nature genetics

Article

https://doi.org/10.1038/s41588-025-02400-1

# Proteome-wide model for human disease genetics

Received: 26 February 2025

Accepted: 8 October 2025

Published online: 24 November 2025

Check for updates

Rose Orenbuch[1], Courtney A. Shearer[1], Aaron W. Kollasch[1], Aviv D. Spinner[1], Thomas Hopf[2], Lood van Niekerk[1], Dinko Franceschi[1], Mafalda Dias[3,4,5] ✉, Jonathan Frazer[3,4,5] ✉ & Debora S. Marks[1,6,7] ✉

Missense variants remain a challenge in genetic interpretation owing to their subtle and context-dependent effects. Although current prediction models perform well in known disease genes, their scores are not calibrated across the proteome, limiting generalizability. To address this knowledge gap, we developed popEVE, a deep generative model combining evolutionary and human population data to estimate variant deleteriousness on a proteome-wide scale. popEVE achieves state-of-the-art performance without overestimating the burden of deleterious variants and identifies variants in 442 genes in a severe developmental disorder cohort, including 123 novel candidates. These genes are functionally similar to known disease genes, and their variants often localize to critical regions. Remarkably, popEVE can prioritize likely causal variants using only child exomes, enabling diagnosis even without parental sequencing. This work provides a generalizable framework for rare disease variant interpretation, especially in singleton cases, and demonstrates the utility of calibrated, evolution-informed scoring models for clinical genomics.

Even if every human were sequenced and their phenotypes recorded, the space of disease-causing genetic variation may be too large to be studied by population variation or disease-relevant experimental assays alone. Patients with unique combinations of symptoms and genotypes would still go without a genetic diagnosis[1,2]. The biodiversity of life on Earth provides a deeper view of genetic variation across billions of years of evolution, presenting a unique opportunity to uncover complex genetic patterns preserved to maintain fitness. Thus, models that can distill such information accelerate our ability to leverage genetics for diagnosis, prevention and treatment.

For severe genetic disorders, the task is to identify the causal variant among millions of mutations in a patient. One powerful approach is the sequencing of trios—patient and their parents—which can narrow down the pool of candidate variants to those arising de novo when the parents are unaffected or to inherited variants from an affected parent[3,4]. Despite the impressive analysis of large rare disease cohorts[4–9], genetic diagnostic yield is relatively low; in some cases, only 25% of probands receive a genetic diagnosis[5]. There is a need for alternative strategies to identify candidate causal variants directly from a patient's sequencing data, without relying on the frequency of observations in large cohorts. In this work, we present how probabilistic modeling of diverse sequencing, in both humans and across diverse species, can potentially aid clinical interpretation of never-before-seen variation.

Recent work using deep unsupervised models trained only on evolutionary sequences has shown strong promise for clinical variant effect prediction[10–15] and have demonstrated comparable accuracy to experimental approaches[11]. Given these models do not depend on functional or clinical labeling, they can generalize to variants in genes without previous annotation. However, although these models often perform well in terms of separating *Benign* from *Pathogenic* clinical labels in known disease genes, they are not calibrated well across the

[1]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [2]Scientific Consulting, Erding, Germany. [3]Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain. [4]Barcelona Collaboratorium for Modelling and Predictive Biology, Barcelona, Spain. [5]University Pompeu Fabra, Barcelona, Spain. [6]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7]Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University, Boston, MA, USA. ✉e-mail: mafalda.dias@crg.eu; jonathan.frazer@crg.eu; deboramarks@gmail.com

https://doi.org/10.1038/s41588-025-02400-1

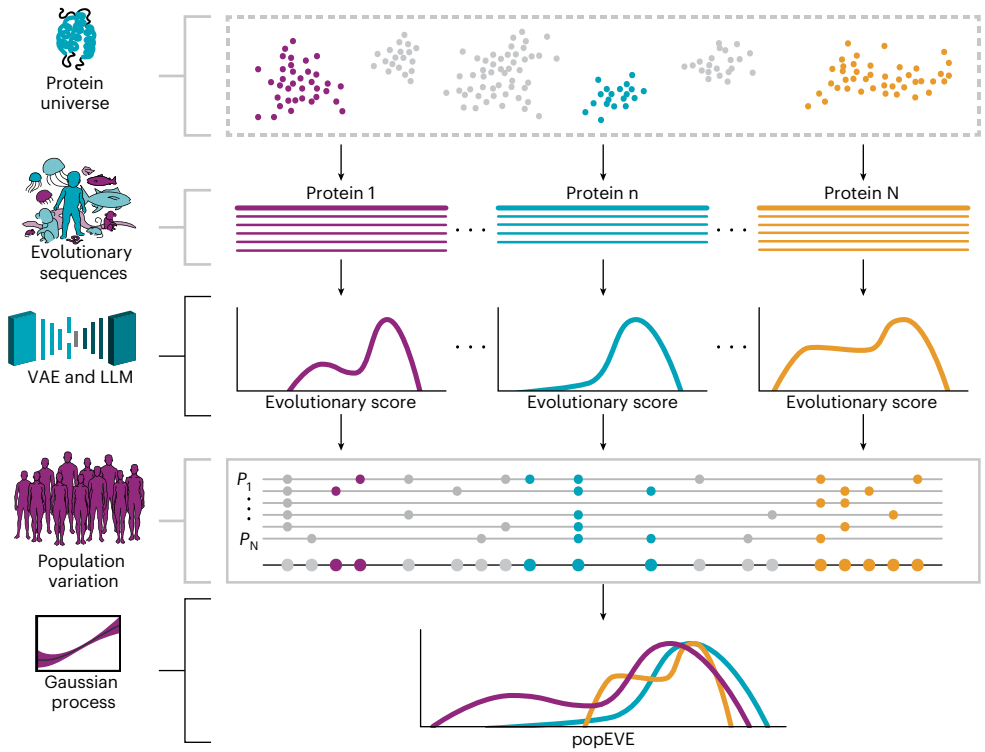



**Fig. 1 | popEVE combines deep evolution and human variation.** popEVE combines variation from across evolutionary sequences, modeled with EVE and ESM-1v, with variation within the human population (UKBB[17] or GnomAD[18]), using a Gaussian process to learn the relationship between evolutionary scores and missense constraint.

entire human proteome; that is, they are not designed for comparing how deleterious a variant is in one gene versus a variant in another. Consequently, previous methods excel at identifying variants that disrupt the function of the resulting protein but do not necessarily predict whether it is detrimental at the organismal level[16].

Variant severity lies on a spectrum: for instance, disruption of function in one protein could have modest effects late in life, while the disruption of another protein can be lethal in childhood. Both can be considered 'pathogenic' and correctly identified as such by a model, but when attempting to find the genetic cause of a severe disorder, it is imperative to be able to distinguish between these two scenarios. Current state-of-the-art variant effect prediction models have not been developed with this spectrum of severity in mind. To overcome this problem, we developed popEVE, a model that places variants on a proteome-wide scale of deleteriousness, enabling us to predict if a variant seen in one gene is more detrimental to human health than a variant seen in another. popEVE leverages deep evolutionary data to achieve missense-resolution variant effect prediction and shallow variation across the UK Biobank[17] (UKBB) or Genome Aggregation Database (GnomAD) (v.2)[18] population to transform the score to reflect human-specific constraint. Analyzing a metacohort[8] of patients with severe developmental disorders (SDDs), we find evidence for 123 candidate novel genetic disorders from their de novo missense mutations (DNMs), which is 4.4× more than previously identified in the same cohort, and yet significantly similar in function to known developmental disease genes. For cases with whole-exome sequencing (WES), we are able to identify the likely causal DNM knowledge of its inheritance pattern. Thus, popEVE provides valuable information for genetic diagnosis, even in the absence of trio sequencing, increasing the scope of genetic analysis.

## Results

### Method development

**A unified model of population and evolutionary sequences.** For a computational model to be broadly useful in human genetics, the model scores should be continuous, have residue resolution and have the same quantitative meaning across different proteins. Previous state-of-the-art computational methods have excelled in various tests of accuracy; for instance, correct classification of pathogenic and benign labels from curated clinical databases and reasonable correlations with high-throughput experiments on specific proteins[11,19–23]. However, these benchmarks can result in overestimated accuracy in and generalizability to real-world scenarios in which thousands of missense variants, including hundreds of rare variants, must be ranked across a single person's genome. This drawback has resulted in the understandable caution of the clinical use of computational methods, not least from the observation of an overprediction of deleterious variants[21,23–26]).

**Converting gene-level scores to proteome-wide scores.** popEVE is designed to provide a human-specific, continuous measure of variant deleteriousness that enables comparison across different proteins (Fig. 1 and Supplementary Fig. 1). To achieve a score that reflects constraint within humans and distinguishes the relative impact of individual variants, we reason that a model would need to not only learn from deep evolutionary variation but also from shallow variation observed in the human population. While deep evolutionary variation from across life can inform us about what is allowed for a protein to function, models trained solely on this information cannot necessarily learn the relative importance of one protein versus another. We build a unified model that predicts the effect of a variant in the population, conditioned on the underlying evolutionary scores using a latent Gaussian process prior, similar in spirit to gene-level and region-level estimates of missense constraint[27–30]. The model trains on the universe of sequences across evolution together with summary statistics of human variation from human population data. For the deep evolutionary sequence analysis, we combine a state-of-the-art alignment-based model, EVE[11], and a large language model, ESM-1v[31]. Although the two models have comparable performance on clinical and deep mutational scan benchmarks, variant scores are not particularly well correlated[22]

https://doi.org/10.1038/s41588-025-02400-1

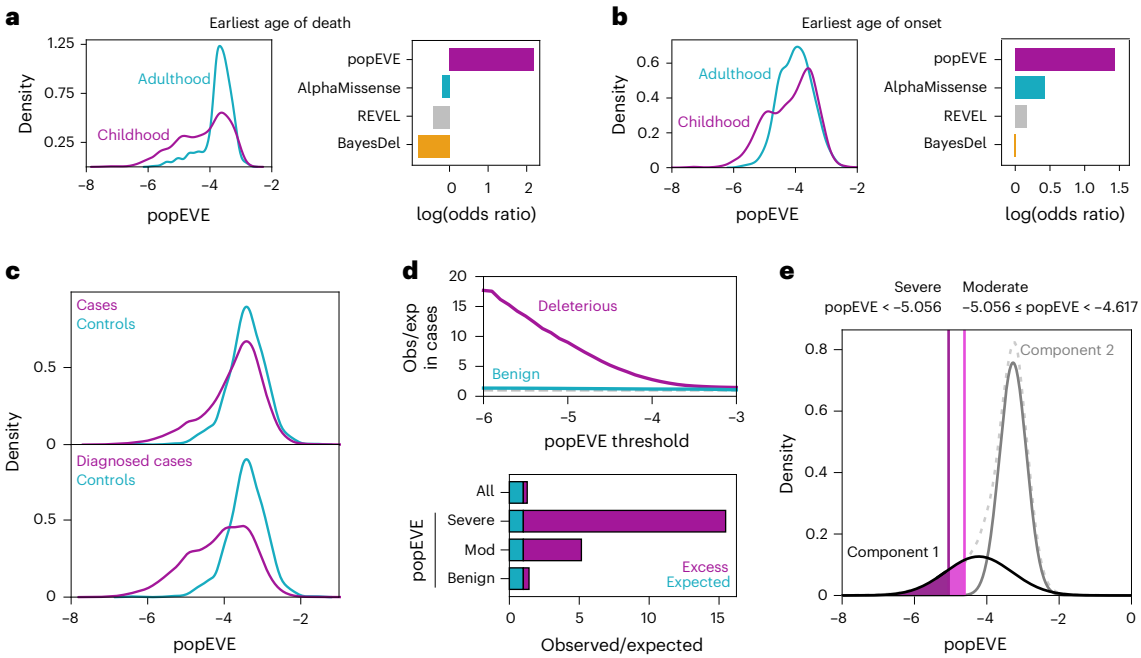



**Fig. 2 | popEVE captures variant severity and pathogenicity. a**, ClinVar pathogenic variants in phenotypes associated with premature death in childhood have more deleterious popEVE scores than those associated with death after maturation (left). Death labels were acquired from OrphaNet. At the fifth percentile of ClinVar benign variants, popEVE has a significantly larger odds ratio than any other method (right). **b**, Variants associated with onset in childhood have more deleterious popEVE scores than those associated with onset later in life (left), and popEVE has a greater odds ratio than other methods (right). **c**, popEVE scores for DNMs in SDD cases (top) and diagnosed cases (bottom) are shifted towards the deleterious end compared to controls (unaffected siblings from autism spectrum disorder family cohorts). **d**, Using DNMs from both SDD cases and controls, we define a severely and moderately deleterious score threshold by fitting a two-component Gaussian mixture model and finding the 99.99% and 99% likelihood of being in the more deleterious distribution. **e**, With increasingly pathogenic thresholds, de novo mutations in the SDD metacohort are significantly enriched (top). At our severely pathogenic threshold, popEVE pathogenic variants exhibit over 15-fold enrichment, while popEVE benign variants are in line with expectation (bottom). Moderately pathogenic variants are enriched fivefold. The expected number of variants is quantified using a background mutation rate based on the number of individuals in the metacohort.

(Extended Data Figs. 1–3), indicating useful orthogonal evidence of variant fitness. popEVE scores are enriched in haploinsufficient genes compared to loss-of-function (LoF) tolerant ones, and, to a lesser extent, in genes with dominant versus recessive inheritance patterns—consistent with its adjustment for constraint (Supplementary Fig. 2). These scores correlate more strongly with missense-based constraint metrics than those based on LoFs (Spearman's ρ, 0.52 Missense-Z[29], 0.44 pLi[18], $P < 0.001$; Extended Data Fig. 5 and Supplementary Tables 2 and 3). To be broadly useful, a variant scoring method should generate missense-resolution scores across the genome that reflect not only pathogenicity, but also the magnitude of the effect on protein fitness and the resulting phenotype. Our framework leverages population variation to calibrate scores across genes, while performing competitively with leading methods on within-gene benchmarks, namely, ClinVar classification and DMS correlation tasks (Extended Data Figs. 2 and 3 and Supplementary Table 1). Importantly, because population data is only used to re-rank variants across genes, internal rankings within genes remain largely unchanged. This allows the population adjustment to be safely incorporated into annotation pipelines that treat allele frequency as an independent evidence source. Standard benchmarks typically emphasize binary classification: determining whether a variant is benign or pathogenic within a single gene. While this is useful for some clinical decisions, it fails to capture variation in disease severity.

**popEVE shows limited to no population bias.** A disadvantage of using population data is that it can introduce population structure bias[32]. To mitigate this limitation, we use a coarse measure of missense variation ('seen' or 'not seen') rather than using allele frequencies. As such, the presence of a rare variant in a single person in the training population is treated the same as the presence of a common variant in the vast majority of people. We find that popEVE score distributions of rare variants (minor allele frequency (MAF) < 0.01) are similar across various ancestries in GnomAD (v.2)[18] (Extended Data Fig. 4). Our results are supported by an independent analysis of ancestry bias in variant scoring methods, which found that popEVE shows minimal bias towards European ancestries, in line with population-free methods[32]. By contrast, state-of-the-art competitors, including AlphaMissense[19], BayesDel[33] and REVEL[34], show significant bias towards these populations[32].

