Supplementary Materials

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

Zhao et al

Table of Contents

Supplementary Notes

Poisson-Inverse-Gaussian model for approximating allele count likelihood

In MisFit, the probability of observed allele counts given a selection coefficient in samples is described by a Poisson-Inverse-Gaussian distribution optimizing recent population growth. The choice of Inverse-Gaussian for population allele frequency is heuristic, but it takes account of the long tail, 0 density at 0 and conjugacy to Poisson distribution. In comparison, Nei's model uses a Gamma distribution to approximate the allele frequency, and the density at 0 would be infinity when optimized for a small population size, which leads to a distorted distribution of the resulting Negative Binomial distribution of allele counts. Recent work enables efficient multiplication of the transition matrix of the Discrete-Time Wright-Fisher model[1] to better picture the likelihood, but the PIG model is much simpler with just a few parameters and can be applied to any choice of mutation rate and selection coefficient in a continuous range, which is more applicable to the scale of missense variants. We showed that our method is accurate in estimating relatively strong selection ($s > 0.01$) by simulations.

MisFit estimates selection coefficient with amino acid resolution

The main output of MisFit is estimated selection coefficient of an individual missense variant, MisFit_S. We show an example of the phosphatase tensin-type domain in the gene *PTEN* (Supplementary Fig. 9). *PTEN* is a well-known disease risk gene, depleted of missense variants (missense z-score 3.49, o/e = 0.33) and protein truncating variants (pLI 0.26, o/e=0.24). MisFit_S follows secondary-structure patterns, and correlates well with conservation. None of the gnomAD[2] constraint metrics can describe the selection of these sub-gene features. The regional constraint[2], which is calculated by o/e for each 1kb genomic region, provides limited information and low resolution.

MisFit for *de novo* and inherited variants

Theoretically, the ratio of *de novo* mutations among all observed variants given a s equals to s. This only holds for a new generation that has not gone through any selection, so that a proportion of mutation rate ν of variants are newly mutated and $(1-s)\nu/s$ are inherited, reaching an equilibrium allele frequency of ν/s. In reality, selection is more complicated as a long process starting before birth through post-reproductive age, and there are overlapping of generations. During MisFit training, we assume the samples are purely post-selection samples of relatively old age (e.g., 40 to 69 years old in UKBB), which affects the allele counts distribution for very large s (Fig. 1a, the part for $s > 0.1$).

In our analysis, we show that *de novo* ratio approximates s in the autism cases more than in the controls. In the unaffected siblings, inherited variants with large s are not significantly depleted as *de novo* variants. The vast majority of inherited missense variants may not play a role in autism based on previous overall burden analysis[3].

MisFit performance compared with EVE

We compared our model with several popular computational methods. EVE[4] is also a state-of-art methods, which is an Bayesian variational autoencoder using multiple sequence alignments as inputs. As EVE only generates scores for 3,219 known disease genes with good MSA, we also limit the deep mutational scan data to a subset of 14 genes with all prediction scores available for fair comparison (Supplementary Fig. 20). Generally, MisFit_D has a similar performance with EVE. Here we used the EVE-predicted posterior probability of damaging component from their mixture model, which is already adjusted specifically for each gene.

In *de novo* variants analysis, variants without EVE annotations are regarded as least damaging. This largely impairs sensitivity, but the top risk variants are less affected as they are likely to come from known disease associated genes with good coverage of EVE.

MisFit model on predicting ClinVar variants

We also analyzed the concordance with ClinVar labelled variants. ClinVar variants were processed from the version of Dec 2022. Variants with 'criteria provided' (1 review star) were collected, and the ones with conflicting labels or annotated as 'variant of uncertain significance' were removed. Pathogenic / likely-pathogenic variants are regarded as positive and benign / likely-benign variants are regarded as negative. To eliminate the gene-level bias, we selected genes with at least one positive label and one negative label, and down-sample the variants to make equal number of positive and negative label in each gene. Finally, this gives out 33,046 variants in 3,246 genes. While AlphaMissense has a best performance in this analysis, MisFit_D is also reasonably good (Supplementary Fig. 21).

Although this analysis is informative in clinical applications, we still lack totally independent data for testing. Supervised methods are usually trained on these labels from ClinVar, or from similarly curated databases. Additionally, some methods (such as CADD and increasingly REVEL) are commonly used as one of the criteria for annotating pathogenicity in ClinVar.

Point estimate for selection coefficient

Under the simplest Poisson assumption, where $m \sim Pois(nv/s)$, we can easily use $s_{MLE} = nv/m$ as a point estimate by maximum likelihood estimation. However, as s is in the denominator of the Poisson mean parameter, s_{MLE} is naturally a biased estimator of s. We prove it here.

$$E(s_{MLE}) = E\left(\frac{nv}{m}\right) = nvE\left(\frac{1}{m}\right) \geq \frac{nv}{E(m)} = \frac{nv}{nv/s} = s$$

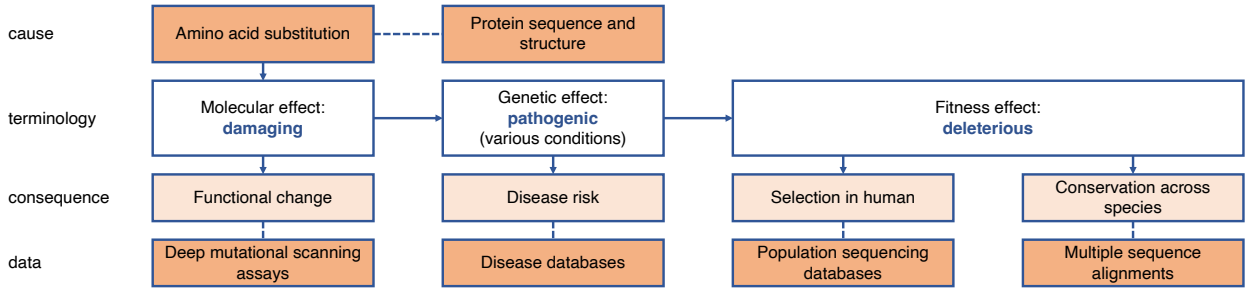
Here, $E\left(\frac{1}{m}\right) \geq \frac{1}{E(m)}$, because $f(x) = 1/x$ is a convex function, and thus $E\big(f(x)\big) \geq f(E(x))$ (Jensen's inequality).

When considering a more realistic population genetics model, the situation can be more complicated, but $s_{MLE} = argmax(p(data|s))$ is still biased because the relationship as a denominator holds. Adding a prior distribution may alter the estimation. Although those Bayes approaches enable describing a whole posterior distribution of s, we usually still need to derive a point estimate for easier downstream analysis. As there are no standard criteria of doing this, several previous studies[5-7] used posterior mean as point estimate, defined as $E(s|data)$. In this study, because every s in the model is transformed in logit scale with $s' = \log\left(\frac{s}{1-s}\right)$, we directly used posterior mean in the logit scale $E(s'|data)$, and transformed it back to original scale, which gives $sigmoid(E(s'|data))$. This value is different with $E(s|data)$ when limited data is provided, but could be potentially less biased. (Supplementary Fig. 22)

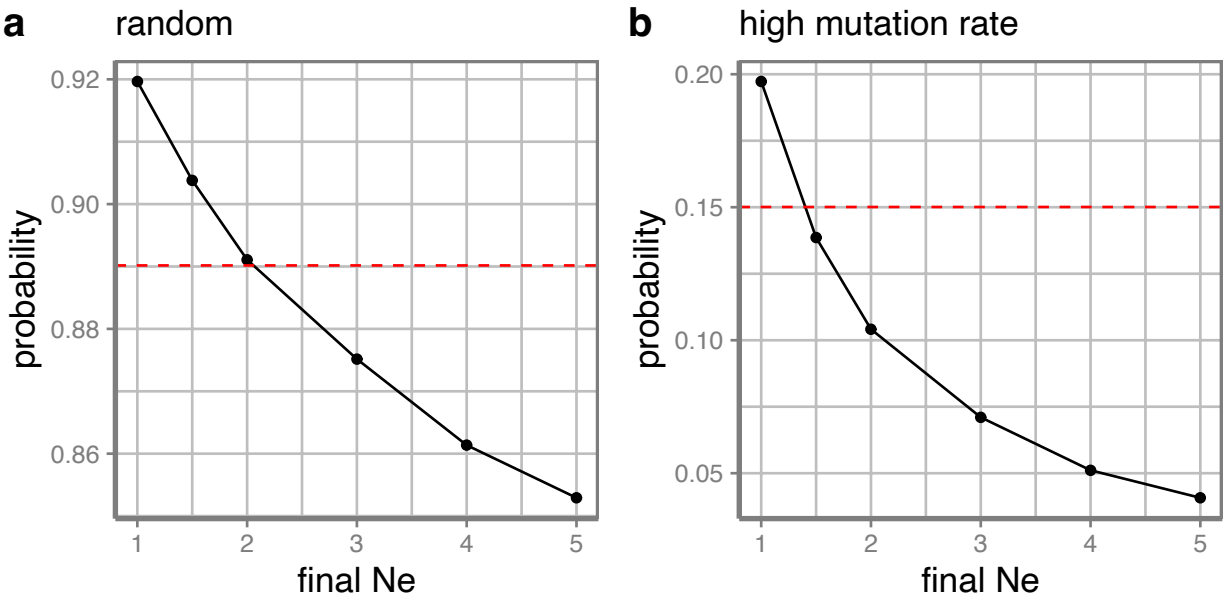
Improve estimation of selection by using genomes from different populations

