nature communications

Article

https://doi.org/10.1038/s41467-025-66607-w

Expanding the utility of variant effect predictions with phenotype-specific models

Received: 19 February 2025

Accepted: 11 November 2025

Published online: 28 November 2025

Check for updates

David Stein[1,2,3], Meltem Ece Kars[4], Baptiste Milisavljevic[5], Matthew Mort[6], Peter D. Stenson[6], Jean-Laurent Casanova[5,7,8,9,10], David N. Cooper[6], Bertrand Boisson[5,7,8], Peng Zhang[5,7,8], Avner Schlessinger[2] ✉ & Yuval Itan[1,3,4,11,12] ✉

Current methods for variant effect prediction do not differentiate between pathogenic variants resulting in different disease outcomes and are restricted in application due to a focus on variants with a single molecular consequence. We have developed Variant-to-Phenotype (V2P), a multi-task, multi-output machine learning model to predict variant pathogenicity conditioned on top-level Human Phenotype Ontology disease phenotypes (n = 23) for single nucleotide variants and insertions/deletions throughout the human genome. V2P leverages a unique approach for the modeling of variant effect that incorporates resultant disease phenotypes as output and during training to improve the quality of variant disease phenotype and effect predictions, simultaneously. We describe the architecture, training strategy, and biological features contributing to V2P's output, revealing initial characteristics underlying the relationship between disease genotype and phenotype. Moreover, we demonstrate the benefit of incorporating disease phenotypes for variant effect predictions by comparing V2P with several variant effect predictors across various high-quality evaluation datasets from manually curated databases and functional assays. Finally, we examine how V2P's predictions result in the successful identification of pathogenic variants in real and simulated patient sequencing data, outperforming other tested methods in initial comparisons. V2P offers a complete mapping of human genetic variants to disease-phenotypes, offering a uniquely conditioned set of variant effect characterizations.

The increasing accessibility of high-throughput sequencing technologies has precipitated the proliferation of genetic data, including observed human sequence variants[1–4]. Although significant efforts have been expended to interpret this data, the substantial majority of variants remain uncharacterized[4]. To aid in the interpretation of these variants of uncertain significance, myriad computational approaches for the rapid, automated prediction of variant effect have been developed. Incremental improvements to these methods employing more sophisticated model architectures and improvements in data curation have allowed substantial progress to be made over the past decades[5–7].

Despite the continued generation of novel genetic data and advancements in detection methodology, several key limitations impede computational tools for variant assessment. Firstly, most methods cannot generate interpretations across variant types, i.e., for both coding and non-coding variants, or for both single nucleotide variants (SNVs) and insertions/deletions (indels)[8]. Secondly, most methods consider pathogenic variants as a homogeneous class, and

A full list of affiliations appears at the end of the paper. ✉e-mail: avner.schlessinger@mssm.edu; yuval.itan@mssm.edu

hence may systematically underperform for certain genes or for variants with particular molecular mechanisms or disease presentations[9–11]. Recent efforts have been made to investigate more granular characteristics of pathogenic variants, such as their mechanism of effect[12] or mode of inheritance[13]. However, many qualities of pathogenic variants, such as their specific phenotypic effects, remain to be addressed by computational methods.

Currently, most efforts to predict relationships between pathogenic genotypes and phenotypes have been carried out at the gene or protein level. For example, several methods exist for the prediction of phenotypes, as defined by the Human Phenotype Ontology[14] (HPO), associated with genes[15–22]. Similarly, a variety of methods have attempted to predict gene-disease relationships[23,24]. More recently, at the variant level, efforts to design pathogenicity prediction methods tailored to specific diseases or phenotypes have reported enhanced classification performance as compared to methods predicting variant pathogenicity in general[25–30]. However, these methods are limited to a single or a small number of disease(s)/phenotype(s). As a result, methods exploring the specific phenotypic effects—as opposed solely to binary pathogenicity—of variants with diverse molecular consequences genome-wide are needed.

Here, we present Variant-to-Phenotype (V2P), a multi-task, multi-output machine learning model to jointly predict variant pathogenicity and the broad phenotypic effect of SNVs and indels throughout the human genome, that exploits relationships between pathogenic variants and top-level disease phenotypes from the Human Phenotype Ontology (HPO) to enhance pathogenicity predictions. We describe V2P's architecture and training strategy, and thoroughly assess V2P's performance on a variety of high-quality datasets. We compare V2P to state-of-the-art variant effect prediction methods, illustrating the benefit of considering variant phenotypic outcomes during training. Moreover, we demonstrate V2P's utility for the identification of causal disease variants in real and simulated patient sequencing data. Finally, we identify the features of pathogenic variants underlying each human phenotype group, including features that occur only in specific human phenotypes and features that are common across various human phenotype groups. We have precomputed V2P scores for every possible SNV in the human hg38 reference genome and for all indels from gnomAD to provide a complete mapping of variants to disease phenotypes. We offer these along with a simple framework to generate V2P phenotype-specific prediction scores for user-specified variants at www.v2p.ai.

Results

V2P architecture and dataset

V2P is an ensemble of multi-label, machine-learning models developed for the classification of pathogenic variants and their resulting phenotypes. V2P accepts as input SNV or indel variants that have been annotated with a variety of gene-level features describing a gene's function in terms of associated diseases, pathways, and more, as well as protein-level features derived from protein sequence and structures, network features describing protein interactions, and variant-level features such as conservation (Fig. 1a and Supplementary Data 1). Contrary to methods solely estimating pathogenicity in general, V2P's output comprises 24 values, each ranging between zero and one, indicating a given variant's likelihood of being pathogenic or benign as well as the variant's likelihood to result in one or more of the 23 first-level disease phenotypes from the phenotypic abnormality sub-ontology of the HPO[14]. These 23 classes, which include designations such as abnormalities of the nervous system and neoplasms, encompass the extent of inherited human disease phenotypes (Fig. 1b).

V2P was developed using 252,125 pathogenic variants from the Human Gene Mutation Database[3] (HGMD) and 244,231 putatively benign variants from gnomAD v2.1[31] exome sequences—spanning 6,620 genes. Of the pathogenic variants, 202,514 were associated with HPO phenotypes. Notably, the distribution of phenotypes resulting from these variants is skewed with substantially greater numbers of variants resulting in the most abundant phenotypes, such as abnormalities of the nervous system or the musculoskeletal system, than the least represented phenotypes such as abnormalities of the thoracic cavity (Fig. 1b). Similarly, the phenotypic classes comprise varying proportions of variants that result in one or more phenotypes, likely due in part to the structure of the ontology. For example, variants resulting in abnormalities of the nervous system were most often unique to this phenotype, whereas variants resulting in neoplasms most often also resulted in additional phenotypes (Fig. 1b).

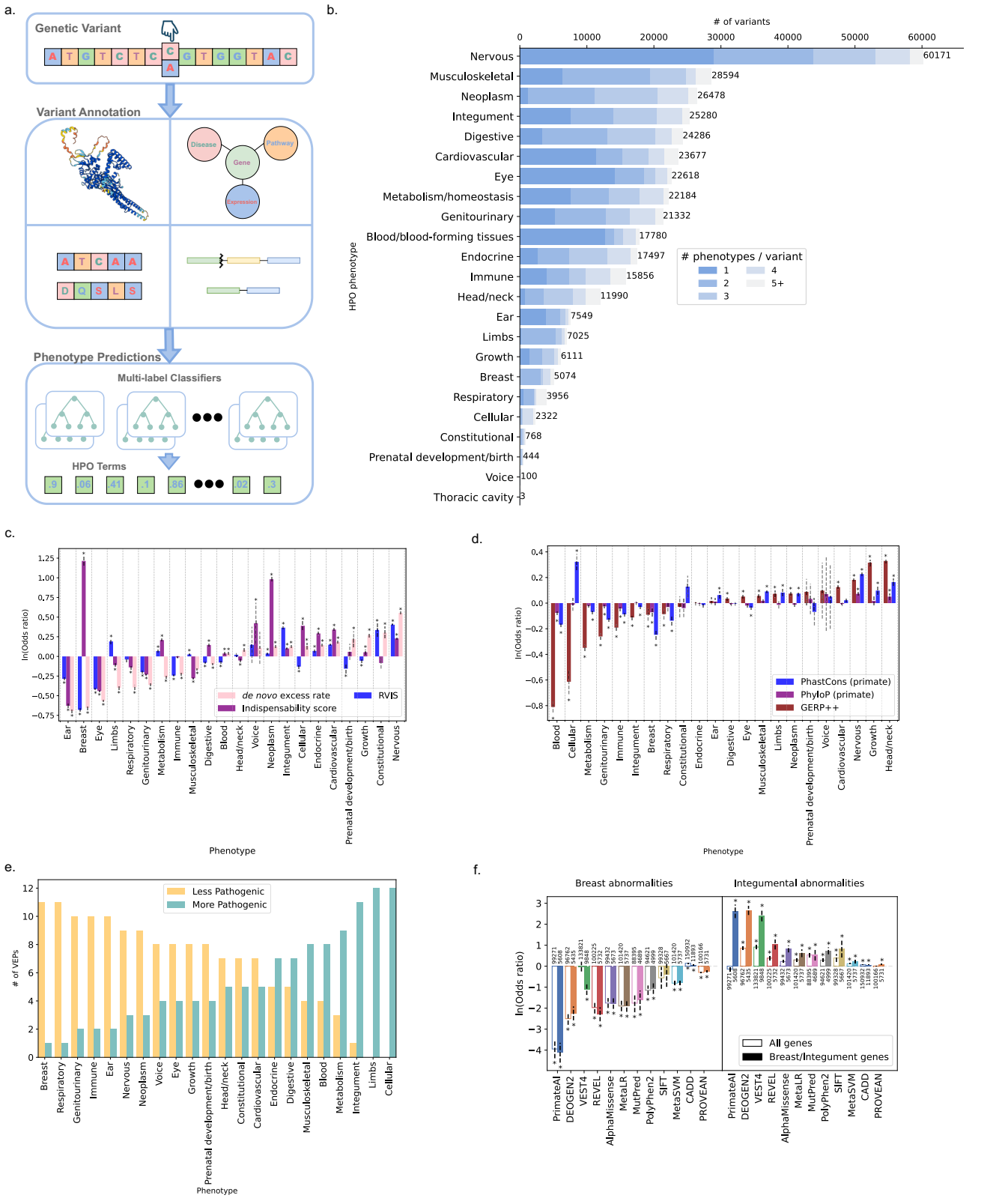
Many variant effect predictors employ genic and positional conservation as a primary component of their estimation of variant effect. Thus, we investigated the tendency of variants resulting in particular pathogenic phenotypes to occur in essential genes, i.e., those intolerant to loss of function, and at variably conserved sites. We observed a distinct tendency between variants resulting in different phenotypes to impact genes of varying essentiality as indicated by the Residual Variation Intolerance Score (RVIS)[32], the de novo excess rate[33], and the indispensability score[34] (Fig. 1c). These three measurements estimate gene essentiality via differing approaches and thus may not always agree. Nevertheless, for abnormalities of the ear, eye, and genitourinary system, we noted that each essentiality measure examined indicated a significant tendency of variants resulting in the phenotype to be less likely to inhabit essential genes (Fig. 1c). Conversely, variants resulting in neoplasms, abnormalities of the integumentary, endocrine, cardiovascular, and nervous systems, exhibited a greater tendency to inhabit essential genes according to each of the included measures (Fig. 1c). Similarly, when examining measures of conservation, we observed variability between phenotypes (Fig. 1d).