## popEVE captures variant severity and pathogenicity

**Distinguishing pathogenic variants based on phenotype severity.** First, we tested whether popEVE can distinguish variants causing severe clinical outcomes—such as childhood-onset or fatal disorders[35]—from those with more moderate effects. popEVE scores significantly separate childhood death-associated variants from adult death variants better than all other methods ($P < 0.001$; Fig. 2a, Extended Data Fig. 6a, Supplementary Fig. 3 and Supplementary Table 4). A similar, albeit weaker, pattern holds for age of onset (Fig. 2b, Extended Data Fig. 6b and Supplementary Fig. 4). This suggests popEVE captures variant severity in disease. Notably, it outperforms models tuned to allele frequency (for example, AlphaMissense, BayesDel) and those trained on clinical labels (for example, REVEL, Vest4). While those methods correctly classify most variants as potentially pathogenic, they lack the resolution that popEVE provides for distinguishing severity (Supplementary Figs. 3 and 4).

**Deleterious scores are enriched in SDD.** To evaluate how well popEVE captures variant severity, we compare de novo missense variants in SDD cases ($n = 31{,}058$) to those in unaffected controls from Autism

https://doi.org/10.1038/s41588-025-02400-1

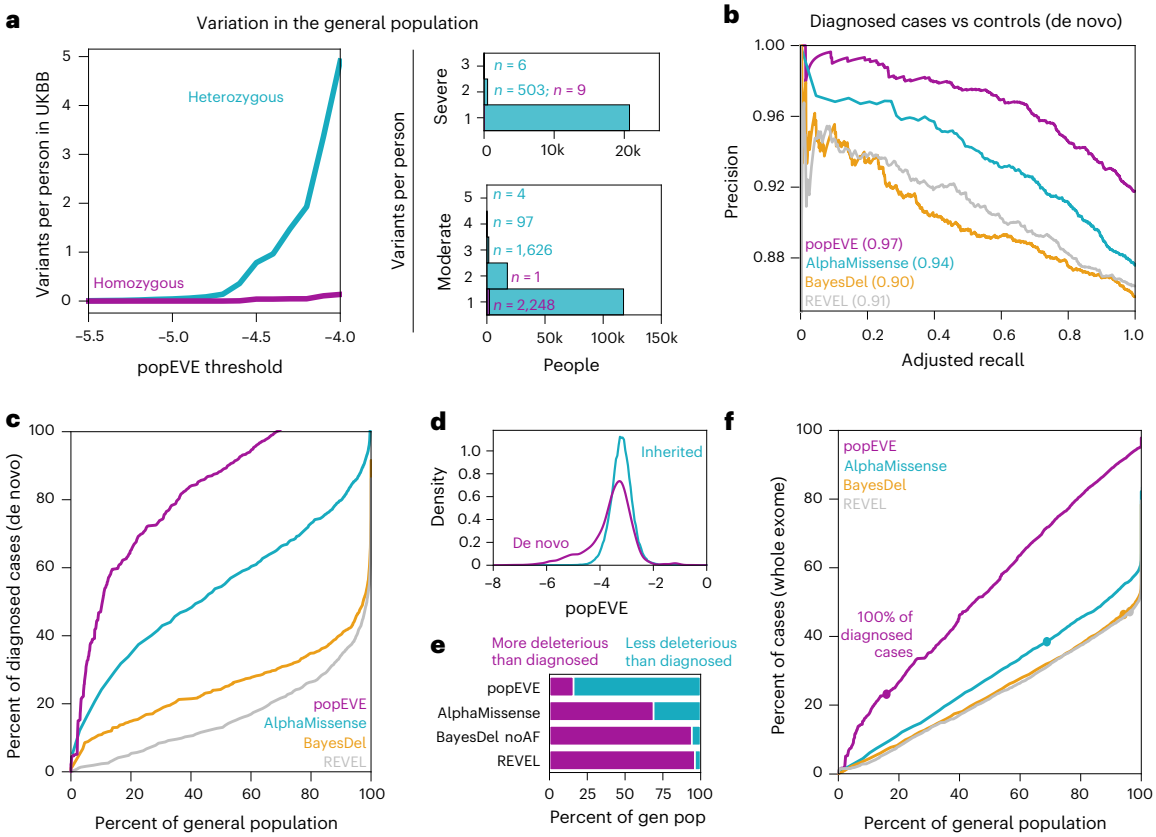



**Fig. 3 | popEVE recalls severe genetic disorder cases without overpredicting pathogenicity in the general population. a**, In the UKBB, individuals have at most one homozygous and up to three heterozygous severely deleterious variants; 96% of the 500k individuals have no severely pathogenic missense variants (left). Approximately 72% of UKBB individuals have no severely or moderately deleterious variants and at most five moderately deleterious variants (right). **b**, popEVE is better at separating diagnosed DD cases from controls based on DNMs than other state-of-the-art variant effect predictors with an average precision of 97%. Recall is adjusted based on the expected number of these cases to have a causal missense DNM (Methods). **c**, popEVE recalls more SDD diagnosed cases based on their DNMs without overpredicting pathogenicity in WES from relatively healthy controls from UKBB (gnomAD-trained popEVE). **d**, DNMs in SDD cases from the DDD Study are enriched for pathogenic variants in comparison to their rare inherited variants (MAF < 0.01) (two-sided Kolmogorov–Smirnov = 0.24, $P < 0.0001$). **e**, To recall 100% of de novo missense-diagnosed cases using their WES, popEVE predicts that far less of the general population will have a similarly deleterious variant than any other model. **f**, When applied to WES from a subset of the SDD cases, popEVE recalls more cases than other models without overpredicting pathogenicity in the general population of UKBB (using gnomAD-trained popEVE). Additionally, popEVE recalls 100% of cases expected to have a causal missense DNM for only 15% of the remaining cases and 16% of the general population (circles). Other models find that >78% of the UKBB has a variant as deleterious as these cases (inset).

Spectrum Disorder cohort trios ($n = 5{,}764$)[36] and the UKBB[17] ($n \approx 500$k) (Supplementary Table 5). popEVE scores in cases were consistently shifted toward higher predicted deleteriousness (Fig. 2c, top). These DNMs showed increasing enrichment at more severe scores, exceeding expectations based on background mutation rates (Fig. 2d, top). Among diagnosable SDD cases ($n = 2{,}982$, per a previous publication[8]), this shift is even more pronounced (Fig. 2c, bottom). Using a label-free two-component Gaussian mixture model on all variants, we set a high-confidence severity threshold at −5.056, where variants below this threshold have a 99.99% of being highly deleterious (Fig. 2e). Variants below this threshold are 15-fold enriched in the SDD cohort—five times higher than other methods like PrimateAI-3D as reported in Gao et al. (2023)[20] (Fig. 2d, bottom). Moderate-scoring variants also show fivefold enrichment. Both severe and moderate case variants are absent from UKBB and gnomAD, while severe-scoring UKBB variants are extremely rare (Fig. 3a).

**Distinguishing SDD cases from controls.** To assess performance at ranking variants across the proteome, we tested our model's ability to separate DNMs from missense-diagnosed SDD cases from

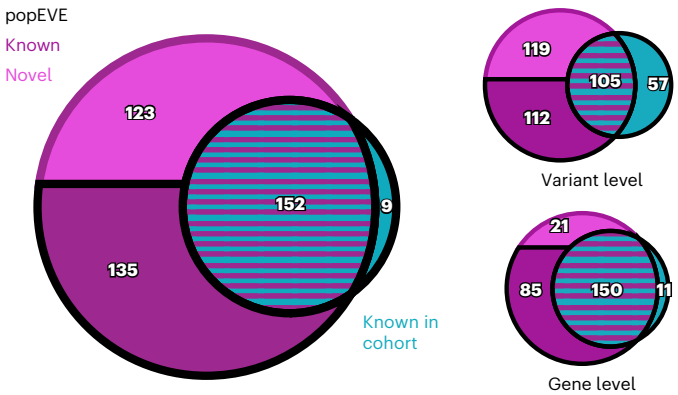



**Fig. 4 | popEVE finds evidence for 123 novel candidate genes in SDDs.** Both popEVE gene and variant-association methods achieve 94% recall of genes previously discovered in the cohort[8] with solely missense variation. There is a greater overlap between popEVE gene collapsing and this previously discovered set than the thresholding approach, owing to the similarity in their approach.

https://doi.org/10.1038/s41588-025-02400-1

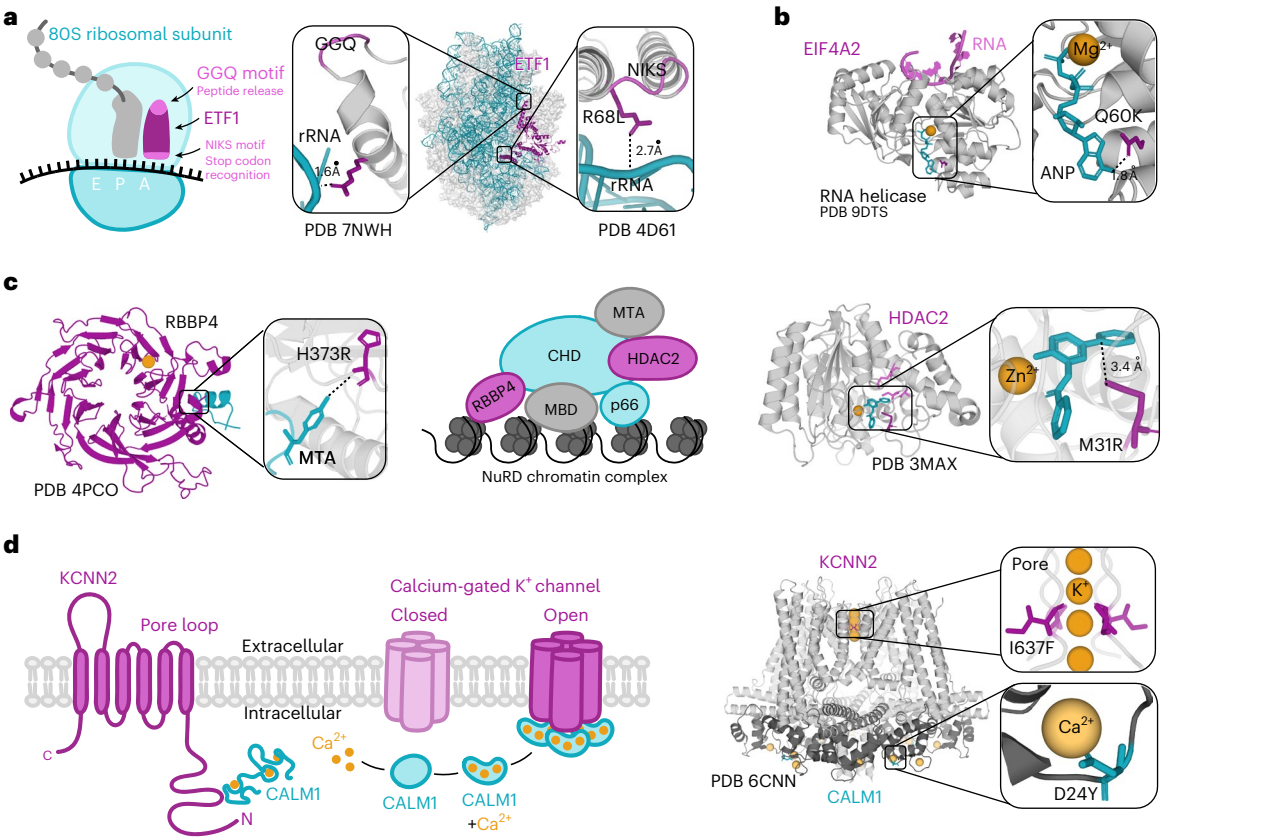



**Fig. 5 | Deleterious scoring variants lie in 3D interaction sites of candidate genes.** A total of 91% of our defined deleterious variants are within 8 Å (72% are within 5 Å) of an interaction partner. **a**, ETF1 (eRF1), a gene crucial for protein synthesis, contains our two most deleterious scoring variants (R192C and R68L), both close (<3.2 Å) to the ribosomal phosphate backbone (PDB 6D90 and PDB 7NWH) and are proximal to known functional motifs; R68 is part of the NIKS motif that determines stop codon recognition, and R192 to the GGQ motif that triggers the hydrolysis of the peptidyl-tRNA ester bond. **b**, Q60 in EIF4A2 (DDX2B) is <2 Å from the N6 of the adenine of phosphoaminophosphonic acid-adenylate ester (ANP). **c**, Many deleterious scoring variants are in the NuRD chromatin complex, such as M31R in HDAC2, which is 3.4 Å from the histone mimic inhibitor in 3MAX, and H37R in RBBP4, which is 3.8 Å from MTA1 in 4PC0. **d**, The calcium-gated ion channel complex contains deleterious scoring variants in key interaction sites, I637F in KCNN2 in the highly conserved T(V/I)GYG $K^+$ pore motif, and D24Y in CALM1, which chelates the $Ca^{2+}$ in the wild type (homologous complex structure PDB 6CNN).

those in unaffected controls. popEVE performs better than all other state-of-the-art models at distinguishing diagnosed cases from healthy controls, improving average precision by 3.2% over the next best model (Fig. 3b). Notably, popEVE differentiates diagnosed cases from controls better than variant scoring methods that train directly on clinical labels that likely include diagnostic variants from these cases (Fig. 3b, Extended Data Fig. 7a,b and Supplementary Table 1). Indeed, an independent analysis supports popEVE as the leading variant scoring method in identifying likely causal variants in SDD cases[16].

**Recovering cases without overpredicting severity.** A severity-aware model should rank variants in severe disorder cases as more deleterious than those in individuals with milder, complex conditions. As such, we compared the models' abilities to distinguish developmental disorder (DD) cases likely to be caused by a single missense DNM from generally healthy individuals from the UKBB[17] (Fig. 3c and Extended Data Fig. 7c). At increasingly stringent thresholds, popEVE recovers far more diagnosed cases without overpredicting severity in the general population. For example, popEVE can recall 50% of diagnosed cases while predicting only 11% of UKBB individuals to have equally severe variants. By contrast, AlphaMissense can identify 50% of cases but predicts 44% of the general population carries such variants, averaging five 'pathogenic' hits per person, compared to far less than one for popEVE.

As a final assessment, we tested how well models differentiate SDD cases from controls using their WES, including both inherited and de novo variants. Inherited variant scores resemble those in UKBB participants, while DNMs are shifted toward higher predicted severity (Fig. 3d). For cases with WES, popEVE shows a near 1:1 case–control separation, except at the most deleterious thresholds (Fig. 3e,f and Extended Data Fig. 7d). Other models overpredict severity in the general population, classifying nearly all UKBB individuals as harboring variants as severe as half of the SDD cases. Once again, popEVE outperforms others by recovering more true cases with fewer false positives.

