nature portfolio

Peer Review File

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

Corresponding Author: Dr Yufeng Shen

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The manuscript by Shen and colleagues proposed a novel computational method, MisFit, to infer fitness effects of protein-coding variants. The manuscript correctly pointed out that existing methods for predicting "pathogenicity" of variants are confounded by gene level effects. The fitness effect of a variant, as the manuscript explained, should depend on both the damaging effect of the variant on the gene function, and the strength of natural selection at the gene level. Separating these two effects can be important in some contexts, e.g. when studying mutational effects on a particular gene in mutational scanning. Importantly, when estimating the fitness of a variant, jointly studying the two effects allows one to borrow information from other variants in the same gene to improve the estimation. The authors validated the new method by various analyses. Using variant frequency data, de novo mutations (DNM), and deep mutation scanning (DMS) data, the authors show that variants predicted by MisFit to have large damaging effects on genes (D-score), or large selection coefficients (S-scores), are highly enriched with likely functional variants. In the case of DNM data, in particular, MisFit predicted variants by S-scores are substantially more enriched in likely risk variants of autism and developmental disorders than those from all existing methods.

Overall, this is a technically novel and solid method with potentially broad impact on the field. There are several key technical innovations that make the method possible. The first is the use of protein language model based embeddings in the method. This allows one to have an informative prior for estimating D, the damaging effects of variants. The second is several computational innovations that reduce the computational burden, including the Poisson-Inverse-Gaussian (PIG) approximation of variant count distribution, and the Variational Inference method in training model parameters. In addition to the method itself, these computational innvoations are likely interesting to researchers developing related methods.

While the paper overall is well-written, some parts are not explained clearly. Below are specific comments:

Major comments:

(1) The section "MisFit model architecture and parameters" in Methods is somewhat difficult to follow. First, the section actually talked about two models: one full model, and one baseline model where the D and S values of variants depend only on global parameters. It would make it easier to follow if the authors can separately explain the two models. Secondly, in statistical literature, one usually starts with the description of the probablistic models, and then explains the inference procedure. But in this section, the two are mixed together without a clear boundary. For example, Equation (19) should occur after the modeling part, Equations (20) - (22). Thirdly, it would be useful to add some explanation of the rationale of some key steps. Importantly, the method has two Neural networks: NN1 and NN2. The first one, NN1, is a function linking embedding to D values, and is relatively easy to understand. But the idea behind NN2 is not explained clearly. The language here is quite vague, "in order to retrieve mu, sigma, in one forward pass ...". My understanding is that in the probablistic model, the D and S values are latent variables, so the Variational Inference uses another NN to approximate their posterior distribution of D, given embedding and allele counts. But this point should be made more explicit. I also feel it may be helpful to draw an analogy with Variational Autoencoders (VAE) here to faciliate understanding.

(2) The model training process is hard to follow. In both the main text, and the "Model training" section in Methods, the

manuscript described the two stages of training, with limited explanations. Why is the inference done in two stages? What is the rationale of using only constrained genes in the first stage? In the first stage of training, only NN1 related parameters are learned. But how does training happen when NN2 is unknown?

(3) The method infers several scores, D, S and S.gene. These are all point estimates. But as the authors pointed out, it can be difficult to estimate selection for individual variants with relatively mild effects. So it may be interesting to quantify uncertainty of predicted effects. A user may, for example, focus only on likely functional variants with high confidence. This can be done with posterior intervals of the estimated parameters. It may be helpful to provide such functionality in the software and discuss this in the paper.

(4) Statistical methods are often validated through simulations. The manuscript has used simulations in the part about PIG approximation of allele counts. Given that the method has several approximations, including PIG and Variational approximation, it may be helpful to assess the inference procedure under a full simulation setting. So one can start with embeddings, a function linking embeddings to D values, and then sample gene S-values, and then sample variant S-values and allele counts. How accurate are the MisFit estimated parameters? What are the power and false positive rates in finding variants with high fitness costs? I understand though, the significant amount of extra work involved for the simulations. I think this would improve the quality of the paper, but is not absolutely required.

(5) This may be a difficult point, but since the method learns a function linking embeddings to D values, it would be interesting to examine this relationship. Can we learn about what features of variants are predictive of their damaging effects? It is true that embeddings are often difficult to interpret, but perpahs there is some existing work in the literature about the embeddings that the authors can leverage here?

Minor comments:

- The manuscript justifies the PIG model by showing that it fits the AF distribution from simulations (Suppl. Fig. 4). However, the figure shows only the results of variants with AF in a range 1E-4 to 1E-3. Would it be possible to show full Site Frequency spectrum?

- In Fig. 1c, how is accuracy defined?

- In the section "Comparison of gene-level constraint", two methods were compared. Citations should be added here.

- Supplementary Fig. 11: no statistical quantification of the difference of the two distributions.

- Fig. 4b is not cited in the main text.

