dense layers to generate a distribution of d (Supplementary Fig. 6b). Rescaling and normalization of d by a gene-level, species-averaged selection strength gives out probabilities of each amino acid at the position. The heterozygous selection coefficient (s) depends on d and the gene-level selection strength in the human population. Here we modeled s in the logit scale as linear to d. We set a global prior for the maximum missense selection coefficient for each gene (s_{gene}, the value of s when d equals 1). (Methods). Finally, probability of generating allele count n given s is given by the PIG model as previously described.

In the first stage, we trained the model to estimate parameters in transformer and dense layers (denoted as NN^1 in Methods), to maximize the log likelihood with allele counts, amino acid in orthologues[52,53], and ESM-2 zero-shot prediction. In other words, we attempt to approximate $p(d|x)$ by maximizing $p(m|x,\nu,n)$, as this gives out estimation of d across genes. During this training stage all possible missense SNVs in 18,708 genes were used, although for most epochs, only a subset of 4073 constrained genes (missense z score >2 or gnomAD[6,7] pLI >0.5) well covered with both mammal sequence alignment and human population sequence were included, because damaging variants in these genes are expected to be under relatively strong selection, and thus the difference in molecular effect can result in a broad range of selection coefficient for training. Allele counts in 236,017 samples are used in training, including 145,103 UKBB[8] unrelated individuals of European ancestry and 90,914 gnomAD[7] individuals of Non-Finnish European ancestry. Finally, s_{gene} for all genes were updated by maximum a posteriori (MAP).

In the second stage, with the estimated d and s_{gene}, we performed variational inference to approximate the posterior distribution of s for each missense SNV. During this stage, s is a hidden variable while its prior is regarded as known with the optimized NN^1, and thus the posterior distribution $p(s|m;x,\nu,n)$ is determined but with no simple analytical form. Sampling-based methods are a solution but time-consuming considering the amount of individual missense variants. Therefore, we treat posterior s as functions of d and population data, which is modeled by another dense neural network (denoted as NN^2 in Methods) to enable efficient variational inference in one forward-pass. (Supplementary Fig. 6b, Methods)

Comparison of gene-level constraint

MisFit-estimated s_{gene} quantifies gene-level selection strength on missense variants. Commonly used metrics for such information include gnomAD missense z score and o/e. Though s_{gene} for each gene generally correlates well with both metrics (Supplementary Fig. 7), they represent different aspects. Missense z score is effectively the significance level assuming a Poisson distribution from the expected number of variants. Thus, when a gene is short, missense z score tends to have a small absolute value. s_{gene} and o/e directly represent the degree of constraint, although the uncertainty for short genes might be large.

Previous studies[44–46] estimated s for protein truncating variants (PTVs) in each gene, assuming PTVs in a gene have the same s. We compared s_{gene} with a sampling-based method[45] for PTVs. PTVs mainly decrease protein levels by nonsense mediated decay. As most of the missense variants are hypomorphs with partial loss of function, s_{gene} and s_{PTV} are highly correlated (Fig. 3a). However, some variants can be damaging through mechanisms other than loss of function. We highlighted risk genes with known genetic modes[54] (Fig. 3b-e, Supplementary Data 1). Autosomal recessive genes are least intolerant of PTVs compared with other genes associated with dominant inheritance. Haploinsufficient genes are under strong selection on PTVs. Genes with dominant negative effects are likely to be under strong selection

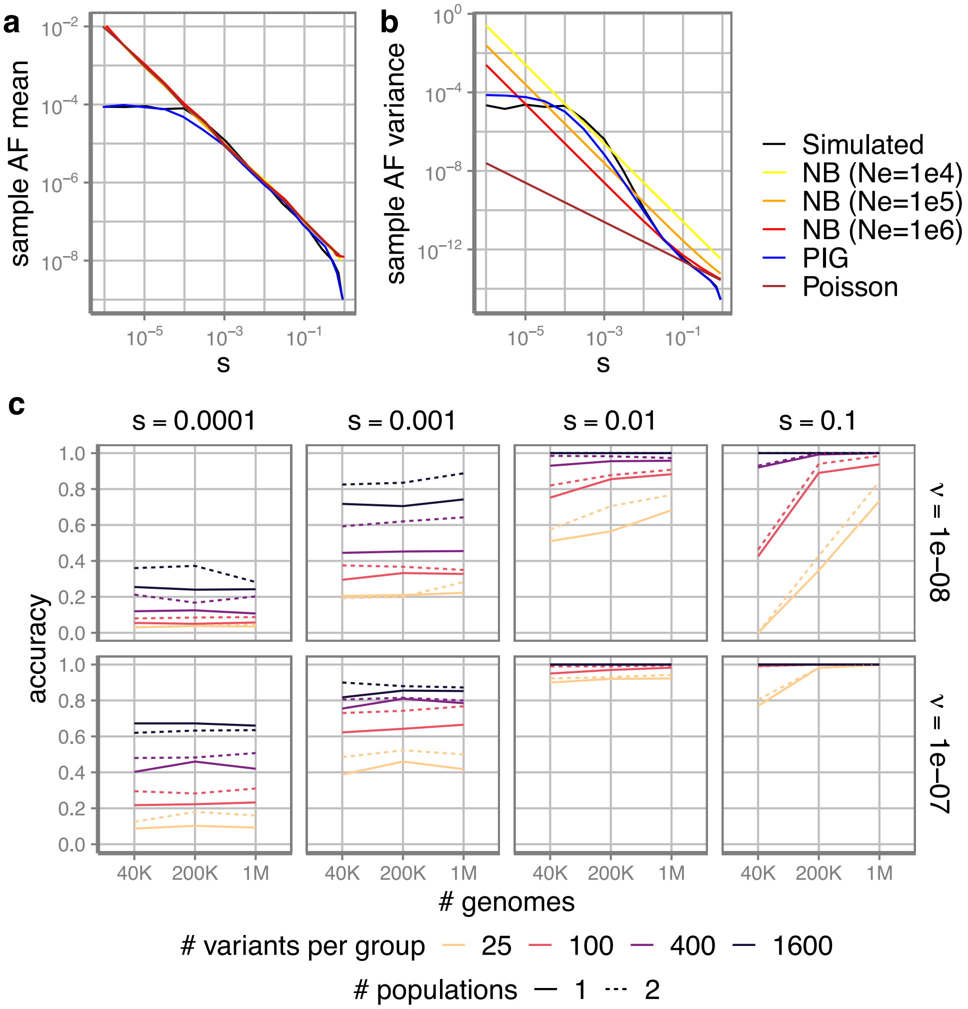

Fig. 1 | Poisson-Inverse-Gaussian (PIG) distribution with adjusted parameters to approximate allele count distribution. **a** mean and (**b**) variance of sample allele frequency (AF) under different population genetics models, including our PIG model and Negative Binomial (NB) model with different effective population size. Diploid sample size is 200 K. Mutation rate is 10^{-8}. **c** The accuracy of maximum likelihood estimation (MLE) of s. Here s is a categorical variable of 0.00001, 0.0001, 0.001, 0.01, 0.1, 1. Accuracy is measured by the proportion that the estimated categorical s equals the simulated in 400 simulated groups. Each group contains a certain number of variants (x-axis) with same s. Solid lines are samples from a single population, while dashed lines are samples from two populations (half of the indicated number for each population).