We showed that adding sequencing samples from the same population does not help with estimating small select coefficient, but adding samples from another population improves the estimation. Assuming the second population evolves totally independently from the first population, such improvement is comparable to doubling the integrated number of variants with same selection coefficient. If the two populations split very recently, the allele frequencies are highly correlated and provide limited additional information, and the result should be the same as adding samples from the same population. We simulated the allele frequencies in two hypothetic populations assuming different length of independent evolution (Supplementary Fig. 23). 2,000 generations (approximating split time of Europeans and Africans) is already long to give out less correlated data, except for very common variants with allele frequencies larger than 0.1.

Supplementary Figures

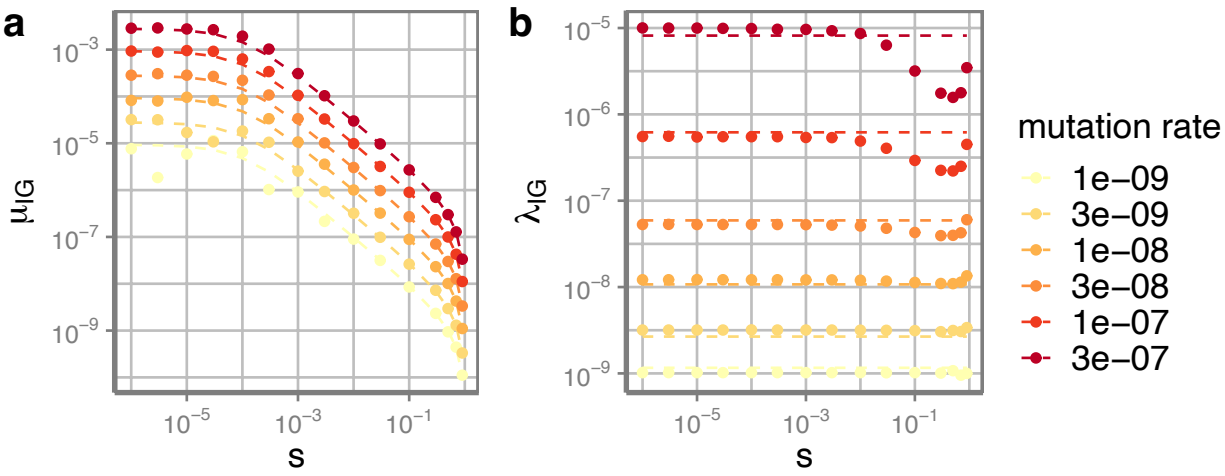


Supplementary Fig 1 Missense variant effect from different aspects

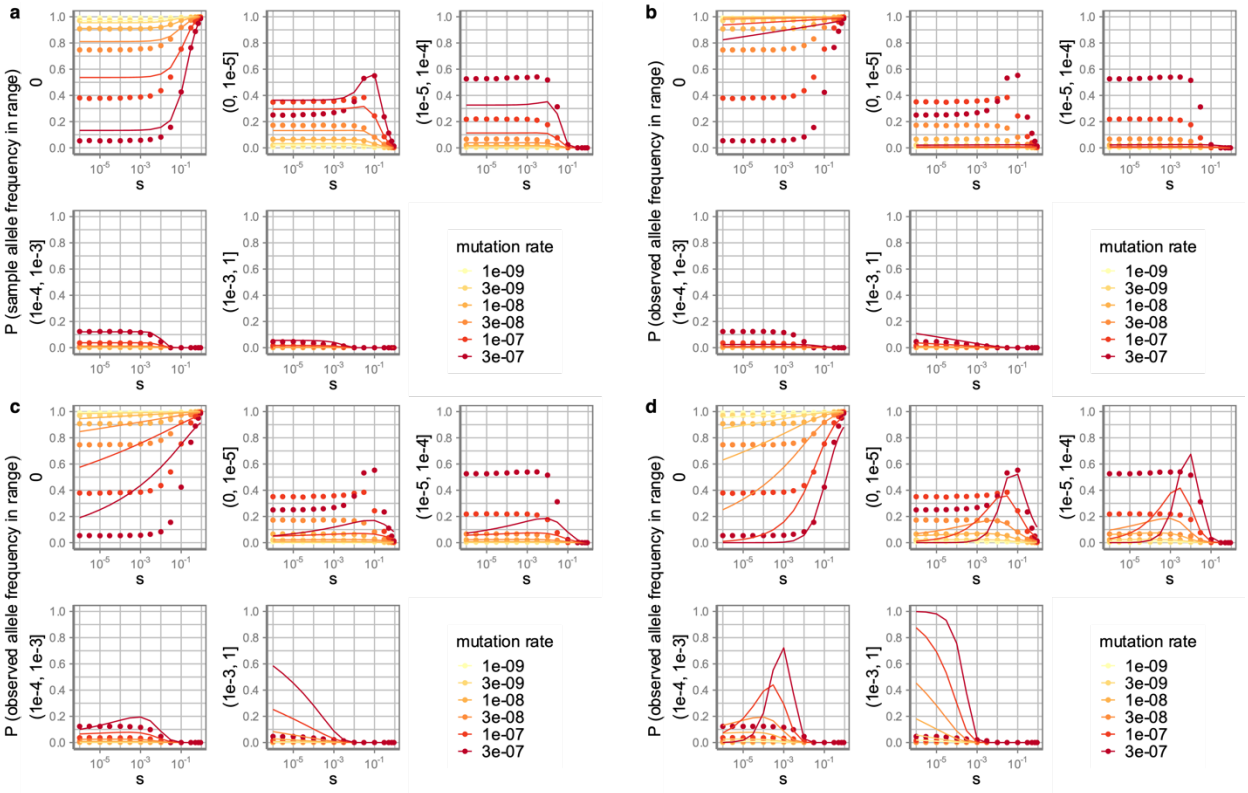


Supplementary Fig 2 Adjusted European final effective population size in simulation

Synonymous variants are simulated based on European effective population size history with different final population size in the latest generation. Probabilities of zero-count allele in simulation (black) and in UKBB plus gnomAD NFE samples (red) are shown. **a** synonymous variants are randomly selected with an average mutation rate of 1e-8. **b** Only C-to-T synonymous variants in CpG sites with mutation rate larger than 1e-7 are selected.

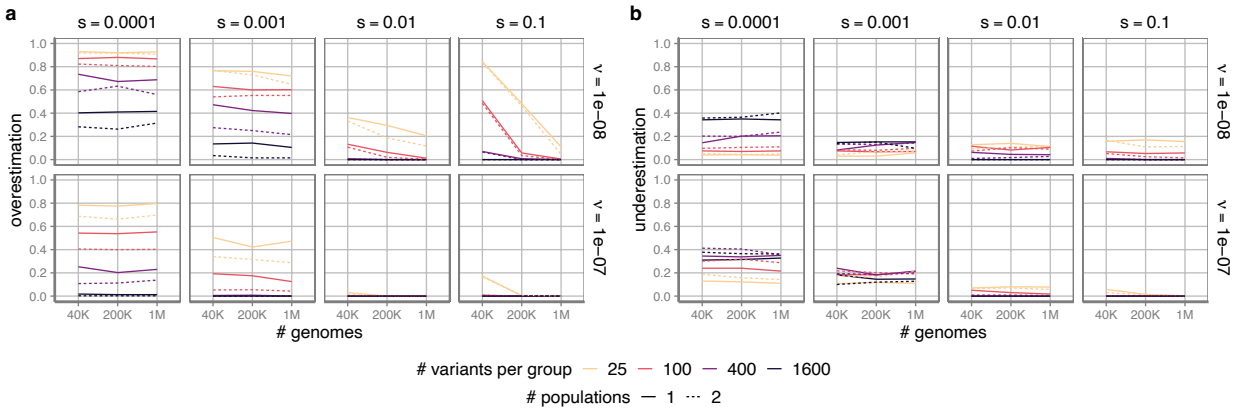


Supplementary Fig 3 Parameters of Poisson-Inverse-Gaussian model
μ_{IG} and λ_{IG} are mean and shape parameters for an Inverse Gaussian distribution for modeling population allele frequency. Dots are the best fits for each simulation condition, while dashed lines are from optimized functions of mutation rate and selection coefficient.