- In Discussion, Line 346, the sentence "resolution of estimating relatively large s is poor", is a bit hard to follow.

- In Methods, "Population sequence data". My understanding is that in training, all possible variants across all positions of a gene, including those with no observed variants in the training data, would be used. Is this true? It would be useful to clarify.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
In this manuscript, Zhao et al. develop a method, MisFit, to estimate the selection coefficient of individual missense variants. Previously, it has been possible to estimate the selection coefficient of PTVs, at least in genes with a high enough combined mutation rate, and it has also been possible to estimate heuristic metrics of missense variant deleteriousness using a variety of methods including ESM-2, which is an input to MisFit. This manuscript combines these scores with population allele frequency data of the type that is used to estimate PTV selection coefficients and shows that they can thus be converted into selection coefficients. This is advantageous for two reasons: (1) it is calibrated across genes, combining variant-level and gene-level properties; and (2) it is interpretable from a population genetic perspective.

I have one major comment and some minor suggestions.

Major comments
1. Arguably the most important validation analysis is that of Figure 4b, which shows that for sites with Mis_S = s, the fraction of alleles observed de novo is indeed close to s. I think this is the only analysis that directly validates the calibration of the selection coefficients, although it is extensively validated that Mis_S correlates with the true selection coefficient both within genes and across them. What are the implications of autism case/control status in this analysis? I think that the s = P(de novo) argument is typically applied to unascertained individuals, and ascertainment may or may not affect the ratio, because both the numerator and the denominator may be affected.

Is it expected that the ratio should be different in cases vs. controls vs. unascertained individuals? Is the apparent difference between cases and controls in Fig 4b something that is expected, a cause for concern, or neither?

Minor comments
2. Is it surprising that MisFit s_gene can vary so widely vs. s_PTV? To what extent should this be interpreted as evidence that the esitmators are noisy, vs. that missense and PTV variants have genuintely different patterns of constraint across genes?
3. In general, what types of applications would most benefit from the cross-gene comparability and the selection coefficient calibration offered by MisFit? For example, when assessing a candidate causal variant in a known disease gene, it is unimportant to compare it across genes; on the other hand, maybe MisFit would be quite beneficial when ranking candidate mutations for a disease whose causal gene(s) are not known. Similarly, when would it be beneficial to have selection coefficients specifically?

Signed,
Luke J. O'Connor

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have done an excellent job of addressing my comments. In particular, they have added additional simulation to demonstrate the accuracy of MisFit. I have no more concerns, and would like to recommend this excellent article for publication.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I read the author's response with interest, and my comments are addressed. Regarding case ascertainment, I think the correct intuition is that this has a large effect on the total frequency of causal alleles in a class, but that if all such alleles have equal effect sizes, then the ratio of inherited vs. de novo alleles is unaffected. For a heterogeneous class containing both causal and null alleles, it seems that cases might be enriched for the causal ones, and therefore also enriched for de novo vs. inherited alleles. I suggest including some discussion of this in the text (alongside the derivation from the authors' response letter), but I'm not so concerned about it that I think additional analyses are needed.

(Remarks on code availability)

Responses to Reviewers' comments

Reviewer #1:

The manuscript by Shen and colleagues proposed a novel computational method, MisFit, to infer fitness effects of protein-coding variants. The manuscript correctly pointed out that existing methods for predicting "pathogenicity" of variants are confounded by gene level effects. The fitness effect of a variant, as the manuscript explained, should depend on both the damaging effect of the variant on the gene function, and the strength of natural selection at the gene level. Separating these two effects can be important in some contexts, e.g. when studying mutational effects on a particular gene in mutational scanning. Importantly, when estimating the fitness of a variant, jointly studying the two effects allows one to borrow information from other variants in the same gene to improve the estimation. The authors validated the new method by various analyses. Using variant frequency data, de novo mutations (DNM), and deep mutation scanning (DMS) data, the authors show that variants predicted by MisFit to have large damaging effects on genes (D-score), or large selection coefficients (S-scores), are highly enriched with likely functional variants. In the case of DNM data, in particular, MisFit predicted variants by S-scores are substantially more enriched in likely risk variants of autism and developmental disorders than those from all existing methods.

Overall, this is a technically novel and solid method with potentially broad impact on the field. There are several key technical innovations that make the method possible. The first is the use of protein language model based embeddings in the method. This allows one to have an informative prior for estimating D, the damaging effects of variants. The second is several computational innovations that reduce the computational burden, including the Poisson-Inverse-Gaussian (PIG) approximation of variant count distribution, and the Variational Inference method in training model parameters. In addition to the method itself, these computational innvoations are likely interesting to researchers developing related methods.

We appreciate the positive comments and summary of our work.