## Evidence of 123 novel candidate DD genes

Given its performance across the various benchmarks and lack of biases, popEVE appears uniquely suited for use in clinical genetics settings to identify candidate variants. We first investigated popEVE's utility in discovering novel variants and genes in the SDD cohort, comprising 31,000 trios in total (Supplementary Table 5).

**Candidate discovery.** We used two approaches to identify associations: thresholding variants with a >99.99% likelihood of falling within the low-fitness distribution; and gene collapsing, comparing observed variant scores to expectations based on background mutation rates given the spectrum of scores within and across proteins ($P < 2.71 \times 10^{-6}$; Methods). This yielded 410 genes, including 152 previously reported by DeNovoWEST[8] (Fig. 4 and Supplementary Table 6). popEVE recovers 94% of missense-identified genes previously found in this cohort, and over half (135) are supported by the Developmental Disorder Gene

https://doi.org/10.1038/s41588-025-02400-1

**Table 1 | Top 25 most deleterious novel candidates**

| Gene | Mutant | Score | PDB ID | Interacting Partner | Distance (Å) |
|---|---|---|---|---|---|
| *ETF1* | R192C, R68L | −7.2, −6.8 | 7NWH, 5A8L | 18S ribosomal RNA | 1.6, 2.7 |
| *RBBP4* | H373R | −6.8 | 4PC0 | Metastasis-associated protein MTA1 | 3.8 |
| *WDR5* | S62N | −6.8 | 2GNQ | | |
| *UBE2D3* | S105Y | −6.7 | 7AHZ | Ubiquitin | 2.2 |
| *EIF4A2* | Q60K | −6.6 | 6B4K | ANP | 1.8 |
| *ANP32A* | L80R | −6.4 | 6XZQ | | |
| *UBE2H* | D120V | −6.2 | 6ZHS | Ubiquitin-activating enzyme *UBA1* | 2.0 |
| *XPO1* | T448K | −6.2 | 4HB3 | Ran (GTP-binding nuclear protein Ran) | 2.9 |
| *AMIGO1* | L112P | −6.1 | 2XOT | | |
| *COPS2* | F69C | −6.1 | 6A73 | Inositol hexakisphosphate ($InsP_6$) | 6.5 |
| *RBBP4* | T155I | −6.0 | 2XU7 | Zinc finger protein *ZFPM1* | 2.9 |
| *RBBP7* | N325D | −6.0 | 5FXY | Metastasis-associated protein *MTA1* | 3.3 |
| *DDX17* | V484M | −5.9 | 3EX7 | ATP | 6.4 |
| *SPIN1* | Y170C | −5.9 | 7OCB | Histone tail | 2.4 |
| *WARS1* | G163V | −5.9 | 4J75 | TRP-AMP | 2.7 |
| *MAT2A* | S206F | −5.9 | 7KCF | SAM (*S*-adenosylmethionine) | 2.5 |
| *KCNN2* | I637F | −5.8 | 6CNM | $K^+$ ion | 2.7 |
| *ZMYND8* | R333G | −5.8 | 5Y1Z | Actin-binding protein Drebrin (*DBN1*) | 6.6 |
| *ACTC1* | S340F | −5.8 | 7TJ7 | Fragmin (actin-binding protein) | 4.7 |
| *RBBP4* | R131C | −5.8 | 4PSX | Sulfate ion | 2.6 |
| *PSMA2* | G125D | −5.8 | 5L4G | Proteasome subunit *PSMA6* | 2.9 |
| *MAP2K4* | S262N | −5.8 | 7JUY | ANP | 2.6 |
| *NFKB2* | W270R | −5.8 | 7VUQ | | |
| *CALM1* | D24Y | −5.8 | 6B8Q | $Ca^{2+}$ ion | 2.0 |

to Phenotype (DDG2P) database[37]. We highlight 123 of the genes as novel candidates, 119 of which were identifiable at the single-variant level (Supplementary Table 7). None of these variants were observed in UKBB or GnomAD individuals. Notably, during the review process of this publication, 25 of these candidates have since been added to the DDG2P (accessed 4 September 2025). A total of 31 genes were recovered using missense variants alone that previously required LoF data. Of the 50 known genes recovered only via collapsing, many had moderate scores, underscoring the value of this combined approach. To assess false positives, we applied gene collapsing to unaffected controls—no significant genes were found. Among controls, 18 variants were predicted as severe, including one linked to Long QT and Brugada syndrome[38,39], which can cause sudden death in midlife (rs199473072). Variant thresholding flagged 7% of missense DNMs in cases to be severe (4.5% of patients), compared to just 0.5% in controls (0.2% of individuals). These results suggest that variant scores alone are informative and support using both methods when possible.

**Case variants lie in 3D interaction sites.** Since our method pinpoints individual variants that may be causal, it allows us to explore their 3D context where protein structures are available (85 out of 100 unique proteins[40]). Although we do not use any 3D structures as part of the modeling, we find that these candidate variants are close to interacting biomolecules, thus plausibly affecting the protein's function. We found that 91% are within 8 Å (72% within 5 Å) of an interaction partner, such as another protein, a metal, ligand, cofactor or nucleic acid, and are >90% closer to an interacting partner than any other random position in their respective protein (Methods). For example, the two candidate variants scored most deleterious by popEVE are in ETF1, a protein that mediates translation termination in the ribosome. R192C and R68L are both close (<3.2 Å) to the phosphate backbone of RNA in the 80S-eRF1-eRF3-GTP ternary complex (Fig. 5a and Table 1), PDB 6D90 (ref. 41). Both these residues are proximal to known functional motifs: R192 to the GGQ motif that triggers the hydrolysis of the peptidyl-tRNA ester bond, terminating protein synthesis[42], and R68 is part of NIKS motif, crucial for stop codon recognition[43]. Many other deleterious variants are associated with translation, such as Q60 in EIF4A2, which lies <2 Å from the ANP (Fig. 5b). Other top-scoring variants are in members of the NuRD complex[44] include H373 in RBBP4, which is 3.78 Å from MTA1; and M31 in HDAC2, which lies directly in the 'foot pocket' of the acetylase active site[45], <2.5 Å (Fig. 5c). Finally, we also see highly deleterious variants in two interacting proteins; I637 in the T(V/I)GYG motif essential for ion transporter KCNN2 and D24 that binds calcium in CALM1 (ref. 46) (Fig. 5d). The enrichment of variants close to the ligands and the numerous examples in the top candidates of common complexes suggests their plausibility, examined in more statistical depth below.

**Functional analysis supports candidate genes.** Three lines of evidence provide support for new candidates. Firstly 70% of the 410 genes identified using popEVE in the SDD cohort are already known to be associated with DDs ($P < 0.001$ compared to random; Methods, Fig. 4) and the score distribution of variants seen in cases in the candidate genes is nearly identical to those in known genes (Fig. 6a). Secondly, the remaining 123 newly identified proteins are hugely enriched for direct physical interactions with the 285 previously identified from the same cohort[8] (two-sided *t*-test, $P = 0$; Fig. 6b, Extended Data Fig. 9 and Supplementary Table 10). This includes 25 variants in 15 proteins from chromatin complexes (for example, NuRD and Sin3a), including HDAC2/5, RBBP4/7 and IKZF1 (Fig. 6c).

https://doi.org/10.1038/s41588-025-02400-1

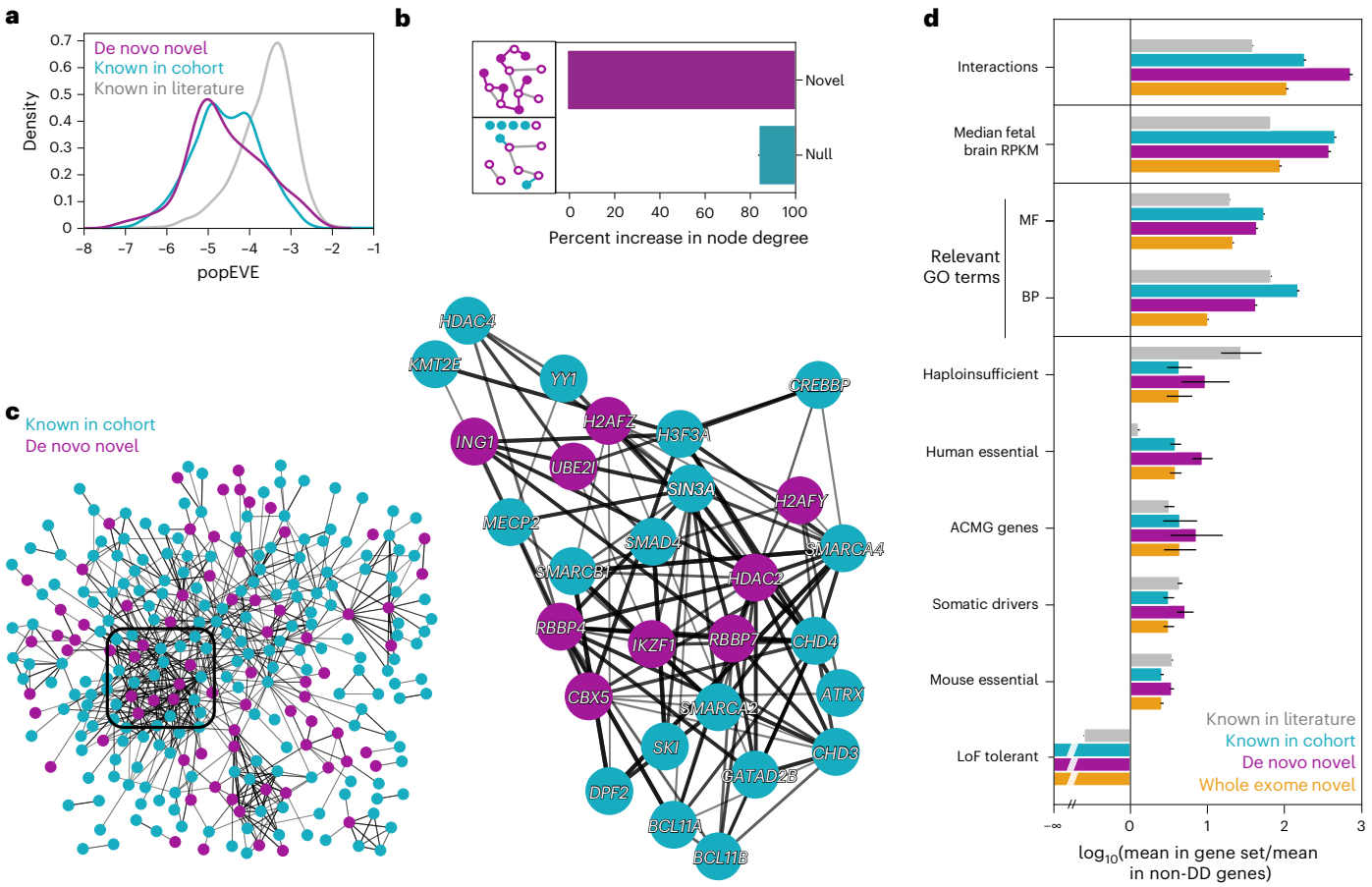



**Fig. 6 | popEVE novel candidates are functionally similar to known DD genes.** **a**, Novel candidate genes and genes previously identified in the same SDD cohort have a similar distribution of case DNMs compared to DD genes identified elsewhere. **b**, Novel genes have many biochemical interactions with genes previously identified in the same cohort—inclusion of novel genes increases node degree 100% compared to random sets (two-sided *t*-test, $P < 0.0001$). **c**, The densest portion of the network includes many genes involved in chromatin modeling. **d**, Novel ($n = 123$) and known DD genes (known in cohort $n = 280$, known in literature $n = 2,235$) show similar enrichment in properties known to differentiate known DD genes from non-DD genes (bar values show log ratio of mean in gene set of interest to mean of non-DD genes; error bars show 95% CI from bootstrapping with 1,000 simulations). RPKM, reads per kilobase of transcript per million mapped reads; GO, Gene Ontology; MF, molecular function; BP, biological processes; Extended Data Fig. 8.

Thirdly, the candidates are functionally similar to known DD genes across a wide range of features that are known to distinguish these genes from those not associated with DDs (Fig. 6d, Supplementary Table 8 and Extended Data Fig. 8). For example, candidate genes are expressed significantly more in the developing fetal brain compared to non-DD genes, even those already known to be associated with DDs ($P < 0.001$); they have similar enrichment for molecular function and biological processes[47,48] as known DD genes, such as chromatin organization (GO:0006325) and nervous system development (GO:0007399) (Supplementary Table 9); 50 out of 123 novel genes are associated with a complex involved in development and survival of neurons (NRTK1); 16 out of 123 with the SWI/SNF chromatin remodeling complex associated with neurodevelopmental disorders[49,50]; and another 15 are in ion channel complexes[51]. Of the 24 remaining genes with no connectivity, two-thirds are significantly enriched in annotations for neuronal development and differentiation (Supplementary Table 9). Additionally, novel candidate genes are enriched for essential genes measured both by homology to mouse experiments[52], by large-scale CRISPR screens[53], somatic driver genes[54] and haploinsufficient genes[55], and are significantly depleted of genes that are tolerant to homozygous LoF.

These lines of evidence, taken with the low potential to overpredic severity in the general population (above), support these novel genes' candidacy for their involvement in SDDs.

## Pinpointing likely causal de novo variants without parental genomes

Finally, we tested whether popEVE can identify likely causal variants from the child's genome alone, without parental data. We analyzed rare (MAF < 0.01) inherited and de novo variants in 9,859 individuals from the Deciphering Developmental Disorders[6] (DDD) cohort. For 2,700 of these cases, a causal missense DNM is expected. Among 513 individuals with a popEVE-severe de novo missense, 98% had this variant ranked as their most deleterious. Selecting the top-scoring variant per person still recovers 95% of genes identified by thresholding de novo variants alone. Compared to other models, popEVE more reliably ranks causal de novo mutations above all rare inherited missense variants in the same patient (Fig. 7a and Extended Data Fig. 7f). This highlights popEVE's clinical utility: when a likely causal de novo variant is present, it will more often be ranked as the most deleterious, outperforming all other models across the proteome.

With respect to new candidates that may be inherited, in addition to identifying DNMs without parental data, we found 409 inherited variants across 209 genes predicted to be severely deleterious; only one appears in the UKBB (Fig. 3f). These genes show strong enrichment for physical interactions (two-sided *t*-test, $P = 0$; Extended Data Fig. 9c and Supplementary Table 10) and functional similarity to known DD genes (Fig. 6d). Among them, 36 are already associated with DD, and

https://doi.org/10.1038/s41588-025-02400-1

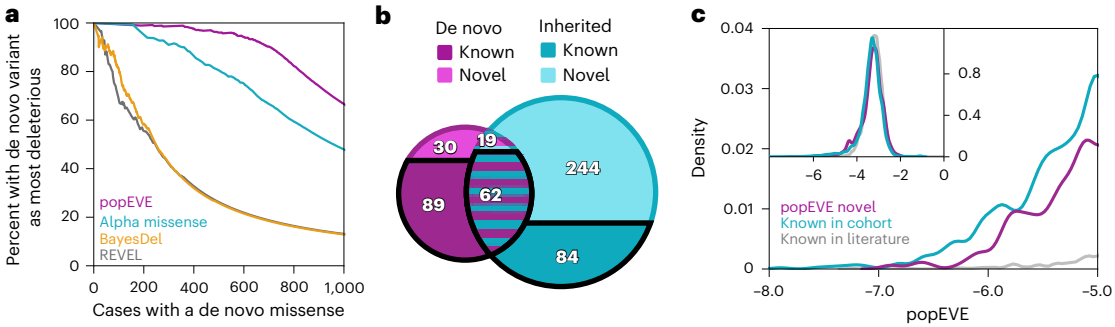



**Fig. 7 | popEVE recalls candidates without parental genomes. a**, popEVE ranks missense DNMs in known DD genes as most deleterious compared to their inherited variants in diagnosed SDD cases, better than any other model. **b**, Genes identified using DNMs compare with those identified using inherited variants. **c**, Novel candidate genes and genes previously identified in the same SDD cohort have a similar distribution of whole-exome case variants as compared to DD genes identified elsewhere, particularly at the deleterious end (inset shows entire distribution of scores).