We explored predictions from several variant effect predictors for variants stratified by resultant phenotypes, noting that variants resulting in certain phenotypes were consistently associated with lower (e.g. breast, respiratory) or higher (e.g. cellular, limbs) estimates of pathogenicity compared to variants from other phenotype groups, as measured by the continuous scores output by the examined methods (Fig. 1e). To investigate if these observed differences were driven solely by gene-level effects, we selected variants from genes associated with both abnormalities of the breast and of the integument, phenotypes associated with lower and higher confidence predictions of pathogenicity, respectively (Fig. 1f). Again, we found that abnormality of the breast variants were generally associated with lower predicted pathogenicity whereas abnormality of the integument variants were associated with higher predicted pathogenicity across examined methods (Fig. 1f), suggesting these methods may produce variable predictions between pathogenic phenotypes even within genes.

Identifying biological features associated with human disease groups

To further investigate mechanisms of pathogenicity and the relationship between features and phenotypes, we employed the Boruta[35] all-relevant feature selection algorithm. On average, 282 features were selected for the task of distinguishing between variants resulting in each respective disease phenotype and the other pathogenic variants, indicating a non-random association between each respective feature-phenotype pair (Supplementary Data 2). These features pertained to a variety of biological qualities, including gene-disease associations[36–38], tissue-specific expression and transcription factor co-expression[39], phenotypes associated with homologous mouse genes[40], characterizations of gene conservation and function, protein interaction[41], characteristics of variants and their impacted nucleotides and amino acids, pathways[42–44], and epigenetics[45] (Supplementary Data 1). Features commonly identified as discriminant across phenotypes included those pertaining to gene function, variant impact on protein

https://doi.org/10.1038/s41467-025-66607-w



structure and function, variant conservation and location, and more. For instance, several features estimating gene essentiality were selected for many phenotypes. Similarly, a variety of protein structural features were frequently selected, such as those indicating an effect on binding residues, disordered regions, and buried residues. Features associated with the presence of variants in regulatory or epigenetically modified regions, proximity to known common or rare variation, and location within cDNA and coding regions were also flagged as important for many phenotypes. Other features were selected as discriminant for a single phenotype only. The majority of these features were in regard to the association between genes and disease, tissue-specific gene expression and co-expression of genes and transcription factors, the involvement of genes in pathways or biological processes, and the subcellular localization of protein products, as well as the

Fig. 1 | Examining conservation, essentiality, and variant effect prediction stratifying variants by phenotype. a V2P workflow. A single nucleotide variant or indel is annotated with gene-, protein-, and variant-related features and input to a multi-task, multi-output ensemble of gradient-boosted decision trees. This method outputs 24 scores ranging between 0 and 1 corresponding to the predicted probability that the input variant results in each of the 23 top-level Human Phenotype Ontology (HPO)[14] phenotypic abnormality sub-ontology terms, respectively, and that the variant is pathogenic/benign. **b** The number of phenotype-labeled samples for each considered HPO phenotype in V2P's development dataset. Distinctly colored portions of each bar indicate the proportion of variants labeled with a given phenotype that result in different numbers of phenotypes, where the darkest blue portion indicates single-phenotype variants and the lightest blue portion indicates variants resulting in five or more phenotypes. **c** The log odds ratios indicating relationships between phenotypes, compared to each other, and three measures of gene essentiality, namely the de novo excess rate[33] (n = 212,545), the RVIS[32] (n = 210,361), and the indispensability score[34] (n = 220,521). (RVIS scores inverted). **d** The log odds ratios indicating relationships between phenotypes, compared to each other, and three measures of sequence conservation, PhastCons[75] (Primate) (n = 221,651), PhyloP[75] (Primate) (n = 221,651), and GERP++[76] (n = 221,651). **e** The number of variant effect predictors (PrimateAI[77], DEOGEN2[78], VEST4[79], REVEL[80], AlphaMissense[6], MetaLR[81], MetaSVM[81], MutPred[82], PolyPhen2[83], SIFT[84], CADD v1.7[7], and PROVEAN[85]) for which variants were associated with greater or lesser probability of pathogenicity for each phenotype. **f** The log odds ratios for tools indicating the association between variants resulting in abnormalities of the breast and abnormalities of the integument from genes harboring variants resulting in both phenotypes (solid bars), and for all variants resulting in either of the phenotypes. Dashed lines indicate confidence intervals of the log odds ratio around the estimated regression coefficient, asterisks indicate significance determined via a two-sided Wald test after Bonferroni correction for multiple hypothesis testing. Variant effect predictor versions are detailed in the dbNSFP[65] 4.7a.

association of orthologous mouse genes with various phenotypes and descriptions of surrounding chromatin states. While none of the included features is directly correspondent to the phenotypes predicted by V2P, these results suggest they may provide a signal for this task.

For example, variants resulting in abnormalities of the musculoskeletal system were enriched in genes associated with a variety of disorders including arthrogryposis and myopathy, among other features (Fig. 2a). Conversely, these variants were depleted in genes with up-regulated expression in non-musculoskeletal tissues (Fig. 2a). Similarly, variants resulting in abnormalities of the integument were enriched for a variety of skin tumors and other dermatological disorders as well as higher expression in both sun-exposed and non-sun-exposed skin and were negatively associated with up-regulation in several other tissue types (Fig. 2b). These findings emphasize the need to consider a wide range of sources of biological information when attempting inference of variant function with a view to elucidating the biological qualities of variants resulting in distinct phenotypes.

Across examined features, we noted a moderate correlation between certain phenotypes as indicated by the features selected during feature selection (Fig. 2c). In particular, abnormalities of the musculoskeletal system and abnormalities of the limbs were most highly associated; indeed, 30% of their selected features were shared and had the same direction of effect (Fig. 2c). This correlation is likely driven to an extent by the definition of the HPO in which certain phenotypes may be more likely to co-occur. Indeed, in several cases, the correlation between phenotypes based on selected features was analogous to the pairwise correlation between phenotypes according to the proportion of variants resulting in both phenotypes in a pair. For instance, abnormalities of the digestive system and neoplasms, as well as abnormalities of the musculoskeletal system and abnormalities of the limbs, were among the most correlated phenotypes as measured by variant co-occurrence (Fig. 2d). Nonetheless, correlation between phenotypes according to shared variants was not observed to guarantee equally strong correlation of selected features, potentially indicating that the included features capture signal unique to variants resulting in the investigated phenotypes.

Using the LightGBM gradient-boosted decision tree implementation as a proxy for V2P, we calculated Shapley values, treating pathogenic variants resulting in a given phenotype as the positive class and benign variants as the negative class, for each phenotype separately (Supplementary Fig. 1). Thus, the identified important features reflect those most important for the model to separate pathogenic variants of the phenotype and benign variants. For all pathogenic variants vs. benign variants, the top twenty features include relative placement in the coding/protein sequence, predicted impact on protein stability and other protein structural qualities, amino acid substitution matrix scores, and measurements of conservation. While some of the sources of this information may be unique to V2P, the content—e.g., conservation, amino acid substitution impact—is common to many variant effect prediction methods. We do not observe features pertaining to particular phenotypes in the top 20, as we might expect, since this model must distinguish between pathogenic variants of all types and benign variants.

Conversely, as we investigate the top features for particular phenotypes, we note that phenotype-specific features have a substantial impact on the model. For example, while the top features for separating pathogenic nervous system variants and benign variants contain some of the same features as those identified for general pathogenicity, such as placement in the coding/protein sequence and conservation, we also observe several phenotype specific features such as gene expression in different brain tissues, gene association with nervous system diseases, and genes associated with nervous system phenotypes in mice. Further, we observe that features describing the protein-protein interactome have a substantial impact, suggesting that protein proximity, as expressed by direct or indirect interaction, may have bearing on pathogenic variant phenotype. Curiously, we also observe that gene association with cardiovascular phenotypes in mice tends to decrease the model's confidence that a variant causes a pathogenic nervous system phenotype. Concordantly, nervous system and cardiovascular abnormalities are one of the least commonly co-occurring phenotype pairs for variants in our dataset. We note the inverse for cardiovascular phenotypes in which gene association with intellectual disability decreases model confidence.

Similar trends are noted for most phenotypes—in particular, several features relating to pathogenicity in general will be important, along with features relating to expression in phenotype-related tissues, association with phenotype-related diseases, and association with different portions of the protein interactome.

V2P's phenotype-conditioned output effectively estimates variant effect

V2P's multi-task architecture produces scores estimating whether a given variant is pathogenic or benign, alongside the phenotypes, if any, resulting from that variant. To assess the quality of V2P's predictions for these complementary tasks, we assembled three distinct evaluation datasets comprising a total of 391,856 variants spanning 4135 genes. Specifically, we assessed V2P on 49,106 pathogenic and benign variants that were held out for testing from the original dataset assembled for the development of V2P, 37,767 pathogenic and benign variants from a more recent release (2023.4) of the HGMD database and from gnomAD, and 304,893 pathogenic and benign variants from ClinVar that were not present in V2P's development dataset. Of the pathogenic variants, 20,116 from the held-out dataset, 17,503 from the HGMD 2023.4 dataset, and 62,312 from ClinVar were assigned phenotypic effects, respectively (Fig. 3a).

