nature biomedical engineering

Article

https://doi.org/10.1038/s41551-026-01636-4

Phenotypic prediction of missense variants via deep contrastive learning

Received: 13 June 2024

Accepted: 13 February 2026

Published online: 14 April 2026

Check for updates

Jun Wen[1,2,3], Sihang Zeng[4], Clara-Lea Bonzel[1,2], Shilpa Nadimpalli Kobren[1], Jiangchuan Du[5], Yi Chai[6], Hao Wang[7], Meng Zhu[8], Siwei Chen[9,10], Fangwei Leng[11], Harrison G. Zhang[12], Katherine P. Liao[1,2,13], Kelly Cho[2,13], Isaac S. Kohane[1], Marinka Zitnik[1,9,13,14], Alexandre C. Pereira[15], Jun S. Liu[16,17] & Tianxi Cai[1,2,18]

Missense variants (MVs) influence clinical phenotypes, but our understanding of their phenotypic consequences remains constrained. Existing computational approaches to interpret MVs predominantly assess their pathogenicity, without considering phenotypic heterogeneity. We present a machine-learning-based method, PheMART, to predict the clinical phenotypic consequences of MVs. PheMART integrates comprehensive variant and phenotype characterizations by leveraging a robust combination of multiple resources involving protein language models, protein–protein interactions, protein domains, medical knowledge graphs and electronic health records. Exploiting contrastive learning, PheMART establishes connections between MVs and 4,179 phenotypes by jointly projecting them into a cohesive low-dimensional metric space where proximity signifies relevance. Besides substantially outperforming existing models, PheMART aids in diagnosing individuals with rare diseases by effectively pinpointing clinical diagnoses and causative MVs. As a resource to the community, we provide a database of phenotypic predictions for 5.1 million putative pathogenic amino acid alterations.

Missense variants (MVs) are critical contributors to monogenic disorders[1,2], imposing notable burdens of morbidity, mortality and economic costs on both paediatric and adult populations[3,4]. Understanding the precise clinical phenotypic implications of distinct genetic variants within the human genome is a fundamental aspect of human genetics research. This comprehension provides a foundation for understanding the genetic underpinnings of diseases and lays the groundwork for therapeutic development[5]. Recent advancements in sequencing technologies, in conjunction with the availability of extensive biobank databases, have facilitated the identification of hundreds of millions of MVs[6]. However, the phenotypic impacts of most of these MVs remain unclear, as evidenced by the prevalence of MVs classified as variants of uncertain significance (VUS) in the ClinVar database[7].

There are currently two major approaches to the interpretation of MVs: genome-wide association studies (GWAS)[8] and variant functional-effects assessment[9,10]. While GWAS have been useful in

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [2]VA Boston Healthcare System, Boston, MA, USA. [3]Biological and Life Sciences Division, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. [4]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA. [5]Department of Statistics, University of Chicago, Chicago, IL, USA. [6]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [7]Department of Computer Science, Rutgers University, Piscataway, NJ, USA. [8]Department of Genetics, Harvard Medical School, Boston, MA, USA. [9]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [10]Department of Human Genetics, The University of Chicago, Chicago, IL, USA. [11]Howard Hughes Medical Institute and Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA, USA. [12]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [13]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Boston, MA, USA. [14]Harvard Data Science Initiative, Cambridge, MA, USA. [15]Brigham and Women's Hospital, Boston, MA, USA. [16]Department of Statistics and Data Science, Tsinghua University, Beijing, China. [17]Department of Statistics, Harvard University, Cambridge, MA, USA. [18]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. e-mail: tcai@hsph.harvard.edu

unravelling the associations involving common phenotypes, their applicability to MVs is hindered by several challenges. Due to selective pressures, MVs are generally rare in the population, with ~99% of them having a global minor allele frequency below 0.5% (ref. 11), and are predominantly implicated in rare monogenic phenotypes[1]. These low frequencies present difficulties for GWAS in achieving statistically significant results. In addition, linkage disequilibrium or the correlation between neighbouring genetic variants further complicates the identification of causal variants in GWAS analyses[12].

The second approach involves assessing the functional impacts of MVs through wet-lab experiments or in silico predictions. Laboratory-assay-based tools, such as deep mutational scanning, offer reliable results by measuring molecular and cellular phenotypes across thousands of variants simultaneously[9]. While these results are robust, they are imperfect proxies for relevant clinical phenotypes and are challenging to scale genome-wide[13]. In silico computational methods could theoretically encompass all MVs by designing various features or machine learning models to analyse biophysical properties or evolutionary constraints of protein sequences[10,14–18]. Yet, existing tools primarily aim to differentiate 'deleterious' from 'neutral' effects at the protein level, without providing detailed characterization of impacts on clinical phenotypes. Given the ubiquity of pleiotropy[19], where deleterious variants in a single gene affect distinct phenotypes, the absence of specific phenotypic information in these predictions restricts their use to inform clinical decisions.