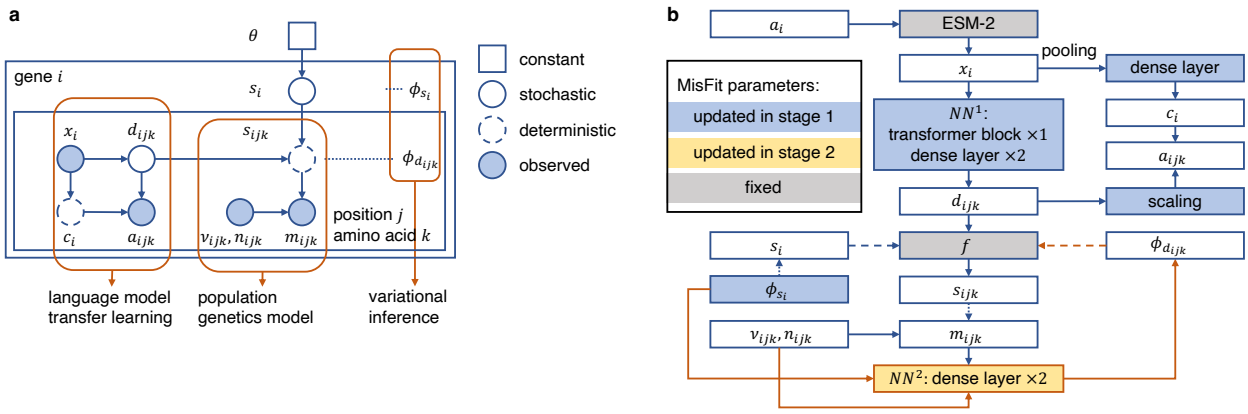
Supplementary Fig 4 Distribution of sample allele frequency under different population genetics model

a PIG model used in MisFit **b-d** Negative Binomial distribution with effective population size of **b** 10,000 **c** 100,000 **d** 1,000,000. Sample size is 200K diploid genomes.



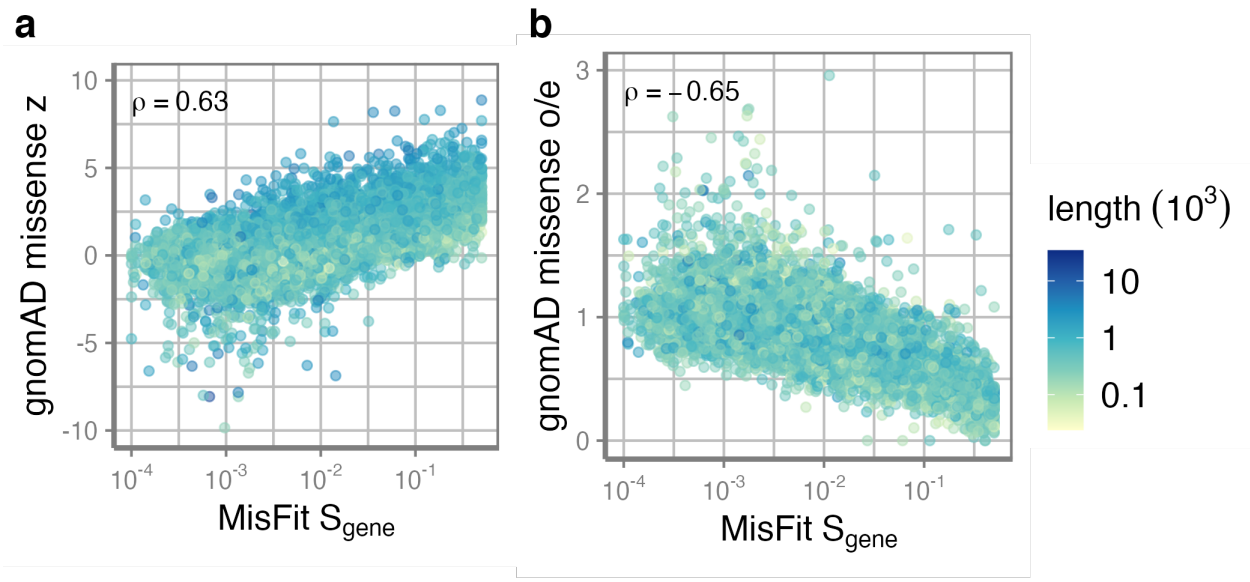
Supplementary Fig 5 Evaluation of MLE estimation of *s*

Probabilities of **a** overestimation **b** underestimation in 400 replications for each simulation condition are shown. Here s is a categorical variable of [0.00001, 0.0001, 0.001, 0.01, 0.1, 1]. 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).