https://doi.org/10.1038/s41467-025-66607-w

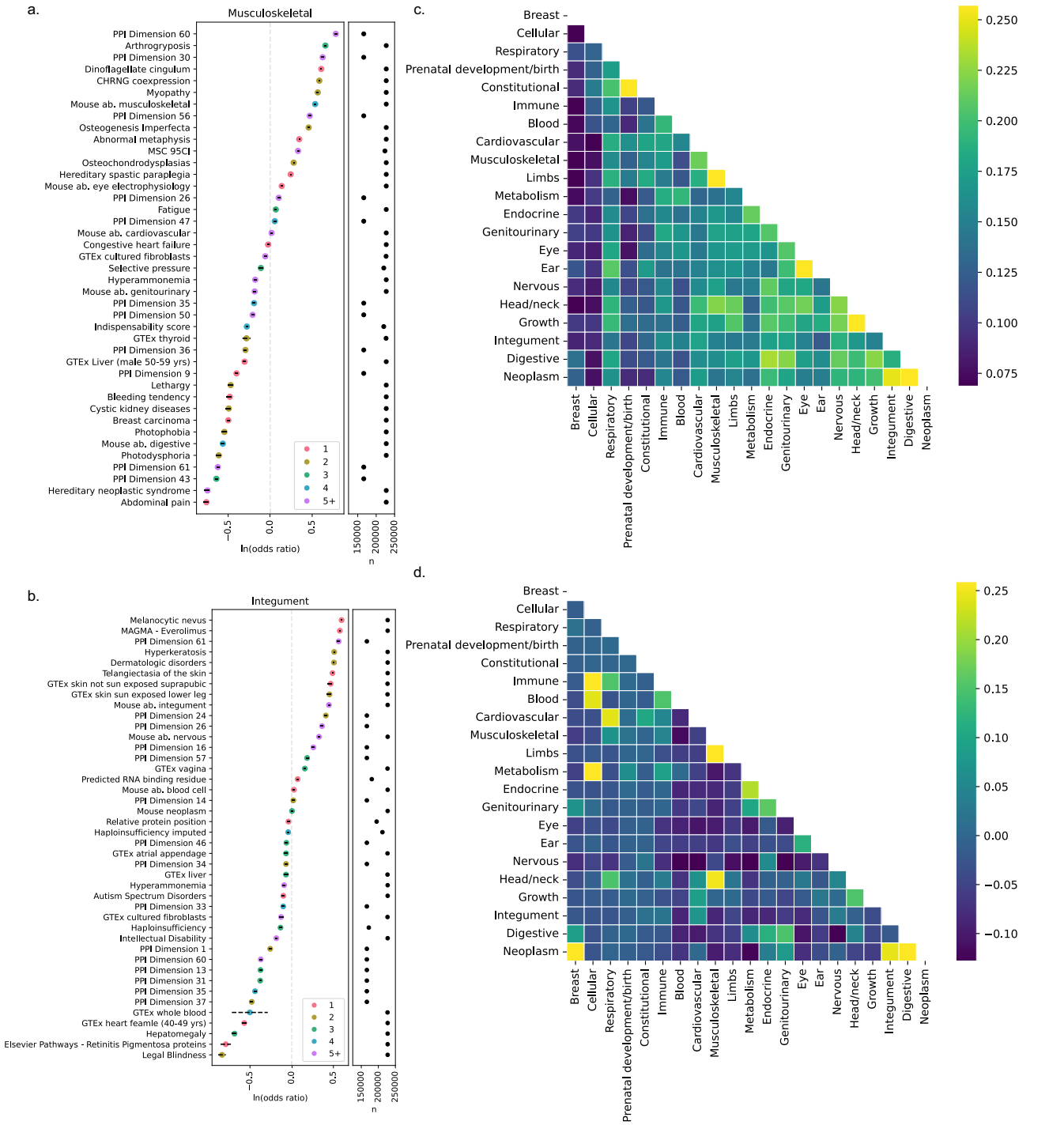


Fig. 2 | Associations between phenotypes and biological features. a Log odds ratios for Boruta[74] selected features for abnormality of the musculoskeletal variants with magnitudes of effect in the 90th percentile or above as measured by Shapley values. **b** Log odds ratios for Boruta selected features for abnormality of the integument system variants with magnitudes of effect in the 90th percentile or above. **c** Pairwise correlation, expressed as the number of shared selected features with concordant direction of effect over the total number of selected features, between phenotypes. **d** Pairwise correlations between disease phenotypes as indicated by the number of variants resulting in both phenotypes simultaneously. Dashed lines indicate confidence intervals of the log odds ratio around the estimated regression coefficient.

V2P consistently demonstrated effective performance for phenotype prediction across the evaluation datasets. In particular, the method improved over the random baseline, often substantially so, for 21 out of the 23 top-level HPO phenotypes in the held-out testing dataset (Fig. 3b). As expected, for the two least represented phenotypes in our set, abnormalities of the voice and thoracic cavity, which account for only 100 and 3 of the labeled samples, respectively, V2P was unable to provide substantial improvement (Figs. 1b, 3b). For the remaining phenotypes, V2P's performance was not solely dependent on the number of training samples. For instance, V2P achieved sizable performance gain for less represented classes such as abnormalities of the ear and cellular phenotypes (Fig. 3b). Furthermore, V2P's performance was consistently strong across evaluation datasets, achieving micro-averaged Average Precision (AP) scores of 0.53, 0.79, and 0.67

https://doi.org/10.1038/s41467-025-66607-w

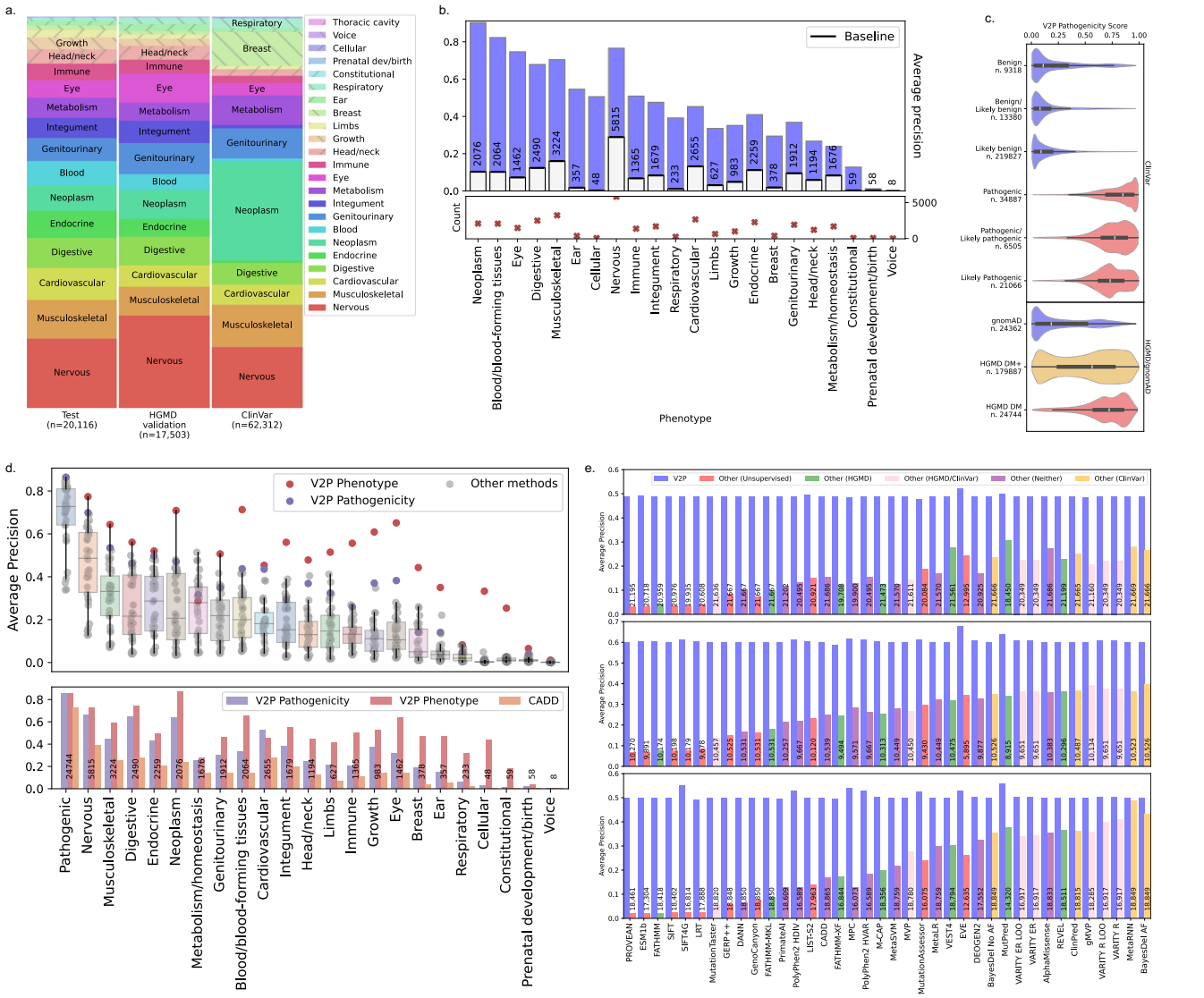


Fig. 3 | Assessment and comparison of V2P performance on clinical variation. **a** Relative proportion of pathogenic variants belonging to each phenotype in each dataset. **b** (Top) V2P's one-vs.-rest Average Precision (AP) on phenotype-labeled variants in the held-out testing dataset for each of the 23 top-level phenotypic abnormality Human Phenotype Ontology (HPO)[14] phenotypes. Horizontal black lines bisecting each bar indicate the baseline performance for variants resulting in each respective phenotype. (bottom) Number of labeled variants resulting in the denoted phenotype in the held-out testing dataset. **c** (Top) Distributions of V2P pathogenicity scores across ClinVar clinical significance labels. (Bottom) Distributions of V2P pathogenicity scores for gnomAD (benign), HGMD DM+ (low-confidence pathogenic), and HGMD DM (high-confidence pathogenic) variants. **d** (Top) Box-and-whisker plots indicating the distribution of variant effect predictor AP's for separating pathogenic missense variants resulting in each phenotype from benign missense variants in the held-out testing dataset. Existing methods' AP's denoted by gray circles, V2P pathogenicity score AP denoted by blue dots, and V2P phenotype-specific score denoted by red dots. (LRT[86], MutationTaster[87], MutationAssessor[72], MetaSVM[81], MetaLR[81], MetaRNN[88], M-CAP[89], MVP[90], gMVP[91], MPC[92], PrimateAI[77], BayesDel[93], ClinPred[94], VARITY[95], ESM1b[96], DANN[97], GenoCanyon[98], GERP++_RS[76], DEOGEN2[78], FATHMM[99], PROVEAN[85], PolyPhen2[83], REVEL[80], SIFT[84], VEST4[79], AlphaMissense[6], CADD[7]. Low coverage methods excluded). (n. left-to-right−9570, 2649, 1386, 750, 850, 640, 845, 786, 883, 801, 459, 393, 325, 603, 291, 377, 173, 104, 53, 11, 42, 24, 6). (bottom) AP's achieved by V2P's pathogenicity score (blue), V2P's phenotype-specific scores (red), and CADD[46] v1.7(orange) for the separation of variants resulting in each phenotype and benign variants from the held-out testing dataset. **e** Weighted average of APs achieved by each variant effect predictor for separating pathogenic missense and benign missense from the held-out testing (top), HGMD 2023.4 (middle), and ClinVar (bottom) datasets, respectively. Numerical values on bars indicate the number of variants for which the compared variant effect predictor had a prediction in the dbNSFP. Variant effect predictor versions are detailed in the dbNSFP 4.7a. Boxes represent the quartiles of the data. Whiskers extend to points that lie within 1.5 times the interquartile range of the lower and upper quartiles. Violins extend to and are clipped at data minima and maxima.

for separating variants resulting in each phenotypic class from the others in the held-out testing, HGMD 2023.4, and ClinVar datasets, respectively (Supplementary Figs. 2a, 3 and 4).

For each of the three evaluation datasets, the distribution of V2P phenotype predictions in genes with variants resulting in a given phenotype was compared to those without. This analysis was performed separately for genes associated with a single phenotype and those associated with multiple phenotypes. In the held-out test dataset, for 11 of 12 phenotypes in genes resulting in a single phenotype, the relevant V2P phenotype-score was significantly higher (Mann–Whitney *U*, one-sided, Bonferroni-corrected). In all genes, the relevant V2P phenotype-score was significantly higher for 19 out of 22 phenotypes. In the HGMD 2023.4 dataset, in single-phenotype genes, the V2P score was significantly higher for 17 out of 17 phenotypes. In all genes, it was significantly higher for 20 of 22 phenotypes. Finally, for the ClinVar dataset, in single-phenotype genes, it was significantly

higher for all 16 phenotypes, and in all genes, it was significantly higher for all 20 phenotypes (Supplementary Data 3).