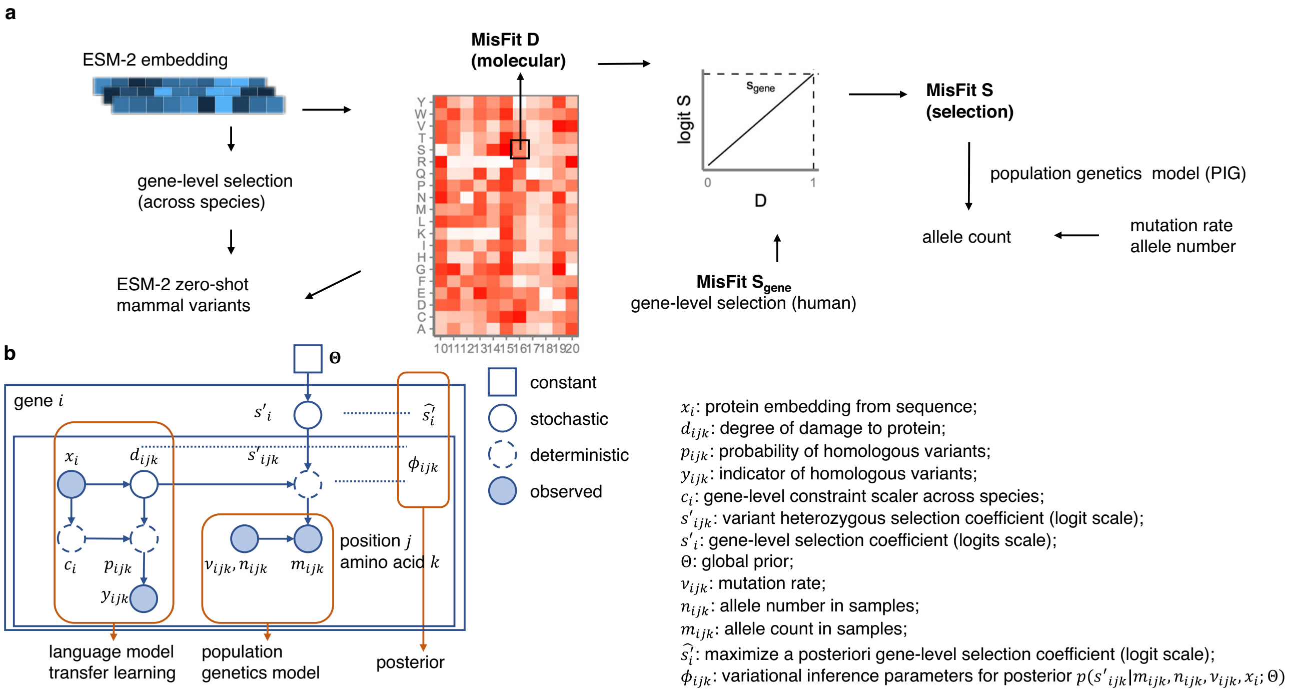
a
ESM-2 embedding
gene-level selection
(across species)
ESM-2 zero-shot
mammal variants
MisFit D
(molecular)
Y W V T S R Q P N M L K I H G F E D C A
10 11 12 13 14 15 16 17 18 19 20
logit S
s_{gene}
0
1
D
MisFit S
(selection)
population genetics model (PIG)
allele count
mutation rate
allele number
MisFit S_{gene}
gene-level selection (human)
b
Θ
gene i
s'_i
$\hat{s}'_i$
x_i
d_{ijk}
s'_{ijk}
ϕ_{ijk}
c_i
p_{ijk}
y_{ijk}
v_{ijk}, n_{ijk}
m_{ijk}
position j
amino acid k
language model
transfer learning
population
genetics model
posterior
constant
stochastic
deterministic
observed
x_i: protein embedding from sequence;
d_{ijk}: degree of damage to protein;
p_{ijk}: probability of homologous variants;
y_{ijk}: indicator of homologous variants;
c_i: gene-level constraint scaler across species;
s'_{ijk}: variant heterozygous selection coefficient (logit scale);
s'_i: gene-level selection coefficient (logits scale);
Θ: global prior;
v_{ijk}: mutation rate;
n_{ijk}: allele number in samples;
m_{ijk}: allele count in samples;
$\hat{s}'_i$: maximize a posteriori gene-level selection coefficient (logit scale);
ϕ_{ijk}: variational inference parameters for posterior $p(s'_{ijk}|m_{ijk}, n_{ijk}, v_{ijk}, x_i; \Theta)$

Fig. 2 | MisFit model for estimating molecular and fitness effect. a Overview of MisFit model. MisFit_D is learned from ESM-2 (650 M) protein embedding, and generates probability of amino acid in orthologues. Heterozygous s is linear to d in logit scale, with gene-level maximum from a global prior. MisFit_S is a point estimate of s per variant, which maximizes the allele count likelihood in population samples. **b** MisFit model in view of a probabilistic generative process.

on missense and PTVs. Notably, for gain-of-function, a subset of genes are only constrained on missense but not on PTVs (Supplementary Fig. 8). For example, several germline missense variants in oncogene *KRAS* lead to Noonan syndrome by hyperactivation of the protein[55]. The gene-level selection on missense variants is significantly higher than PTVs ($s_{gene} = 0.37$, $s_{PTV} = 0.00020$).

MisFit is predictive of allele counts of ultrarare variants in different populations

As MisFit_S is able to predict s with amino acid resolution (Supplementary Fig. 9), we asked how informative MisFit_S is to predict allele frequency of rare variants in a population of different ancestries. We extracted 215,138 positions without observed missense variants or with ultra-rare (sample allele frequency $<5 \times 10^{-6}$) missense variants and high mutation rate ($\nu > 10^{-7}$) in 4073 constrained genes of the training set (UKBB and gnomAD NFE, 236,017 samples, thus allele count ≤ 2 for most highly covered sites). We binned the variants by estimated MisFit_S and analyzed the counts in a second population of a different ancestry, which is gnomAD African/African American (AFR) with 28,872 individuals. Putative variants in these positions would have emerged very recently, and their allele frequencies are relatively independent between the two populations. As expected, the proportion of variants with 0-count in gnomAD AFR samples is positively correlated with MisFit_S (Supplementary Fig. 10a). The opposite trend is observed for the percent of variants with 10 times higher allele frequency in AFR (Supplementary Fig. 10b). To assess which part of the model helps with prediction of s, we built several models with fewer and simpler components. In the baseline model (model 0), s is estimated from only the mutation rate and allele counts with a global prior. For this chosen set of variants of high mutation rate, allele count is informative as shown in the stepwise curve caused by allele counts of 0, 1, 2 in the training set. However, the difference in absolute value of selection is subtle (Supplementary Fig. 10c). Adding the gene-level selection (model 1) in the model largely improves and smooths the estimation and outputs a wider range of s. Using ESM-2 zero-shot score to infer probability of damage (model 2) further helps the prediction, indicated by a greater slope of the monotonic increase, but is not as good as the full MisFit model, which uses the ESM-2 embedding.

Comparison of selection coefficient with de novo fraction

Next, we evaluate whether MisFit_S approximates heterozygous selection coefficient s in absolute scale. We obtained missense de novo (16,876 cases, 5750 controls) and inherited variants (6507 cases, 2992 controls) from an autism spectrum disorder study[1] (2). MisFit_S of de novo variants are significantly higher than inherited variants (Supplementary Fig. 11). We binned the variants based on MisFit_S and normalized the counts as per individual (Fig. 4a). The difference between cases and controls is significant for de novo missense variants for strongly deleterious variants (MisFit_S > 0.01), but is subtle for inherited variants, even if limiting the data to known autism genes (Supplementary Fig. 12).

In a new generation when selection has not occurred, de novo variants are expected to take up a proportion that is equivalent to s when s is relatively large[44]. Such relationship holds for randomly selected samples regardless of their own phenotypes, but not for samples specially chosen with known family backgrounds (Supplementary Note). We aggregated the variants by their selection coefficient and calculated the fraction from the de novo variants. The de novo ratio in autism cases is consistent with MisFit_S, indicating the accuracy of estimated s in absolute scale. In controls, which were unaffected siblings in families ascertained by cases, highly deleterious de novo variants are lower than s as expected.