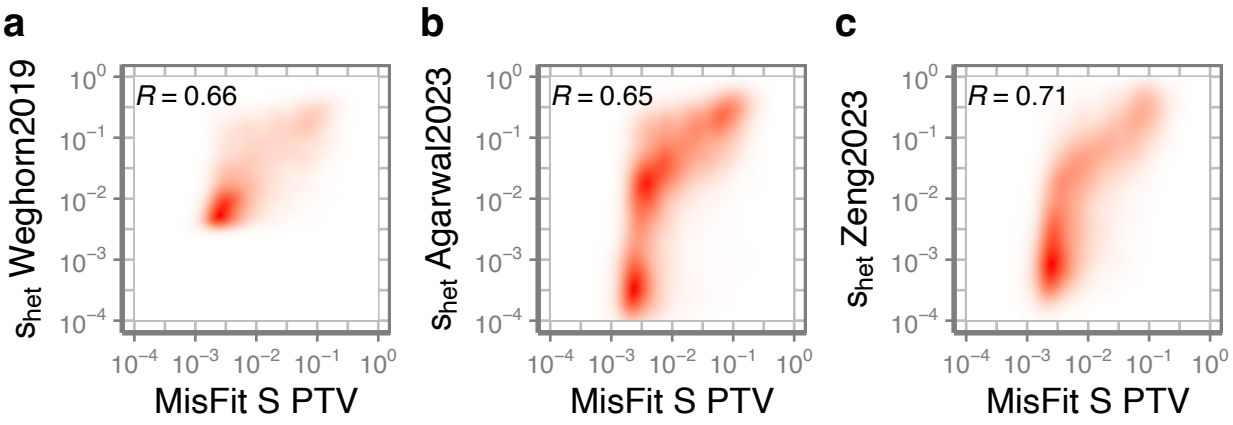


Supplementary Fig 6 Overview of MisFit model

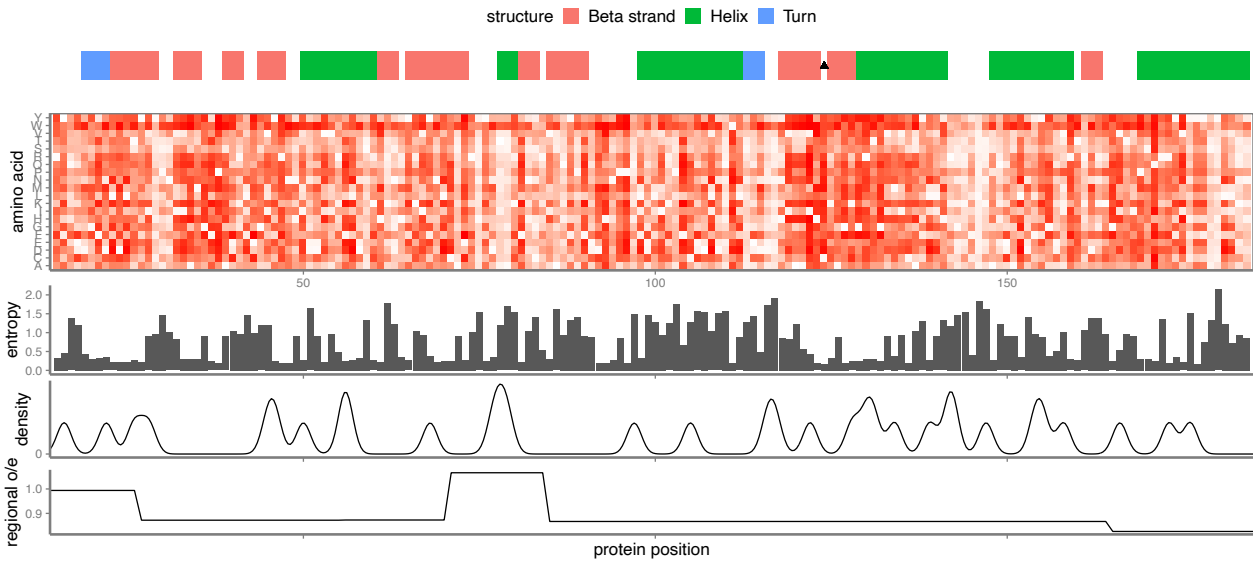
a MisFit model in view of a probabilistic graphical model. **b** Full structure of MisFit model and training stages.



Supplementary Fig 7 MisFit estimated gene-level missense selection correlates with gnomAD missense z score or o/e

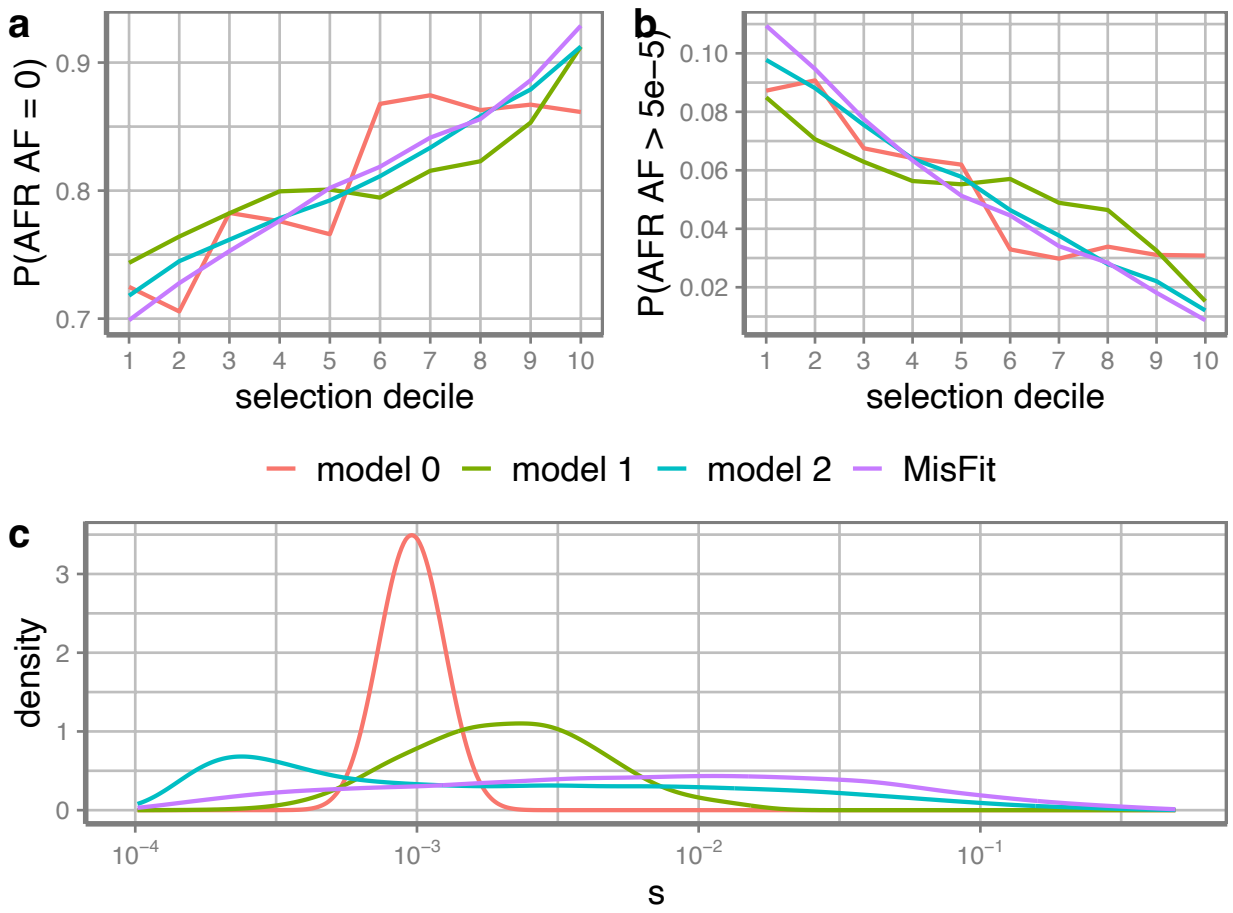


Supplementary Fig 8 MisFit estimated s for protein-truncating variants correlates with previous estimations.



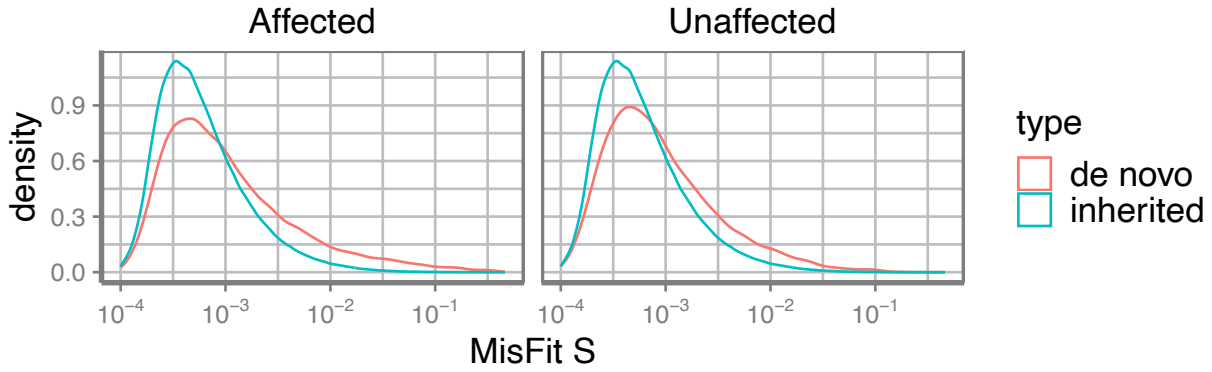
Supplementary Fig 9 Predicted selection coefficient for each amino acid substitution in PTEN

Secondary structures from UniProt are shown at the top. Entropy is calculated by amino acid distribution across Ensembl homologues, lower the value means more conservation. Missense variant density combines UKBB and gnomAD NFE. Regioal o/e is extracted from gnomAD-3 1kb window.

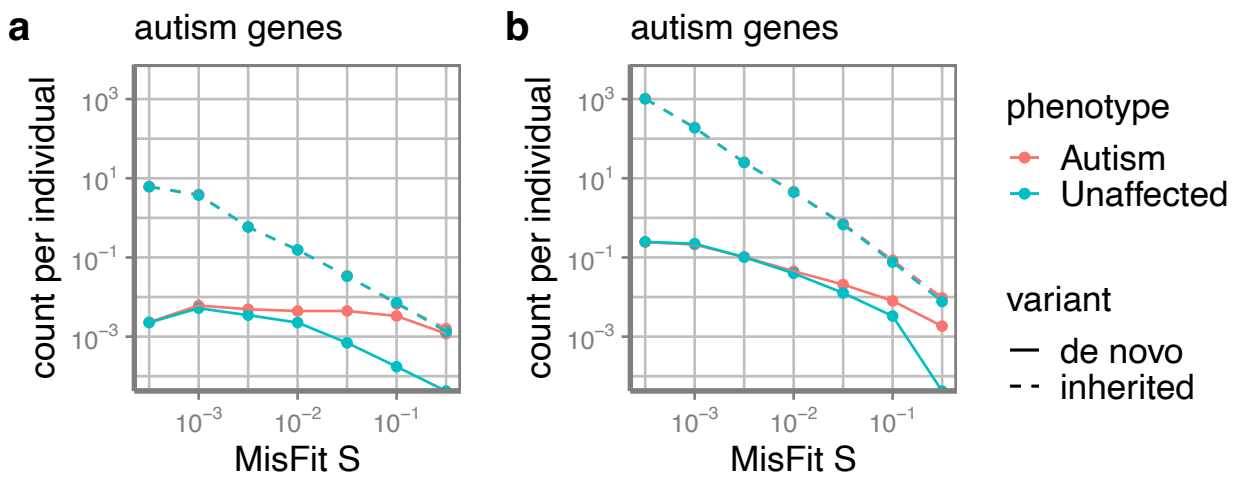


Supplementary Fig 10 MisFit estimated selection coefficient s predict allele frequency in a second population better than baseline models