Similarly, V2P achieved notable success in its discrimination between pathogenic and benign variants in general. Across the three evaluation datasets, V2P's pathogenicity score resulted in considerable improvement over the respective random baselines. In particular, V2P achieved APs of 0.86, 0.93, and 0.94 for separating pathogenic and benign variants in the held-out, HGMD 2023.4, and ClinVar datasets, respectively (Supplementary Fig. 2b). Likewise, the distribution of V2P's pathogenicity scores demonstrated substantial divergence between pathogenic and benign variants in both ClinVar and the HGMD (Fig. 3c and Supplementary Figs. 2–4). Further, V2P successfully identified de novo variants in cases vs. controls (Supplementary Fig. 5).

To assess V2P's phenotype-specific strategy in comparison with methods considering pathogenic variants homogeneously, pathogenic variants in the held-out testing dataset were split into distinct groups according to their resultant phenotype(s). For each of the predicted phenotypes, each compared method was assessed for its ability to separate variants resulting in the given phenotype from the benign variants in the held-out testing dataset, as measured by the AP. Remarkably, for missense variants, for 21 out of the 22 phenotypes included, V2P's phenotype-specific predictions surpassed not only its own general pathogenicity predictions, but also all of the compared methods, for the examined metrics (Fig. 3d). For variants of any molecular consequence, V2P's phenotype predictions also outperformed CADD[46]—a leading method able to predict variants beyond coding regions—across all phenotypes except abnormalities of the thoracic cavity for which no assessment variants were available (Fig. 3d). For missense variants in each of the three evaluation datasets, averaged across phenotypes, V2P's phenotype-specific scores resulted in the highest AP scores (Fig. 3e). Similarly, for missense variants from ClinVar submissions last reviewed after V2P's training and in genes with ten or fewer pathogenic/likely pathogenic variants, V2P's phenotype-specific scores resulted in the highest AP scores (Supplementary Fig. 6).

V2P is highly concordant with functional evidence of variant effect

Deep mutational scanning (DMS) is a technology that systematically maps genetic variations to phenotypic variations by measuring protein fitness upon mutation, providing an efficient and affordable alternative to in vivo studies. To compare V2P and other variant effect predictors in their alignment with DMS outputs, we examined 66 Deep Mutational Scanning (DMS) assays of 52 proteins. For data from each assay, the correlation of the output of each method with the assay's continuous output was assessed along with each method's AP score for predicting the binarized assay outputs, where the assessed mutations were either functional or non-functional (Fig. 4a and Supplementary Figs. 7, 8). Notably, V2P's pathogenicity predictions were comparable to the top-performing methods—as measured by the AP and the Spearman correlation coefficient (pairwise Mann–Whitney U, all $p > 0.05$) (Fig. 4a and Supplementary Figs. 7, 8).

For example, V2P's pathogenicity score achieved a Spearman's rank correlation coefficient (ρ) of 0.59 with DMS output for the PRKN protein—a component of an E3 ubiquitin ligase complex that is known to cause Parkinson's disease. Notably, despite this high concordance, V2P surpassed the DMS assay for classification of PRKN variants in ClinVar, achieving an AP score of 0.87 compared to an AP score of 0.84. Examining the distribution of the predictions, V2P appears to be sensitive to structural context, correctly classifying five missense variants (K211N, T240R, M434T, G430D, A46T) in functional domains of PRKN that were experimentally misclassified as per ClinVar (Fig. 4b). Conversely, ClinVar and the experimental data were contrary to V2P's predictions in only two cases (R42C, G359D) (Fig. 4b). Interestingly, for two pathogenic/likely pathogenic variants (R33Q, T415N) and one benign/likely benign variant (R366W) from ClinVar, both V2P and the DMS assay agreed on the inverse classification (Fig. 4b).

In other cases, both the DMS assay results and the V2P scores aligned strongly with ClinVar. For example, in P53, for which V2P's score correlated strongly with the DMS output ($\rho = 0.67$), V2P and DMS yielded AP scores of 0.992 and 0.979, respectively, according to ClinVar labels (Fig. 4c, d). Again, V2P proved sensitive to the structural and functional context of the protein, yielding significantly higher pathogenicity scores ($p < 10^{-5}$) for variants in the DNA-binding domain between residues 102–292—a region with a high density of pathogenic ClinVar variants—than variants in residues outside this region (Fig. 4c, e). Similarly, variants occurring in the TADI, TADII, Bipartite nuclear localization signal, Nuclear export signal, [KR]-[STA]-K motifs had significantly higher pathogenicity scores ($p < 10^{-5}$) than variants not located in a motif or in the DNA-binding domain (Fig. 4c, e).

We further assessed the quality of V2P's predictions for non-coding variants by comparing V2P's output with that from 16 massively parallel reporter assays (MPRA) of distinct regulatory elements. While none of the variant effect prediction methods reached the performance of Enformer[47], a specialized method for predicting non-coding variant effect on expression, we found that, compared to CADD and FATHMM, for 7 out of the 16 assays, V2P achieved the highest correlation with the experimental data. CADD also had the highest correlation for 7 assays, and FATHMM[48] had the highest for 2 assays (Fig. 4f). Across MPRAs, V2P had the highest median correlation of $\rho = 0.34$, followed by CADD with $\rho = 0.29$, and FATHMM with $\rho = 0.19$ (Fig. 4f).

V2P aids in causal variant identification in patient exomes

Whilst most variant effect predictors have focused solely on the pathogenicity of missense variants, V2P provides predictions of disease phenotypes for both SNVs and indels in both coding and non-coding regions, genome-wide. Thus, it is well-suited for the analysis of patient exome or genome sequencing data containing variants of diverse effects, leveraging the patients' pathology and the corresponding V2P phenotype score(s). To investigate V2P's utility for the prioritization of 73 causal disease variants (Supplementary Data 4) in patient sequencing data, variants from 116 exomes of patients from the Human Genetics of Disease laboratory at Rockefeller University with rare immune disorders were ranked by the V2P phenotype-specific score(s) associated with each patient's disease, CADD, and Capice[49]—another method capable of scoring variants genome-wide. The positions of the known pathogenic variants for each patient were compared. Remarkably, the median causal variant ranking was two when using V2P compared to 5.5 when using CADD and 10 when using Capice (Fig. 5a, b). We observed the distribution of the ranks of causal variants to be significantly lower when using V2P than when using CADD ($p = 0.004$) or Capice ($p < 10^{-5}$) (Fig. 5a, b).

To further assess the effectiveness of V2P in prioritizing causal variants with a broader range of phenotypic effects, we selected, at random, 100 patient exomes sequenced as part of the diverse cohort curated for Mount Sinai's BioMe BioBank. Pathogenic variants from each of the three evaluation datasets—held-out, HGMD 2023.4, and ClinVar—were introduced one at a time into each exome. Again, the variants from each patient's exome, including the introduced pathogenic variant (spiked-in), were ranked according to corresponding V2P phenotype score(s), CADD, and Capice. The rankings of each spiked-in variant were averaged across the 100 exomes. Overall, the average rankings of spiked-in variants from each dataset, as ranked by V2P's phenotype-specific scores, were substantially lower than the rankings of spiked-in variants as ranked by CADD ($p < 10^{-5}$) and Capice ($p < 10^{-5}$) (Fig. 5c and Supplementary Fig. 9). Notably, 49%, 70%, 80%, 87%, and 95% of pathogenic variants were ranked within the top 1, 5, 10, 20, and 100 by V2P, respectively, compared with 22%, 49%, 62%, 75%, and 95% for CADD and 18%, 48%, 68%, 80%, and 95% for Capice (Fig. 5d). Examining variants resulting in a single phenotype, we found that