In this paper, we present an in silico Phenotypic predictor for Missense vARianTs, named PheMART, to elucidate the clinical phenotypic consequences of MVs. PheMART integrates variant and phenotype characterizations by leveraging a combination of multiple resources including a pretrained protein language model (PLM)[20], established biological knowledge such as protein–protein interactions (PPIs)[21–23] and protein domains[24], medical knowledge graphs[20,25] and large-scale electronic health records (EHR) data[26]. Specifically, for variants, a 'variant encoding module' (VEM) built on PLM is developed to assess their biological effects by contrasting them with the corresponding wild-type proteins in a phenotype-aware context, thereby differentiating the nuanced phenotypic impacts of single substitutions of amino acids on the same wild type. On the phenotype side, to enable their integrated interpretations, a 'phenotype encoding module' (PEM) that models their clinical interconnectedness is designed by synthesizing large-language-model (LLM) embeddings, condensed from the expert-curated medical knowledge graph in the Unified Medical Language System (UMLS)[20,25], with pretrained phenotype representations summarized from large-scale EHR data[26–28]. On the basis of the variant–phenotype annotations derived from ClinVar (Fig. 1a), PheMART employs semisupervised contrastive learning to jointly map MVs and phenotypes into a cohesive low-dimensional embedding space, positioning each MV near the phenotypes it is implicated in. As a result, PheMART elucidates the clinical impact of each variant across 4,179 phenotypes consolidated from ClinVar reports (Fig. 1a), thus generating a predictive phenotypic map for comprehensive interpretations of MVs.

PheMART is a computational model designed to comprehensively assess the phenotypic impacts of MVs, demonstrating substantially improved performance over existing methods. Its phenotypic predictions align closely with the most recent ClinVar reports on MVs. Leveraging the patient-level data from the Undiagnosed Disease Network (UDN)[29], PheMART highlights its clinical promise by effectively pinpointing clinical diagnoses and causal variants for individuals with rare disorders that pose diagnostic challenges. Serving as a resource to the community, we provide phenotypic predictions for all ClinVar VUS and 5.1 million single-amino-acid substitutions, identified as 'likely pathogenic' by AlphaMissense[10]. This effort not only aids in clinical diagnostics but also contributes insights into understanding the molecular mechanisms underpinning these phenotypes.

Results

Overview of PheMART

PheMART is a computational framework that combines LLMs, EHR-derived phenotype representations, PLMs and established biological knowledge to predict the phenotypic effects of pathogenic missense variants. This is achieved by jointly embedding phenotypes and variants into a low-dimensional metric space where the proximity between entities signifies their clinical relevance (Fig. 1b). PheMART leverages recent advances in self-supervised protein sequence embedding to capture proteins' biological effects and structural nuances without relying on annotations. In addition, it uses contrastive learning to establish meaningful connections between variants and phenotypes in the metric space.

PheMART addresses three major challenges in extrapolating to large-scale VUS sets when trained from a relatively limited set of genetic variants with currently available phenotypic annotations. The first challenge is modelling the nuanced biological effects of single-amino-acid substitutions to differentiate their varied phenotypic effects, especially those sharing the same wild type. The second involves deciphering the intricate clinical semantic relationship between phenotypes. This includes understanding synonyms, the semantic hierarchy and other nuanced clinical semantics, thus facilitating their integrated interpretations. The third is to effectively learn from the non-exclusive and sparsely reported associations between variants and phenotypes amid the immense potential combinations across millions of variants and thousands of phenotypes.

To address the first challenge, PheMART incorporates a VEM. This module takes both variant and corresponding wild-type protein sequences as input and distinguishes the phenotypic impacts of variants sharing the same wild type by contrasting them against the wild type in a phenotype-aware context. For variants of different wild types, they are bridged by leveraging established knowledge including PPIs[21–23], protein domains[24] and gene pathways[30]. To address the second challenge, PheMART integrates the phenotype semantic information from the UMLS-guided LLM[25] and the EHR-derived embedding vectors[26] (Fig. 1b). Leveraging the semantic relationships between phenotypes facilitates inference involving phenotypes with few variant annotations. To address the final challenge, PheMART employs a contrastive learning strategy, which jointly projects variants and phenotypes into a unified metric space. In this space, the proximity between elements signifies their clinical relevance, effectively linking genetic variants to specific clinical phenotypes. PheMART is trained using the annotated phenotype–variant associations extracted from the ClinVar database[7] (Fig. 1a). During inference, PheMART predicts the relevance of each variant to the 4,179 clinical phenotypes consolidated from ClinVar reports.

PheMART accurately predicts variants' phenotypic effects

To comprehensively assess PheMART's generalizability to VUS, we conducted four distinct experimental evaluations: (1) 10-fold per-variant cross-validation on ClinVar variants, (2) temporal validation using variants with the most recent ClinVar annotations, (3) external validation with extra unseen variants from the Human Gene Mutation Database (HGMD)[31] and (4) cross-protein-domain validation, where we evaluated PheMART's generalizability to variants residing in unseen protein domains.