Variants of high mutation rate (>1e-7) with sample AF < 5e-6 in training set (UKBB + gnomAD NFE) are selected for analysis. Variants are separated into 10 groups by estimated s in each model. The proportions of variants with **a** gnomAD AFR sample AF = 0 or **b** gnomAD AFR sample AF > 5e-5 are shown. **c** Distribution of estimated s of these variants. Model 0: baseline model learned only from NFE allele number, NFE allele count, and mutation rate; model 1: model 0 + gene-level selection; model 2: model 1 + ESM-2 zero shot as d; MisFit: model 1 + ESM-2 embeddings as d.

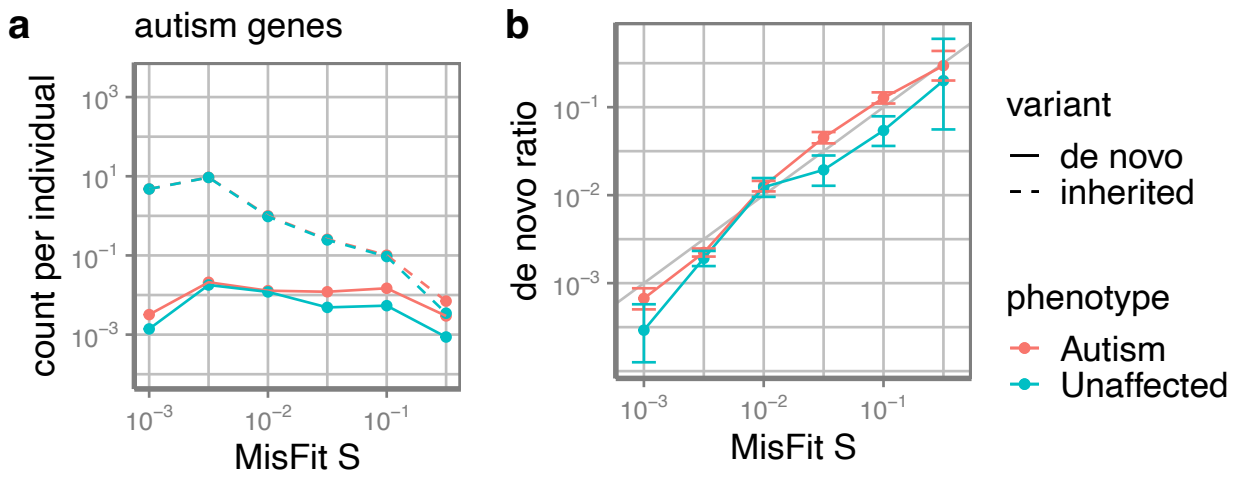


Supplementary Fig 11 Distribution of inherited or de novo missense variants in autism dataset



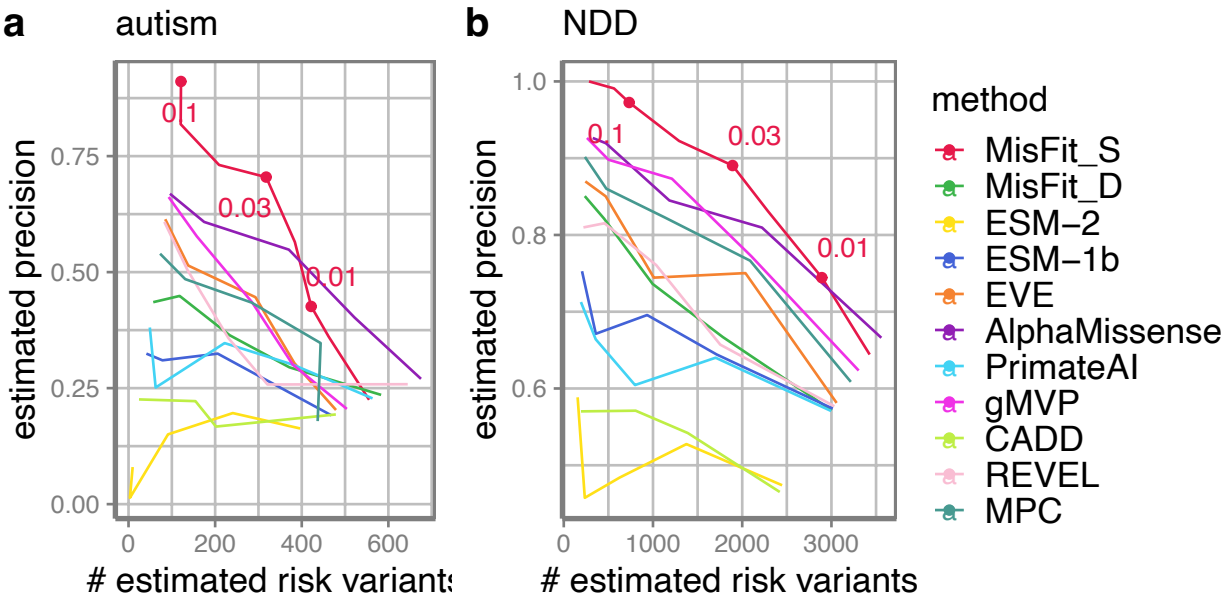
Supplementary Fig 12 Count of de novo or inherited variants binned by of MisFit_S in different gene sets

Dashed line indicates the inherited, solid line indicates the *de novo*. **a** known 162 autism genes from SPARK. **b** other genes.



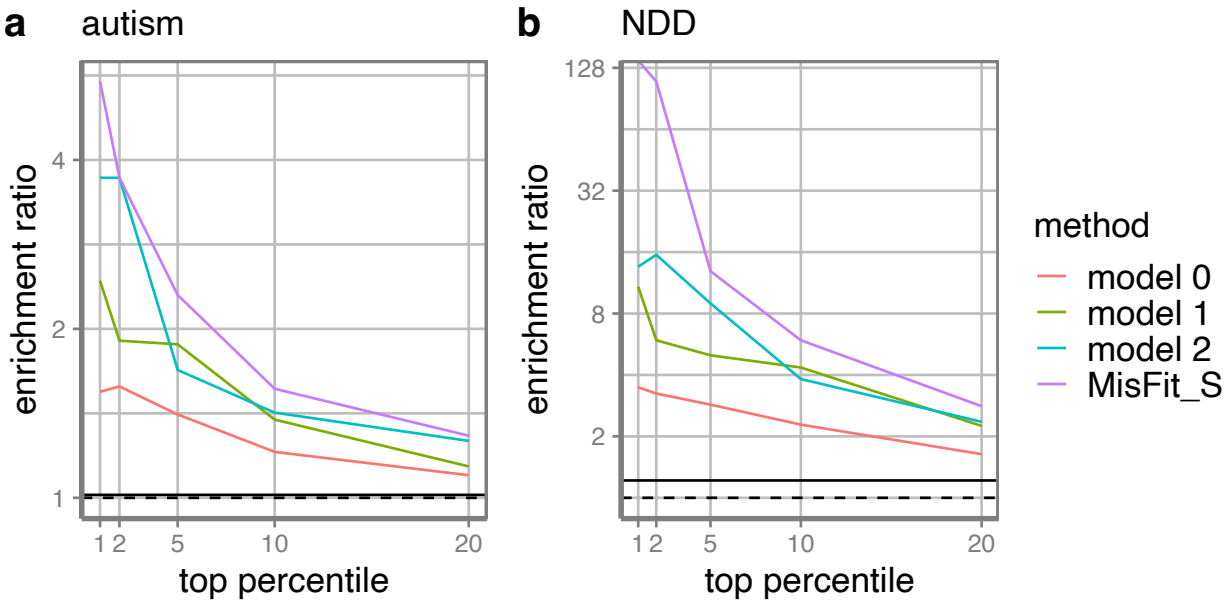
Supplementary Fig 13 de novo or inherited protein-truncating variants binned by of MisFit_S

a Count of *de novo* or inherited missense variants. Dashed line indicates the inherited, solid line indicates the *de novo*. **b** The proportion of *de novo* to all variants in autism dataset. Error bars show 95% confidence intervals.