29 overlap with novel genes from the full SDD cohort. Case variants in these candidates show popEVE score distributions comparable to those in known DD genes (Fig. 7c). While many cases are likely explained by missense DNMs, inherited variants may also contribute. Notably, in the original trio analysis, 84% of flagged variants were inherited, the majority being missense mutations[56].

## Discussion

As patient sequencing becomes standard, with growing accessibility worldwide, there is increasing demand for broadly applicable variant interpretation tools—even for cases involving diseases as rare as one patient. While standard burden analyses work when enough individuals share a rare disease, many ultra-rare conditions lack sufficient cases. This work introduces a model designed to support genetic diagnosis in such cases. Recent years have seen a rise in models predicting whether variants are benign or pathogenic, but most overlook differences in severity and penetrance. Here, we propose that treating pathogenicity as a spectrum can be more informative in certain contexts. Capturing this spectrum requires a model that ranks variants both within and across genes, that is, a true proteome-wide model. While several models offer genome-scale predictions, popEVE is, to our knowledge, the first designed specifically to calibrate scores to be comparable across genes, making it the first, albeit simple, model of the human proteome. Advancing whole-proteome modeling requires several key developments. A natural next step is to incorporate protein–protein interactions, just as protein-level models evolved from position-independent to interaction-aware frameworks. Another clear limitation of current models, including popEVE, is their inability to evaluate nonsense or truncating mutations and, thus, are unable to compare their severity to missense variants. To our knowledge, no unified model of LoF and missense variants with sufficient predictive power currently exists. However, popEVE's modular design makes it compatible with such extensions, as its human proteome calibration is agnostic to the variant type and can be easily expanded. Despite the simplicity of popEVE, it presents multiple opportunities for diagnosis and broader exploration of disease genetics. We identify novel DD gene candidates undetectable by enrichment-based methods in a cohort of this size; 104 have flagged variants in only one or two individuals. Functional, structural and network analyses show these genes are closely linked to known DD genes, and their variants often occur in functionally critical regions, providing further evidence that these variants potentially give rise to genetic disorders. More broadly, the model predicts that a large number of genes are capable of causing severe phenotypes, implying that there are still many genetic disorders yet to be identified or even seen. A similar conclusion is reached in Kaplanis et al.[8] through a distinct approach. Here, we go further by identifying specific genes and high-risk variants. Finally, we note the detrimental impact of building large-scale proteome or genome models; we are reaching a point where the energy and computational consumption of developing and training these models is costly, both financially and environmentally[57]. In this work, we sought to use a modular approach, enabling us to repurpose previous models and update components of the model with future developments at a minimal computational cost. Deep learning strategies with these properties are currently scarce, and we urgently need more techniques that lend themselves to reducing computational costs or have components that can be readily reused or recycled.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-025-02400-1.

## References

1. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nat. Genet.* **27**, 234–236 (2001).
2. Shirts, B. H., Pritchard, C. C. & Walsh, T. Family-specific variants and the limits of human genetics. *Trends Mol. Med.* **22**, 925–934 (2016).
3. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
4. Wright, C. F. et al. Genomic diagnosis of rare pediatric disease in the United Kingdom and Ireland. *N. Engl. J. Med.* **388**, 1559–1571 (2023).
5. 100,000 Genomes Project Pilot Investigators. 100,000 genomes pilot on rare-disease diagnosis in health care—preliminary report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).
6. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
7. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
8. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
9. Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
10. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).

11. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
12. Notin, P. et al. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. Preprint at https://doi.org/10.1101/2022.12.07.519495 (2022).
13. Notin, P. et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *Proc. 39th Int Conf on Machine Learning* (eds. Chaudhuri, K. et al.) 16990–17017 (PMLR, 2022).
14. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
15. Shin, J.-E. et al. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 2403 (2021).
16. Finucane, H.K. Variant scoring performance across selection regimes depends on variant-to-gene and gene-to-disease components. Preprint at https://doi.org/10.1101/2024.09.17.613327 (2024).
17. 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).
18. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
19. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, 7492 (2023).
20. Gao, H. et al. The landscape of tolerated genetic variation in humans and primates. *Science* **380**, 8153 (2023).
21. Livesey, B. J. & Marsh, J. A. Variant effect predictor correlation with functional assays is reflective of clinical classification performance. *Genome Biol.* **26**, 104 (2025).
22. Notin, P. et al. Proteingym: large-scale benchmarks for protein fitness prediction and design. *Adv. Neural Inf. Process. Syst.* **36**, 64331–64379 (2023).
23. Pejaver, V. et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am. J. Hum. Genet.* **109**, 2163–2177 (2022).
24. Dias, M., Orenbuch, R., Marks, D. S. & Frazer, J. Toward trustable use of machine learning models of variant effects in the clinic. *Am. J. Hum. Genet.* **111**, 2589–2593 (2024).
25. Harrison, S. M., Biesecker, L. G. & Rehm, H. L. Overview of specifications to the ACMG/AMP variant interpretation guidelines. *Curr. Protoc. Hum. Genet.* **103**, e93 (2019).
26. Tejura, M. et al. Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes. *Am. J. Hum. Genet.* **111**, 2031–2043 (2024).
27. 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).
28. 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 Genetics* **9**, 1003709 (2013).
29. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
30. Traynelis, J. et al. Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res.* **27**, 1715–1729 (2017).
31. 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).
32. Pathak, A.K. et al. Pervasive ancestry bias in variant effect predictors. Preprint at https://doi.org/10.1101/2024.05.20.594987 (2024).
33. Feng, B.-J. PERCH: A unified framework for disease gene prioritization. *Hum. Mutat.* **38**, 243–251 (2017).
34. 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).
35. Schaefer, J., Lehne, M., Schepers, J., Prasser, F. & Thun, S. The use of machine learning in rare diseases: a scoping review. *Orphanet J. Rare Dis.* **15**, 145 (2020).
36. 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).
37. Firth, H. et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
38. Kapplinger, J. D. et al. Spectrum and prevalence of mutations from the first 2,500 consecutive unrelated patients referred for the FAMILION® long QT syndrome genetic test. *Heart Rhythm* **6**, 1297–1303 (2009).
39. Meregalli, P. G. et al. Type of *SCN5A* mutation determines clinical severity and degree of conduction slowing in loss-of-function sodium channelopathies. *Heart Rhythm* **6**, 341–348 (2009).
40. Armstrong, D. R. et al. PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.* **48**, 335–343 (2020).
41. Pisareva, V. P., Pisarev, A. V. & Fernández, I. S. Dual tRNA mimicry in the cricket paralysis virus IRES uncovers an unexpected similarity with the hepatitis C virus IRES. *Elife* **7**, e34062 (2018).
42. Frolova, L. Y. et al. Mutations in the highly conserved GGQ motif of class 1 polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis. *RNA* **5**, 1014–1020 (1999).
43. Frolova, L., Seit-Nebi, A. & Kisselev, L. Highly conserved NIKS tetrapeptide is functionally essential in eukaryotic translation termination factor eRF1. *RNA* **8**, 129–136 (2002).
44. Alqarni, S. S. et al. Insight into the architecture of the NuRD complex: structure of the RbAp48–MTA1 subcomplex. *J. Biol. Chem.* **289**, 21844–21855 (2014).
45. Bressi, J. C. et al. Exploration of the HDAC2 foot pocket: synthesis and SAR of substituted *N*-(2-aminophenyl)benzamides. *Bioorg. Med. Chem. Lett.* **20**, 3142–3145 (2010).
46. Lee, C.-H. & MacKinnon, R. Activation mechanism of a human SK-calmodulin channel complex elucidated by cryo-EM structures. *Science* **360**, 508–513 (2018).
47. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
48. Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).
49. Santen, G. W. E., Kriek, M. & Attikum, H. SWI/SNF complex in disorder: SWItching from malignancies to intellectual disability. *Epigenetics* **7**, 1219–1224 (2012).
50. Sokpor, G., Xie, Y., Rosenbusch, J. & Tuoc, T. Chromatin remodeling BAF (SWI/SNF) complexes in neural development and disorders. *Front. Mol. Neurosci.* **10**, 243 (2017).
51. D'Adamo, M. C., Liantonio, A., Conte, E., Pessia, M. & Imbrici, P. Ion channels involvement in neurodevelopmental disorders. *Neuroscience* **440**, 337–359 (2020).
52. Blake, J. A. et al. The mouse genome database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* **39**, 842–848 (2010).
53. Hart, T. et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 (Bethesda)* **7**, 2719–2727 (2017).
54. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **173**, 1823 (2018).

55. Rehm, H. L. et al. Clingen—the clinical genome resource. *New Eng. J. Med.* **372**, 2235–2242 (2015).
56. Wright, S. J. Coordinate descent algorithms. *Math. Program.* **151**, 3–34 (2015).
57. Lannelongue, L. et al. Greener principles for environmentally sustainable computational science. *Nat. Comput. Sci.* **3**, 514–521 (2023).

# Methods

## Statistics and reproducibility

This study is based on analysis of large-scale sequencing and variant annotation datasets (UniRef[58], UKBB[17], gnomAD[18], ClinVar[59], ProteinGym[22] and DD cohorts, including DDD[6], GeneDx[8], Radboud[8], SPARK[36] and SSC[36]). No statistical method was used to predetermine sample size; all available data from each cohort or resource were included in the analyses. No data were excluded from the analyses unless explicitly stated in the Methods (for example, sibling pairs with shared de novo variants, or genes with insufficient coverage in UKBB). The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. Reproducibility was assessed by benchmarking across multiple independent datasets (ClinVar, deep mutational scans, population sequencing cohorts and DD cohorts) and by comparing results with previously published models. All code and trained models are publicly available ('Code availability'), ensuring that analyses can be reproduced.

## Data acquisition

**Multiple sequence alignments.** Following previously published protocols[60,14], the EVCouplings pipeline[61], which builds on the profile HMM homology search tool Jackhmmer[62], was used to build multiple sequence alignments (MSAs), in which sequences were obtained from the UniRef100 database of non-redundant protein[58], downloaded in March 2022.

## Human variation data

Variants from the UKBB[17] 500k release were annotated using VEP GRCh38 RefSeq and a custom RefSeq annotation built from NCBI genebank files to maximize the number of variants for pre-existing models. Variants were filtered for genotyping quality across all samples, and annotations were filtered based on matching between the RefSeq reference sequence and transcript sequences. When analyzing variants seen in the UKBB outside of training, we removed genes in which less than 95% of UKBB participants had at least 10× coverage[63].

**DD cohorts.** All cohorts included in this study obtained written, informed consent from all participants or, if the participants were minors or lacked capacity, from their parents or legal guardians, in accordance with relevant institutional and national ethical guidelines.

**SDD metacohort.** De novo mutations from a metacohort composed of subjects from the DDD study, GeneDx and Radboud Medical Center were acquired from a previous publication[8] ($n$ = 31,058). Quality filtering was performed by the respective centers as described in the supplement of the original publication[8]. The variants were re-annotated with VEP using GRCh37 RefSeq and custom mapping based on NCBI RefSeq assembly mapping files.

**Autism spectrum and unaffected siblings metacohort.** De novo mutations from SFARI's SPARK and SCC cohorts (the other two cohorts) were acquired from previously published work[36] ($n$ = 5,764). The variants were re-annotated with VEP using GRCh37 RefSeq and custom mapping based on NCBI RefSeq assembly mapping files. Sibling pairs with shared de novo variants were discarded.

**DDD cohort.** Variants from WES for the DDD cohort, a subset of the SDD metacohort ($n$ = 9,859), were re-annotated with VEP using GRCh37 RefSeq and custom mapping files. Variants were filtered by quality based on the filters used in the previously published SDD metacohort[8].

**ClinVar benign and pathogenic variants.** To assess predictive performance, we used two sets of clinically labeled variants from the ClinVar public archive[59]: the 2019 and 2020 sets curated in a previous publication[23].

**Deep mutational scans from ProteinGym.** For assessing the predictive performance based on correlation with high-throughput functional assays (otherwise known as deep mutational scans or multiplexed assays of variant effects), we consider the human subset of ProteinGym[22], which is thought to be clinically relevant. As reference sequences must have mappings to the human reference genome GrCH38, we do not have sequence matches for all available assays. Thus, the resulting test set consists of 23 assays across 18 proteins, so a modest expansion of the set considered in the previous work[11].

## Model building

**Overview of modeling strategy.** From a methodological perspective, our goal is to rank the severity of genetic variant effects across an individual's proteome. To achieve this goal, we developed a probabilistic model trained on protein sequence data from both diverse species (UniRef100) and human populations (UKBB or gnomAD). These datasets offer complementary advantages: cross-species sequences reflect millions of years of evolution, revealing conserved patterns linked to structure and function[11], while human exome data captures population-specific constraint at the gene level. By combining both, our model aims to accurately predict variant impact across the proteome at single-residue resolution.

In the following sections, we first introduce the models (referred to here as evo models) used for identifying patterns of conservation across diverse organisms. These models provide a 'fitness' score for a given sequence of interest by obtaining an estimate for the log odds:

$$\sigma = \log\left(\frac{p(\boldsymbol{x})}{p(\boldsymbol{x}^{\text{ref}})}\right), \tag{1}$$

where $\boldsymbol{x}$ represents the sequence of interest and $\boldsymbol{x}^{\text{ref}}$ is the reference sequence.

We introduce popEVE, a model that predicts the presence or absence of a variant in the human population based on input fitness scores from underlying models. It produces a calibrated score that effectively rescales and ensembles these inputs, enabling comparison of variant effects across different proteins.

**Modeling individual proteins using evolutionary data.** Recent work has shown that unsupervised models trained on protein sequence distributions across diverse species can distinguish benign from pathogenic variants in known disease genes, performing comparably to functional assays[11,64]. We use two subtypes of such models: an alignment-based model, which is a variational autoencoder, trained on MSAs of individual genes; and an alignment-free model, inspired by large language models, trained on a full protein database (UniRef90). Below, we summarize each approach.