State-of-the-art performance across variants. We performed 10-fold per-variant cross-validations based on the curated variant–phenotype pairs from ClinVar. Specifically, we randomly selected 90% of the variants for training and reserved the remaining 10% for performance evaluation. For each variant, phenotypes documented in ClinVar were considered positive, while 30 phenotypes not associated with the variant were randomly chosen as negative controls. Performance was evaluated using three metrics: the area under the receiver operating characteristic curve (auROC); the mean reciprocal rank (MRR), which

https://doi.org/10.1038/s41551-026-01636-4

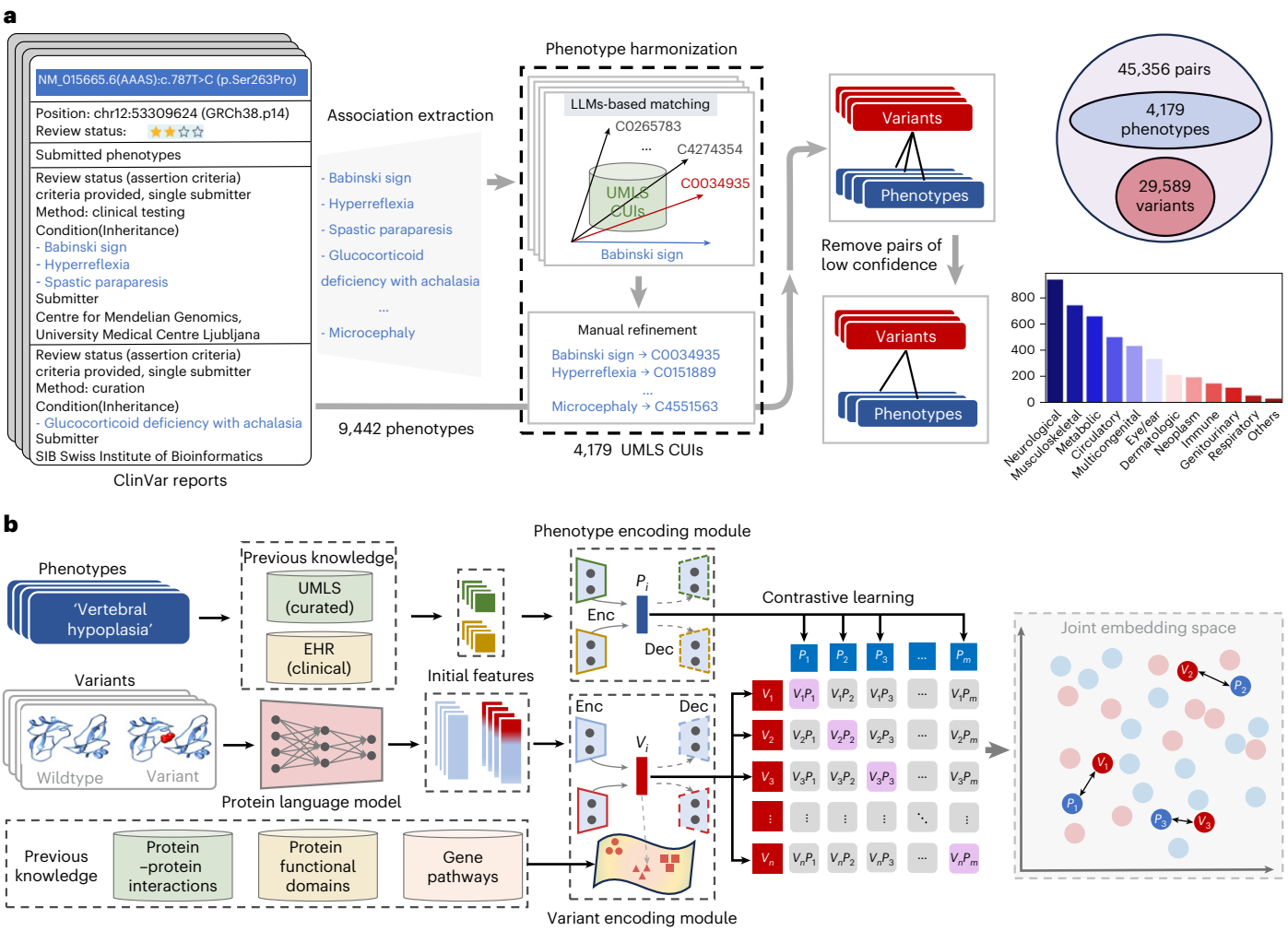

Fig. 1 | Overview of PheMART. a, Annotated data from ClinVar. We obtained variant–phenotype annotations based on the phenotypic reports of variants from the ClinVar database[7]. The phenotypes were standardized into CUIs of the UMLS, in which their semantic relationships were documented[20]. By removing reports of low confidence or those with unspecified phenotypes, we obtained 45,356 variant–phenotype pairs, covering 29,589 variants and 4,179 phenotypes. **b**, Computational flow of PheMART. It features phenotypes using knowledge from two sources: the UMLS knowledge graph, condensed using LLM embeddings[25], and large-scale EHR data summarized into clinical phenotype representations based on co-occurrences among clinical concepts[26]. The two sources were integrated by a two-stream encoder–decoder-based PEM designed to encapsulate the comprehensive interconnectedness of phenotypes. Variants were featurized using a PLM, which captures biologically informative features of proteins, such as structure and function, by learning from millions of protein sequences[83,94]. These features were further processed through a VEM with a two-stream encoder–decoder architecture. This module contrasted the variants against their corresponding wild-type proteins, enabling the differentiation of phenotypic effects arising from single-amino-acid substitutions that shared the same wild type. The generalization of variant encodings was enhanced by leveraging established biological knowledge, including PPIs, protein domains and gene pathways. Variants and phenotypes were jointly projected into a low-dimensional metric space where distances signify their clinical relevance. Consequently, PheMART elucidates the clinical significance of each variant across 4,179 clinical phenotypes, offering comprehensive phenotypic interpretations.