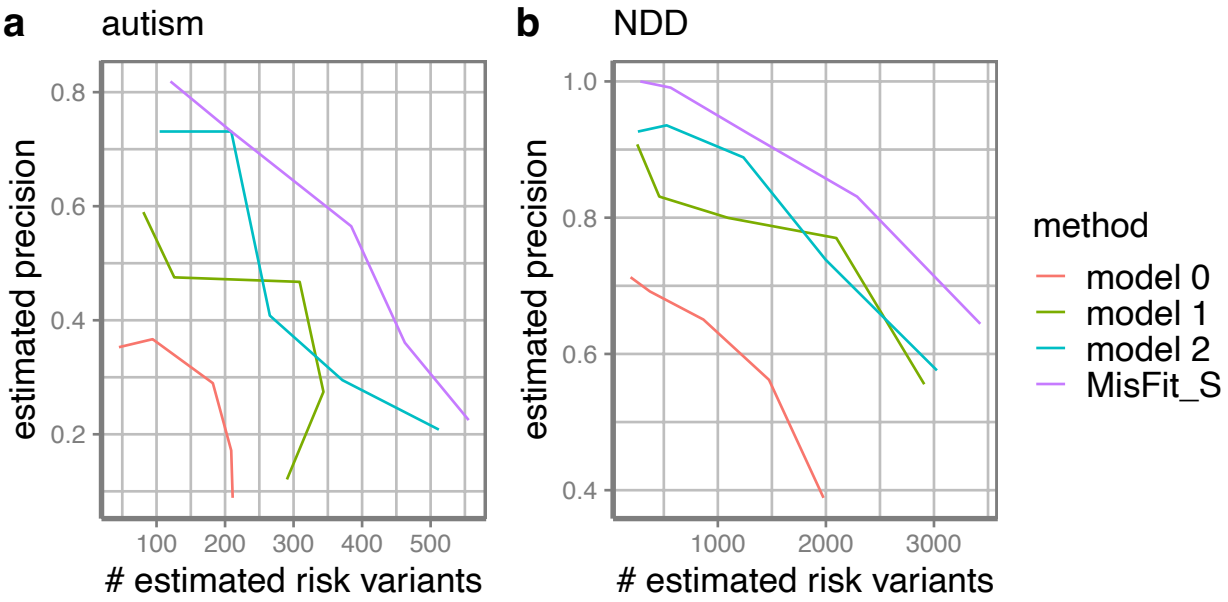


Supplementary Fig 14 Precision-recall-proxy curve for de novo missense variants

Thresholds of MisFit_S are annotated.

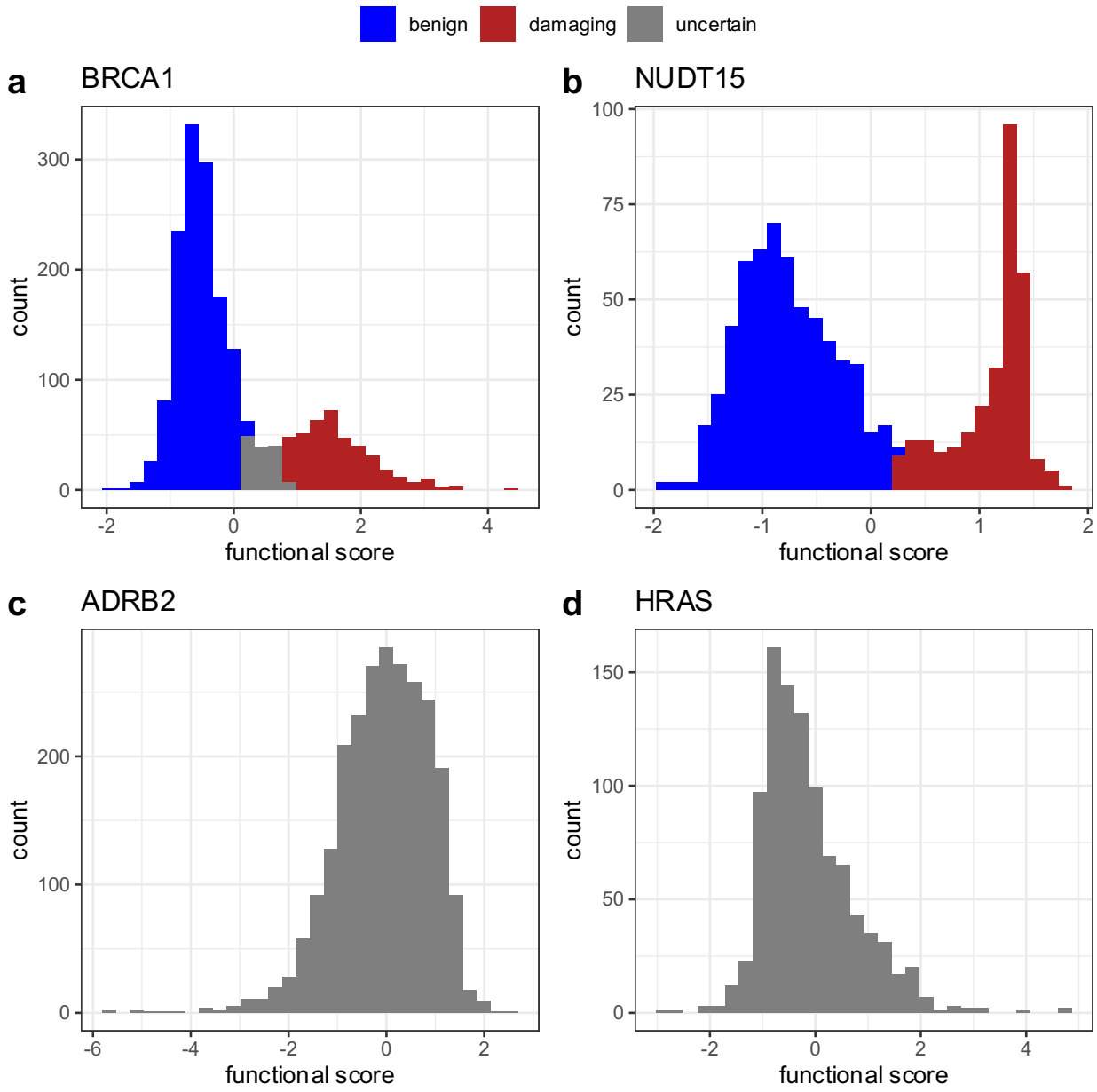


Supplementary Fig 15 Enrichment of de novo variants in baseline models
Model 0: baseline model learned only from NFE allele number, NFE allele count, and mutation rate; model 1: model 0 + gene-level selection; model 2: model 1 + ESM-2 zero shot as d; MisFit: model 1 + ESM-2 embeddings inferred d.