*The Bayesian variational autoencoder (EVE).* Variational autoencoders (VAEs)[65,66] are a class of latent variable models that have been shown to be effective at capturing high-dimensional distributions in computer vision[67,68], natural language processing[69] and more. The assumption underlying a VAE is that the observed high-dimensional distribution is generated by a much smaller number of hidden (also known as latent) variables $z_i$. The generative story is thus:

$$\begin{aligned} \boldsymbol{z} &\sim \mathcal{N}(0, \boldsymbol{I}_D) \\ p(x_i^{\alpha}|\boldsymbol{z}, \boldsymbol{\theta}) &= \text{softmax}((f^{\boldsymbol{\theta}}(\boldsymbol{z}))_i^{\alpha}), \end{aligned} \tag{2}$$

where $x_i^{\alpha}$ is an indicator function for the presence of amino acid $\alpha$ at position $i$, and the 'decoder' $f^{\boldsymbol{\theta}}(\boldsymbol{z})$ is modeled with a fully connected neural network, with spherical Gaussian prior for the parameters $\boldsymbol{\theta}$. In words, the VAE models the conditional probability of seeing the amino acid $\alpha$ at position $i$, given the latent variables $\boldsymbol{z}$. Parameter inference is

achieved by the use of amortized inference, where we model the distribution $q(\boldsymbol{z}|x_{ij}, \boldsymbol{\theta})$ with another fully connected neural network, often referred to as the encoder. In previous work[11] we found a symmetric relationship between encoder and decoder to work well with three layers, consisting of 2,000–1,000–300 and 300–1,000–2,000 nodes, respectively.

To score a sequence, we use the evidence lower bound (ELBO), which is a lower bound on the log-marginal likelihood $p(\boldsymbol{x})$:

$$\mathrm{ELBO}(\mathbf{x}) = N_{\mathrm{eff}} \cdot \mathbb{E}_{p(\boldsymbol{x})} \left[ \mathbb{E}_{q(\theta_{\mathbf{p}}), q(\mathbf{z}|\boldsymbol{x})} \left( \log p(\boldsymbol{x}|\mathbf{z}, \theta_{\mathbf{p}}) \right) - D_{KL}(q(\mathbf{z}|\boldsymbol{x}, \boldsymbol{\phi}_{\mathbf{p}}) \parallel p(\mathbf{z})) \right] - D_{KL}(q(\theta_{\mathbf{p}}) \parallel p(\theta_{\mathbf{p}})) \tag{3}$$

where $N_{\mathrm{eff}} = \sum_{n=1}^{N} w_{x_n}^{C}$ and $w_{x_n}^{C}$ is defined in equation (5). The fitness score is then simply

$$\sigma = \log \frac{p(\boldsymbol{x}|\theta_{\mathbf{p}})}{p(\boldsymbol{x}^{\mathrm{ref}}|\theta_{\mathbf{p}})} \approx ELBO(\boldsymbol{x}) - ELBO(\boldsymbol{x}^{\mathrm{ref}}) \tag{4}$$

*Sequence reweighting*. All models used in this work make the false assumption that the training data is independently and identically distributed. This independently and identically distributed assumption breaks down owing to phylogenetic and ascertainment biases. The fact that the VAE is trained on aligned data presents an opportunity to correct for these two biases with sequence reweighting. Following the approach described in previous work[70], we re-weight each protein sequence $\boldsymbol{x}_i$ from a given MSA according to the reciprocal of the number of sequences in the corresponding MSA within a given Hamming distance cutoff, $T$.

$$w_{\boldsymbol{x}_n}^{C} = \left( \sum_{\substack{m=1 \\ m \neq n}}^{N} \mathbb{1}[\mathrm{Dist}(\boldsymbol{x}_n, \boldsymbol{x}_m) < T] \right)^{-1} \tag{5}$$

where $N$ is the number of sequences in the MSA, and bold, lowercase $\boldsymbol{x}$ represents a protein sequence, indexed by subscript Latin indices. As in previous work[60], we set $T = 0.2$ for all human proteins.

*Masked language model (ESM-1v)*. The transformer architecture has enabled the training of single, alignment-free models of essentially all known proteins. In this work, we make use of ESM-1v[31], which is trained on UniRef90.

ESM-1v[31] is a high-capacity 650 million parameter language model that uses a form of self-supervision known as masking. During training, each sequence has a randomly sampled fraction of its amino acids replaced with a 'mask' token, and the network is then trained to predict the amino acids that have been masked. For each masked amino acid, the negative log likelihood of the missing amino acid, conditioned on the sequence context, is independently minimized.

$$\mathcal{L} = \mathbb{E}_{x \sim X} \mathbb{E}_{M} \sum_{i \in M} -\log p(x_i | x_{/M}). \tag{6}$$

Hence, for the model to successfully perform this task, the dependencies between the masked amino acid and the unmasked sequence context must be learned.

## Estimating variant-level constraint in humans using Gaussian processes

The models described above perform well for ranking variants within a given gene but are not effective at comparing variants across genes (Fig. 1 and Supplementary Table 1). This limitation is expected, particularly for alignment-based models, which are trained independently for each coding region. Although several models provide genome-wide scores (for example, Fig. 2e), none have been explicitly designed to rank variant severity across the proteome. To address this gap, we introduced popEVE, the first model aimed at proteome-wide comparison of missense variant effects.

Similar to above, we define the evo score from one of the evo models, which we index $A$, with $A \in \{1, 2\}$, as the log odds between the sequence of interest $\boldsymbol{x}$ and some reference sequence $\boldsymbol{x}_{\mathrm{ref}}$:

$$\sigma^A = \log \left( \frac{p^A(\boldsymbol{x})}{p^A(\boldsymbol{x}^{\mathrm{ref}})} \right). \tag{7}$$

In what follows, sequences that differ from the reference sequence by a single amino acid substitution have a special role, so it is convenient to define $(\sigma_i^{\alpha})_n^A$ as the score from model $A$ for a protein sequence, which differs from the reference $\boldsymbol{x}_n^{\mathrm{ref}}$ sequence for protein $n$ solely by having amino acid $\alpha$ at position $i$.

We expect the probability of observing a sequence in the population to depend, in a fairly simple manner, on the score from the underlying evo models. We adopt a simple Bayesian, non-linear, non-parametric approach to modeling this relationship, with the use of a Bernoulli likelihood and a latent Gaussian process. Specifically, we model the presence or absence of the variant in the UKBB as:

$$p_n^A(y_i^{\alpha} | \sigma_{in}^{\alpha A}) = \mathrm{Ber}\left( y_{in}^{\alpha} | \varphi(f_n^A(\sigma_{in}^{\alpha A})) \right) \tag{8}$$

where $y_{ijk} \in \{1, 0\}$ indicates the presence or absence of a variant in the UKBB, the link function $\varphi(\cdot)$ is the inverse logit function (also referred to as the logistic function) $\varphi(z) = \exp(z)(1 + \exp(z))^{-1}$ and the function $f_n^A(\sigma)$ is drawn from a Gaussian process prior:

$$f(\sigma) \sim GP(m(\sigma), \mathcal{K}(\sigma, \sigma')), \tag{9}$$

with zero mean function $m(\sigma) = 0$ and radial basis function kernel

$$\mathcal{K}(\sigma, \sigma') = \exp\left( -\gamma(\sigma - \sigma')^2 \right). \tag{10}$$

The inferred function $f_n^A(\sigma)$ can be thought of as a new fitness score. The intuition is that by modeling the amount of variation seen per protein in the UKBB or gnomAD, $f_n^A(\sigma)$ it essentially rescales the evo score $\sigma_n^A$ to account for the degree of constraint acting on a per-variant basis in the population, and how that constraint depends on $\sigma_n^A$; thus resulting in a score that can rank the pathogenicity of variants across different coding regions.

**Efficient function inference by restoring conjugacy with Pólya-gamma data augmentation.** For each protein of interest, indexed $n$, and each underlying evo model, indexed $A$, we seek to infer the functions $f_n^A$. To do so, we consider the scores of all possible single amino acid substitutions in that protein and their corresponding labels $y_{in}^{\alpha}$, indicating if that variant has been observed, or not, in the UKBB (we also provide a version of the model trained on gnomAD instead of the UKBB). Dropping the indices $n$ and $A$ for compactness, we denote the training data as the set of scores $\boldsymbol{\sigma} = [\sigma_1^1, \ldots, \sigma_L^{19}] \in \mathbb{R}^N$ and $\boldsymbol{y} = [y_1, \ldots, y_N] \in \{0, 1\}^N$, where $L$ is the number of amino acids in the protein and $N = 19L$ is the total number of possible single amino acid substitutions. Let $\boldsymbol{f} = [f_1, \ldots, f_N]$ be the function values corresponding to the input $\boldsymbol{\sigma}$, then equation (8), together with the Gaussian process prior for $f$, implies:

$$p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{\sigma}) \propto p(\boldsymbol{y}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{\sigma}), \tag{11}$$

where $p(\boldsymbol{f}|\boldsymbol{\sigma}) = \mathcal{N}(\boldsymbol{f}|0, K_{NN})$, with $K_{NN}$ denoting the kernel matrix evaluated at the training points. Therefore, in contrast to models with a Gaussian

process prior and a Gaussian likelihood, inference is analytically intractable because the Gaussian prior is not conjugate to the Bernoulli likelihood.

One appealing approach to overcoming this issue is to introduce additional latent variables that restore conjugacy. Following previous work[71], we introduce the auxiliary variables $\boldsymbol{\omega}$ and define the augmented likelihood to factorize as:

$$p(\boldsymbol{y}, \boldsymbol{\omega}) = p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{\omega})p(\boldsymbol{\omega}) \tag{12}$$

The goal then is to find a prior $p(\omega)$ to satisfy two properties: that when marginalizing out $\omega$, the original model is recovered; and the Gaussian prior $p(\boldsymbol{f})$ is conjugate to the likelihood $p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{\omega})$. These conditions are satisfied by the Pólya-gamma distribution, which may be thought of as an infinite convolution of Gamma distributions; that is, $\omega \sim \mathrm{PG}(b, c)$, where

$$\omega = \frac{1}{2\pi^2}\sum_{m=1}^{\infty}\frac{g_m}{\left(m-\frac{1}{2}\right)^2+\left(\frac{c}{2\pi}\right)^2}, \tag{13}$$

and $g_m \sim \Gamma(b, 1)$. Alternatively, we can define the Pólya-gamma distribution in terms of its moment generating function (these two definitions are related by the Laplace transform):

$$\mathbb{E}_{\mathrm{PG}}[\exp(-\omega t)] = \frac{1}{\cosh^b\left(\sqrt{\frac{t}{2}}\right)}. \tag{14}$$

This second definition is useful, as it suggests that the logistic link function may be expressed in terms of Pólya-gamma variables

$$\begin{aligned}\varphi(z_i) &= \frac{\exp(z_i)}{(1+\exp(z_i))}\\ &= \frac{\exp\left(\frac{z_i}{2}\right)}{2\cosh\left(\frac{z_i}{2}\right)}\\ &= \frac{1}{2}\int \exp\left(\frac{z_i}{2}-\frac{z_i^2}{2}\omega_i\right)p(\omega_i)\mathrm{d}\omega_i.\end{aligned} \tag{15}$$

Hence, substituting $z_i = y_i f(\sigma_i)$, we obtain

$$p(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{f}) \propto \exp\left(\frac{1}{2}\boldsymbol{y}^{\top}\boldsymbol{f}-\frac{1}{2}\boldsymbol{f}^{\top}\Omega\boldsymbol{f}\right), \tag{16}$$

where $\Omega = \mathrm{diag}(\boldsymbol{\omega})$ is the diagonal matrix of the Pólya-gamma variables. This augmented likelihood is conjugate to $p(\boldsymbol{f})$ as required.

Developed in prior work[72], once conditional conjugacy is restored, it is possible to derive closed-form updates for variational inference with natural gradients and a learning rate close to one, enabling highly efficient inference of $\boldsymbol{f}$.

**Making the models scale with inducing points.** Inference in GPs with a Gaussian likelihood, while exact, take $\mathcal{O}(N^3)$ time, and hence additional methods are required to perform inference when the training data are large. One such method is to learn a 'summary' of the data with $M \ll N$ pseudo inputs, otherwise known as inducing points[73], and hence reduce the complexity to $\mathcal{O}(M^3)$. Following the prior work[72], we introduce $M$ additional variables $\boldsymbol{u} = [u_1, \ldots, u_M]$, where the function values of the GP $\boldsymbol{f}$ are related to $\boldsymbol{u}$ by

$$\begin{aligned}p(\boldsymbol{f}|\boldsymbol{u}) &= \mathcal{N}(\boldsymbol{f}|K_{NM}K_{MM}^{-1}\boldsymbol{u}, \tilde{K})\\ p(\boldsymbol{u}) &= \mathcal{N}(\boldsymbol{u}|0, K_{MM}),\end{aligned} \tag{17}$$

where $k_{MM}$ is the kernel matrix resulting from evaluating the kernel at the $M$ inducing points, $K_{NM}$ is the kernel between the training points and the inducing points and $\tilde{K} = K_{NN} - K_{NM}K_{MM}^{-1}K_{MN}$.

Hence, the complete joint distribution of our model is given by

$$p(\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{f})p(\boldsymbol{\omega})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u}). \tag{18}$$

**Implementation.** This model is implemented in GPyTorch[74] and is publicly available through a dedicated GitHub repository.

**Ensembles of models with only partially intersecting domains.** Ensembles of models can often achieve a performance similar to, and sometimes even stronger than, the strongest constituent model[75]. Our setup provides a novel opportunity to build a highly performant ensemble model by incorporating the scores from multiple evo models. By training a separate Gaussian process model for each evo model, we naturally create directly comparable scores between models, thereby enabling the typical, but potentially problematic, standardization step to be bypassed entirely. We define the popEVE score $\bar{\sigma}$ to simply be the mean of the means of the posteriors of each GP, for each evo model whose domain contains the variant of interest.

$$\bar{\sigma}_{ij}^{\alpha} = \frac{1}{Q_{ij}^{A}}\sum_{A=1}^{Q_{ij}^{\alpha}}\mathbb{E}\left[f_j^A(\sigma_i^{\alpha})\right], \tag{19}$$

where $Q_{ij}^{A}$ is the number of evo models capable of making a prediction for the amino acid substitution $\alpha$ at position $i$ in protein $j$.

## Performance assessments

We evaluated key properties of our model with a number of clinically relevant tasks and compared its performance with pre-existing models, whose scores were downloaded from dbNSFP (v.4.7)[76] (we use the same set of models that were analyzed in a previous publication[23]).

**Comparing model performance within proteins.** To assess model performance at ranking the pathogenicity of variants within the same gene, similar to prior work[11], we consider two tests: correlation of model predictions with deep mutational scans and ability to predict benign and pathogenic labels in ClinVar.

To assess concordance with deep mutational scans, we compute the Spearman's correlation between the model score and reported experimental fitness (for example, expression) (Extended Data Fig. 2). To assess the ability to separate benign from pathogenic variants, we made use of two curated sets of ClinVar labels as described in 'ClinVar benign and pathogenic variants'. We computed the area under the receiver operating characteristic curve for all genes with at least five benign and five pathogenic labels. With these sets, we were able to assess the performance across 50 and 31 proteins and in 2019 and 2020 datasets, respectively (Extended Data Fig. 1).