summarizes the ranks of the variants' annotated phenotypes among the 4,179 distinct ones; and the average sensitivities of variants' top-k phenotypic predictions (sensitivity@k). We compared PheMART to five established methods designed for pairwise relation learning or semisupervised training, including Deep Drug–Drug Interactions (DeepDDI)[32], Semisupervised Deep Generative Models (SemiDGM)[33], Multimodal Drug–Disease Relation Learning (M2REMAP)[28], Contrastive Drug–Target Interactions (CCL-DTI)[34] and Contrastive Language-Image Pre-Training (CLIP)[35] (their implementation details are provided in Section 1.8).

As shown in Fig. 2a, contrastive representation learning plays a crucial role in improving performance, with CLIP and CCL-DTI substantially outperforming DeepDDI, SemiDGM and M2REMAP, particularly in the rank-based metrics, MRR and sensitivity@k. For example, compared with M2REMAP, which surpasses DeepDDI and SemiDGM, CLIP achieves a relative improvement of 337.5% in MRR. Although PheMART shares a contrastive learning framework similar to CLIP, it demonstrates substantial advantages over CLIP across all three evaluation metrics by integrating multisource previous knowledge to enhance variant and phenotype representations. Specifically, PheMART achieves an auROC of 0.972, an MRR of 0.538 and a sensitivity@k of 0.696, representing relative gains of 5.9%, 40.4% and 67.2%, respectively, over CLIP, which is the second best-performing model. These gains in MRR and sensitivity@k have important clinical implications as they accelerate variant interpretation and improve the efficiency of genetic diagnostics.

Phenotypic predictions align with the latest ClinVar reports. To assess PheMART's generalization capabilities, we evaluated its phenotypic predictions using the most recent variant–phenotype annotations in ClinVar. Specifically, PheMART was trained on variants updated up to year 2022 and then evaluated on variants with updates in year 2023. The evaluation dataset includes 1,912 variant–phenotype