While the paper overall is well-written, some parts are not explained clearly. Below are specific comments:

Major comments:

(1) The section "MisFit model architecture and parameters" in Methods is somewhat difficult to follow. First, the section actually talked about two models: one full model, and one baseline model where the D and S values of variants depend only on global parameters. It would make it easier to follow if the authors can separately explain the two models. Secondly, in statistical literature, one usually starts with the description of the probablistic models, and then explains the inference procedure. But in this section,

the two are mixed together without a clear boundary. For example, Equation (19) should occur after the modeling part, Equations (20) - (22). Thirdly, it would be useful to add some explanation of the rationale of some key steps. Importantly, the method has two Neural networks: NN1 and NN2. The first one, NN1, is a function linking embedding to D values, and is relatively easy to understand. But the idea behind NN2 is not explained clearly. The language here is quite vague, "in order to retrieve mu, sigma, in one forward pass ...". My understanding is that in the probablistic model, the D and S values are latent variables, so the Variational Inference uses another NN to approximate their posterior distribution of D, given embedding and allele counts. But this point should be made more explicit. I also feel it may be helpful to draw an analogy with Variational Autoencoders (VAE) here to faciliate understanding.

Thanks for the constructive comments. We have reorganized a few parts in the Results and Methods sections to improve clarity (Line 160-180, 518-590). The revised paragraphs in the Results are copied here (m as allele count, n as allele number, d as damage, s as selection, v as mutation rate):

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, and ESM-2 zero-shot prediction. In other words, we attempt to approximate $p(d|x)$ by maximizing $p(m|x,v,n)$, 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 4,073 constrained genes (missense z score > 2 or gnomAD 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 unrelated European samples and 90,914 gnomAD[7] Non-Finnish European population samples. 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 $p(s|m;x,v,n)$ is determined but with no simple analytical form. Sampling-based methods are a solution but time-consuming considering the amount of 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)

We also split the illustration of NN1 and NN2 in Supplementary Figure 6 with the simple sampling process and KL divergence in NN2 representing the variational inference.

(2) The model training process is hard to follow. In both the main text, and the "Model training" section in Methods, the manuscript described the two stages of training, with limited explanations. Why is the inference done in two stages? What is the rationale of using only constrained genes in the first stage? In the first stage of training, only NN1 related parameters are learned. But how does training happen when NN2 is unknown?

In MisFit model, $\boldsymbol{d}$ represents effect of the mutations at molecular level, i.e. the effect on the encoded proteins, which is assumed to have a consistent scale across proteins. $\boldsymbol{s}$ represents the effect at population level, which reflects both the impact on protein function and how important the protein is in all biological processes involved in reproductive fitness. Therefore, the mutations from different genes can have same $\boldsymbol{d}$ values but vastly different $\boldsymbol{s}$ values. In MisFit's graphical model, $\boldsymbol{d}$ is a function ($P(d|x)$) of a representation (x) of protein sequences with parameters shared across all genes, whereas $\boldsymbol{S_{gene}}$ (the prior of $\boldsymbol{s}$) is a function of $\boldsymbol{d}$ with gene-specific parameters. That's why we setup a two-stage training process: in the first stage, we estimate the parameters for $P(d|x)$ across genes and $\boldsymbol{S_{gene}}$, in the second stage, we estimate posterior of S for each missense in individual genes.

Technically, in stage 1, NN1 maps embedding x to $P(d|x)$, and we sample d from $P(d|x)$ to maximize the marginal probability of allele counts in population $P(m|n,x)$ (and probability of observing homologues variants across species as well). These properties actually can be derived from NN1 alone. NN1 only has x as inputs, while m, n, etc. are only used for calculating loss. Then in stage 2, we infer variant-level $P(s|m,n,x)$ by NN2. NN2 takes $P(d|x)$, m, n, S_{gene} all as input, where $P(d|x)$ comes from NN1 and is regarded as known. We added more details in the text (Line 160-180, 518-590).

Regarding the question of the role of constrained genes in stage 1 training: in this stage we estimate parameters in $P(d|x)$ through maximizing $P(m|x,v,n)$, in which m is allele counts in human. In non-constrained genes, allele counts in human is not very informative, as most of the variants are under weak selection. Thus, we emphasize more on constrained genes for efficiency and stability in training. However, non-constrained genes are also used for the first and last few epochs, because we need to estimate S_{gene} for all genes as priors of $\boldsymbol{s}$.

(3) The method infers several scores, D, S and S.gene. These are all point estimates. But as the authors pointed out, it can be difficult to estimate selection for individual variants with relatively mild effects. So it may be interesting to quantify uncertainty of predicted effects. A user may, for example, focus only on likely functional variants with high confidence. This can be done with posterior intervals of the estimated parameters.