**Comparing model performance across proteins.** *Ranking clinical pathogenic variants by severity.* To evaluate how well models rank clinical pathogenic variants by severity, we assembled a set of ClinVar 1+ star variants linked to phenotypes with childhood or adult onset or death, based on OrphaNet annotations[35]. For each model, we identified the 5th percentile score threshold for benign ClinVar variants and used this as a reference. We then compared the log odds of childhood-versus-adult-associated pathogenic variants falling below this threshold. Pairwise $z$-scores and $P$ values were computed to assess whether model differences were statistically significant. Models focused solely on pathogenicity are expected to perform poorly, whereas those capturing variant fitness and phenotypic severity should distinguish between childhood and adult onset variants.

*Distinguishing SDD cases from unaffected controls.* We next evaluated the model's ability to rank variant severity across the proteome and across individuals. To do so, we constructed a test set from de novo

variants found in patients with SDDs and unaffected siblings of individuals with autism spectrum disorder. The goal was to determine whether the model could distinguish cases carrying likely pathogenic de novo variants from controls with variants not expected to contribute to disease.

We defined 'cases' as individuals with at least one de novo variant in a known DD gene (DDG2P, downloaded January 2023), following the approach of a previous publication[8]. However, only a subset of these variants is expected to be truly pathogenic. Based on an observed excess of 2,982 de novo missense variants among 5,625 cases, we estimate that approximately 53% carry a causal variant. Accordingly, we adjusted the maximum achievable recall to reflect this expected upper bound.

This test set (Supplementary Table 5) allows us to benchmark model performance in a realistic clinical setting. popEVE outperforms all current state-of-the-art models for pathogenicity prediction, including the underlying evolutionary models it builds upon (Fig. 3b, Extended Data Fig. 9a and Supplementary Table 1).

*Assessing overprediction of deleterious variants in the general population.* We evaluated each model's ability to recover SDD cases from de novo mutations or WES while minimizing false positives in the general population. For each model and score threshold, we computed the percentage of individuals in the UKBB and SDD metacohort (or DDD sub-cohort) with at least one variant as or more pathogenic (Fig. 2h,k and Supplementary Fig. 10d,f), comparing the cumulative distributions between cases and controls. For the SDD metacohort, we used individuals with de novo missense variants in genes previously implicated by DeNovoWEST[8], representing high-confidence diagnoses.

To assess prioritization of causal variants, we calculated the fraction of DDD individuals whose de novo variant was ranked more pathogenic than all inherited variants at each threshold. At −5.056, 513 individuals had a qualifying variant, and in 98% of cases, it ranked as the most deleterious. The rate of decay reflects each model's genome-wide ranking ability.

## Analysis of patient data

We explored two approaches to analyzing de novo mutations from a metacohort composed of subjects from the DDD study, GeneDx and Radboud Medical Center in the hope that popEVE may provide novel evidence for the genetic diagnosis of currently unsolved cases: burden testing of variants per gene across the full cohort, and per-patient direct case–variant association.

**Direct case–variant association.** Based on the set of de novo variants from cases and controls, we constructed a Bayesian Gaussian mixture model to determine a score cutoff as:

$$\begin{aligned} \boldsymbol{\mu}_1 &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ \boldsymbol{\mu}_2 &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ \lambda_1 &\sim \text{Lognormal}\,(\mu_\lambda, \sigma_\lambda) \\ \lambda_2 &\sim \text{Lognormal}\,(\mu_\lambda, \sigma_\lambda) \\ \boldsymbol{\pi} &\sim \text{Dirichlet}\,(\boldsymbol{\alpha}) \\ \text{For } i = 1, \ldots, N: & \\ a_i &\sim \text{Categorical}\,(\boldsymbol{\pi}) \\ \mathbf{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}_{a_i}, \lambda_{a_i}) \end{aligned} \tag{20}$$

where $\boldsymbol{\mu}_0 = -3.6$ and $\boldsymbol{\Sigma}_0 = 0.7$. We then identified an uncertainty cutoff corresponding to a greater than 99.99% likelihood of being in the lower fitness distribution, $\bar{\sigma} \leq -5.056$. We found that constructing the model with solely the de novo variants from the cases resulted in a similar threshold.

Using the threshold $\bar{\sigma} \leq -5.056$, we searched the full DD cohort for individuals with at least one de novo missense variant below this score and no predicted LoF variants. For these individuals, we consider the identified missense variant a strong candidate for being causal. In addition to the high accuracy of the Gaussian mixture model, these variants show strong enrichment relative to the background mutation rate (Fig. 3d), further supporting their likely pathogenicity.

**Gene-collapsing model.** To compare with previous methods, we implemented gene-collapsing models for the SDD cohort following the DeNovoWEST framework[8]. For each gene, we estimated the probability of observing a total score $\geq x_{\text{obs}}$ by testing mutation counts (0 to N) until the Poisson likelihood, based on expected de novo mutation rates, fell near zero. Context-dependent mutation rates from Samocha et al. (2014)[29] were used to estimate expected counts and sample variants. We ran 10,000 simulations per gene. To adjust for multiple testing, we also assessed the likelihood of observing a score $\geq x_{\text{obs}}$ anywhere in the proteome.

$$p(\text{ gene }) = \sum_{n=0}^{N} P(n = n_{gene}|\lambda_{gene}) P(x_{gene} \leq x_{obs}|n) P(x_{proteome} \leq x_{obs}|n) \tag{21}$$

We selected our significance cutoff by dividing 0.05 by the total number of tests (or total number of genes or proteins we have modeled): $p < 0.05 / 18{,}395$. We also performed gene collapsing on the unaffected controls using the same method.

## Structural and functional analysis of deleterious mutations

**Functional similarity of novel and known DD genes.** We compared the functional properties of our novel genes with comparison to known DD genes[37] across features known to differentiate DD genes from genes not associated with DDs, similar to previous work[8]. We calculated the enrichment of these variables in either our novel genes or known DD genes compared to non-DD genes (Fig. 6d, Extended Data Fig. 8 and Supplementary Tables 8 and 9).

For interactions, we selected protein–protein interactions (BioGRID[77]) enriched in known DD genes compared to non-DD genes. To ensure that these interactions are generalizable, we selected those that are significantly enriched in known DD genes ($P < 0.05$ after Benjamini–Hochberg correction) and present in at least 10% of known DD genes ($n > 223$). Median expression, measured in reads per kilobase of transcript per million mapped reads, in the fetal brain was determined across samples from the Allen Brain Atlas[78]. For the relevant Gene Ontology terms Molecular Function and Biological Processes, we selected terms enriched in known DD genes compared to non-DD genes using DAVID[79]. To ensure these terms are generalizable, we selected those that are significantly enriched in known DD genes ($P < 0.05$ after Benjamini–Hochberg correction) and present in at least 10% of known DD genes ($n > 223$). Haploinsufficient is a binary variable of genes with evidence of haploinsufficiency (dosage pathogenicity level three) from the ClinGen Dosage Sensitivity Map[55]; human essential is a binary variable of whether a gene was deemed essential in human cell lines based on CRISPR screens[53]; ACMG genes is a binary variable indicating whether a gene is a clinically actionable gene according to the American College of Medical Genetics and Genomics (v.2.0)[80]; somatic drivers is a binary variable of whether the gene is known to be a somatic driver gene[81]; mouse essential is a binary variable of whether a gene is essential in mice; that is, homozygous knockouts of that gene resulted in lethality[52,82]; and LoF tolerant is a binary variable of whether a gene is tolerant of homozygous LoF mutations in humans[18].

**Functional network.** We created a functional gene network with 123 de novo novel genes from popEVE (with a 99.99 threshold) and significant genes from DeNovoWEST[8] using STRING[83], as shown in

Extended Data Fig. 9a and Supplementary Table 10. We show edges from medium-confidence (0.4) experiments. Genes were denoted as previously discovered if they were already observed in DDG2P[37] or DeNovoWEST.

Additionally, we used medium-confidence (0.4) experiment annotations to calculate the average node degree of DDG2P genes and DeNovoWEST genes with and without our significant popEVE genes included, both from de novo analysis and no-trios analysis (Extended Data Fig. 9b). To determine the relative difference that the significant genes make versus a random set of genes, we performed *t*-tests 100,000 times with random samples of genes from the whole human genome (with the known and popEVE genes excluded).

**Manual structure analysis.** From the 131 novel popEVE mutations, we individually investigated structures for the top 20 most predicted deleterious. We only analyzed cryo-electron microscopy or crystallographic structures where our mutation is included in the resolved protein structure. To enhance our analysis, we prioritized structures that exhibited interactions with other proteins and/or ligands. This allowed us to capture and understand the potential consequences of these interactions. All distances listed were calculated with the distance function in PyMol.

**Comparative analysis of protein–ligand interactions using a null model.** We compared de novo variants from the SDD metacohort and inherited variants from the DDD subset against a null model, which calculated the distance from each position on the variant-containing chain to the nearest ligand. Resolved crystal structures were processed using the Evcouplings PDB reader, with distances computed using the Evcouplings compare package[61]. *Z*-scores were computed by subtracting the mean chain-wide distance (excluding the variant site) from the variant-to-ligand distance. Variants were excluded if the chain length differed from the full protein length by more than two standard deviations. Full results are in Supplementary Table 11.

**Functional interactions in 3D structures for high-scoring pathogenic variants.** 3D structures of proteins with high-scoring pathogenic variants were retrieved by alignment to SIFTS database sequences[84] using EVcouplings[61] with one HMMER iteration[62] and a 0.2 bits per residue threshold. A variant was considered to have evidence of functional interaction if, in any matched structure, its position contacted a non-self PDB entity within 8 Å, excluding water and common crystallographic additives (entity info from PDBe REST API[40]).

### Reporting summary

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

## Data availability

Interactive web viewer and downloads for popEVE scores are available at pop.evemodel.org. De novo variants from SDD cases and unaffected controls were sourced from previous publications[8,36]. WES data from the DDD subset of the SDD cohort are available under controlled access from the European Genome–Phenome Archive (EGA; accession EGAC00001000282). Access is managed by the DDD Data Access Committee to ensure use is consistent with participant consent and ethical approvals. Researchers wishing to obtain the data should apply through the EGA portal, providing details of their research project, institutional affiliation and ethical approval. Population variants are from GnomAD (v.2) and UKBB.

## Code availability

Code is available at Github (https://github.com/debbiemarkslab/popEVE) and Zenodo (https://doi.org/10.5281/zenodo.17055823)[85].

## References


58. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
59. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, 1062–1067 (2018).
60. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
61. Hopf, T. A. et al. The EvCouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019).
62. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, 1002195 (2011).
63. Wang, Q. et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
64. Livesey, B. J. & Marsh, J. A. Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol. Syst. Biol.* **19**, e11474 (2023).
65. Kingma, D.P. & Welling, M. Auto-encoding variational Bayes. Preprint at https://doi.org/10.48550/arXiv.1312.6114 (2014).
66. Rezende, D.J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. Preprint at https://doi.org/10.48550/arXiv.1401.4082 (2014).
67. Vahdat, A. & Kautz, J. NVAE: a deep hierarchical variational autoencoder. Preprint at https://doi.org/10.48550/arXiv.2007.03898 (2020).
68. Ramesh, A. et al. Zero-shot text-to-image generation. Preprint at https://doi.org/10.48550/arXiv.2102.12092 (2021).
69. Bowman, S.R. et al. Generating sentences from a continuous space. Preprint at https://doi.org/10.48550/arXiv.1511.06349 (2016).
70. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E* **87**, 012707 (2013).
71. Polson, N. G., Scott, J. G. & Windle, J. Bayesian inference for logistic models using Pólya-gamma latent variables. *J. Am. Stat. Assoc.* **108**, 1339–1349 (2013).
72. Wenzel, F., Galy-Fajou, T., Donner, C., Kloft, M. & Opper, M. Efficient Gaussian process classification using Pòlya-gamma data augmentation. In *Proc. AAAI Conf. on Artificial Intelligence* 5417–5424 (AAAI Press, 2019).
73. Snelson, E. & Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems* Vol. 18 (eds. Weiss, Y. et al.) 1257–1264 (MIT Press, 2005).
74. Gardner, J.R., Pleiss, G., Bindel, D., Weinberger, K.Q. & Wilson, A.G. Gpytorch: blackbox matrix–matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems* (eds. Bengio, S. & Wallach, H. M.) 7587–7597 (Curan Associates, 2018).
75. Zhou, Z.-H. *Ensemble Methods: Foundations and Algorithms* (CRC Press, 2012).
76. 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).
77. Oughtred, R. et al. The biogrid database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).
78. Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
79. Sherman, B. T. et al. David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, 216–221 (2022).

80. Kalia, S. S. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (acmg sf v2. 0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
81. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
82. Georgi, B., Voight, B. F. & Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.* **9**, 1003484 (2013).
83. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, 638–646 (2022).
84. Velankar, S. et al. SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.* **41**, D483–D489 (2013).
85. Marks Lab. popEVE: population-based variant effect prediction models. *Zenodo* https://doi.org/10.5281/zenodo.17055823 (2025).

## Acknowledgements

We thank all members of the Marks Lab, Dias and Frazer Lab and Sander Lab for valuable discussions. We also thank J. Nicoludis and Invitae for their assistance with training some of the EVE models. R.O., A.W.K., C.A.S., M.D., J.F. and D.S.M. are supported by a Chan Zuckerberg Initiative Award (Neurodegeneration Challenge Network, CZI2018-191853). H.S. and D.S.M. are supported by a National Institutes of Health Transformational Research Award (TR01 1R01CA260415). C.A.S. is supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE-2146755. M.D. and J.F. are supported by the Spanish Ministry of Science and Innovation (PID2022-140793NA-I00) funded by MCIN/AEI/10.13039/501100011033/FEDER, UE) and acknowledge the support of the Spanish Ministry of Science and Innovation through the Centro de Excelencia Severo Ochoa (CEX2020-001049-S, MCIN/AEI/10.13039/501100011033) and the Generalitat de Catalunya through the CERCA programme.

## Author contributions

R.O., J.F., M.D. and D.S.M. conceived the end-to-end approach. R.O., J.F. and M.D. built the models. R.O. and C.A.S. compiled and annotated the clinical and genomic data for analysis. A.W.K. and L.v.N. supported model training. R.O., T.H., D.S.M., A.D.S., D.F. and C.A.S. performed the structural and functional analysis. T.H. developed the interactive web application. R.O., D.S.M., M.D. and J.F. wrote the manuscript. D.S.M., J.F. and M.D. led and supervised the project.

## Competing interests

The authors declare no competing interests.