https://doi.org/10.1038/s41551-026-01636-4

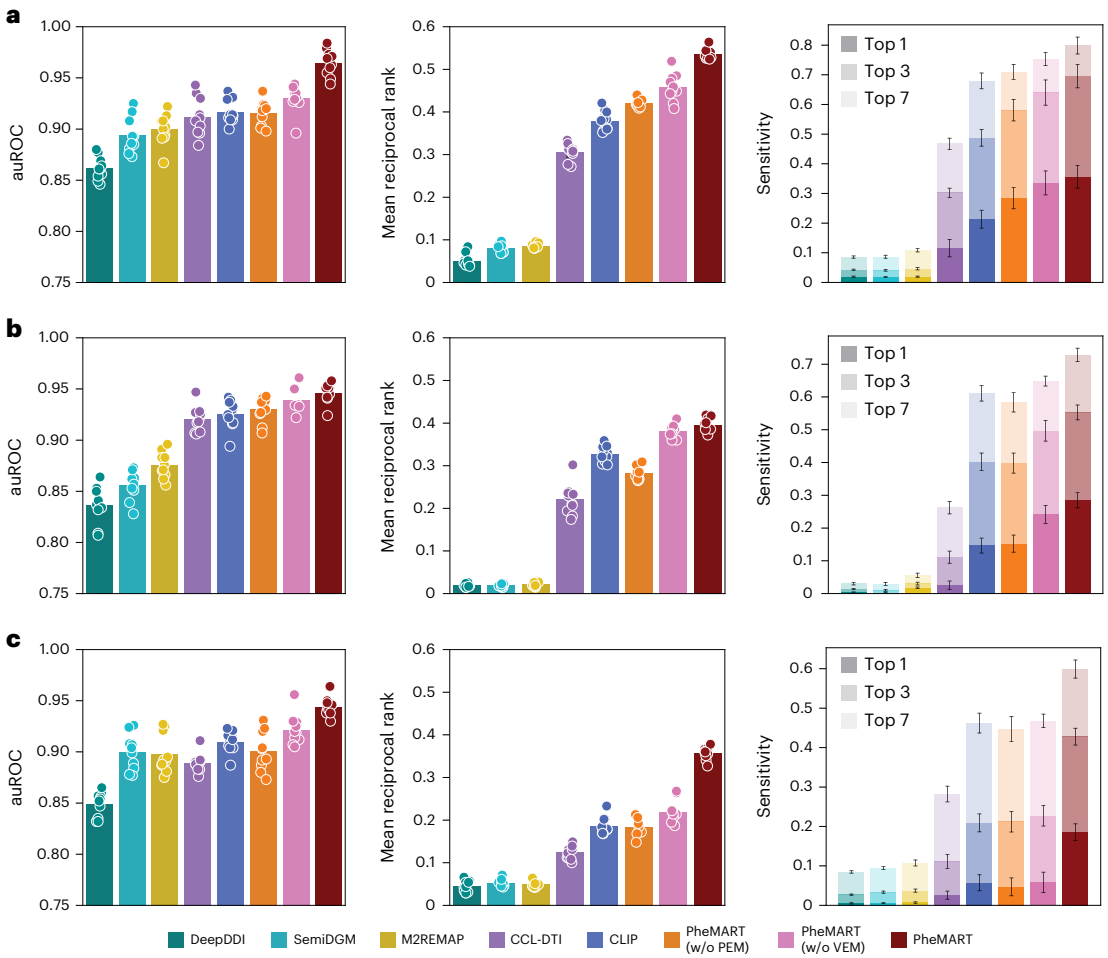

Fig. 2 | Comprehensive evaluation of PheMART's phenotypic predictions. The performances were evaluated using auROC, MRR and sensitivity@*k*. Error bars show the 95% confidence interval of 1,000 bootstrap resamples. **a**, Results from 10-fold per-variant cross-validation on ClinVar[7]. PheMART substantially outperforms the established methods across the three metrics. The contributions of the proposed PEM and VEM were underscored by the notable performance declines when they are removed (indicated as 'w/o PEM' and 'w/o VEM'). PheMART effectively prioritizes the associated phenotypes among the 4,179 investigated phenotypes. It achieves sensitivities of 68.6% and 80.8% for the top-3 and top-7 phenotypic predictions per variant, respectively. **b**, Phenotypic predictions for variants most recently updated in ClinVar (independent testing set). The models were trained on data until year 2022, and the predictions were evaluated on variants with phenotypic reports updated since year 2023. For 28.5% of the variants, the reported phenotypes are ranked the top predictions by PheMART, and for an additional 44.3%, they fall within the top 2–7 predictions. **c**, External validations against unseen variants annotated in HGMD[31]. PheMART accurately predicts the phenotypic effects of those HGMD variants, with 42.8% of reported phenotypes ranking within the top-3 predictions and an additional 17.1% ranking within the top 4–7 predictions.

pairs, covering 952 unique phenotypes and 897 distinct variants. The reported variant–phenotype pairs receive substantially higher prediction scores than the unreported pairs, in which a randomly sampled phenotype was assigned to the involved variant, thus achieving an auROC of 0.948 (Fig. 2b). Notably, PheMART predicts the annotated phenotypes as the top candidate for the associated variants for 28.5% of these pairs. Expanding the evaluation to the top-3 and top-7 phenotypic predictions for each variant leads to increased sensitivities of 55.3% and 72.9%, respectively. These results demonstrate PheMART's potential for generalizing phenotypic predictions to large-scale VUS.

Phenotypic predictions validated by external variant annotations. We further assessed PheMART's generalizability to MVs beyond those catalogued in ClinVar by incorporating additional MVs with phenotypic annotations from HGMD[31]. Our analysis focused on MVs within the 84 genes recommended as clinically actionable by the American College of Medical Genetics and Genomics[10,36] due to their well-established disease associations and highly penetrant mutations. To integrate HGMD annotations into our validation framework, we mapped the reported phenotypes to UMLS Concept Unique Identifiers (CUIs), retaining only those within the 4,179 phenotypes considered in our study. This resulted in 22,124 variant–phenotype pairs, involving 229 phenotypes and 22,124 variants not included in ClinVar. PheMART then predicted the relevance of each of these 22,124 variants to the 4,179 phenotypes under investigation.

PheMART consistently demonstrates superior performance over existing methods, achieving an auROC of 0.947 and an MRR of 0.359, thus outperforming the second-best models by 3.77% and 89.87%, respectively (Fig. 2c). PheMART accurately predicts the phenotypic effects of those variants in HGMD, with 42.8% of the reported phenotypes ranked within the top-3 predictions and an additional 17.1% ranked in the top 4–7 predictions. Notably, for one of the top-performing genes, *CASR*, PheMART predicts that His1834Arg and Ser913Phe are implicated in 'myofibrillar myopathy' and 'hypertrophic cardiomyopathy', respectively, which aligns with the recent findings in ref. 37 and ref. 38 curated in the HGMD.