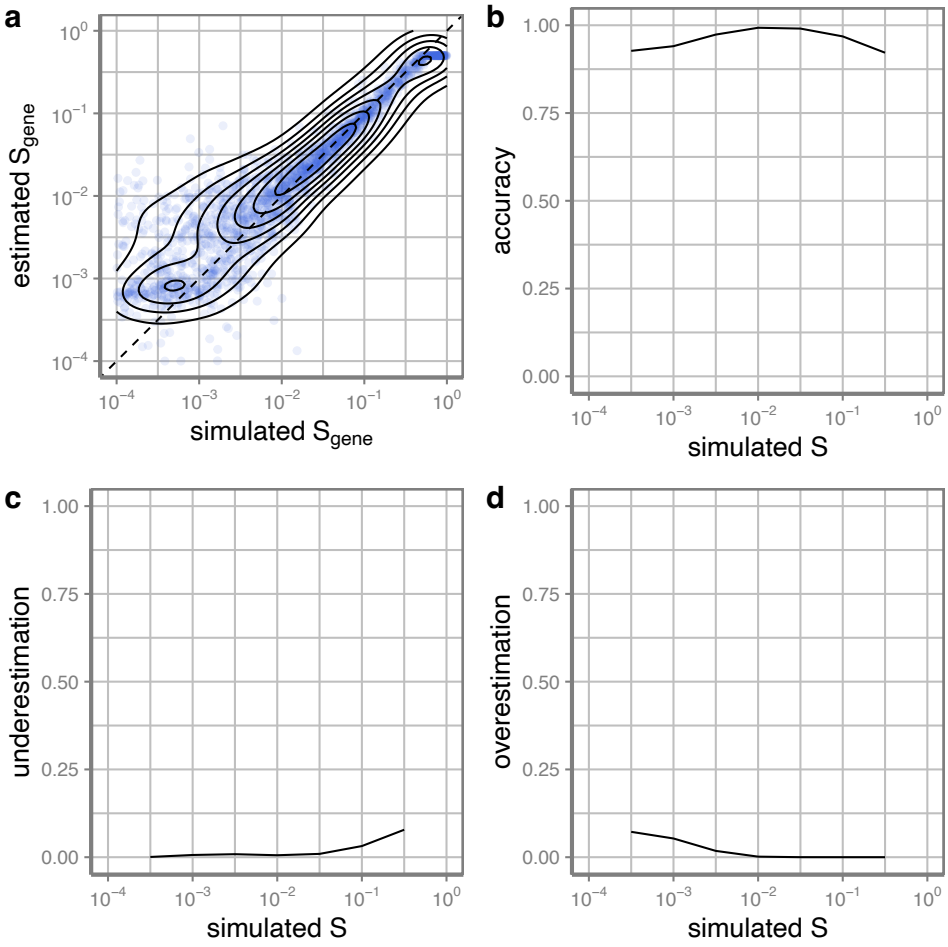
It may be helpful to provide such functionality in the software and discuss this in the paper.

We agree that reporting uncertainty would be helpful in general. It was indeed feasible to estimate credible interval or standard deviation from the variational inference procedure. However, in additional extensive simulations (in response to comment #4 below), estimated values from maximum a posteriori (MAP) is more accurate than variational inference (in revised Supplementary Figure 24). In the revision, we have decided to switch to MAP, which is a point estimate without an assumption on posterior distribution family, and thus the standard deviation cannot be easily provided. Although we can provide a standard deviation for variant selection coefficient given estimated $\boldsymbol{S_{gene}}$, this does not fully describe the total uncertainty of selection estimation.

(4) Statistical methods are often validated through simulations. The manuscript has used simulations in the part about PIG approximation of allele counts. Given that the method has several approximations, including PIG and Variational approximation, it may be helpful to assess the inference procedure under a full simulation setting. So one can start with embeddings, a function linking embeddings to D values, and then sample gene S-values, and then sample variant S-values and allele counts. How accurate are the MisFit estimated parameters? What are the power and false positive rates in finding variants with high fitness costs? I understand though, the significant amount of extra work involved for the simulations. I think this would improve the quality of the paper, but is not absolutely required.

This is an excellent suggestion. In the revision, we performed simulations to assess the accuracy of Bayesian estimation component in MisFit. In particular, we simulated S_{gene} for 2000 randomly chosen genes and simulated $\boldsymbol{d}$ for missense variants, then we sampled $\boldsymbol{s}$ for the missense variant based on MisFit assumptions about $d \rightarrow s$ within each gene, and then given $\boldsymbol{s}$, we simulated allele counts based on effective population size history by Wright-Fisher process. Finally, we adopted the basic structure (the stage 2, NN^2 part) of MisFit to estimate selection coefficients ($\boldsymbol{S_{gene}}$ and $\boldsymbol{s}$) from simulated $\boldsymbol{d}$ and allele counts, regarding $\boldsymbol{d}$ as known. We show our results in **Supplementary Figure 23** (also shown on next page). Under model assumptions, the estimation on both gene-level prior and variant-level selection coefficient is very accurate.