Analytical utility of selection coefficient for de novo variants in developmental disorders

In addition to the autism data, we obtained de novo variants from studies of neurodevelopmental disorders (NDD, most individuals have global developmental delay or intellectual disability)[56] (31,565 cases) (Supplementary Data 2). Previous studies[1,3,56] have shown that a substantial fraction of de novo missense variants in these cases are risk variants for NDD. Autism and NDD are relatively common conditions with early-onset phenotypes. Autism has a prevalence approaching

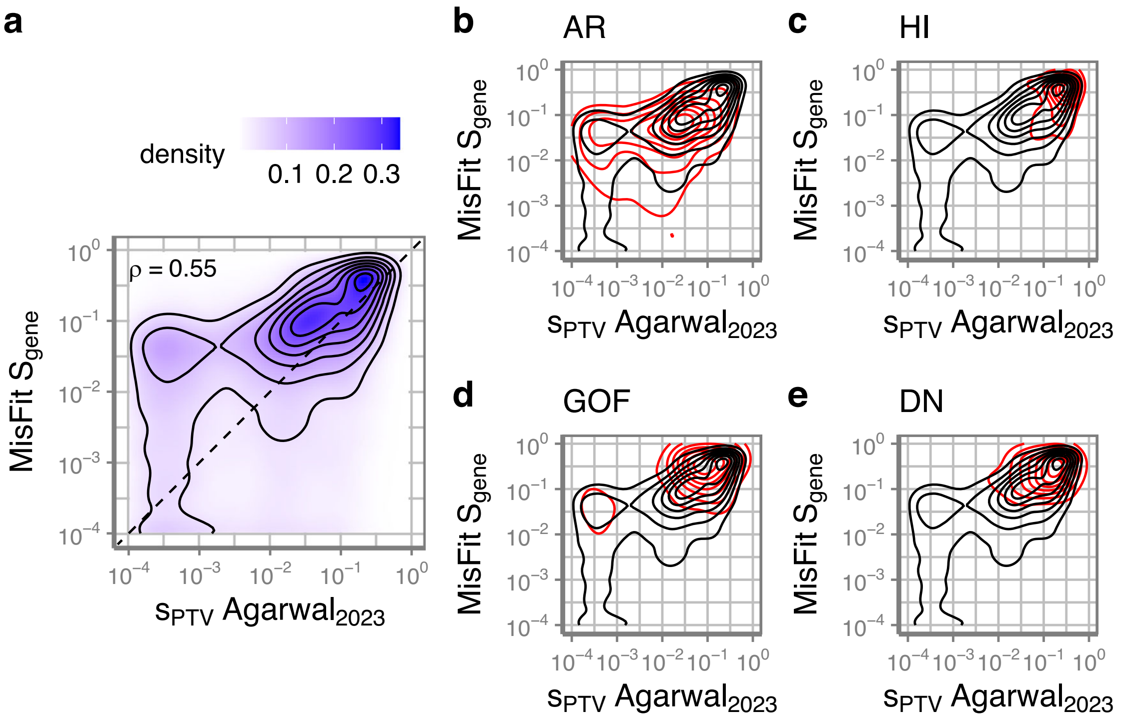

Fig. 3 | Gene-level missense selection S_{gene} compared with selection coefficient of protein-truncating variants (PTV). **a** the distribution of all genes (black contour) with Pearson correlation coefficient. **b**–**e** in genes harboring variants of known mechanisms (red contours): **b** autosomal recessive (*n* = 641) **c** happloinsufficient (*n* = 66) **d** gain-of-function (*n* = 69) **e** dominant negative (*n* = 54).

0.028[57], and selection on autism is around 0.7[58]. Thus, highly penetrant risk variants are not likely to be transmitted into the next generation, resulting in a high selection coefficient. As expected, de novo variants in cases have a higher MisFit_D and MisFit_S than controls (Fig. 5). We compared our results with other missense variant effect prediction methods[34–37,51,59–61]. Although there is no ground truth to know which variants actually increase disease risk, we could calculate the enrichment of variants under different thresholds, which is the ratio of number of variants in cases to what is expected in controls (Methods). Among variants ranked in the top 10 percentiles by multiple methods, MisFit_S reached a higher enrichment ratio (Fig. 6) than any other method.

We then derived the precision-recall-proxy curves (Supplementary Fig. 13, Methods) by the excess number of variants under thresholds. MisFit_S outperforms other methods in high precision range, reaching a precision of 0.67 and 0.87 for autism and NDD, respectively, at MisFit_S = 0.1. The next best methods are AlphaMissense[61] and gMVP[37]. The estimated precision can serve as weights or informative priors in statistical methods like DeNovoWEST[56] or extTADA[62–64] to improve the power in risk gene discovery. The selection coefficients estimated by baseline methods with fewer components are also informative in enrichment of de novo variants but are inferior to MisFit (Supplementary Fig. 14-15).

MisFit identifies damaging variants consistent with deep mutational scan data

MisFit-estimated *d* (MisFit_D) is about the molecular effect of missense variants, which can be partly measured by deep mutational scanning (DMS) experiments. We compared MisFit_D with published methods[34–37,51,59–61] on predicting damaging variants in DMS for individual genes. First, we collected functional readout scores from 32 DMS assays in 26 genes[11–17,19–31] with 44,100 single amino acid substitutions (Supplementary Data 3). We calculated the Spearman correlation between the functional scores and computational scores (Fig. 7a). MisFit_D has a similar performance with ESM and AlphaMissense.

As raw functional readouts from these experiments could be noisy, we further restricted the sets to variants in 13 genes with DMS annotated binary labels or with a bi-modal functional score distribution (Supplementary Fig. 16). For the latter, we labeled damaging variants by two-component Gaussian Mixture models for each assay independently (Methods, Supplementary Data 4). For genes with multiple DMS assays, we combined these datasets and label a variant as damaging if it is damaging in any one of the assays. The average area under ROC curve (AUROC) for MisFit_D still approaches the state-of-art performance (Fig. 7b).

In some genetic analysis, we often set a heuristic and fixed threshold across all genes when selecting possibly damaging variants. To evaluate the performance under this setting, we combined the DMS assays across genes, and tested the performance in the combined dataset. Since the labels are unbalanced, we define the optimal threshold as that which achieves the highest Matthew's correlation coefficient (MCC) in the combined dataset. When setting this optimal threshold for classification, we calculated the MCC in each individual gene. MisFit_D remains effective, meaning that the prediction is consistently informative across genes (Fig. 7c). MisFit_D is intended to quantify the degree of damage solely based on variant-level property, and we expect it to be distributed similarly across genes. In contrast, selection coefficient (MisFit_S) is by nature determined by both variant- and gene-level properties and should not have the same range in different genes. Supplementary Fig. 17 shows gene-specific score distribution and optimal threshold.

Finally, we investigated the distribution of sensitivity in different genes (Supplementary Fig. 18). Sensitivity is only related to the damaging variants in the dataset. Deep mutational scanning assays are usually designed to evaluate only one aspect of gene function, so the identified damaging variants could be more reliable, while benign ones may disrupt the protein in some other ways not evaluated by the assays. Under a threshold achieving a global sensitivity of 0.5, MisFit_D has a low variance across genes (Fig. 7d). Overall, unsupervised methods (MisFit,

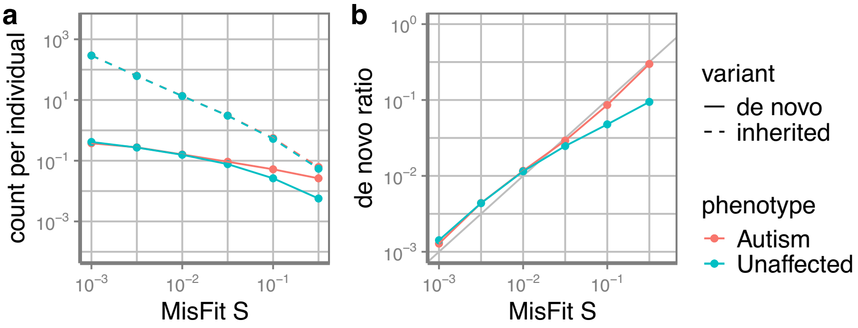

Fig. 4 | de novo or inherited missense variants binned by of MisFit_S. a Count of de novo or inherited missense variants. **b** The proportion of de novo to all variants in autism dataset. Error bars show 95% confidence intervals.