Ablating PheMART components to reveal key drivers of performance. We systematically investigated the contributions of PheMART's key components that underpin its performance, including PEM, VEM and

contrastive variant–phenotype embedding (CVPE), which bridges variants and phenotypes. First, the integration of knowledge from LLM and EHR data enhances PheMART's ability to comprehensively grasp the semantic interconnectedness of phenotypes (example visualizations are provided in Supplementary Fig. 9a,b), which is evidenced by performance drops after removing PEM (Fig. 2). Neither LLM nor EHR data alone can fully encapsulate the intricacies of phenotype relationships (as evidenced in Supplementary Fig. 2).

Second, VEM enables nuanced discerning of variants' biological impacts (evidenced by the performance drop of PheMART (w/o VEP) in Fig. 2). By contrasting variant features against those of the wild-type protein in a semisupervised manner, VEM accentuates the specific phenotypic effects attributable to variants and allows for greater genotype–phenotype specificity (example visualization is provided in Supplementary Fig. 9c,d). Within the VEM, we have also shown that both the PPIs and protein domain annotations are important for enhancing model generalizations (Supplementary Fig. 2). Furthermore, the choice of PLM is critical for representing missense variants. RGN2 embeddings, which explicitly capture residue-level structural perturbations and are sensitive to subtle conformational changes, offer substantial advantages over alternative embeddings, including ESM-1v[39], ESM2_t30 and ESM2_t33 (ref. 40), which primarily excel at capturing global protein structure and evolutionary constraints (Supplementary Fig. 2).

Lastly, effective contrastive learning is essential to align variants and phenotypes in the metric embedding space in a generalizable manner. In comparison to a cross-entropy loss and the standard contrastive training loss[41], the advantages of the adopted CLIP-derived contrastive loss are substantial (Supplementary Fig. 2). Both cross-entropy loss and the standard contrastive loss are susceptible to label noise. However, since existing reported variant–phenotype pairs are non-exclusive, either random sampling or knowledge-guided negative sampling[28] may inadvertently introduce false-negative pairs, compromising model learning. Overall, these results underscore the importance of nuanced characterizations of both variants and phenotypes and the effective strategy to establish their clinical relevance.

Calibrated PheMART predictions across phenotype groups. To ensure balanced performance across the varied phenotypes, PheMART employs phenotype-aware calibrations. This involves using a univariate logistic regression model that incorporates both the variant–phenotype similarity prediction and the phenotype embeddings (as detailed in Methods). This calibration process finetunes PheMART's predictive capability to accommodate the diversity of phenotypes. The calibrated PheMART scores, which range from 0 to 1, represent approximate probabilities of variants being implicated in specific phenotypes. This probabilistic interpretation provides users with a nuanced understanding of a given variant's likelihood of involvement in particular phenotypes, thereby enhancing the applicability and reliability of PheMART predictions.

Performances of the 12 phenotype groups are mostly balanced in both auROC and MRR, except for the 'Others' group, which exhibits comparatively poorer performance (Supplementary Fig. 1a). This outcome is reasonable given the smaller annotation size and the diverse range of phenotypes in this group. The 'musculoskeletal', 'metabolic' and 'circulatory' groups perform the best, while the 'immune', 'genitourinary' and 'neurological' groups perform relatively poorer. The overall performance across the 12 phenotype groups is effectively calibrated (Supplementary Fig. 1b), resulting in an average estimated calibration error of 0.07 (ref. 42). This indicates the reliability of PheMART's predictions across diverse phenotype categories.

PheMART's phenotypic predictions resonate with established biological knowledge

To understand the properties of amino-acid substitutions identified by PheMART when linking variants to phenotypes, we analysed the phenotypic predictions of a large set of missense variants across diverse functional domains.

Overall associations between protein functional domains and phenotypes. We analysed PheMART's phenotypic predictions on the 583,722 missense variants catalogued in ClinVar[7]. Focusing on high-confidence predictions (positive predictive value = 0.9 after calibration), we identified 159,960 variant–phenotype pairs. The protein domain information of the variants was obtained from UniProt[24], if available. To explore the distribution of these high-confidence variants across functional domains, we considered the top 50 domains containing the greatest number of these variant–phenotype paris and then visualized their normalized contributions across the 12 phenotype groups (see Supplemenary Section 1.9 for details).

PheMART's predictions resonate well with the established biological understandings (Fig. 3a). For example, variants in the ion transport domain contribute predominantly to neurological and circulatory disorders, aligning with the canonical roles of ion channels in generating action potentials for nervous system conduction and cardiac contractions[43–45]. In the study of respiratory phenotypes, we found an enrichment of pathogenic variants within the AAA+ ATPases domain, which is essential for protein quality control, particularly within the mitochondria[46], and contributes to the maintenance of cellular health in respiratory tissues. This aligns with emerging evidence that mitochondrial dysfunction is increasingly recognized as a crucial factor in the pathogenesis of several lung diseases, including chronic obstructive pulmonary disease, asthma and lung cancer[47].