https://doi.org/10.1038/s41467-025-66607-w

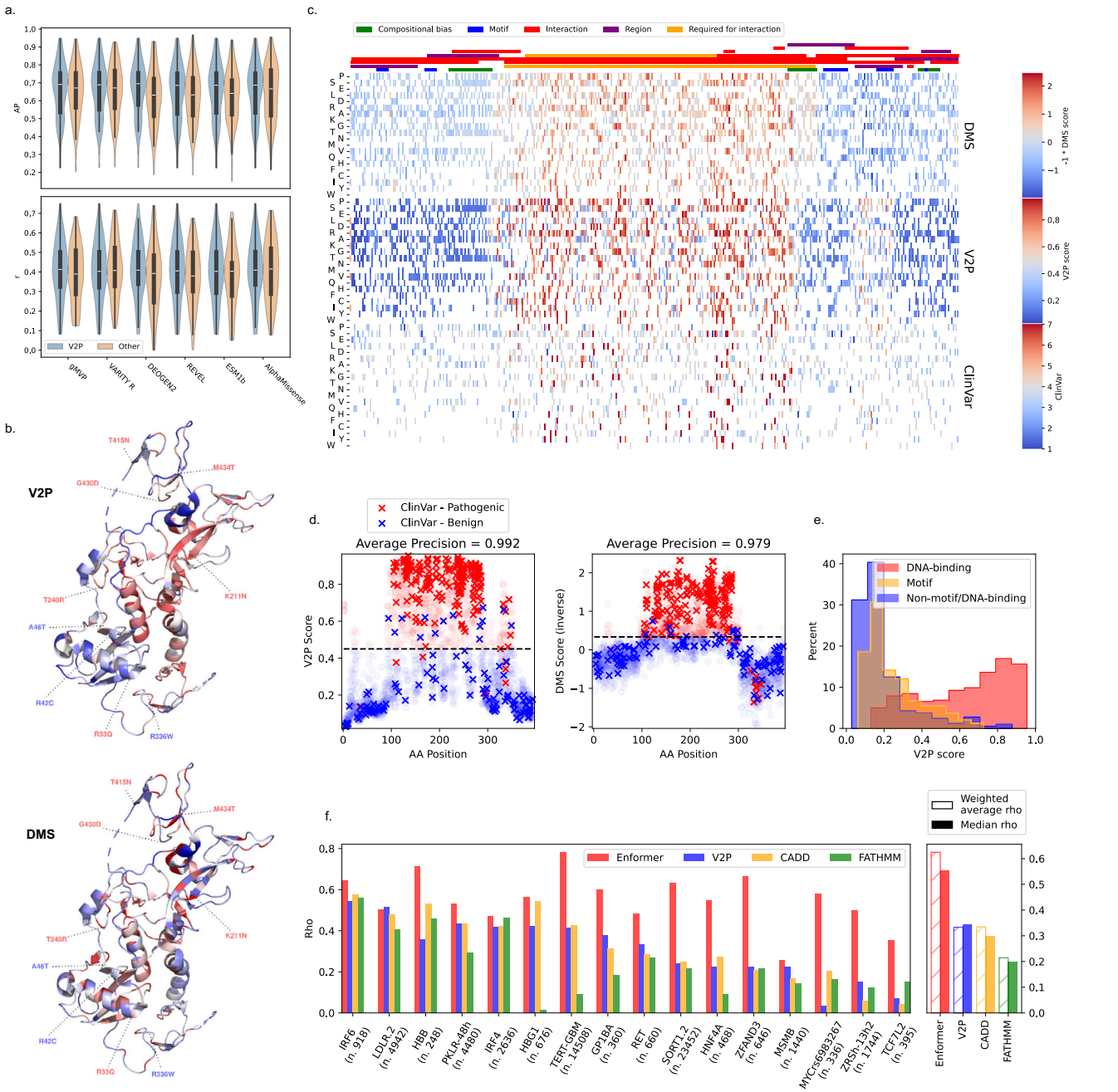


Fig. 4 | Assessment and comparison of V2P performance on functional characterization of variant effect. a Comparison of V2P's pathogenicity predictions with six previously published methods on variants from 66 deep mutational scanning (DMS) assays of 52 proteins, for which each pair of methods provided predictions. (Top) Distribution of average precision scores per assay. (Bottom) Spearman's rank correlation coefficient (ρ) per assay. **b** V2P (top) and inverse DMS output (bottom) averaged at each amino acid for single-nucleotide variant (SNV) missense variants in the PRKN protein (PDB 5C1Z). Highlighted variants colored according to ClinVar[4] classification: pathogenic (red) and benign (blue). **c** (Top) P53 families and domains. (Bottom) Inverse of DMS output, V2P pathogenicity scores, and ClinVar labels, respectively, for SNV missense variants in P53. ClinVar label key. 1: Benign, 2: Benign/Likely benign, 3: Likely Benign, 4: Uncertain significance, 5: Likely pathogenic, 6: Pathogenic/Likely pathogenic, 7: Pathogenic. **d** V2P and DMS scores for pathogenic and benign P53 variants from ClinVar. **e** Distribution of V2P scores in regions of P53. **f** Comparison of V2P's pathogenicity predictions with Enformer, CADD v1.7, and FATHMM on 16 massively parallel reporter assays (MRPA) of distinct regulatory elements. (left) ρ for each MPRA. (right) Median ρ and average ρ weighted by the number of variants per assay across MPRAs. Variant effect predictor versions are detailed in the dbNSFP 4.7a. Boxes represent the quartiles of the data. Whiskers extend to points that lie within 1.5 times the interquartile range of the lower and upper quartiles. Violins extend to and are clipped at data minima and maxima.

median V2P rankings were lower for 13 out of 18 phenotypes in the held-out dataset, 16 out of 18 phenotypes in the HGMD 2023.4 dataset, 12 out of 17 in the ClinVar dataset (Supplementary Fig. 9), and 15 out of 19 across the datasets (Fig. 5e). Considering all variants, including those resulting in multiple phenotypes, V2P's median ranks of causal variants were lower for 267 out of 501 phenotypes in the held-out dataset (CADD = 159, Capice = 75), 237 out of 349 phenotypes in the HGMD 2023.4 dataset (CADD = 60, Capice = 52), 58 out of 96 in the ClinVar dataset (CADD = 24, Capice = 14), and 411 out 690 across datasets (CADD = 180, Capice = 99). Thus, V2P's phenotype-specific

https://doi.org/10.1038/s41467-025-66607-w

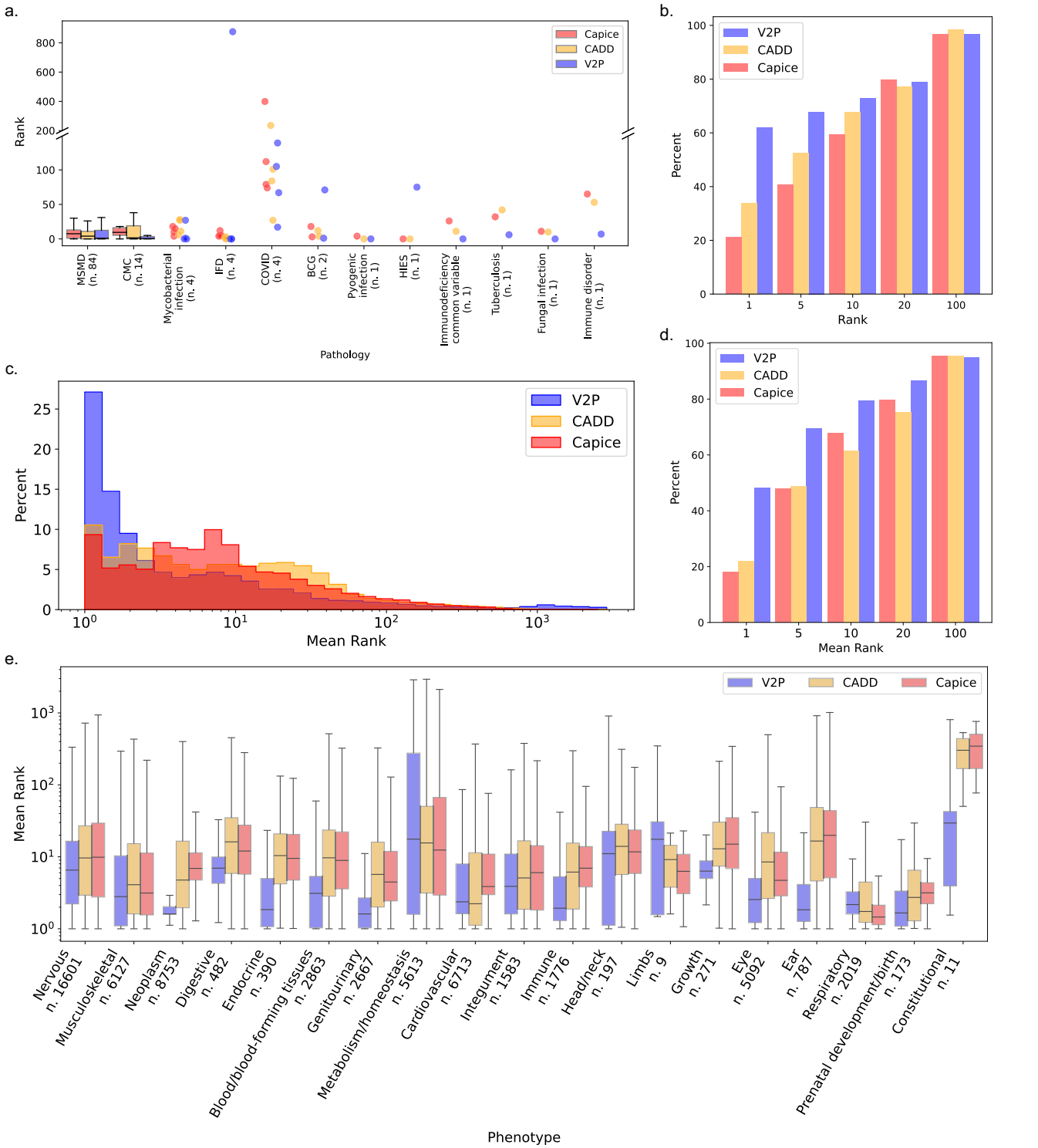


Fig. 5 | Assessment of V2P for phenotype-driven causal variant identification in sequencing data. a Distribution of the rankings of causal variants in 116 immune disorder patients, where variants are ordered by their relevant V2P phenotype score(s) (blue), CADD v1.7 (orange), and Capice 5.1.2 (red). **b** The proportion of causal exome variants (percentage) within the top *n* rankings for causal immune disorder patient variants. **c** The distribution of pathogenic variants from the held-out testing, HGMD 2023.4, and ClinVar datasets according to their average rankings in patient exomes. **d** The proportion of causal exome variants (percentage) within the top *n* rankings for variants from the held-out testing, HGMD 2023.4, and ClinVar datasets. **e** The distribution of rankings for variants resulting in a single phenotype from the held-out testing, HGMD 2023.4, and ClinVar datasets. Boxes represent the quartiles of the data. Whiskers extend to points that lie within 1.5 times the inter-quartile range of the lower and upper quartiles.

prediction strategy, in conjunction with observed patient phenotypes, allows for significantly improved resolution of causal disease variants in sequencing data as compared to other genome-wide prediction methods on these data.

Discussion

We have developed V2P, a high-throughput, multi-task predictor of variant pathogenicity and top-level HPO phenotype ($n = 23$). V2P's unique architecture, which exploits commonalities between the related pursuits of general and phenotype-specific pathogenicity prediction, allows for the exploration of pathogenic variant functional effect throughout the human genome. Due to the broad body of evidence upon which V2P's predictions are conditioned, V2P may be employed for both SNVs and indels in both coding and non-coding regions of the genome, extending its usefulness beyond the assessment of missense variants to other pertinent undertakings such as the prioritization of causal variants in patient sequencing data. V2P allows for the mapping of human SNVs and indels to plausible phenotypes, providing evidence that may help guide novel investigations of disease etiology and variant effect throughout the genome. Our investigation has yielded several notable findings.

Varying disease severity and presentation within a single gene may be masked by binary representations of variant effect, i.e., pathogenic vs. benign, resulting in less effective predictive methods[50]. Often, this performance variability is not transparent to the user, as most variant effect predictors assess performance on variants with varying phenotypes. V2P seeks to address this limitation by generating distinct predictions for different disease phenotypes. Crucially, for our benchmarking data presented in Fig. 3d, e, V2P's phenotype-specific models outperform existing methods and its own general pathogenicity prediction. Across the 23 examined phenotypes, V2P's phenotype-specific scores yielded on average a 0.16 improvement in AP score compared to the next best method for the given phenotype and an average increase of 0.38 over the median AP score of the compared methods across phenotypes in the three evaluation datasets (Fig. 3d, e). Further, we observe that V2P does not lose performance, compared with other methods, for underrepresented phenotypes. Indeed, for cellular phenotypes, abnormalities of the ear, prenatal development and birth, and constitutional symptoms—several of the least represented phenotypes in our dataset—we observe that V2P's phenotype-specific score outperforms the compared methods on the examined data for the assessed metrics. Together, these results indicate that V2P's phenotype-specific approach may have utility for the identification of pathogenic variants in the context of their phenotypic effects. For the investigation of a particular phenotype or disease, V2P may offer a unique perspective on variant effect.

Circularity in training and testing data can result in misleading results when assessing variant effect predictor performance[51]. Thus, it is essential to include evaluation data from diverse sources for accurate benchmarking. Notably, V2P generalizes beyond clinical characterizations of variant effect per the data examined in this study, aligning well with experimentally-derived evidence of pathogenicity in both coding and non-coding regions of the genome. In particular, despite the additional complexity of V2P's learning task in relation to that of most compared methods, V2P performs favorably in comparison with previously published variant effect predictors in its correlation with outcomes from DMS assays across a broad array of proteins (Fig. 4a). For non-coding variants, V2P was comparable to CADD[46] and achieved better correlation with outputs from MPRAs of several promoter and enhancer regulatory elements than FATHMM[48] (Fig. 4f).

Moreover, initial investigations indicate V2P may be efficiently employed for the automated identification of causal variants in simulated and actual patient sequencing data across phenotypes. Often, patients suffering from rare diseases do not harbor a known causative pathogenic variant[11]. In such cases, investigators and clinicians must leverage alternative techniques for filtering variants in whole-exome (WES) or whole-genome sequencing results, rendering variant effect predictors essential to the diagnostic process[11]. To simulate this process, we introduced known pathogenic variants into WES data from patients in the Mount Sinai BioMe biobank, and, for each patient separately, ranked each variant according to scores from V2P, CADD, and Capice. Across the three evaluation datasets, we found V2P's phenotype-specific scores to result in a significantly lower average ranking than when ranking with other methods. We observed similar phenomena with V2P for 116 patients suffering rare immune disorders with previously identified causal variants.