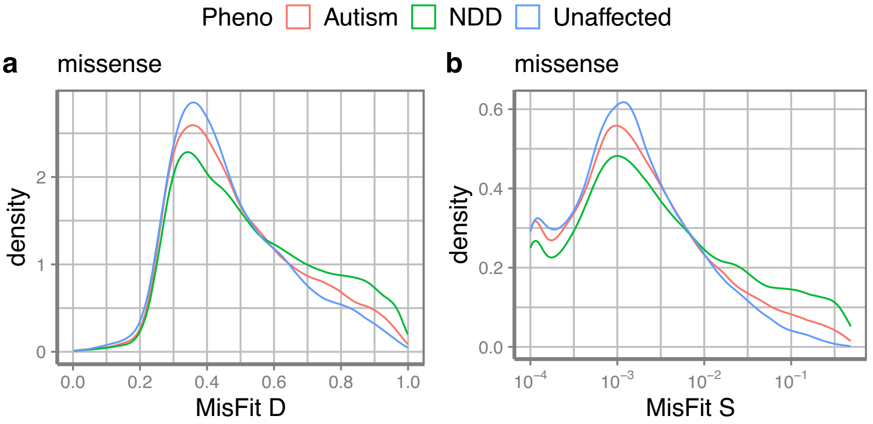

Fig. 5 | Distribution of MisFit_D and MisFit_S for de novo variants in autism and neurodevelopmental disorders (NDD) datasets. Two-sided Kolmogorov-Smirnov tests are used between case and control. **a** Missense variants MisFit D (Autism: $p = 5.5 \times 10^{-6}$; NDD: $p = 3.2 \times 10^{-42}$); **b** Missense variants MisFit S (Autism: $p = 2.6 \times 10^{-6}$; NDD: $p = 3.7 \times 10^{-54}$).

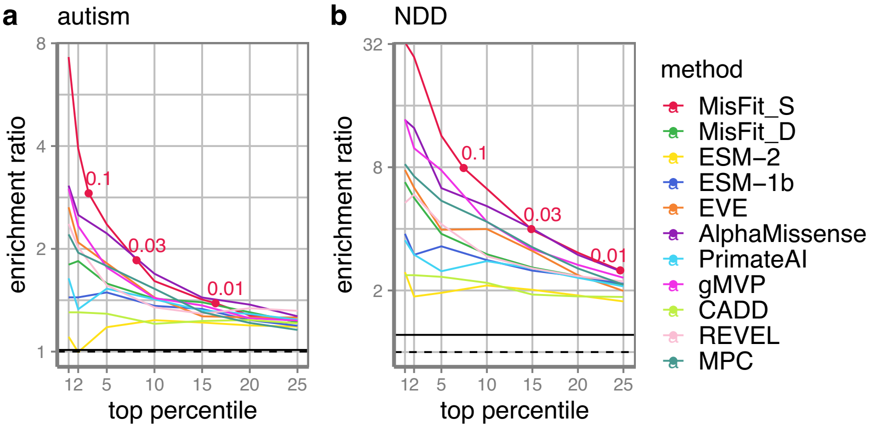

Fig. 6 | Enrichment ratio of de novo missense variants under different thresholds for multiple prediction methods. Thresholds of MisFit_S are annotated. The solid horizontal line is the overall enrichment ratio including all variants, the dashed line stands for no enrichment. **a** in autism spectrum disorder dataset (**b**) in neurodevelopmental disorders dataset.

ESM1b and ESM2) have lower variance of sensitivity across genes than supervised methods (gMVP, REVEL, and AlphaMissense).

Discussion

We developed a probabilistic graphical model, MisFit, to estimate the fitness effect of missense variants using large population sequencing data. Selection coefficient (s) is a quantitative measurement of fitness effect that can be informed by allele frequency in human populations, but it is very difficult to estimate for individual variants. MisFit addresses this issue by modeling it as a sigmoid-shaped function of the molecular effect d of a variant with a gene-specific prior, and jointly modeling d as a non-linear function, approximated by deep neural networks, of its protein sequence context. We trained the model using large sets of population sequencing data without any label of pathogenicity. The estimated s is highly correlated with frequency of ultra-rare variants in an independent population. Its value is consistent with

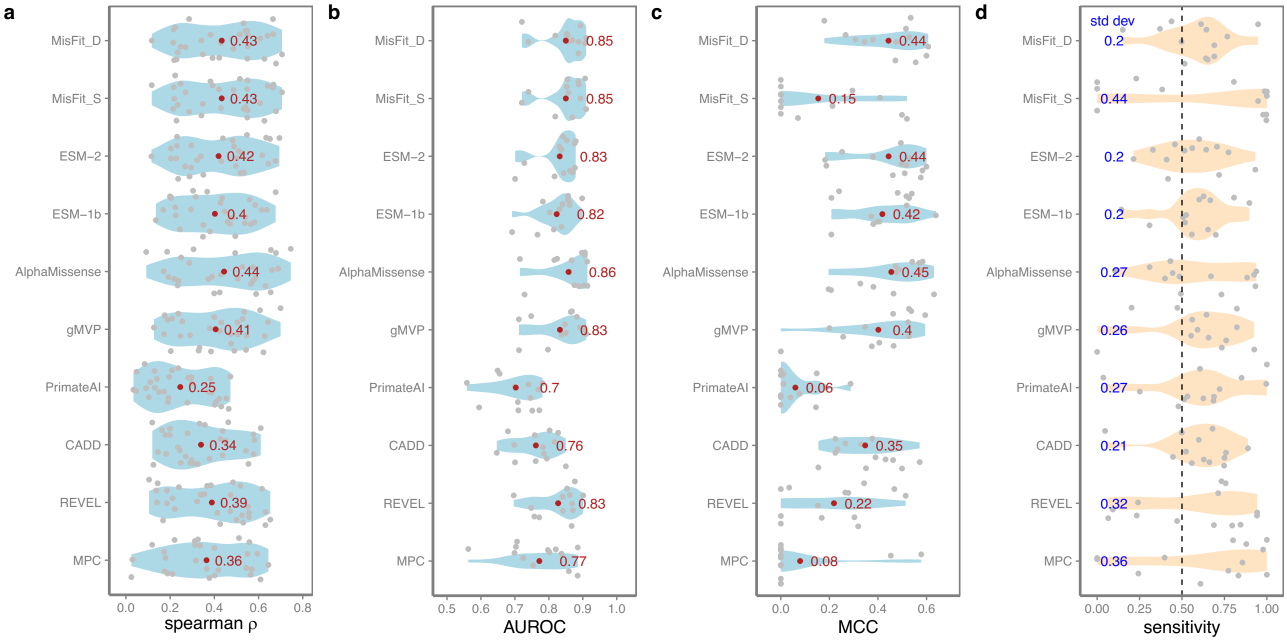

Fig. 7 | Performance in predicting damaging variants in deep mutational scanning assays and cross-gene consistency. a Spearman correlation coefficient of predicted scores with functional scores from deep mutational assays. Mean is annotated in red. **b** Area Under the Receiver Operating Characteristic Curve (AUROC) of predicting confidently labeled damaging or benign variants in deep mutational assays. Mean is annotated in red. **c** Matthew correlation coefficient (MCC) in each gene with a global threshold that achieves best MCC in the combined dataset. Mean is annotated in red. **d** Sensitivity in different genes when setting a threshold to achieving a global sensitivity of 0.5 (dashed) in the combined dataset. Standard deviation is annotated in blue. For **b**–**d** different assays of same gene are combined so that variants with a damaging label in any of the assays will be regarded as damaging.

theoretical expectation of the proportion of de novo mutations among observed variants in a population.