Our analysis also revealed a notable connection between mutations in the myosin motor domain and circulatory diseases. Given the fundamental role of myosin in cardiac muscle contraction and its involvement in cardiac remodelling over time, this discovery is supported by its known associations with several heart conditions, such as familial cardiomyopathies, left ventricular non-compaction[48] and electrophysiological anomalies[49]. In addition, for immune-related phenotypes, an enrichment of pathogenic variants was found in the SPRY/B30.2 domain, which plays a vital role in managing both innate and adaptive immune responses[50].

PheMART provides phenotype-specific pathogenicity predictions. We conducted a detailed analysis of phenotype-specific pathogenicity predictions at the gene level, revealing strong mechanistic correlations between mutated domains and their associated phenotypes (Supplementary Fig. 4). For example, on cyclin-dependent kinase-like 5, a protein kinase essential for neurological development, we identified a predominance of mutations within its kinase domain, which were predicted to contribute to neurological disorders and epilepsy. This finding underscores the sensitivity of neurological pathways to disruptions in kinase activity[51]. Similarly, the Sodium Voltage-Gated Channel Alpha Subunit 5 (SCN5A) exhibited high pathogenicity densities in its voltage-gated ion channel domain, a key regulator of cardiac electrical activity. Mutations in this domain are strongly associated with arrhythmias and sudden cardiac arrest, hallmark features of Brugada syndrome[52]. Comparable patterns were also observed in its homologous family members, SCN9A and SCN8A. Developmental disorders exhibited a high concentration of pathogenic mutations in critical developmental domains. For Twist Family BHLH Transcription Factor 1 (TWIST1), mutations within the bHLH domain were associated with Saethre–Chotzen syndrome, a craniofacial disorder[53]. Similarly, mutations in the homeobox domain of Meis homeobox 2 (MEIS2) were linked to congenital anomalies, including cleft lip, microcephaly and cardiac malformations[54,55]. Pathogenic mutations in the homeobox domain of Short Stature Homeobox (SHOX) were also associated with Leri–Weill dyschondrosteosis and hereditary neuropathy. Metabolic disorders were predominantly linked to mutations in key metabolic genes. For instance, mutations in the peripheral subunit-binding