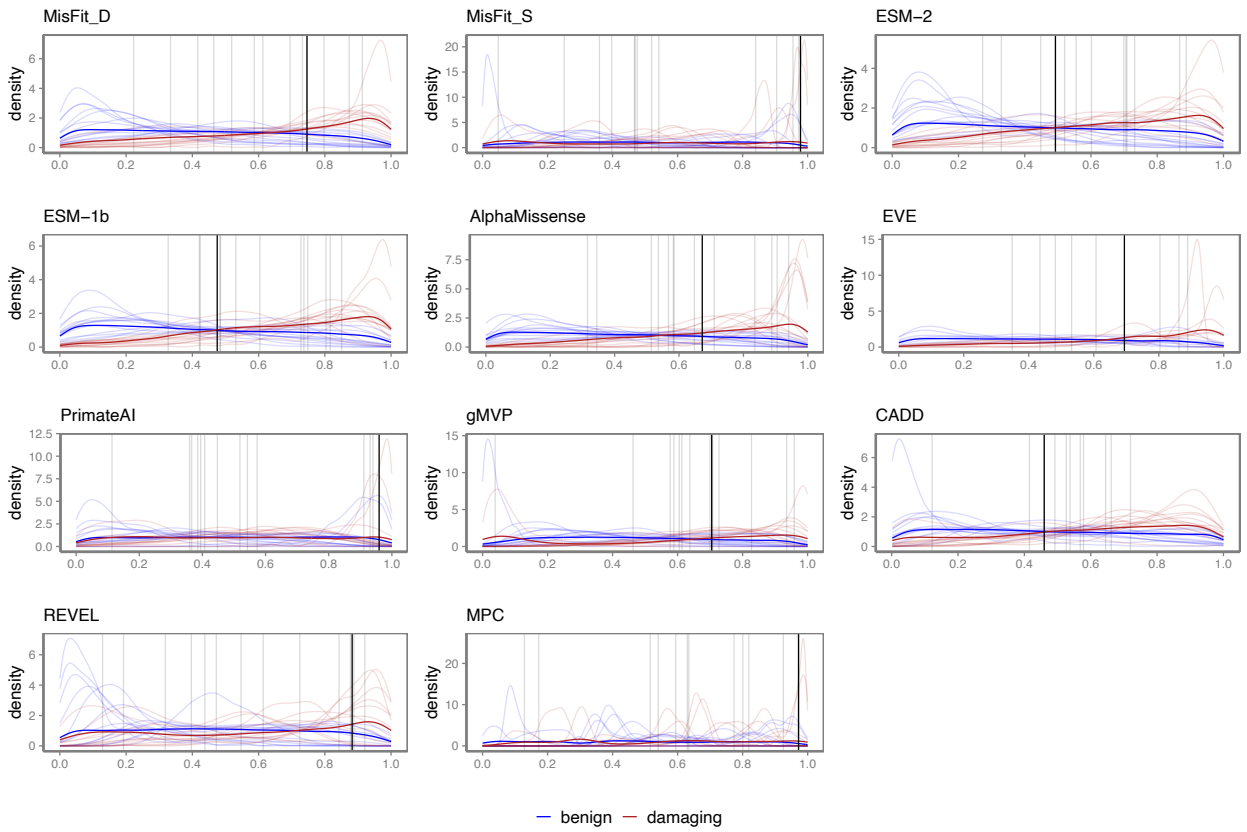
Supplementary Fig 16 Precision-recall-proxy curve of de novo variants in baseline models

Model 0: baseline model learned only from NFE allele number, NFE allele count, and mutation rate; model 1: model 0 + gene-level selection; model 2: model 1 + ESM-2 zero shot as d; MisFit: model 1 + ESM-2 embeddings inferred d.



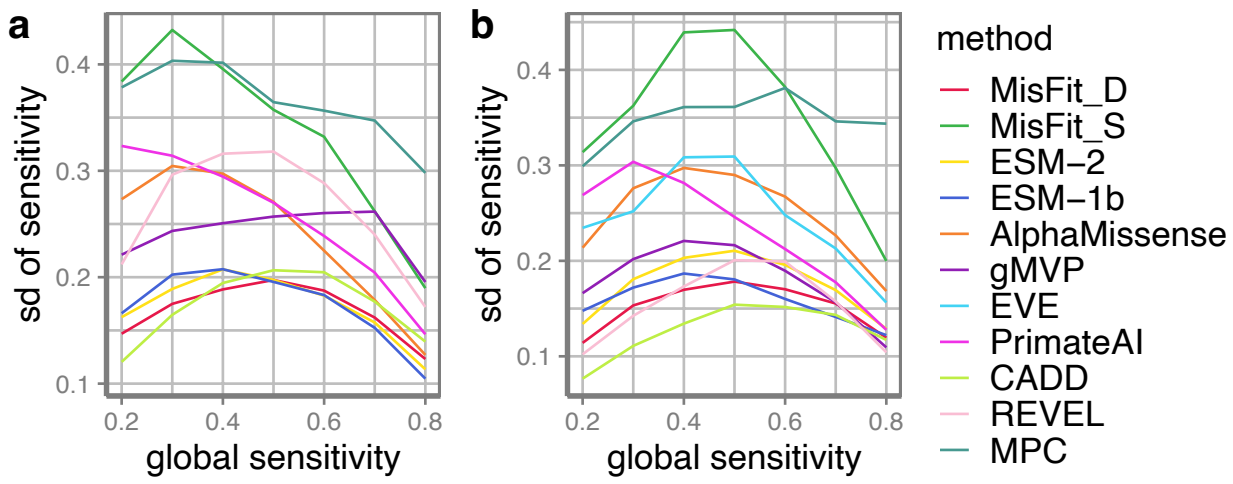
Supplementary Fig 17 Various functional score distributions in different deep mutational scanning experiments

a-b Examples of bimodal distribution with originally annotated labels. **c-d** Examples of unimodal distribution.



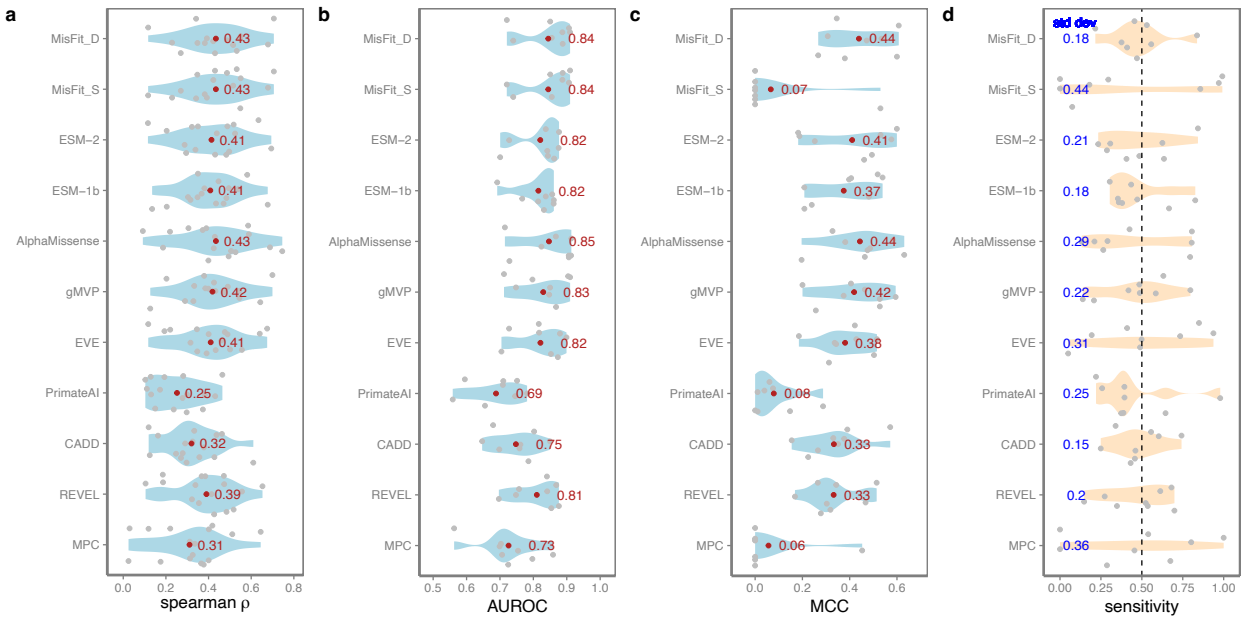
Supplementary Fig 18 Distribution of scores (normalized by rank) across genes

The red and blue curves show the distribution of damaging and benign variants, respectively. Dark curves are the distribution of scores in the combined data, while light curves show each gene separately. The black lines are the optimal threshold achieving highest MCC in the combined dataset, while the grey lines are the optimal threshold for each gene.