Previous efforts in estimating gene-specific[6,7,65] or sub-genic[41,66] regional constraints of missense variants showed the feasibility of using human population data to identify coding regions that are under strong selection, but these methods are heuristic and do not estimate the effect of individual variants. MisFit is based on population genetics models, representing an improved approach to using large-scale population sequencing datasets for estimating variant effect. Additionally, the effect of a variant at the organism and population levels is a combination of how the variant alters the protein and how the protein is involved in key biological processes relevant to human traits and diseases. Two variants with the same degree of damage to protein function may have different effects at the organism and population levels if they occur in different genes. Methods that predict pathogenicity by supervised learning are confounded by gene-level properties, as shown by the large variance of classification accuracy across genes given a fixed threshold evaluated by deep mutational scan data. MisFit's graphical model is designed to untangle the gene-variant confounding. As a result, MisFit_D has a more consistent scale across genes assessed by mutational scan data; and MisFit_S, as a natural combination of variant and gene properties, has superior performance in prioritizing de novo variants in studies of developmental disorders that have strong negative consequence in fitness.

In a longer timescale across species, negative selection is manifested in conservation among homologous sites. Some unsupervised models, such as ESM[51,67,68] and EVE[60], predict amino acid probabilities using representation learning based on massive amounts of protein sequences or multiple sequence alignment of homologous proteins. Those alleles used in training are effectively neutral to become nearly fixed in the corresponding species[69]. When further taking phylogenic history into account, observed sequences are correlated, but the distribution may deviate from the stationary distribution of fitness landscape. Although these models are empirically effective[70,71], for relatively large s, all such deleterious variants are likely to be depleted from the collection of wild-type sequences which are mostly fixed alleles in each species, so theoretically their difference in s cannot be easily estimated from MSA alone.

Such resolution in estimating relatively large s is especially important for analysis of rare variants in genetic studies of early onset conditions. If we assume an early onset condition is the main trait under selection for a risk variant, then the selection of the variant could be approximated as prevalence × relative risk × selection of the condition. Thus, risk variants in conditions with high prevalence and low fecundity, such as intellectual disability and autism[58], tend to have large selection coefficient. As the affected population are enriched of risk variants with higher selection coefficients, they usually show higher overall proportion of de novo variants in genetic analysis, because this proportion approximates selection coefficients, which we have shown theoretically and empirically. This explains why MisFit shows superior performance in prioritizing de novo variants in autism or NDD datasets. Additionally, the fitness effects of missense variants estimated by MisFit are directly comparable to the estimated heterozygous selection coefficient of protein truncating variants by previous methods in a quantitative way. This could improve the power of identifying new risk genes and help characterize genetic etiology of human diseases.

We used the embeddings from a protein language model (ESM-2) to represent protein sequence context as the input for the non-linear function that predicts the effect at molecular level. ESM-2 embeddings implicitly capture protein structure information[51]. Explicitly representing protein structure features as input[53,61] may improve prediction by better capturing residue interactions and critical sites.

Finally, based on the simulation results, the accuracy of MisFit in estimating mildly deleterious variants ($s < 0.001$) is limited. Random drift of these variants causes significant dispersion of allele frequency. Merely increasing the sample size of the same population does not help with the estimation. On the other hand, including diverse populations with different continental ancestry in training would improve the accuracy, as ancestral effective population size increases and variance from genetic drift decreases. We expect that the sample size individuals of non-European ancestry with genome sequencing will increase substantially in the near future from ongoing efforts such as gnomAD[7], All of Us[5], GenomeAsia[72], and the Three Million African Genomes project[73]. We will be able to use these data to improve estimation of fitness effect of variants under moderate selection in the future.

Methods

Simulation based on European effective population size history

We simulated the distribution of allele frequency based on the history of effective population size of European population for 10,000 generations. We obtained the European effective population size history from the Schiffels and Durbin model[49]. We smoothened the data by setting a growth rate for each period and adjusted the final effective population size to 1.5 million, which is most consistent with distribution of observed allele counts of rare synonymous variants with high roulette mutation rate ($\nu > 10^{-7}$). We assume no linkage and the same background mutation rate (using the average mutation rate), no positive selection effect, and each locus obtains one type of mutation at most. We simulated the evolution of alleles with dense grids of mutation rates $\nu \in [10^{-9}, 3\times10^{-7}]$, and selection coefficients $s \in [10^{-6}, 1]$.

For a given mutation rate and selection coefficient, we simulated 100,000 independent sites. The simulation follows the Wright-Fisher process considering mutation, drift and selection. We set a backward mutation rate of $\nu_0 = 10^{-8}$. Suppose the effective population size at t^{th} generation is N_t, we have

$$q_{t-1} = \frac{(1-s)(1-f_{t-1})f_{t-1} + (1-2s)f_{t-1}^2}{(1-f_{t-1})^2 + 2(1-s)(1-f_{t-1})f_{t-1} + (1-2s)f_{t-1}^2} \tag{1}$$

$$f'_t = q_{t-1}(1-\nu_0) + (1-q_{t-1})\nu \tag{2}$$

f_{t-1} and q_{t-1} are the pre-selection and post-selection allele frequency in the previous generation, and f'_t is allele frequency in zygotes after introducing new mutations. Here, $2s$ (clipped at 1 if $s > 0.5$) is homozygous selection coefficient by fixing a dominance factor of 0.5. Then we sample population allele counts in the new generation by a binomial distribution:

$$m_t \sim Binomial(2N_t, f'_t) \tag{3}$$

$$f_t = \frac{m_t}{2N_t} \tag{4}$$

In the latest generation, sample allele counts m within sample allele number n drawn from population could be regarded as a Hypergeometric distribution.

$$m \sim Hypergeometric\left(2N_{final}, m_{final}, n\right) \tag{5}$$

As we have $m_{final} \sim Binomial(2N_{final}, f'_{final})$, this is equivalent to

$$m \sim Binomial\left(n, f'_{final}\right) \tag{6}$$

Considering the age of sequencing samples (UK Biobank and gnomAD) are relatively old, the observed alleles are already subject to selection. We therefore used the adjusted post-selection allele frequency for training the model.

$$q_{final} = \frac{(1-s)\left(1-f'_{final}\right)f'_{final} + (1-2s)f'^2_{final}}{\left(1-f'_{final}\right)^2 + 2(1-s)\left(1-f'_{final}\right)f'_{final} + (1-2s)f'^2_{final}} \tag{7}$$

To investigate how a second population with a different genetic ancestry can help with estimation, we simulated a pseudo-population with the same European population size history. Here, q is kept same for both populations at the beginning, and then evolves independently for the recent N_r generations. We set N_r to be 2000 based on the split time of European and Africa population. In this way, the final q in two populations are partially correlated.