## Additional information

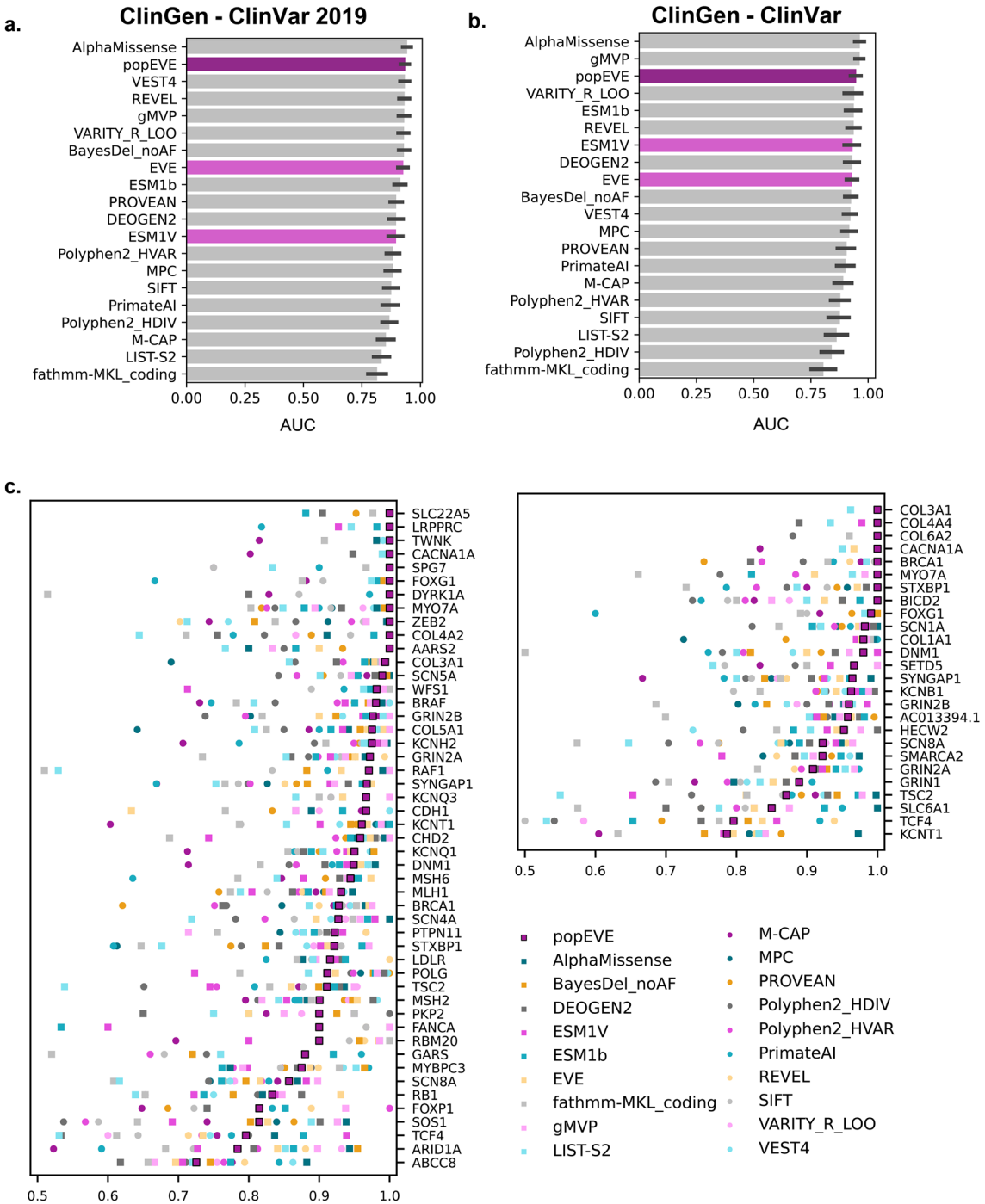
**Extended data** is available for this paper at https://doi.org/10.1038/s41588-025-02400-1.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02400-1.

**Correspondence and requests for materials** should be addressed to Mafalda Dias, Jonathan Frazer or Debora S. Marks.

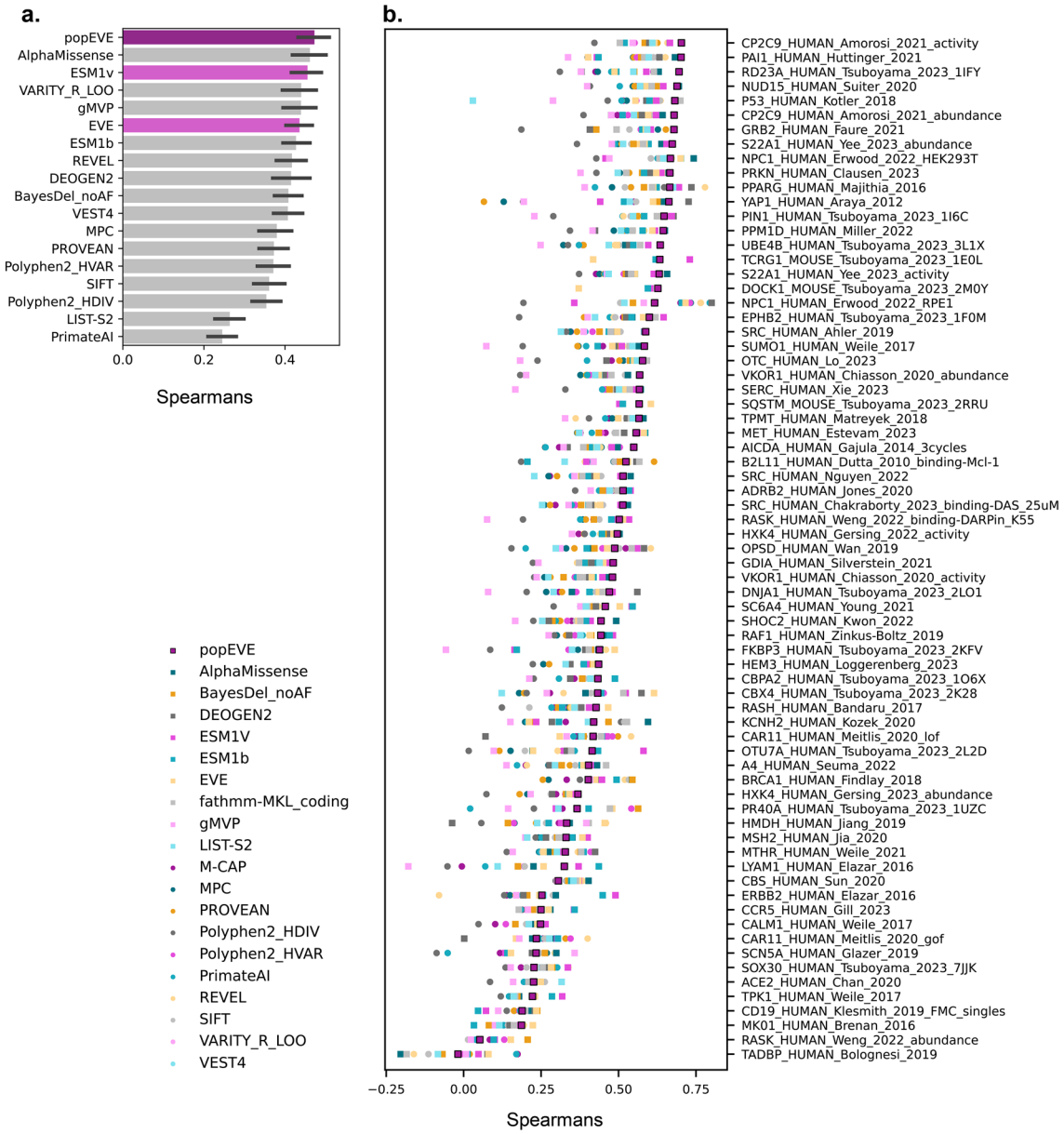
**Peer review information** *Nature Genetics* thanks Ryan Dhindsa and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-025-02400-1



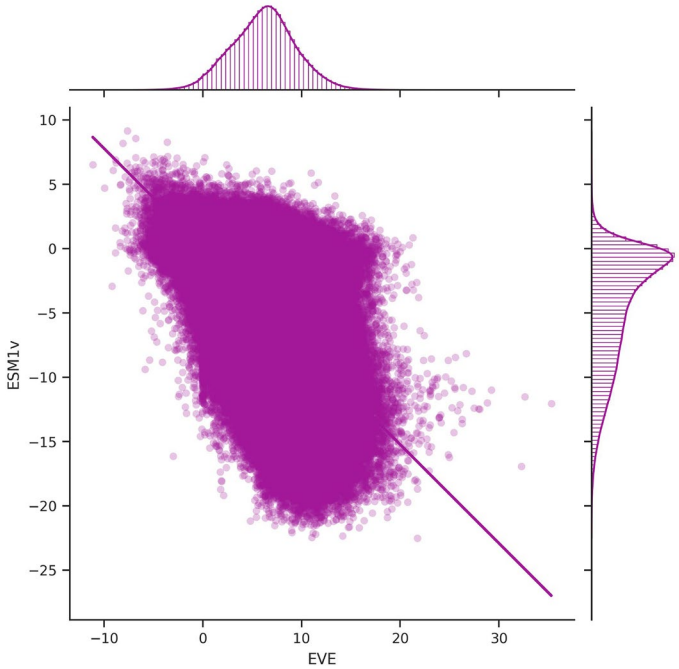

**Extended Data Fig. 1 | Performance summary for separating Benign/Likely Benign from Pathogenic/Likely Pathogenic ClinVar labels.** Assessing the performance of popEVE and popular supervised and unsupervised variant effect prediction models on individual genes that have at least 5 benign and 5 pathogenic variants from the ClinGen curation of **a**. ClinVar 2019 and **b**. ClinVar 2020, using the area under the receiver-operating curve. The ClinGen dataset attempts to address data leakage in the estimation of performance of supervised methods by removing ClinVar variants used in training. This test lacks the resolution to distinguish state-of-the-art models. This is highlighted by the fact the ranking of AUCs in the ClinGen 2020 and ClinGen 2019 significantly changes. c. Breakdown of performance by gene.

https://doi.org/10.1038/s41588-025-02400-1



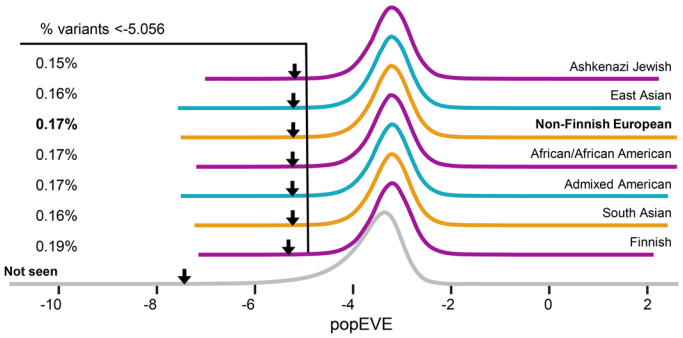

**Extended Data Fig. 2 | Correlation of computational variant effect predicting models with high-throughput experimental assays.** Assessing the performance of popEVE compared to popular supervised and unsupervised variant effect prediction models when compared to high-throughput functional assays on human genes (from ProteinGym), averaging a and across individual assays b. On average popEVE outperforms other models.

https://doi.org/10.1038/s41588-025-02400-1



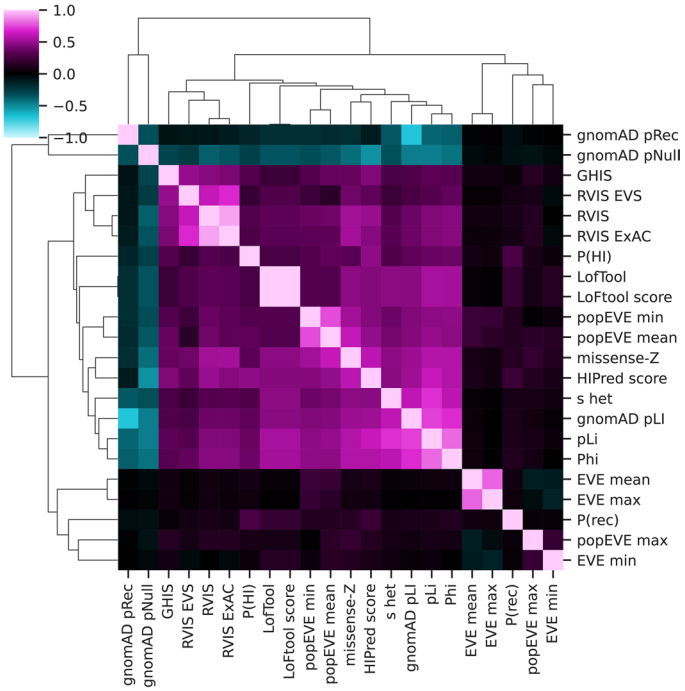

**Extended Data Fig. 3 | Correlation between EVE and Esm1v scores.** Scores for a random sample of 1 million variants across the human proteome. Pearson correlation is low - 0.55 (p-value=0.0).

https://doi.org/10.1038/s41588-025-02400-1



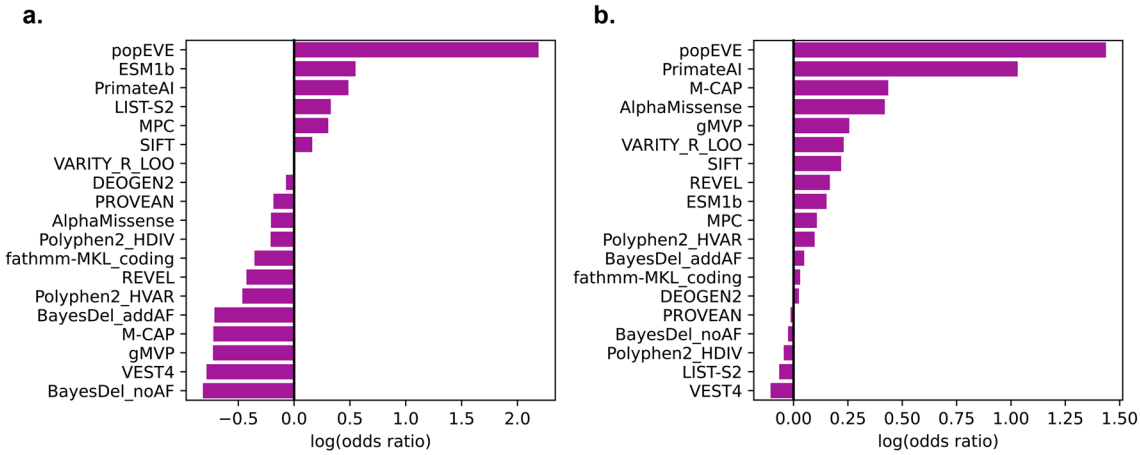

**Extended Data Fig. 4 | popEVE shows minimal population bias across diverse ancestries.** The distribution of popEVE scores for rare variants (AF<0.01) is consistent across populations found in gnomAD, indicating that despite using primarily non-Finnish European subjects for score adjustment there is no population bias. Variants not seen any gnomAD population are in grey. The 99.9% percentile for each distribution is marked with an arrow.