Therefore, in the revision, we used maximum a posteriori (MAP) estimation for S_{gene} to replace variational inference in the previous version. One explanation about the problem of variational inference is that empirically, the posterior distribution of logit S_{gene} is skewed with an asymmetrical long-tail distribution, which is not well modeled by normal distribution as variational family. As a result, the variational inference distribution mode is a biased estimator. MAP estimation which directly estimates the mode can be more accurate.



Supplementary Fig 23 Estimated model performance in simulation. Selection coefficients and allele frequencies of variants in 2000 genes are simulated based on MisFit probabilistic model assumptions. (a) Estimated S_{gene} (prior of max s in a gene) versus simulated. (b-d) Accuracy/underestimation/overestimation rate of variant selection coefficient. An accurate estimation is defined as ($10^{-0.5}$×simulated S) ≤ estimated S ≤ ($10^{0.5}$× simulated S).

(5) This may be a difficult point, but since the method learns a function linking embeddings to D values, it would be interesting to examine this relationship. Can we learn about what features of variants are predictive of their damaging effects? It is true that embeddings are often difficult to interpret, but perpahs there is some existing work in the literature about the embeddings that the authors can leverage here?

This is an interesting and important question. By design, such protein masked language model estimates probability of amino acid given its surrounding sequence context, which is exactly the outcome of average 'damage'. From the original ESM2 paper (Lin et al 2023), the ESM2 latent embeddings pretrained as a masked language model can predict protein structure from sequence without multiple sequence alignments. It implicitly learned information related to amino acid properties and positional correlation that reflects 3D structural contacts, which are fundamentally important for structure prediction. However, connecting specific embedding

dimensions to variant effect is very challenging. There are recent methods such as Sparse Autoencoders (https://www.biorxiv.org/content/10.1101/2024.11.14.623630v2.full) that may enable us to identify interpretable structural and functional features from ESM's embeddings. This is an ongoing effort, but beyond the scope of this manuscript.

Minor comments:

- The manuscript justifies the PIG model by showing that it fits the AF distribution from simulations (Suppl. Fig. 4). However, the figure shows only the results of variants with AF in a range 1E-4 to 1E-3. Would it be possible to show full Site Frequency spectrum?

We showed the probability within different AF ranges. For example, in Supplementary Figure 4a, y-axis shows probability mass of sample allele frequency (allele count/allele number) within a specific range that is annotated vertically for each panel. Thus there are probability mass of AF ranges [0, 0], (0, 1e-5], (1e-5, 1e-4], (1e-4, 1e-3], (1e-3, 1). We added a sentence for clarification.

- In Fig. 1c, how is accuracy defined?

In Fig. 1c, we assume that s is categorical ($\boldsymbol{s} \in \{0.00001, 0.0001, 0.001, 0.01, 0.1, 1\}$). Given simulated data, if the categorical s which gives out highest likelihood of $p(m|s)$, it is regarded as categorical MLE estimation of that group. If s_{MLE} matches the simulation condition, it counts as one 'accurate' estimation.

- In the section "Comparison of gene-level constraint", two methods were compared. Citations should be added here.

Citations about gnomAD metric have been added in revision (Line 190). References 44-46 were already included.

- Supplementary Fig. 11: no statistical quantification of the difference of the two distributions.
KS tests are added to Supplementary Figure 11 and main Figure 5.

- Fig. 4b is not cited in the main text.
Added to line ~240.

- In Discussion, Line 346, the sentence "resolution of estimating relatively large s is poor", is a bit hard to follow.
Variants with relatively large s are all depleted in wildtype sequences, their difference in $\boldsymbol{s}$ may not be well distinguished if only use MSA. Modified Line ~346.

- In Methods, "Population sequence data". My understanding is that in training, all possible variants across all positions of a gene, including those with no observed variants in the training data, would be used. Is this true? It would be useful to clarify.

Yes, this is true. All possible missense single nucleotide variants are considered even if they are not observed in population sequencing data (will be noted as allele count = 0, allele number ~ 500K). However, variants within genomic regions of poor mapping quality are excluded (allele count = 0, allele number = 0). Clarified in Methods part Line 474-482.

Reviewer #2:

In this manuscript, Zhao et al. develop a method, MisFit, to estimate the selection coefficient of individual missense variants. Previously, it has been possible to estimate the selection coefficient of PTVs, at least in genes with a high enough combined mutation rate, and it has also been possible to estimate heuristic metrics of missense variant deleteriousness using a variety of methods including ESM-2, which is an input to MisFit. This manuscript combines these scores with population allele frequency data of the type that is used to estimate PTV selection coefficients and shows that they can thus be converted into selection coefficients. This is advantageous for two reasons: (1) it is calibrated across genes, combining variant-level and gene-level properties; and (2) it is interpretable from a population genetic perspective.
We appreciate the summary and positive comments.

I have one major comment and some minor suggestions.