Modeling allele counts

Assuming infinite effective population size, allele frequency q at the equilibrium state is deterministic given the mutation rate ν and heterozygous selection coefficient s.

$$q = \frac{\nu}{s} \tag{8}$$

Therefore, the allele count m in samples with allele number n follows a Poisson distribution:

$$m \sim Poisson(nq) \tag{9}$$

Although the formula gives us an overview of the relationship between expected m and s, there is a substantial overdispersion of m caused by the random drift effect. Taking random drift into account, Nei's model[48] describes q as a Gamma distribution.

$$q \sim Gamma\left(4N_e\nu, 4N_e s\right) \tag{10}$$

N_e is the effective population size. Then we have a Negative Binomial distribution for m.

$$m \sim NegBinom\left(4N_e\nu, \frac{n}{4N_e s + n}\right) \tag{11}$$

However, the real N_e is not constant. There has been exponential population growth in all major continental populations. We used an Inverse Gaussian model with adjusted parameters μ_{IG}, λ_{IG} to describe the distribution of allele frequency. Inverse Gaussian distribution can model a very long tail while keeping the probability density at 0 to be 0 (In contrast, Gamma distribution may give out infinity density at 0). More importantly, the likelihood function $p(m|s;\nu,n)$ should have a tractable gradient to s. Then m follows a Poisson Inverse Gaussian (PIG) distribution:

$$q \sim InvGaussian(\mu_{IG}, \lambda_{IG}) \tag{12}$$

$$m \sim PoisInvGaussian(n\mu_{IG}, n\lambda_{IG}) \tag{13}$$

μ_{IG} and λ_{IG} are Inverse Gaussian mean and shape respectively. For each setting of ν, s, we used the simulated allele frequency q_{sim} to estimate $\mu_{IG} = mean\left(q_{sim}\right)$, $\lambda_{IG} = 1/mean(\frac{1}{q_{sim}} - \frac{1}{\mu_{IG}})$. Then we fit functions $\mu_{IG} = f_1(\nu, s)$ and $\lambda_{IG} = f_2(\nu, s)$. Specifically, $\log\mu_{IG}$ is a softminus over s′ = logit(s) and linear over $\log\nu$, while $\log\lambda_{IG}$ is quadratic to $\log\nu$ (Supplementary Fig. 3). The likelihood of PIG distribution is calculated by Bessel function of second kind.

Data used in training and testing

Proteins and variants. We limit the gene set to 18,708 protein-coding genes. One protein sequence is selected per gene (from Ensembl v104[74]), based on the following order: 1. Uniprot[75] canonical isoform; 2. Corresponding to the transcript of 'MANE select'; 3. Corresponding to Ensembl canonical transcript (usually the longest). Among them, 18,605 have available population sequencing data for missense variants and 16,623 for protein-truncating variants. All possible single nucleotide variants in the coding region +−2 bp of the selected transcripts are annotated using bcftools v1.17[76]. For protein-truncating variants, 'stop-gained', 'splice_donor' and 'splice_acceptor' variants are further annotated by LOFTEE[6], and only high-confidence (HC) ones are used in training or genetic analysis.

Population sequence data. We used the allele counts from UKBB[8] unrelated individuals of European ancestry (145,103 exomes from November 2020 release) and gnomAD[6,7] Non-Finnish European ancestry (56,885 exomes of v2.1.1 plus 34,029 genomes of v3.1.2) sequencing data. These datasets only contain observed variants in these individuals, but the vast majority of possible but not yet observed variants are also important for estimation. We include all possible missense variants that could result from a single nucleotide substitution. We set the allele number (sample size) for positions without observed variants by the allele number of the nearest position with observed variant in the same exon, to account for sequencing depth variation, and the allele count of these non-observed variants to 0. Same was done for gnomAD African / African American population (8128 exomes plus 20,744 genomes) in analysis. Variants that do not pass RF, InbreedingCoeff or AS_VQSR filtering, or are located in low-complexity-region (annotated by gnomAD), are excluded in training and analysis.

Site-specific mutation rate was mainly obtained from roulette mutation rates. Variants on sex chromosomes do not have available roulette mutation rates, so we used gnomAD[6] mutation rates based on 3-mer context and methylation level, and calibrated them to an average of 10^{-8} in consistent with roulette. During training, mutation rate and allele count are added across all single-nucleotide variants that lead to the same amino acid change.

Protein sequence embeddings. Protein sequence embeddings are extracted from the last layer of ESM-2 (650 M) model for each 600 AA length fragments (overlapping 200 AA if longer than 600 AA). The zero-shot prediction of ESM-2 comes from logits value in the last layer of ESM-2 and further renormalized to 20 amino acids excluding other tokens.

Mammalian homologs. Homologous variants used in training include: a. 21.8 million alternative amino acids in multiple sequence alignment in 465 mammals from Zoonomia Project[52]; b. 2.9 million alternative amino acids in 233 primate species from primateAI-3D[53].

Deep mutational scanning assays. We selected 32 deep mutational scanning assays from literature and MaveDB[18] (Supplementary Data 3). Several experiments provide classification of damaging or benign variants in the publications. For the remaining experiments, we model the functional scores (usually as log enrichment or depletion) by a two-component Gaussian mixture for each experiment. Amino acid substitutions with probability of damaging >0.75 are defined as damaging and that <0.25 are defined as benign. We selected experiments with bimodal score distribution, of which the confident damaging + benign variants make up more than 90% of all variants. In total, 13 genes with damaging / benign labels were selected for evaluating AUROC and MCC (Supplementary Data 4). If there are multiple assays for the same gene (*CYP2C9, PTEN, VKORC1*), we took the union of damaging labels as positives.

MisFit model architecture and parameters in view of a probabilistic graphical model

For a gene i, the maximum heterozygous selection coefficient for missense variants is denoted as s_i. In our model settings, s_i is transformed into logit scale s_i' to facilitate numeric computation. For each variant k at position j, d_{ijk} is assumed to be a random variable of logit-normal distribution.

$$s_i' = logit(s_i) \sim Normal\left(\mu_{s_{global}}, \sigma_{s_{global}}\right) \tag{14}$$

$$d_{ijk}' = logit\left(d_{ijk}\right) \sim Normal\left(\mu_{d_{global}}, \sigma_{d_{global}}\right) \tag{15}$$

Note that in the full MisFit model, distribution of d_{ijk}' is learned by neural networks (NN^1) as functions of the protein embeddings x_i.

$$d_{ijk}' = logit\left(d_{ijk}\right) \sim Normal\left(NN_\mu^1(x_i), NN_\sigma^1(x_i)\right) \tag{16}$$

For variant-level heterozygous selection coefficient s_{ijk}', we assume it's linear to d_{ijk} (ranging from 0 to 1), where the minimum is set to $logit(10^{-4})$ and the maximum is $s_{i'}$.

$$s_{ijk}' = f\left(s_i', d_{ijk}\right) \tag{17}$$

The purpose of our model is to approximate $p(d_{ijk}|x_i)$ and $p(s_{ijk}|x_i, \nu_{ijk}, m_{ijk}, n_{ijk})$, by two parts of MisFit model, NN^1 (functions of only x) and NN^2 (functions of $NN_1(x), m, n, \nu$), respectively. MisFit_D and MisFit_S are the point estimates from these two probabilities, which will be described in the next section.

Model training

The MisFit model contains 4.4 M parameters in total. Training of MisFit involves several stages.

In stage 0, before the construction of full MisFit model, we trained a baseline model (corresponding to model 1 in main text and Supplementary Figs. 10, 14, 15). We estimated $\Theta : \mu_{s_{global}}, \sigma_{s_{global}}; \mu_{d_{global}}, \sigma_{d_{global}}$ by maximizing $\sum_{ijk} \log E_{s_{ijk}} p(m_{ijk}|s\prime_{ijk}, \nu_{ijk}, n_{ijk})$, where $s\prime_{ijk}$ is calculated from samples of d_{ijk} given the global priors (Eqs. 15, 16). Then we set $\widehat{s_i}$ as maximize a posteriori estimation of s_i.

$$\widehat{s_{i\prime}} = argmax_{s\prime_i} \sum_i \left(\sum_{jk} \left(\log \mathrm{E}_{s'_{ijk}} p\left(m_{ijk}|s'_{ijk}, \nu_{ijk}, n_{ijk}, s\prime_i\right)\right) + \log p\left(s\prime_i|\mu_{s_{global}}, \sigma_{s_{global}}\right)\right) \tag{18}$$

This value of $\widehat{s_i}$ is then used to initialize the main MisFit model.