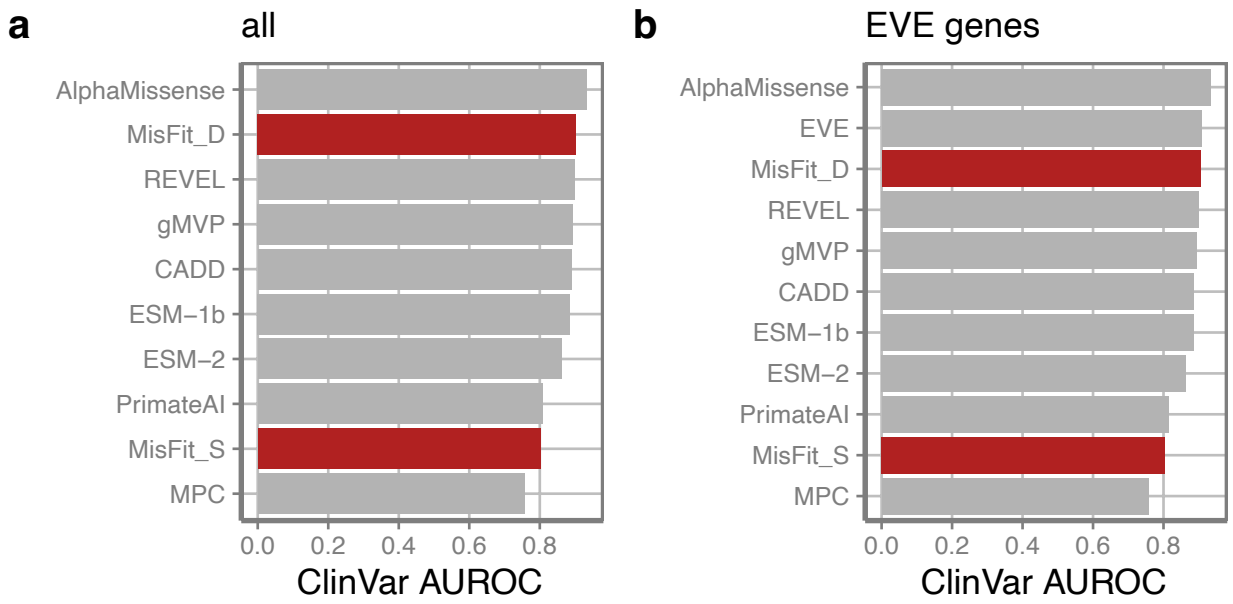
Supplementary Fig 19 Consistency of sensitivity across genes

Thresholds are set to achieve certain global sensitivity (x-axis) in the combined data. Sensitivities in different genes are then evaluated. Y-axis shows the standard deviation. **a** in all genes; **b** in EVE genes.



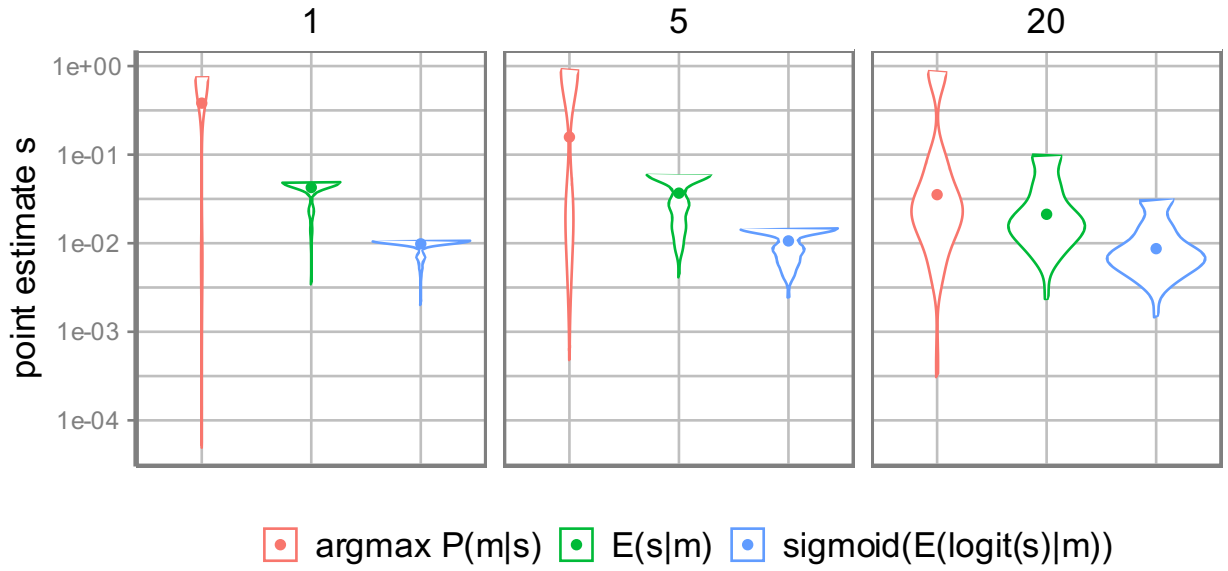
Supplementary Fig 20 Performance in predicting damaging variants in deep mutational scanning assays and cross-gene consistency

A subset of genes from **Fig. 7** with all prediction scores available are used for analysis. **a** Spearman correlation coefficient of predicted scores with functional scores from deep mutational assays. Mean is annotated in red. **b** AUROC of predicting confidently labeled damaging or benign variants in deep mutational assays. Mean is annotated in red. **c** 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.



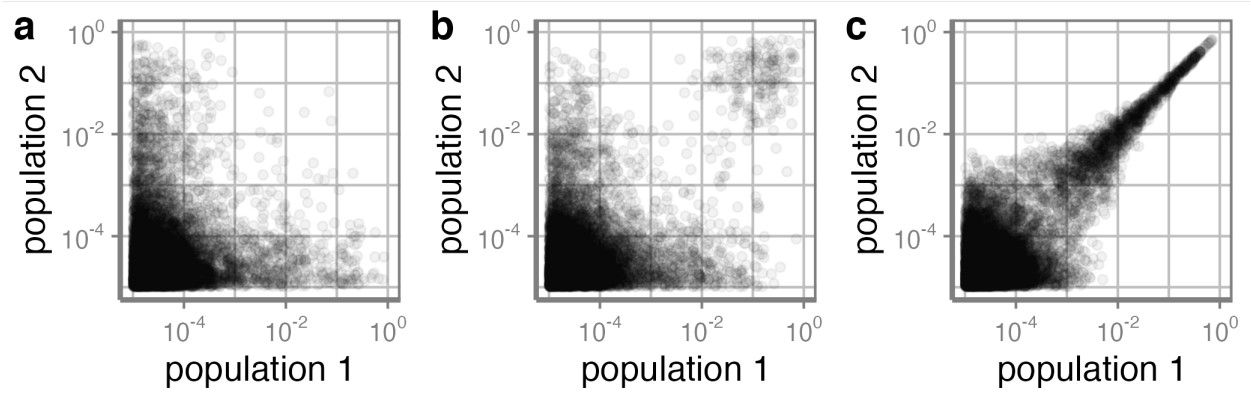
Supplementary Fig 21 AUROC in predicting ClinVar balanced pathogenic / benign variants

a 33,046 variants in 3,246 genes. **b** 26,171 variants in 2,229 genes with all prediction scores available.



Supplementary Fig 22 Bias of point estimate of heterozygous selection coefficient

The number on top shows the aggregated number of variants per group, and y-axis shows the distribution of s estimated for each group in 100 replicates. The simulation condition is $s = 0.01$, so the mean of point estimate (dot) closer to this value meaning the estimator is less biased. Here mutation rate is 1e-8 and sample size is 200K.



Supplementary Fig 23 Correlation of population allele frequency in two populations

The two simulated populations split **a**. 10,000 generations ago **b** 2,000 generations ago **c** 200 generations ago. Mutation rate is 1e-7 and heterozygous selection coefficient equals 1e-4.

Supplementary Tables

Supplementary Table 1

MisFit estimated gene-level selection for 830 genes with known disease mechanisms.

Supplementary Table 2

Genetic variants data used in analysis.

Supplementary Table 3

Deep mutational scanning experiments used in analysis and Spearman correlation coefficient with computational methods.

Supplementary Table 4

A subset of deep mutational scanning experiments that have bimodal distribution of functional scores used in analysis, and AUROC of computational methods.

Supplementary Table 5

MisFit estimated gene-level selection (including missense and protein-truncating variants) for all genes trained in model.

References

1. Jeffrey, P.S., Tony, Z., Hakhamanesh, M. & Jonathan, K.P. Scaling the Discrete-time Wright Fisher model to biobank-scale datasets. *bioRxiv*, 2023.05.19.541517 (2023).
2. Siwei, C. *et al.* A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv*, 2022.03.20.485034 (2022).
3. Zhou, X. *et al.* Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nature Genetics* **54**, 1305-1319 (2022).
4. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91-95 (2021).
5. Cassa, C.A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nature Genetics* **49**, 806-810 (2017).
6. Tony, Z., Jeffrey, P.S., Hakhamanesh, M. & Jonathan, K.P. Bayesian estimation of gene constraint from an evolutionary model with gene features. *bioRxiv*, 2023.05.19.541520 (2023).
7. Weghorn, D. *et al.* Applicability of the Mutation–Selection Balance Model to Population Genetics of Heterozygous Protein-Truncating Variants in Humans. *Molecular Biology and Evolution* **36**, 1701-1710 (2019).