Major comments
1. Arguably the most important validation analysis is that of Figure 4b, which shows that for sites with Mis_S = s, the fraction of alleles observed de novo is indeed close to s. I think this is the only analysis that directly validates the calibration of the selection coefficients, although it is extensively validated that Mis_S correlates with the true selection coefficient both within genes and across them. What are the implications of autism case/control status in this analysis? I think that the s = P(de novo) argument is typically applied to unascertained individuals, and ascertainment may or may not affect the ratio, because both the numerator and the denominator may be affected.

Is it expected that the ratio should be different in cases vs. controls vs. unascertained individuals? Is the apparent difference between cases and controls in Fig 4b something that is expected, a cause for concern, or neither?

This is an important question. The equation (de novo ratio (F) among variants with $\boldsymbol{s}$ = $\boldsymbol{s}$) directly applies to randomly ascertained individuals from population not yet under selection (e.g., zygote). This type of human data in large quantity is not yet available, therefore, in this work we had to use trios/quads genomic data from autism studies

through which we have access to both de novo and inherited variants of large number of individuals. The autism studies enrolled families based on "probands", i.e. autistic individual (cases). The probands can be regarded as nearly random ascertained. After the probands, the study then enrolled parents, additional cases, and unaffected siblings (controls) in each family when feasible. Here we show that the de novo ratio = $\boldsymbol{s}$ still holds for data from *cases* (individuals with autism), and that the ratio is lower than $\boldsymbol{s}$ under strong selection for *controls* (unaffected siblings). The following section is added to the revised Supplementary Notes, page 3-7, titled "*Fraction of de novo mutations among all variants with a selection coefficient in populations ascertained by a clinical condition*":

Under equilibrium or in a new generation when selection has not occurred, among all observed variants under moderate to strong negative selection in a population, *de novo* mutations are expected to take up a proportion ($\boldsymbol{F}$) equal to the selection coefficient of the variant[3]. Here we provide a proof that this equality still holds for populations ascertained by a clinical condition.

Denote A as the alternative ("mutant") allele and a as the reference ("wildtype") allele of a variant V, and the selection coefficient of A as $\boldsymbol{s}$. If A is not associated with the ascertainment condition, then the populations ascertained based on the condition is similar to general population regarding the frequency and proportion of de novos of A ($\boldsymbol{F} = \boldsymbol{s}$).

Now we consider a clinical condition with population prevalence $\boldsymbol{\tau}$ ($\tau < 5\%$), assuming that A is associated to the condition, and that the condition is the only condition through which A is under moderate or strong negative selection. As the variant is rare, homozygous genotypes (AA) are not considered, and in a proband-parents family at most one parent harboring the heterozygous genotype (Aa).

Denote the equilibrium allele frequency of A in a new generation (before selection) as q, mutation rate per allele as $\boldsymbol{\nu}$, the selection of the condition as $\boldsymbol{s_0}$, and the penetrance of the condition as p. By definition:

$$\boldsymbol{p} = \boldsymbol{P}(\text{Affected}|\boldsymbol{Aa}) \quad (\mathbf{1})$$

In a population with a constant n individuals, there are about $2nq$ individuals with genotype Aa, among which $2pnq$ are affected. As a result of the clinical condition, $2pnqs_0$ are negatively selected. Among all individuals with aa genotypes, the number of affected is $2n(1-q)\tau$. Since τ and q are both small, the impact on variant V due to negative selection on the condition from other factors is negligible. Therefore, by definition, selection coefficient of allele A is:

$$s \approx \frac{2pnqs_0}{2nq} = ps_0 \quad (2)$$

We replace s_0 with s/p, and the genotypic composition of post-selected individuals, denoted as generation G1 (they are potential parents of the next generation G2) is:

$$P_{G1}(Aa) \approx 2(1-s)q \quad (3)$$

$$P_{G1}(aa) \approx 1 - 2(1-s)q \tag{4}$$

The probability that one of the G1 parent is heterozygous is:

$$P(\text{G1: } Aa \text{ and } aa) = 2P_{G1}(Aa)P_{G1}(aa) \approx 4(1-s)q \tag{5}$$

$$P(\text{G1: } aa \text{ and } aa) \approx 1 - 4(1-s)q \tag{6}$$

Then, in the next generation G2, the probability of Aa genotypes is:

$$P_{G2}(\text{inherited } Aa) = P(\text{G1: } Aa \text{ and } aa) \times \frac{1}{2} \approx 2(1-s)q \tag{7}$$

$$P_{G2}(\textit{de novo } Aa) \approx 2\nu \tag{8}$$

Assume equilibrium, in G2 we have

$$q = P_{G2}(Aa)/2 = (P_{G2}(\text{inherited } Aa) + P_{G2}(\text{de novo } Aa))/2 \tag{9}$$

therefore,

$$q \approx \frac{2(1-s)q}{2} + \frac{2\nu}{2} = \frac{\nu}{s} \tag{10}$$

Replacing q with ν/s, we rewrite Equation 3 and 4 (the genotypic composition of G1 individuals that would be parents) into:

$$\begin{aligned} P_{G1}(Aa) &\approx \frac{2(1-s)\nu}{s} \\ P_{G1}(aa) &\approx 1 - \frac{2(1-s)\nu}{s} \end{aligned} \tag{11}$$

and the genotypic composition of G2 pre-selection individuals:

$$\begin{aligned} P_{G2}(\text{inherited } Aa) &\approx \frac{2\nu(1-s)}{s} \\ P_{G2}(\textit{de novo } Aa) &\approx 2\nu \\ P(aa) &\approx 1 - \frac{2\nu}{s} \end{aligned} \tag{12}$$