Finally, we undertook a preliminary investigation of the biological qualities underlying pathogenic variants across the spectrum of human phenotypes. Using feature importance to quantify the relationship between phenotypes and features, we identified a variety of biological properties that were predictive of variant phenotype. These features included protein structural characteristics, measures of evolutionary conservation, tissue-specific gene expression, and regulatory and epigenetic qualities, among others (Fig. 2a, b and Supplementary Data 2). While some features were indicated to be relevant across phenotypes, the remaining identified features for each phenotype varied quite widely. These findings revealed biologically meaningful features contributing to the variant disease phenotype predictions.

There are several possible explanations for V2P's observed performance in relation to other variants' effect predictors on our benchmarking data. For instance, the multi-label approach V2P employs allows for increased specificity for each individual classification task. Moreover, each individual task, that is, each phenotype prediction, may be bolstered by the others, since our approach models the correlation between labels. This is particularly important for phenotypes with fewer labeled samples, for which existing VEPs tend to be less effective per our benchmarking. Simultaneously, V2P allows relevant features to be prioritized for each phenotype, and indeed, we observe phenotype-related features to influence the model in our Shapley analysis (Supplementary Fig. 1). Another possible explanation is overfitting, which is challenging to conclusively rule out for any machine learning task. However, our extensive benchmarking—spanning thousands of genes and including our test set that exclusively includes proteins with low sequence similarity to those in our training data—provides evidence that suggests V2P generalizes well.

Whilst V2P offers a unique approach for the investigation and identification of pathogenic variants, its characterization of variant phenotypes is broad. In particular, V2P classifies variants according to the 23 top-level phenotypic categories defined by the HPO[14], e.g., abnormalities of the nervous system and neoplasms. Due to limited sample sizes, supervised learning approaches such as that employed by V2P cannot easily be extended to more granular phenotypes or specific diseases. Future iterations employing strategies such as pre-training and transfer learning with deep learning models may help to ameliorate this limitation. Furthermore, some HPO classes contain phenotypes/diseases that do not necessarily share pathophysiological mechanisms. Because of its training strategy, V2P adopts the assumptions and limitations of the HPO's definition. Similarly, as V2P is trained on data from the HGMD, it may exhibit bias for well-studied genes and the curation strategy employed[51]. Further, databases such as the HGMD and ClinVar are understood to contain false positives, though the proportion of false positives has improved over time[52]. As seen in Supplementary Data 5, certain molecular consequences, e.g., frameshift, are overrepresented in the pathogenic class of the V2P dataset. While the observed disparity likely reflects the underlying biology to some extent rather than solely an issue of data coverage, in future releases of V2P, we may be able to move closer to parity with the incorporation of additional data from gnomAD v4, which may improve performance in some cases. Regardless, V2P demonstrates utility for pathogenic variant detection and provides

indications of variant phenotypes that can be expanded upon in future work. In particular, for the investigations of a particular phenotype and for the filtration of sequencing data, V2P may prove to be of use. Future VEP developers may also consider applying their architectures and approaches in a phenotype-specific manner, as in V2P, to investigate the effects on performance. We anticipate that the novel resources provided by V2P will allow for new insights into the relationship between pathogenic variants and their phenotypic outcomes during future investigations and as these data are explored in detail by the genetics community.

Methods

Model development and dataset assembly

Pathogenic and putatively benign variants used for the development of V2P were retrieved from the HGMD[3] Professional version 2022.4 and from the gnomAD[31] version 2.1 exome database, respectively. Specifically, all high-confidence, pathogenic variants in the disease-causing mutations (DM) category of the HGMD database that were mapped to the GRCh38 human reference genome were selected, for a total of 252,125 pathogenic variants. Of these, 201,969 were associated with one or more HPO[14] phenotypes, which were mapped to one or more of the 23 top-level phenotypes in the phenotypic abnormality sub-ontology. The Sequence Ontology[53] (SO) consequence (Supplementary Data 5) and impacted gene or regulatory region affected by each pathogenic variant was identified using Ensembl's Variant Effect Predictor[54] version 108. All gnomAD variants from the affected genes and regions were collected, and any variants from gnomAD that co-occurred in the HGMD were filtered out. For each gene or region containing HGMD variants, gnomAD variants were selected such that the total number of gnomAD variants from that gene or region was equal to that of the HGMD variants and also such that the SO consequences of the gnomAD variants matched those of the HGMD variants (Supplementary Data 5). When an equivalent number of gnomAD variants resulting in a given SO consequence were not available, other gnomAD variants from the same gene or region were selected randomly so as to achieve parity between the total count of gnomAD and HGMD variants for the gene or region. When fewer gnomAD variants were available than HGMD variants for a gene or region, all gnomAD variants from that gene or region were selected. As a result of this process, 244,231 putatively benign variants were selected from gnomAD

Assigning HPO phenotypes to pathogenic variants

HPO phenotypes were assigned to pathogenic variants by annotation of the diseases associated with variants against the Unified Medical Language System (UMLS) metathesaurus[55], carried out using a simple word permutation-based method[56]. The disease names were mapped to UMLS concept identifiers (CUI) using the open source UMLS-Query module[56]. The UMLS concepts were then cross-mapped onto the HPO. Each HPO phenotype was subsequently mapped to one or more of the top-level phenotypic abnormality sub-ontology phenotypes by traversal of the HPO graph. Crucially, variants were assigned to all top-level phenotypes to which they were related per the mapping procedure rather than a single top-level phenotype.

Training, testing, and validation dataset split

The labeled variants from the HGMD and gnomAD were apportioned into training, validation, and testing sets such that each accounted for roughly 80%, 10%, and 10% of the data, respectively, with the constraint that variants from proteins with greater than 40% sequence identity did not belong to different sets and such that the proportions of variants with each label were similar between the sets. To calculate sequence identity, the Ensembl reference amino acid sequences for each protein in the dataset were retrieved from the Ensembl BioMart[57] version 108. CD-HIT[58] version 4.8.1 was employed to assign each protein to a set such that each protein sequence in the set had no less than 40% sequence identity with all the others in that set.

To ensure that the distribution of labels across the training, validation, and testing datasets reflected the overall distribution of labels, a randomized algorithm was employed to calculate the deviation of the distribution of the label powerset for a given split from an ideal split—i.e. a split in which the distribution of the labels is exactly equivalent to the distribution of the labels in the complete dataset—where the label powerset is defined as every unique, observed combination of labels for every sample in the dataset. Specifically, to obtain the training dataset, for 100,000 iterations, the sets of homologous proteins, calculated as described above, were randomly shuffled with a unique random seed. A total of 80% of these protein sets were selected, and all variants occurring in the proteins/genes belonging to those sets were assigned to the candidate training dataset. The deviation between the candidate training dataset and the ideal training dataset was then measured with $p = 0.2$ as defined in Eq. (1) where d is the total deviance, l is the label powerset, p is the expected proportion of a label set in the remainder of the data after splitting, n_R is the number of occurrences of a label set in the remaining data after allocating the desired data, and n_T is the number of occurrences of a label set in the total dataset.

$$d = \sum_{n}^{|l|} \left| p \times \left(\frac{n_R}{n_T} \right) \right| \times n_T \tag{1}$$

After calculating the deviance for each split, the split with the lowest deviance d was selected. Similarly, after allocating the training data, the remaining ~20% of the data was split into validation and testing datasets using the above procedure for 100,000 iterations with $p = 0.5$.

Model architecture and training

V2P comprises six constituent models, ensembled for more robust performance. Three problem transformation approaches for multi-label classification were employed, namely, binary relevance[59], label powerset[59], and RaKel-D[60] with the LightGBM[61] v3.3.5 implementation of the gradient boosted decision tree (GBDT) algorithm as the base estimator. GBDT was chosen considering it has been consistently demonstrated to outperform other algorithms for classification on tabular data[62]. Two models were created employing each of the problem transformation approaches, respectively. Each constituent model was trained on the combined samples from the training and validation datasets. For one model from each pair using a particular problem transformation technique, samples representing minority phenotypes were over-sampled using random multi-label oversampling[63] to increase the frequency of the minority classes by 25%. Over-sampling was performed to provide a greater signal for the least represented classes during training. The predictions from each constituent model for each class are averaged to obtain the final predictions.

Model evaluation datasets

In addition to samples from the held-out testing set described above, V2P's performance was evaluated on data from the 2023.4 release of the HGMD that was not used in V2P's development. In particular, all DM variants that were not present in the 2022.4 release were extracted, and those associated with HPO phenotypes were labeled with the 23 top-level HPO phenotypes as described above. In total, 17,767 pathogenic variants not present in the original training/validation datasets were obtained. Of those, 17,503 were associated with one or more HPO phenotypes. Putatively benign variants were matched to the pathogenic variants following the same process detailed for the model development dataset. A total of 17,336 putatively benign variants were

accumulated in this manner. An additional 2664 gnomAD missense variants were retrieved from the same genes, bringing the total number of putatively benign variants to 20,000. DM+ (low-confidence pathogenic) variants from the HGMD were also retrieved for examination of V2P score distribution across variant pathogenicity confidence levels.

Additional labeled data were retrieved from ClinVar[4]. In particular, all variants mapped to the GRCh38 human reference genome were retrieved from the ClinVar FTP website release 10/15/2024. Variants mapped to the GRCh38 reference genome were selected, and those with clinical significance of "Pathogenic", "Pathogenic/Likely pathogenic", "Likely pathogenic", "Benign", "Benign/Likely benign", and "Likely benign" were retained. Any of these variants occurring in the HGMD/gnomAD training/validation dataset were removed. Variants were further filtered to retain only those with review status of "criteria provided, single submitter", "criteria provided, multiple submitters, no conflicts", "reviewed by expert panel", "practice guideline", removing those variants with the lowest confidence/assertion criteria. Further, variants with greater than 0.01% allele frequency in gnomAD v4 exome or genome datasets or in the African, Admixed American, East Asian, non-Finnish European, or South Asian subpopulations were removed. Finally, variants were assigned HPO phenotypes via a combination of UMLS mapping and manual labeling (Supplementary Methods, Supplementary Data 6). Those without phenotype information were removed. The resulting dataset contained 242,523 benign variants and 62,460 pathogenic variants. Data from ClinVar were further refined to investigate model performance in genes with fewer known pathogenic variants and for variants reviewed after V2P's training. Specifically, variants mapped to the GRCh38 human reference genome were retrieved from the ClinVar FTP website release 06/15/2025. Variants "LastEvaluated" in 2024 and 2025 and "Pathogenic", "Pathogenic/Likely pathogenic", "Likely pathogenic", "Benign", "Benign/Likely benign", and "Likely benign" variants in genes with ten or fewer pathogenic/likely pathogenic entries were retained. Variants with the lowest confidence/assertion criteria were removed. A total of 17,334 benign/likely benign and 618 pathogenic/likely pathogenic variants were accumulated as a result of this process.