In stage 1, we aimed to optimize the parameters in NN^1 which connects x_i to d_{ijk} (Eq. 16). In brief, we would like to estimate

$$NN^1, \widehat{s_{i\prime}} = argmax_{NN^1, s\prime_i} \sum_i \left(\sum_{jk} \left(\begin{matrix} \log \mathrm{E}_{s'_{ijk}} p\left(m_{ijk}|s'_{ijk}, \nu_{ijk}, n_{ijk}, s\prime_i\right) \\ + \log E_{d_{ijk}} p(y_{ijk}|d_{ijk}) \end{matrix} \right) + \log p\left(s\prime_i|\mu_{s_{global}}, \sigma_{s_{global}}\right)\right) \tag{19}$$

where s_{ijk}' is sampled from Eqs. 16, 17. This stage involves in several periods. Here y_{ijk} is a Bernoulli variable denoting where such amino acid change exists in wildtype homolog sequences, where the Bernoulli logit is simply d_{ijk} transformed by scaler c_i. First, we trained NN^1 using all missense variants 13,406 genes well covered with both mammal sequence alignment and human population data for 30 epochs with initial learning rate as 0.001. $\widehat{s_i}$ was temporarily set as the value in stage 0. Then we trained NN^1 and $\widehat{s_i}$ as well on 4073 constrained genes (gnomAD[6,7] missense z score > 2 or pLI > 0.5) for 50 epochs with initial learning rate 0.0005. Finally, we kept NN^1 and further inferred $\widehat{s_i}$ for all genes for 30 epochs.

In stage 2, we did variational inference on the posterior distribution.

$$p\left(d'_{ijk}|x_i, m_{ijk}, n_{ijk}, \nu_{ijk}, \widehat{s}_i\right) = \frac{p\left(d'_{ijk}|x_i\right)p\left(m_{ijk}|d'_{ijk}, \widehat{s}_i, n_{ijk}, \nu_{ijk}\right)}{p\left(d'_{ijk}\right)} \quad (20)$$

Here, distribution of $s\prime_i$ is simply represented by its point estimate $\widehat{s'_i}$. We used a Normal distribution $Normal\left(\mu_{d_{ijk}}, \sigma_{d_{ijk}}\right)$ as variational family to approximate this distribution. In order to retrieve $\mu_{d_{ijk}}, \sigma_{d_{ijk}}$ in one forward pass, they are modeled as functions in a second neural network (corresponding to dense layers NN^2 in stage 2 in Supplementary Fig. 6).

$$\mu_{d_{ijk}} = NN^2_{\mu}\left(\widehat{s_i\prime}, m_{ijk}, n_{ijk}, \nu_{ijk}, NN^1_{\mu}(x_i), NN^1_{\sigma}(x_i)\right) \quad (21)$$

$$\sigma_{d_{ijk}} = NN^2_{\sigma}\left(\widehat{s_i\prime}, m_{ijk}, n_{ijk}, \nu_{ijk}, NN^1_{\mu}(x_i), NN^1_{\sigma}(x_i)\right) \quad (22)$$

Then like a variational autoencoder, optimizing the evidence lower bound (ELBO) is equivalent to maximizing

$$\begin{aligned} &E_{d'_{ijk} \sim Normal\left(\mu_{d_{ijk}}, \sigma_{d_{ijk}}\right)} logp\left(m_{ijk}|d'_{ijk}, \widehat{s_i\prime}\right) \\ &-KL\left(Normal\left(\mu_{d_{ijk}}, \sigma_{d_{ijk}}\right)|Normal\left(NN^1_{\mu}(x_i), NN^1_{\sigma}(x_i)\right)\right) \end{aligned} \quad (23)$$

KL() represents Kullback-Leibler Divergence.

During this stage, $\widehat{s}_i$ and parameters in NN^1 were fixed and only NN^2 is updated for 20 epochs with initial learning rate of 0.001.

MisFit scores are specifically defined as follows:

MisFit_D: $sigmoid(NN^1_{\mu}(x_i))$, the mean of $p(d'_{ijk}|x_i)$ transformed back to original scale.

MisFit_S$_{\text{gene}}$: $sigmoid(\widehat{s_i\prime})$, the MAP estimation of $p(s'_i|x_i, m_i, n_i, \nu_i)$ transformed back to original scale.

MisFit_S: $sigmoid(f(\widehat{s_i\prime}, sigmoid(\mu_{d_{ijk}})))$, the derived s'_{ijk} when using the point estimate of s'_i, and the point estimate of posterior mean of d_{ijk} given by Eq. 21.

In our model, the random variable s are all represented in logit scale, and our point estimate of s is also inferred in logit scale then transformed back to the original scale. This eases the calculation and potentially limits the systematic bias (Supplementary Note).

The main training stage 1 takes around 10 hours on 2 NVIDIA A40 GPUs.

Enrichment of de novo variants and estimated precision-recall

De novo missense variants in 4 previous genetic studies are used for analysis (Supplementary Data 2). Given a score threshold (to enrich disease risk variants), the number of selected variants is m_1 and m_0 in cases and controls respectively. These numbers are normalized by number of synonymous variants m_1^{syn} and m_0^{syn} to calculate the enrichment ratio.

$$r = \frac{m_1}{m_0} \times \frac{m_0^{syn}}{m_1^{syn}} \quad (24)$$

Sensitivity (recall approximate) is estimated by the total number of excess of variants comparing cases and control.

$$m'_1 = \frac{r-1}{r} m_1 \quad (25)$$

Precision is estimated by

$$\frac{m'_1}{m_1} = \frac{r-1}{r} \quad (26)$$

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data generated by this study are all accessible, including MisFit_D, MisFit_S for missense variants, and MisFit S_{gene} for each gene, and have been deposited in the Zenodo[77] (https://doi.org/10.5281/zenodo.15230898). The data used for training in this study are available at: • gnomAD[6,7] (https://gnomad.broadinstitute.org/) • Zoonomia[52] (https://zoonomiaproject.org/) • primateAI-3D[53] (https://primateai3d.basespace.illumina.com/) • MaveDB[18] (https://www.mavedb.org/) for deep mutational scanning data with accession numbers urn:mavedb:00000001; 00000005; 00000013; 00000035; 00000036; 00000047; 00000048; 00000049; 00000050; 00000054; 00000055; 00000057; 00000059; 00000069; 00000078; 00000095; 00000096; 00000097; 00000108. These raw referenced data can be obtained upon application: • Allele frequency data from UK Biobank[8] (https://www.ukbiobank.ac.uk/) • Autism data from the SPARK for autism study, including all coding variants, can be obtained from *SFARI base*[1]: https://base.sfari.org . Variant prediction results for analysis are collected: • ESM-2[51]and ESM-1b[67] (https://github.com/facebookresearch/esm), AlphaMissense[61] (https://doi.org/10.5281/zenodo.8360242), gMVP[37] (https://www.dropbox.com/s/nce1jhg3i7jw1hx/gMVP.2021-02-28.csv.gz?dl=0) from their original releases. • PrimateAI[59], CADD[34], REVEL[35], MPC[41] from dbNSFP[78] v4.3 (https://www.dbnsfp.org/).

Code availability

Codes and software used during data processing include: • ESM-2-t33_650M_UR50D(https://github.com/facebookresearch/esm) • Bcftools[76] v1.17 (https://samtools.github.io/bcftools/) • LOFTEE[6] v1.0 (https://github.com/konradjk/loftee). The machine learning model is built using *tensorflow*[79] v2.8.0, and statistical analysis is performed by *scipy*[80] v1.8.0 and *scikit-learn*[81] v1.2.2. Other custom codes for model training and data analysis could be found at *Github* (https://github.com/ShenLab/MisFit) and also on *Zenodo*[77] (https://doi.org/10.5281/zenodo.15230898).