Given penetrance p, we get Phenotype ∩ Genotype of G2 individuals with Aa genotypes in 4 categories:

$$\begin{aligned} P_{G2}(\text{Affected} \cap \text{inherited } Aa) &= \frac{2p\nu(1-s)}{s} \\ P_{G2}(\text{Affected} \cap \text{de novo } Aa) &= 2p\nu \\ P_{G2}(\text{Unaffected} \cap \text{inherited } Aa) &= \frac{2(1-p)\nu(1-s)}{s} \\ P_{G2}(\text{Unaffected} \cap \text{de novo } Aa) &= 2(1-p)\nu \end{aligned} \tag{13}$$

From these equations, we can deduce the *de novo* mutation fraction ($\boldsymbol{F}$) in G2 Aa genotypes among individuals with or without the clinical condition:

$$F_{all} = \frac{P_{G2}(\text{de novo } Aa)}{P_{G2}(Aa)} = s \tag{14}$$

$$F_{\text{affected}} = \frac{P_{G2}(\text{de novo } Aa|\text{Affected})}{P_{G2}(Aa|\text{Affected})} = s \tag{15}$$

$$F_{\text{unaffected}} = \frac{P_{G2}(\text{de novo } Aa|\text{Unaffected})}{P_{G2}(Aa|\text{Unaffected})} = s \tag{16}$$

Thus, $F = \boldsymbol{s}$ applies to random samples in the new generation G2, regardless of their condition.

However, our cohort is not randomly sampled from a population. The autism cohorts were ascertained this way: first, affected individuals were (near-) randomly enrolled in the study as the proband cases; then, a much smaller number of affected siblings from the same families were enrolled as additional cases, and a subset of unaffected siblings from the same families as controls (instead of enrolling unaffected children from population). Here we prove that the de novo fraction is lower in the affected and unaffected siblings.

$$\begin{aligned} P(\text{G2 is Affected}|\text{G1: } Aa \text{ and } aa) &= \frac{p}{2} \\ P(\text{G2 is Affected}|\text{G1: } aa \text{ and } aa) &= 2pv \end{aligned} \tag{17}$$

Using Bayes rule, with Equation 5, 6 and 10, we have the genotype frequency of G1 individuals that are parents of an affected G2 individual:

$$\begin{aligned} P(\text{G1: } Aa \text{ and } aa| \text{ one affected individual in G2}) &= \frac{1-s}{1-4(1-s)v} \approx 1 - \text{s} \\ P(\text{G1: } aa \text{ and } aa| \text{ one affected individual in G2}) &\approx \text{s} \end{aligned} \tag{18}$$

This approximation holds when s is a lot greater than $4v$, which is plausible as we are assuming very rare variants ($q \ll 1/4$).

Finally, the probabilities of Aa genotypes in a second child of the ascertained families:

$$P(\text{2nd individual in G2: de novo } Aa| \text{ one affected individual in G2}) = 2v \tag{19}$$

$$P(\text{2nd individual in G2: inherited } Aa|\text{one affected individual in G2}) = \frac{1-s}{2} \tag{20}$$

Therefore, for variants associated with the condition, the fraction of *de novo* carried by a second individual from the ascertained families is:

$$F = \frac{P(\text{2nd individual: de novo } Aa| \text{ one affected individual})}{P(\text{2nd individual: } Aa \text{ |one affected individual})} = \frac{4v}{4v + 1 - s} \tag{21}$$

With Equation 21, $\boldsymbol{F}$ is always smaller than $\boldsymbol{s}$ when $\boldsymbol{s}$ is greater than $4v$. We note v is about 10^{-7} to 10^{-8} in human. Therefore, $\boldsymbol{F}$ is close to 0 (in same magnitude as v) for $\boldsymbol{s}$ in the range of

interest of this study (10^{-4} to 1), except when s is very close to 1. In the autism cohorts, most (>90%) of the cases were the ascertained probands based on which the families were enrolled in the studies, and all of the controls are unaffected siblings conditioned on one affected child in the families. Therefore, among variants that are associated with autism, F will be close to 0 in controls and be below but close to s in cases.