Method comparison

CADD[46] version 1.7 scores were retrieved from the CADD web application, dbNSFP v4.7a, or computed locally with the default parameters. AlphaMissense[6] scores were retrieved from the web resource provided by the authors and dbNSFP v4.7a. EVE[50] scores were obtained from the ProteinGym[64] website and dbNSFP v4.7a. FATHMM-XF[48] scores were retrieved from the FATHMM website. Capice[49] and Enformer[47] scores were computed locally with default parameters. When Capice produced multiple scores for a variant in different transcripts of a gene, these scores were averaged. All other variant effect predictor scores were obtained from dbNSFP[65] version 4.7a. For dbNSFP scores, when scores were available for multiple transcripts, the score for the transcript employed by V2P—i.e., the transcript prioritized by Ensembl's Variant Effect Predictor—was used. When multiple scores were available but could not be matched to the transcript employed by V2P, the average of scores across transcripts was used. The continuous scores output by each method were used for all comparisons, unless otherwise explicitly stated. Methods were compared in terms of their average precision (AP) scores (Eq. (2)) and the Spearman rank correlation coefficients (Eq. (3)). Precision-recall baselines were calculated as the ratio of positive samples to total samples for respective datasets. AP is defined as follows, where P_n and R_n are the precision and recall at the n^{th} threshold, respectively.

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{2}$$

Spearman's rank correlation coefficient is defined as follows, where d_i is the difference in rankings for the i^{th} observation and n is the number of observations.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3}$$

Functional studies of variant effect

DMS data were retrieved from the ProteinGym[64] database. Sequence identity between the assayed amino acid sequence and the Ensembl amino acid sequence employed by V2P was calculated using the BLOSUM62 substitution matrix and the Needleman-Wunsch algorithm[66] implemented in BioPython version 1.81. Assays of sequences with an identity less than 85% were discarded. For each remaining protein, V2P scores were generated for all amino acid substitutions expressible via an SNV. Methods were compared solely on variants for which each included tool yielded a prediction. When multiple assays of a single protein were available, results from each tool were averaged across available assays. MPRA data were retrieved from the MPRA data access portal[67]. Variants with fewer than 10 tags or a p value greater than 0.00001 were removed. Elements MYCrs11986220, BCL11A, FOXE1, UC88 were omitted[7]. For DMS and MPRA data, variants occurring in V2P's training data were removed. Hypothesis testing for V2P scores in varying regions of P53 was conducted via one-tailed Mann–Whitney U tests.

Causal variant identification

To simulate human disease cases, pathogenic variants from each evaluation dataset with known phenotypic outcomes were inserted, one at a time, into 100 randomly selected sequenced exomes from patients in the Mount Sinai BioMe Biobank. Variants were filtered according to several criteria closely matching the procedure employed by Exomiser[68]. First, variants were filtered according to allele frequencies from gnomAD v4. Variants passing calling filters with an exome, genome, or population (African, Admixed American, East Asian, non-Finnish European, South Asian) allele frequency greater than 0.01 were removed. Next, variants with Sequence Ontology molecular consequence 5_prime_UTR_variant, 3_prime_UTR_variant, non_coding_transcript_variant, upstream_gene_variant, downstream_gene_variant, intergenic_variant, or intron_variant were removed. Finally, genes associated with the phenotype(s) caused by the pathogenic variants were predicted using Phen2Gene[69], a phenotype-driven gene prioritization tool to score genes. Variants in genes with Phen2Gene scores in the 95th percentile and above were retained. If the causal variant was filtered out, sets of 100 genes with the next highest Phen2Gene scores were added until the variant was recovered to mimic a scenario in which progressively broader analyses are required to identify a satisfactory candidate variant. If the causal gene was not in the Phen2Gene prioritized genes, all genes were included. After filtration, for each exome, the endogenous variants, excluding structural variations, and the spiked-in variant were annotated with the features employed by V2P for prediction, and their functional impact was predicted by V2P. Variants were sorted into descending order according to the V2P phenotype-specific scores corresponding to the phenotype resulting from the spiked-in variant, the CADD scores, and the Capice scores, separately. For variants resulting in more than one phenotype, z-scores for the corresponding V2P phenotype-specific predictions were summed. For each pathogenic variant, this process was repeated in each of the 100 exomes, and the average rank of the pathogenic variant according to the different scores was determined. Hypothesis testing was conducted via one-tailed Mann–Whitney U tests.

Feature provenance and engineering

Gene model data were retrieved for variants according to their GRCh38 genomic coordinates using Ensembl's Variant Effect Predictor[54] (VEP) version 108. AlphaFold2[70] models of protein structures were employed for the calculation of a variety of protein structural descriptors. A thorough description of the generation and processing of the features calculated from the AlphaFold2 models may be found in ref. 12. Protein-protein interaction features were calculated using node2vec[71] on the human protein interactome data from the STRING[41] database version 11. Node2vec produced 64 output dimensions characterizing the data. Association between mouse phenotypes and top-level HPO phenotypes was calculated as the number of level four terms from the Mammalian Phenotype Ontology[40] associated with a given gene that mapped to a top-level HPO phenotype. Gene functional data, obtained from Enrichr, were converted into sparse binary features. Particularly, for each term in each Enrichr library, if the term was associated with a gene, the gene would have a value of one for that term. Otherwise, the gene would have a value of zero for that term. Other features were retrieved from their respective databases or tools, unless otherwise specified (Supplementary Data 1).

Some included features such as MutationAssessor[72] and Eigen[73] were designed for the prediction of variant effect. Features predicting variant effect were chosen to limit the effects of circularity due to overlapping training/testing datasets. In particular, the included variant effect prediction features only include unsupervised models and those not making use of strongly labeled pathogenicity/neutrality data for training. To further assess V2P, we retrained the model with variant effect prediction features removed, taking a broad definition of variant effect predictor. These features include MaxEntScan_alt, MaxEntScan_diff, MaxEntScan_ref, ada_score, rf_score, Eigen_PC_raw_coding, Eigen_raw_coding, GERP++_NR, GERP++_RS, GM12878_confidence_value, GM12878_fitCons_score, GenoCanyon_score, H1_hESC_confidence_value, H1_hESC_fitCons_score, HUVEC_confidence_value, HUVEC_fitCons_score, LINSIGHT, LIST_S2_score, LRT_Omega, LRT_score, MPC_score, MutationAssessor_score, SiPhy_29way_logOdds, integrated_confidence_value, integrated_fitCons_score, GDI, MSC_95CI, RVIS, Indispensability_score, A3D_SCORE, concavity_score, S_DDG[SEQ], S_DDG[3D], s_het, targetScan, mirSVR-Score, mirSVR-E, mirSVR-Aln, GerpRS, GerpRSpval, GerpN, GerpS, SpliceAI-acc-gain, SpliceAI-acc-loss, SpliceAI-don-gain, SpliceAI-don-loss, MMSp_acceptorIntron, MMSp_acceptor, MMSp_exon, MMSp_donor, MMSp_donorIntron, dbscSNV-ada_score, and dbscSNV-rf_score.

Feature analysis

Relevant features were selected using the training dataset for each phenotype using the Boruta[74] all-relevant feature selection algorithm with Shapley values to estimate feature importance and a LightGBM base estimator. Specifically, for each phenotype, pathogenic variants resulting in a given phenotype were treated as the positive class, and pathogenic variants not resulting in the phenotype were treated as the negative class. The pairwise relationship between phenotypes based on shared selected features was calculated as the number of commonly selected features with the same direction of effect, as indicated by the log odds ratio, over the total number of selected features for each phenotype. Log odds ratios were calculated for the standardized features via logistic regression with the Newton-conjugate gradient solver.

Three measures of genic essentiality were assessed: RVIS[32], the de novo excess rate[33], and the indispensability score[34]. RVIS is a metric derived by comparing common functional genetic variations to neutral variations in a large dataset of human whole-exome sequences. Positive RVIS suggests a prevalence of common functional variations, whereas a negative score suggests intolerance. The de novo excess rate evaluates the frequency of de novo mutations per gene and determines if a gene harbors a greater number of de novo mutations than anticipated by random chance alone. Higher scores indicate greater essentiality. The indispensability score measures a gene's essentiality by analyzing its network and evolutionary characteristics. Again, higher scores indicate greater essentiality.

Three measures of evolutionary conservation were assessed: PhastCons[75] (primate), PhyloP[75] (primate), and GERP++[76]. For each measure, higher values indicate greater conservation of the site.

Log odds ratios were calculated via logistic regression with the Newton-conjugate gradient solver. Hypothesis testing for essentiality and conservation features was carried out via Wald tests of the logistic regression coefficients.

Reporting summary

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

Data availability

The phenotype/variant effect predictions generated in this study are available at https://v2p.ai/downloads/. The annotation data employed by V2P are accessible at https://v2p.ai/downloads/. The provenance of features used for model training is detailed in Supplementary Dataset 1. The Human Gene Mutation Database Professional data are available via license. gnomAD v2.1 lifted to the GRCh38 reference was retrieved from https://storage.googleapis.com/gcp-public-data-gnomad/release/2.1.1/liftover_grch38/vcf/exomes/gnomad.exomes.r2.1.1.sites.liftover_grch38.vcf.bgz. The ClinVar variant_summary.txt.gz data were obtained at the ClinVar FTP portal (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/), release 10/15/2024. dbNSFP v4.7a data were obtained at https://sites.google.com/site/jpopgen/dbNSFP. AlphaMissense scores were obtained at https://zenodo.org/records/10813168/files/AlphaMissense_hg38.tsv.gz?download=1. EVE scores were retrieved from ProteinGym, https://proteingym.org/download. FATHMM-XF scores were obtained at https://fathmm.biocompute.org.uk/fathmm-xf. DMS data were retrieved from https://proteingym.org/download. MPRA data were obtained at https://kircherlab.bihealth.org/satMutMPRA/. De novo mutations from Deciphering Developmental Disorders were retrieved from Supplementary Data S7 of https://doi.org/10.1126/science.adg7492. Autism spectrum do novo mutations were retrieved from Supplementary Table S2 of https://doi.org/10.1126/science.aat6576. Congenital heart disease de novo mutations were collected from Supplementary Table S9 of https://doi.org/10.1038/ng.3970. Sequencing data for causal variant ranking were obtained from the Mount Sinai BioMe BioBank and from the Human Genetics of Disease laboratory at Rockefeller University. A website to query and generate predictions from V2P is available at www.v2p.ai.

Code availability

Code related to the v2p project is available at https://github.com/davidfstein/v2p (https://doi.org/10.5281/zenodo.17316362).

References

1. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
2. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
3. Stenson, P. D. et al. The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **139**, 1197–1207 (2020).
4. Landrum, M. J. et al. ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
5. Garcia, F. A. O., de Andrade, E. S. & Palmero, E. I. Insights on variant analysis in silico tools for pathogenicity prediction. *Front. Genet.* **13**, 1010327 (2022).

6. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
7. Schubach, M., Maass, T., Nazaretyan, L., Röner, S. & Kircher, M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res.* **52**, D1143–D1154 (2024).
8. Livesey, B. J. et al. Guidelines for releasing a variant effect predictor. *Genome Biol.* **26**, 97 (2025).
9. Sevim Bayrak, C. et al. Identification of discriminative gene-level and protein-level features associated with pathogenic gain-of-function and loss-of-function variants. *Am. J. Hum. Genet.* **108**, 2301–2318 (2021).
10. Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat. Commun.* **13**, 3895 (2022).
11. Stenton, S. L. et al. Critical assessment of variant prioritization methods for rare disease diagnosis within the rare genomes project. *Hum. Genomics* **18**, 44 (2024).
12. Stein, D. et al. Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of a diverse feature set. *Genome Med.* **15**, 103 (2023).
13. Petrazzini, B. O. et al. Ensemble and consensus approaches to prediction of recessive inheritance for missense variants in human disease. *Cell Report Methods.* **4**, 12 (2024).
14. Köhler, S. et al. The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
15. Kahanda, I., Funk, C., Verspoor, K. & Ben-Hur, A. PHENOstruct: prediction of human phenotype ontology terms using heterogeneous data sources. *F1000Research* **4**, 259 (2015).
16. Doğan, T. HPO2GO: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences. *PeerJ* **6**, e5298 (2018).
17. Liu, L., Huang, X., Mamitsuka, H. & Zhu, S. HPOLabeler: improving prediction of human protein-phenotype associations by learning to rank. *Bioinformatics* **36**, 4180–4188 (2020).
18. Kulmanov, M. & Hoehndorf, R. DeepPheno: predicting single gene loss-of-function phenotypes using an ontology-aware hierarchical classifier. *PLoS Comput. Biol.* **16**, e1008453 (2020).
19. Gao, J. et al. HPOAnnotator: improving large-scale prediction of HPO annotations by low-rank approximation with HPO semantic similarities and multiple PPI networks. *BMC Med. Genomics* **12**, 187 (2019).
20. Notaro, M., Schubach, M., Robinson, P. N. & Valentini, G. Prediction of human phenotype ontology terms by means of hierarchical ensemble methods. *BMC Bioinforma.* **18**, 449 (2017).
21. Requena, D. et al. CDG: an online server for detecting biologically closest disease-causing genes and its application to primary immunodeficiency. *Front. Immunol.* **9**, 1340 (2018).
22. Itan, Y. et al. The human gene connectome as a map of short cuts for morbid allele discovery. *Proc. Natl. Acad. Sci. USA* **110**, 5558–5563 (2013).
23. Piro, R. M. & Di Cunto, F. Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.* **279**, 678–696 (2012).
24. Ata, S. K. et al. Recent advances in network-based methods for disease gene prediction. *Brief. Bioinform.* **22**, bbaa303 (2021).
25. Zhang, X. et al. Disease-specific variant pathogenicity prediction significantly improves variant interpretation in inherited cardiac conditions. *Genet. Med.* **23**, 69–79 (2021).
26. Fang, M. et al. VIPPID: a gene-specific single nucleotide variant pathogenicity prediction tool for primary immunodeficiency diseases. *Brief. Bioinform.* **23**, bbac176 (2022).
27. Bu, F. et al. DVPred: a disease-specific prediction tool for variant pathogenicity classification for hearing loss. *Hum. Genet.* https://doi.org/10.1007/s00439-022-02440-1 (2022).
28. Kumaran, M. & Devarajan, B. eyeVarP: a computational framework for the identification of pathogenic variants specific to eye disease. *Genet. Med.* **25**, 100862 (2023).
29. Zhan, H. & Zhang, Z. ProPath: disease-specific protein language model for variant pathogenicity. Preprint at http://arxiv.org/abs/2311.03429 (2023).
30. Evans, P. et al. Genetic variant pathogenicity prediction trained using disease-specific clinical sequencing data sets. *Genome Res.* **29**, 1144–1151 (2019).
31. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
32. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
33. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
34. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput. Biol.* **9**, e1002886 (2013).
35. Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
36. Piñero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).
37. Grissa, D., Junge, A., Oprea, T. I. & Jensen, L. J. Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration. *Database* **2022**, baac019 (2022).
38. McKusick, V. A. Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
39. Lachmann, A. et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366 (2018).
40. Smith, C. L., Goldsmith, C.-A. W. & Eppig, J. T. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.* **6**, R7 (2005).
41. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
42. Milacic, M. et al. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res.* **52**, D672–D678 (2024).
43. Huang, R. et al. The NCATS BioPlanet—an integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics. *Front. Pharmacol.* **10**, 445 (2019).
44. Pico, A. R. et al. WikiPathways: pathway editing for the people. *PLoS Biol.* **6**, e184 (2008).
45. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
46. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
47. Avsec, Ž et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
48. Rogers, M. F. et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511–513 (2018).
49. Li, S. et al. CAPICE: a computational method for consequence-agnostic pathogenicity interpretation of clinical exome variations. *Genome Med.* **12**, 75 (2020).
50. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

51. Grimm, D. G. et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).
52. Sharo, A. G., Zou, Y., Adhikari, A. N. & Brenner, S. E. ClinVar and HGMD genomic variant classification accuracy has improved over time, as measured by implied disease burden. *Genome Med.* **15**, 51 (2023).
53. Eilbeck, K. et al. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44 (2005).
54. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
55. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
56. Shah, N. H. & Musen, M. A. UMLS-Query: A Perl Module For Querying the UMLS. *Amia. Annu. Symp. Proc.* **2008**, 652–656 (2008).
57. Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
58. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
59. Tsoumakas, G., Katakis, I. & Vlahavas, I. A review of multi-label classification methods. *Proceedings of the 2nd ADBIS workshop on data mining and knowledge discovery* (2006).
60. Tsoumakas, G., Katakis, I. & Vlahavas, I. Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.* **23**, 1079–1089 (2011).
61. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. in *Proc. 31st International Conference on Neural Information Processing Systems,* Vol. 30, 3149–3157 (Curran Associates Inc., 2017).
62. Shwartz-Ziv, R. & Armon, A. Tabular Data: deep learning is not all you need. *Pattern Recognition*. **118**, https://doi.org/10.48550/arXiv.2106.03253 (2021).
63. Charte, F., Rivera, A. J., del Jesus, M. J. & Herrera, F. Addressing imbalance in multilabel classification: measures and random resampling algorithms. *Neurocomputing* **163**, 3–16 (2015).
64. Notin, P. et al. ProteinGym: large-scale benchmarks for protein design and fitness prediction. *37th Conference on Neural Information Processing System*. (2023).
65. 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).
66. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
67. Kircher, M. et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).
68. Smedley, D. et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* **10**, 2004–2015 (2015).
69. Zhao, M. et al. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genomics Bioinforma.* **2**, lqaa032 (2020).
70. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
71. Grover, A. & Leskovec, J. node2vec: scalable feature learning for networks. *KDD* 855–864 https://doi.org/10.1145/2939672.2939754 (2016).
72. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
73. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
74. Bury, T. All relevant feature selection. (2021).
75. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
76. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP. *PLoS Comput. Biol.* **6**, e1001025 (2010).
77. Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
78. Raimondi, D. et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.* **45**, W201–W206 (2017).
79. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics* **14**, S3 (2013).
80. 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).
81. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* **37**, 235–241 (2016).
82. Li, B. et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**, 2744–2750 (2009).
83. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
84. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
85. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7**, e46688 (2012).
86. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).
87. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–362 (2014).
88. Li, C., Zhi, D., Wang, K. & Liu, X. MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. *Genome Med.* **14**, 115 (2022).
89. Jagadeesh, K. A. et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).
90. Qi, H. et al. MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.* **12**, 510 (2021).
91. Zhang, H., Xu, M. S., Fan, X., Chung, W. K. & Shen, Y. Predicting functional effect of missense variants using graph attention neural networks. *Nat. Mach. Intell.* **4**, 1017–1028 (2022).
92. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at https://doi.org/10.1101/148353 (2017).
93. Feng, B.-J. PERCH: a unified framework for disease gene prioritization. *Hum. Mutat.* **38**, 243–251 (2017).
94. Alirezaie, N., Kernohan, K. D., Hartley, T., Majewski, J. & Hocking, T. D. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *Am. J. Hum. Genet.* **103**, 474–483 (2018).
95. Wu, Y. et al. Improved pathogenicity prediction for rare human missense variants. *Am. J. Hum. Genet.* **108**, 1891–1906 (2021).
96. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA* **118**, e2016239118 (2021).

97. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
98. Lu, Q. et al. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.* **5**, 10576 (2015).
99. Shihab, H. A. et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).

Acknowledgements

We would like to thank the members of the Itan, Schlessinger, Casanova, and Cooper labs for their support for this project. We would also like to thank Professors Vikas Pejaver, Bruce Gelb, and Anne Bowcock for their valuable guidance and discussion of this work. This work was financially supported by the National Institutes of Health (NIH) grants R24AI167802 and P01AI186771, the Fondation Leducq, and the Leona M. and Harry B. Helmsley Charitable Trust grant 2209-05535 (Y.I.). Additional financial support was provided by the NIH grants R01CA277794, R01HD107528, and R01NS145483 (A.S.). This work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. Research reported in this publication was also supported by the Office of Research Infrastructure of the National Institutes of Health under award numbers S10OD026880 and S10OD030463. The Laboratory of Human Genetics of Infectious Diseases is supported by the Howard Hughes Medical Institute, The Rockefeller University, the St. Giles Foundation, the French *Agence Nationale de la Recherche* (ANR) under the France 2030 program (ANR-10-IAHU-01), the Integrative Biology of Emerging Infectious Diseases Laboratory of Excellence (ANR-10-LABX-62-IBEID), the French Foundation for Medical Research (FRM) (EQU202503020018), the Square Foundation, *Grandir - Fonds de solidarité pour l'enfance*, the *Fondation du Souffle*, the SCOR Corporate Foundation for Science, the Battersea & Bowery Advisory Group, William E. Ford, General Atlantic's Chairman and Chief Executive Officer, Gabriel Caillaux, General Atlantic's Co-President, Managing Director and Head of Business in EMEA, and the General Atlantic Foundation, *Institut National de la Santé et de la Recherche Médicale* (INSERM), REACTing-INSERM, Paris Cité University, and the Imagine Institute. D.N.C., P.D.S., and M.M. acknowledge receipt of financial support from Qiagen Inc. through a License Agreement with Cardiff University.

Author contributions

D.S. collected data, designed the model, and contributed to project ideation along with A.S. and Y.I. D.S., M.E.K., A.S. and Y.I. wrote the manuscript, along with editing contributions from D.N.C. P.Z., B.M., B.B. and J.L.C. managed, curated, and helped to analyze sequencing data from immune system disorder patients. M.M., P.D.S., and D.N.C. developed and provided access to curated associations between high-confidence disease variants and relevant Human Phenotype Ontology phenotypes. All authors contributed to and commented on the manuscript and figures.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-66607-w.

Correspondence and requests for materials should be addressed to Avner Schlessinger or Yuval Itan.

Peer review information *Nature Communications* thanks the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

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

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]AI Small Molecule Drug Discovery Center, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]The Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY, USA. [6]Division of Cancer and Genetics, School of Medicine, Cardiff University, Cardiff, UK. [7]Human Genetics of Infectious Diseases Laboratory, Necker Branch, INSERM, U1163 Paris, France. [8]Imagine Institute, Paris Cité University, Paris, France. [9]Howard Hughes Medical Institute, New York, NY, USA. [10]Department of Pediatrics, Necker Hospital for Sick Children, Paris, France. [11]Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [12]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ✉e-mail: avner.schlessinger@mssm.edu; yuval.itan@mssm.edu