References

1. Zhou, X. et al. Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nat. Genet.* **54**, 1305–1319 (2022).
2. Jin, S. C. et al. Contribution of rare inherited and de novo variants in 2871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
3. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
4. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
5. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
6. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
7. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

8. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* **12**, e1001779 (2015).
9. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
10. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
11. Bandaru, P. et al. Deconstruction of the Ras switching cycle through saturation mutagenesis. *eLife* **6**, e27810 (2017).
12. Weile, J. et al. A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957 (2017).
13. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
14. Heredia, J. D. et al. Mapping interaction sites on human chemokine receptors by deep mutational scanning. *J. Immunol.* **200**, 3825–3839 (2018).
15. Kotler, E. et al. A systematic p53 mutation library links differential functional impact to cancer mutation pattern and evolutionary conservation. *Mol. Cell* **71**, 178–190.e8 (2018).
16. Matreyek, K. A. et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
17. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
18. Esposito, D. et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* **20**, 223 (2019).
19. Jiang, R. J. Exhaustive mapping of missense variation in coronary heart disease-related genes. (University of Toronto, CA, 2019). http://hdl.handle.net/1807/98076.
20. Chan, K. K. et al. Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science* **369**, 1261–1265 (2020).
21. Chiasson, M. A. et al. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife* **9**, e58026 (2020).
22. Jones, E. M. et al. Structural and functional characterization of G protein–coupled receptors with deep mutational scanning. *eLife* **9**, e54895 (2020).
23. Suiter, C. C. et al. Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *Proc. Natl Acad. Sci.* **117**, 5394–5401 (2020).
24. Sun, S. et al. A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. *Genome Med.* **12**, 13 (2020).
25. Amorosi, C. J. et al. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *Am. J. Hum. Genet.* **108**, 1735–1751 (2021).
26. Jia, X. et al. Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am. J. Hum. Genet.* **108**, 163–175 (2021).
27. Weile, J. et al. Shifting landscapes of human MTHFR missense-variant effects. *Am. J. Hum. Genet.* **108**, 1283–1300 (2021).
28. Erwood, S. et al. Saturation variant interpretation using CRISPR prime editing. *Nat. Biotechnol.* **40**, 885–895 (2022).
29. Roychowdhury, H. & Romero, P. A. Microfluidic deep mutational scanning of the human executioner caspases reveals differences in structure and regulation. *Cell Death Discov.* **8**, 7 (2022).
30. Gersing, S. et al. A comprehensive map of human glucokinase variant activity. *Genome Biol.* **24**, 97 (2023).
31. van Loggerenberg, W. et al. Systematically testing human HMBS missense variants to reveal mechanism and pathogenic variation. *Am. J. Hum. Genet.* **110**, 1769–1786 (2023).
32. MacArthur, D. G. et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
33. Stenson, P. D. et al. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9 (2014).
34. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
35. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
36. Jagadeesh, K. A. et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).
37. Zhang, H., Xu, M. S., Fan, X., Chung, W. K. & Shen, Y. Predicting functional effect of missense variants using graph attention neural networks. *Nat. Mach. Intell.* **4**, 1017–1028 (2022).
38. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14**, S3 (2013).
39. Dong, C. et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2014).
40. Qi, H. et al. MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.* **12**, 510 (2021).
41. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at *bioRxiv* https://doi.org/10.1101/148353 (2017).
42. Fuller, Z. L., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
43. Cassa, C. A. et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).
44. Weghorn, D. et al. Applicability of the mutation–selection balance model to population genetics of heterozygous protein-truncating variants in humans. *Mol. Biol. Evolution* **36**, 1701–1710 (2019).
45. Agarwal, I., Fuller, Z. L., Myers, S. R. & Przeworski, M. Relating pathogenic loss-of function mutations in humans to their evolutionary fitness costs. *eLife* **12**, e83172 (2023).
46. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat. Genet.* **56**,1632–1643 (2024).
47. Huang, Y.-F. & Siepel, A. Estimation of allele-specific fitness effects across human protein-coding sequences and implications for disease. *Genome Res.* **29**, 1310–1321 (2019).
48. Nei, M. The frequency distribution of lethal chromosomes in finite populations. *Proc. Natl Acad. Sci.* **60**, 517–524 (1968).
49. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
50. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
51. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
52. Christmas, M. J. et al. Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023).
53. Gao, H. et al. The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
54. Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat. Commun.* **13**, 3895 (2022).
55. Schubbert, S. et al. Germline KRAS mutations cause Noonan syndrome. *Nat. Genet.* **38**, 331–336 (2006).

56. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
57. Maenner, M. J. et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 Sites, United States, 2020. *MMWR Surveill. Summ.* **72**, 1–14 (2023).
58. Power, R. A. et al. Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**, 22–30 (2013).
59. Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
60. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
61. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
62. He, X. et al. Integrated model of De Novo and inherited genetic variants yields greater power to identify risk genes. *PLOS Genet.* **9**, e1003671 (2013).
63. Nguyen, H. T. et al. Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Med.* **9**, 114 (2017).
64. Fu, J. M. et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.* **54**, 1320–1331 (2022).
65. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic Intolerance to functional variation and the interpretation of personal genomes. *PLOS Genet.* **9**, e1003709 (2013).
66. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
67. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci.* **118**, e2016239118 (2021).
68. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv. neural Inf. Process. Syst.* **34**, 29287–29303 (2021).
69. Ohta, T. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**, 263–286 (1992).
70. Weinstein, E., Amin, A., Frazer, J. & Marks, D. Non-identifiability and the blessings of misspecification in models of molecular fitness. *Adv. neural Inf. Process. Syst.* **35**, 5484–5497 (2022).
71. Verkuil, R. et al. Language models generalize beyond natural proteins. Preprint at *bioRxiv*, https://doi.org/10.1101/2022.12.21.521521 (2022).
72. Wall, J. D. et al. The genomeAsia 100K project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
73. Wonkam, A. Sequence three million genomes across Africa. *Nature* **590**, 209–211 (2021).
74. Howe, K. L. et al. Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2020).
75. Consortium, T. U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2020).
76. Danecek, P. & McCarthy, S. A. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**, 2037–2039 (2017).
77. Zhao, Y. ShenLab/MisFit: MisFit v1.5, a probabilistic graphical model for estimating selection coefficients of nonsynonymous variants from human population sequence data. *Zenodo*, https://doi.org/10.5281/zenodo.15230825 (2025).
78. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).
79. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. Preprint at *arXiv*, https://doi.org/10.48550/arXiv.1603.04467 (2016).
80. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 352–352 (2020).
81. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

We thank gnomAD and UK biobank investigators for making genomic data available. We thank the SPARK study and especially the thousands of families participating in SPARK. This work is supported by NIH grants: R35GM149527 (Y.S.), R01GM120609 (Y.S.), and P50HD109879 (Y.S. and W.K.C.), Simons Foundation: SFARI #1019623 (Y.S.), and Columbia Precision Medicine Pilot grants program (Y.S.). We thank Itsik Pe'er, Molly Przeworski, Mohammed AlQuraishi, Ben Hu, Adi Garg, Alan Tian, Haicang Zhang, Na Zhu, Chang Shu, and Lu Qiao for helpful discussions.

Author contributions

Y.S. conceived the concepts and guided the study. Y.Z. developed methods and performed the main analysis. T.L. contributed to interpreting autism variant results. G.Z. and J.H. aided in processing genomic datasets. Y.Z., T.L., G.Z., J.H., H.P., W.K.C., Y.S. all contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-59937-2.

Correspondence and requests for materials should be addressed to Yufeng Shen.

Peer review information *Nature Communications* thanks Xin He and Luke O'Connor for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025