Empirically, a substantial fraction of variants under strong selection are associated with autism and neurodevelopmental disorders (see Supplementary Figure 13 and previous publications using metrics correlated with selection strength of protein truncating variants [4,5]), and that the chance of being associated positively correlated with the estimated selection coefficient. F of these variants is close to 0 in controls. Therefore, among all variants under strong selection, F in unaffected siblings (i.e. controls) is smaller than s, leading to deviation from the diagonal of F vs s curve, and the deviation is more pronounced with greater s values (Figure 4b). In cases, the deviation is minor since most of the cases were probands with near random ascertainment.

Minor comments

2. Is it surprising that MisFit s_gene can vary so widely vs. s_PTV? To what extent should this be interpreted as evidence that the esitmators are noisy, vs. that missense and PTV variants have genuintely different patterns of constraint across genes?

In the revision, we changed the method for estimating S_{gene} from variational inference to maximum a posteriori (MAP). The distribution of MAP estimated S_{gene} for most of genes is not wider than s_{PTV} (Figure 3a). S_{gene} is the prior of the maximal value of s of all missense variants in a gene. S_{gene} is highly correlated with s_{PTV}. A common mechanism of effect of PTVs is nonsense mediated decay (NMD), which significantly reduce the expression of functional proteins encoded by the gene. Missense variants affect protein function through a few different mechanisms. A common mechanism is destabilizing the protein, which at maximal can be close to the effect of NMD. Other mechanisms include gain of function (e.g. making the protein hyperactive or constitutively active) and dominant negative (the "mutant" proteins interfere with the function of the "wild types"), both could lead to equal or more severe consequence than NMD effect from PTVs. As a result, S_{gene} as the prior of maximal selection of missense is generally greater than s_{PTV}, especially for genes in which PTVs are not under strong selection. This is supported by the joint distribution of (S_{gene}, s_{PTV}) of genes that are known to be dominant negative or gain of function (Figure 3d and 3e).

We do acknowledge that S_{gene} estimates are noisier than s_{PTV} due to effective sample size difference. s_{PTV} is estimated based on data of nearly all PTVs in a gene under a reasonable assumption that they have similar effect through NMD. S_{gene} is the maximal s of missense variants in a gene. Its estimation noise is a combination of sampling errors and additional uncertainty and a wide range of effect of individual missense variants in a gene.

3. In general, what types of applications would most benefit from the cross-gene comparability and the selection coefficient calibration offered by MisFit? For example, when assessing a candidate causal variant in a known disease gene, it is unimportant to compare it across genes; on the other hand, maybe MisFit would be quite beneficial when ranking candidate mutations for a disease whose causal gene(s) are not known. Similarly, when would it be beneficial to have selection coefficients specifically?

MisFit is designed to have the ability to compare across genes with a clear interpretation: $\boldsymbol{d}$ is a proxy of how much the function of a protein is perturbed, $\boldsymbol{s}$ is the effect in reproductive fitness in the population. We agree that ranking candidate mutations in genetic diagnosis of rare developmental disorders is an important application. A major genetic risk of developmental disorders would have substantial fitness impact, and MisFit $\boldsymbol{s}$ provides a way to rank missense variants together with PTVs (s_{PTV}) by fitness impact. Another application is to use MisFit $\boldsymbol{s}$ to estimate weights or informative priors in novel risk gene discovery for early onset conditions, as shown in **Figure 6** about autism and neurodevelopmental disorders. This can be applied to both *de novo* variants and rare inherited variants. As for MisFit $\boldsymbol{d}$, in principle, it can be useful to improve statistical power in new gene discovery for conditions that are not under strong selection, such as late-onset complex conditions. The consistent scale of MisFit $\boldsymbol{d}$ across genes would make it feasible to calculate more accurate weights or priors. We agree that for a known disease gene, it is unimportant to compare it across genes.

Responses to Reviewers' comments

Reviewer #1 (Remarks to the Author):

The authors have done an excellent job of addressing my comments. In particular, they have added additional simulation to demonstrate the accuracy of MisFit. I have no more concerns, and would like to recommend this excellent article for publication.

Reviewer #2 (Remarks to the Author):

I read the author's response with interest, and my comments are addressed. Regarding case ascertainment, I think the correct intuition is that this has a large effect on the total frequency of causal alleles in a class, but that if all such alleles have equal effect sizes, then the ratio of inherited vs. de novo alleles is unaffected. For a heterogeneous class containing both causal and null alleles, it seems that cases might be enriched for the causal ones, and therefore also enriched for de novo vs. inherited alleles. I suggest including some discussion of this in the text (alongside the derivation from the authors' response letter), but I'm not so concerned about it that I think additional analyses are needed.

We appreciate your comment. As suggested, we add a few sentences in Results line 244 - 246:

Such relationship holds for randomly selected samples regardless of their own phenotypes, but not for samples specially chosen with known family background (Supplementary Note).

and in Discussion (line 364 – 367):

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.