https://doi.org/10.1038/s41588-025-02400-1



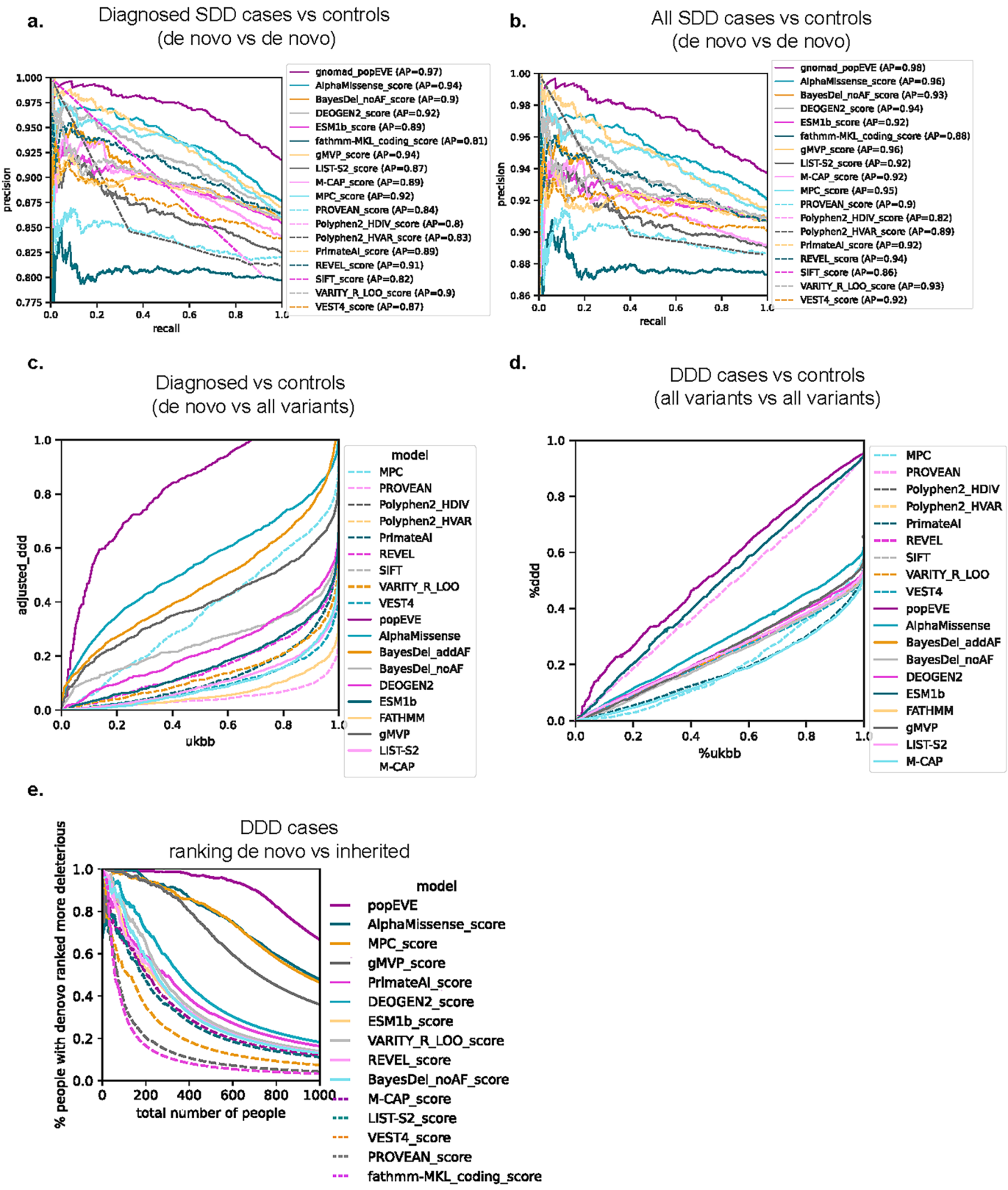

**Extended Data Fig. 5 | Correlation between popEVE gene-level statistics and gene-level constraint measures.** Pearson correlation between gene-level measures of constraint and EVE and popEVE minimum, maximum and mean score per gene. We find poor correlation between popEVE and gene-level constraint measures, except for MissenseZ and popEVE mean, with pearson = 0.61 (p-value = 0.0).

https://doi.org/10.1038/s41588-025-02400-1




**Extended Data Fig. 6 | Odds ratios of ClinVar pathogenic variants in genes associated with premature death and onset.** Odds ratios (threshold for each model set at 5$^{th}$ percentile of benign variants in ClinVar) for various models of ClinVar pathogenic variants (with at least 1 star curation rating) in phenotypes associated with (**a**) death and (**b**) onset in childhood versus adulthood.

https://doi.org/10.1038/s41588-025-02400-1

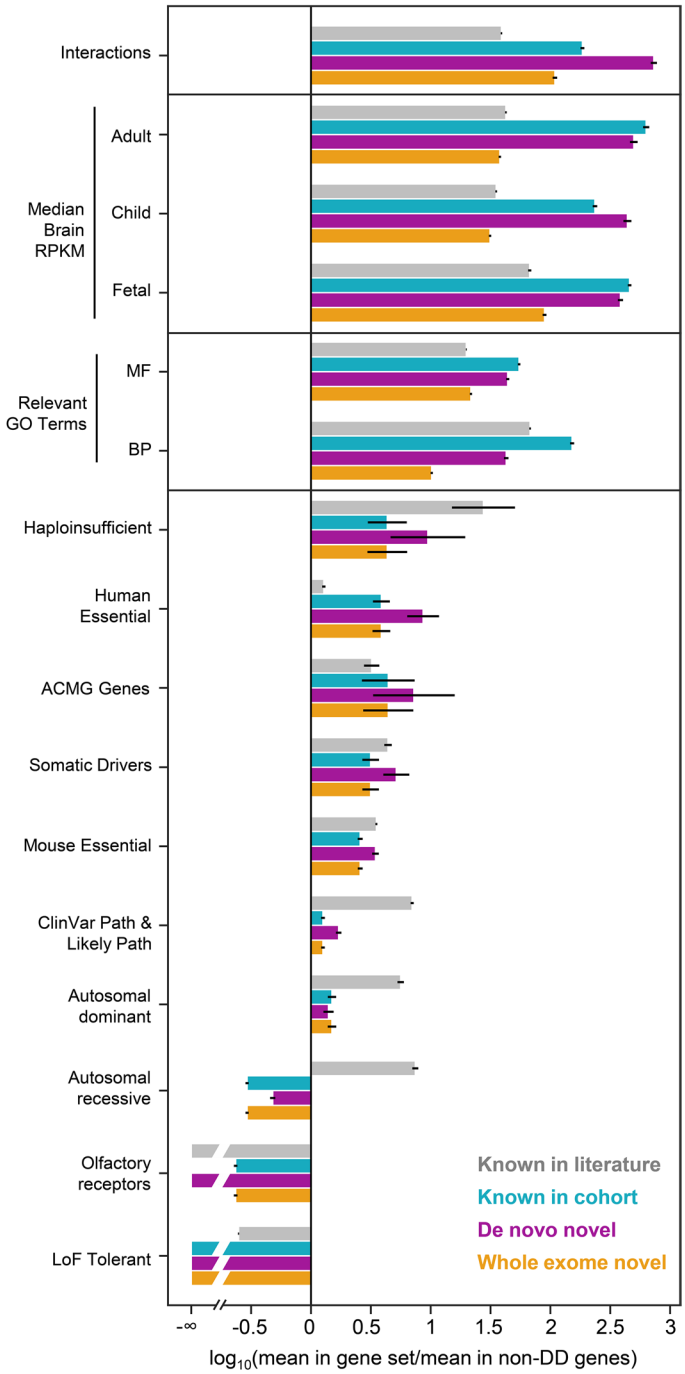



**Extended Data Fig. 7 | See next page for caption.**

https://doi.org/10.1038/s41588-025-02400-1

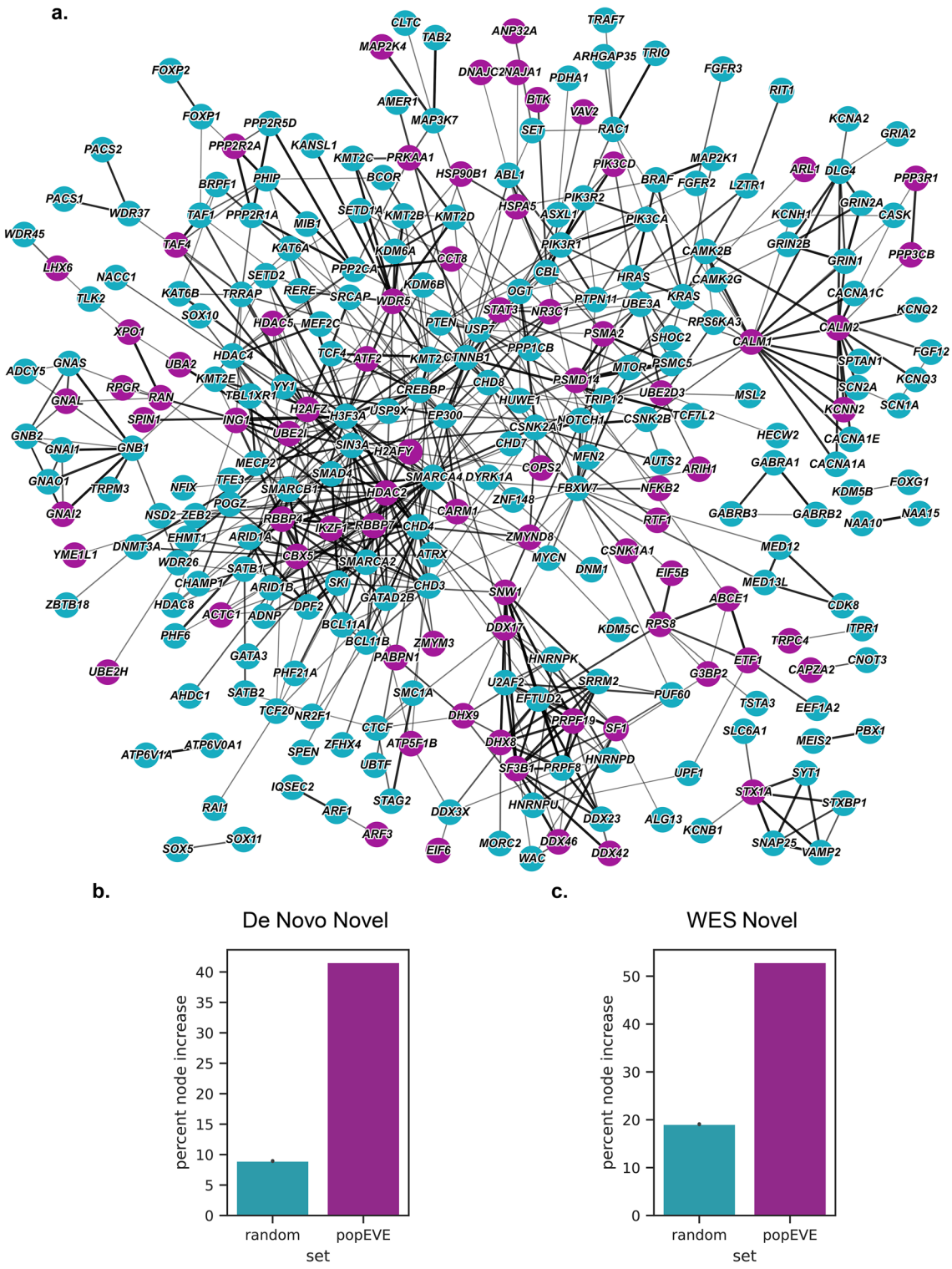
**Extended Data Fig. 7 | popEVE is better at separating developmental disorder cases from healthy controls than other state-of-the-art models. a**. Extension of Fig. 2c with all models - popEVE is better at separating "diagnosed" SDD cases whose disorder is likely to be caused by a de novo missense variant (cases with at least one missense variant in a known developmental disorder gene) from controls than other state of art variant effect predictors with an average precision of 97%. **b**. Precision recall for "high-confidence diagnosed" SDD cases (at least one de novo missense in a gene discovered by DeNovoWEST in the same cohort). **c**. Precision recall for all cases vs controls. **d**. Extension of Fig. 2d with all models - popEVE recalls more SDD cases (with at least one missense variant in DNW-discovered genes) without overpredicting pathogenicity in healthy controls from UKBB. While popEVE recalls 50% of these individuals for only 16% of the UKBB, the next best model, Alpha Missense, predicts 92% of UKBB has a variant as pathogenic as 50% of this SDD subset. **e**. Extension of Fig. 3b with all models - For each score threshold, we plot the percent of individuals with a de novo missense variant ranked as more deleterious than rare inherited variants. In individual cases, popEVE is better at ranking de novo mutations as more deleterious than rare inherited variants (MAF<0.01) than other models.

https://doi.org/10.1038/s41588-025-02400-1




**Extended Data Fig. 8 | Functional enrichment of known and novel genes.** Novel (from de novo SDD case variants and whole exome DDD variants) and known developmental disorder genes (from literature and previously discovered in the SDD cohort) show similar enrichment in properties known to differentiate known DD-genes from non-developmental disorder genes (95% CI from bootstrapping shown).

https://doi.org/10.1038/s41588-025-02400-1




**Extended Data Fig. 9 | Novel genes increase node connectivity of known developmental disorder genes. a**, Novel popEVE discovered genes are embedded into the network of previously-discovered disease associated genes from DDG2P and DeNovoWEST. Taking the set of 99.99 confidence threshold popEVE genes, we built a network using STRINGdb ('experiments' and 'coexpression' at a medium 0.4 score threshold). Colored nodes are novel discoveries and white nodes are known disease-associated genes. These nodes were clustered into four clusters using k-means clustering. **b**, When added to a network of know developmental disorder genes, novel genes from the full SDD meta-cohort had a 42% increase in node degree as compared to random sets of the same number of genes which saw an average of 9% (with $p < 0.000$, t-test). **c**, When added to a network of know developmental disorder genes, novel genes from the DDD sub-cohort had a 53% increase in node degree as compared to random sets of the same number of genes which saw an average of 19% (with $p < 0.000$, t-test).

nature portfolio

Corresponding author(s): Debora Marks, Jonathan Frazer, Mafalda Dias

Last updated by author(s): 09/02/2025

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Multiple sequence alignments were obtained using HMMER (3.1.b2). |
| Data analysis | Models and analysis were written in Python (3.7) and Pytorch (1.4). All code is available through GitHub (https://github.com/debbiemarkslab/popEVE) |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The main data analysed and generated in this study is available in Supplementary Information and at pop.evemodel.org. All other data is available from original references, public repositories or protected repositories described in the text.

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | n/a |
| Reporting on race, ethnicity, or other socially relevant groupings | n/a |
| Population characteristics | n/a |
| Recruitment | n/a |
| Ethics oversight | n/a |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | All experimental data and associated sample sizes were used as published in their respective works. |
| Data exclusions | No data was excluded from this analysis. |
| Replication | The majority of data was obtained from public repositories or published sources. Access to the variants from the UK Biobank and the DDD study must be applied for. |
| Randomization | Randomization is not relevant to the computational analysis of this study, since it is fully unsupervised. |
| Blinding | The same model building process was applied to all genes in this study and required no human interpretation and so no blinding was necessary. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

### Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Plants

| | |
|---|---|
| Seed stocks | *Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.* |
| Novel plant genotypes | *Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.* |
| Authentication | *Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.* |