https://doi.org/10.1038/s41551-026-01636-4

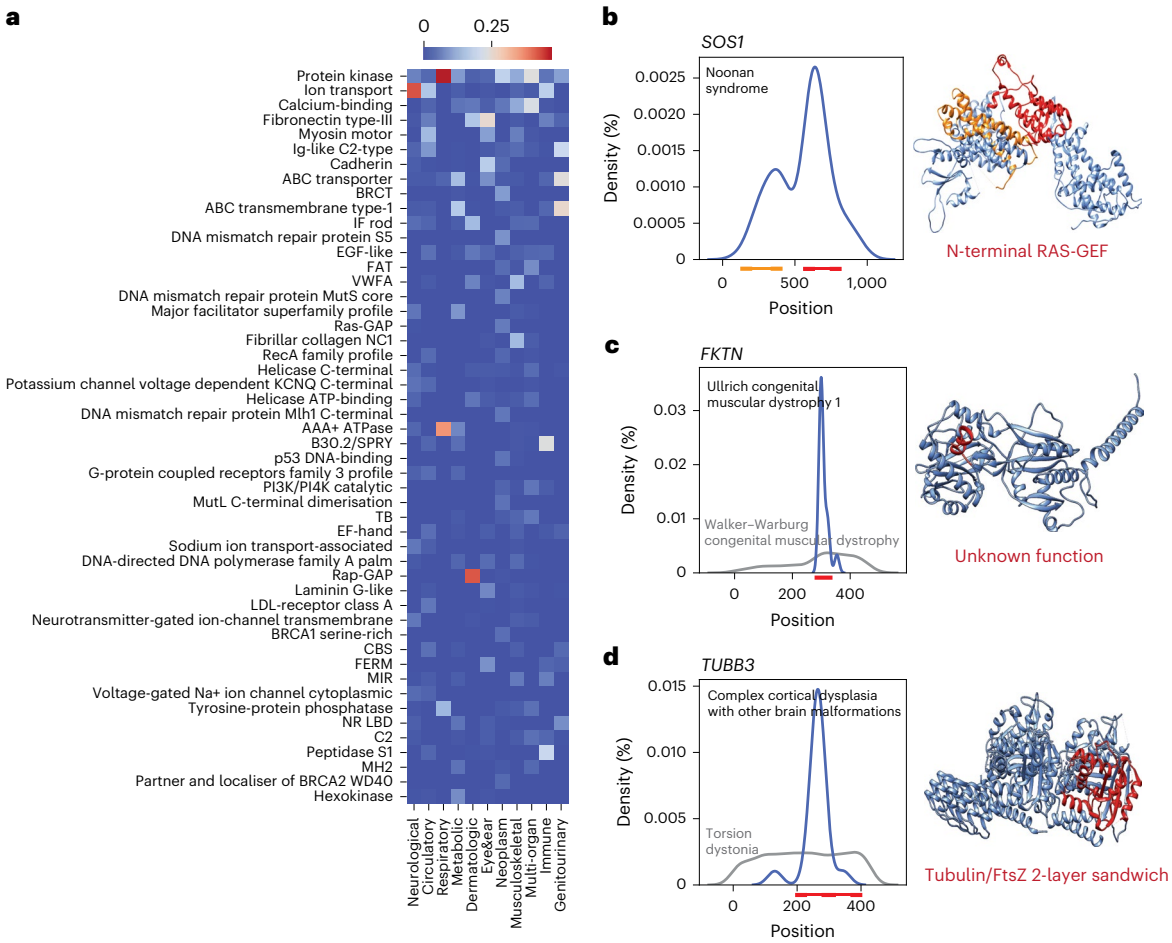

Fig. 3 | PheMART's phenotypic predictions resonate with established biological knowledge. a, PheMART uncovers associations between phenotype categories and protein functional domains. Through the comprehensive prediction of missense variants catalogued in ClinVar, we obtained 159,960 high-confidence variant–phenotype pairs. For each of the 12 phenotype groups, we then visualized the normalized contributions of variants across the top 50 associated protein domains (details are provided in Supplementary Section 1.9), which affirms the biological underpinnings and validity of PheMART's predictions. **b**, PheMART identifies enriched pathogenic variants in the N-terminal Ras-GEF domain of *SOS1* (*Son of Sevenless Homolog 1*), aligning with the existing knowledge on the involvement of the RAS signalling pathway in Noonan syndrome. **c**, On *FKTN*, PheMART identifies regions harbouring a high density of potential pathogenic variants for both Walker–Warburg syndrome and Ullrich congenital muscular dystrophy, suggesting putative novel functional domains. **d**, On *TUBB3*, PheMART reveals distinct patterns of pathogenic density distribution across the two different phenotypes, especially in the Tubulin/FtsZ 2-layer sandwich domain.

domain of Peroxisomal Biogenesis Factor 1 (PDX1), a critical regulator of carbohydrate metabolism, were associated with maturity-onset diabetes. Likewise, pathogenic variants in the forkhead domain of protein Forkhead Box P3 (FOXP3), a key regulator of insulin signalling, were implicated in insulin-dependent diabetes[56].

The mapping of pathogenic missense variants to specific clinical phenotypes could also deepen our understanding of the molecular mechanisms underlying disease development. For example, it identifies enrichments of pathogenic variants associated with Noonan syndrome in the N-terminal Ras-GEF domain (Fig. 3b). This domain is crucial because it interacts with GTPase, an enzyme that plays a pivotal role in the RAS signalling pathway essential for cell division and growth. Variations in this domain can disrupt the normal function of the pathway, leading to the developmental and growth anomalies seen in Noonan syndrome. By pinpointing these variants, PheMART confirms the involvement of the RAS pathway in Noonan syndrome and offers insights into how the signalling disruptions caused by these variants contribute to the disease's development. Meanwhile, the predictions could also facilitate identifying functional protein regions that may not be well characterized in existing databases such as UniProt[24]. For example, PheMART has identified a region within the FKTN (Fukutin) (amino acids 285–310) that harbours a high density of potential pathogenic variants linked to Walker–Warburg syndrome and Ullrich congenital muscular dystrophy (Fig. 3c), although these regions do not have documented functions in the current literature. This discovery suggests new functional domains within FKTN and adds new insights into the pathological mechanisms of the disorder.

The ubiquity of pleiotropy, where a single gene influences multiple disorders, complicates the clinical assessment of variant pathogenicity. PheMART addresses this challenge by providing disease-specific identification of pathogenic variants, enabling more precise and actionable clinical interpretations of patient conditions. For example, PheMART reveals distinct patterns of pathogenic distribution across phenotypes associated with the gene *Tubulin Beta 3 Class III* (*TUBB3*) (Fig. 3d). Variants linked to torsion dystonia are dispersed across the first 400 amino acid positions, whereas pathogenic variants associated with complex cortical dysplasia are predominantly enriched within the FtsZ2-layer sandwich domain. This domain is located in the C-terminal region of tubulin proteins which are vital for the polymerization and formation of microtubules and essential for neuronal development. This finding aligns with existing research indicating that mutations in tubulin genes such as *Tubulin α 1a* (*TUBA1A*), *Tubulin β Class I* (*TUBB*)

and *Tubulin γ* (*TUBG*) are implicated in a spectrum of cortical malformations and neurodevelopmental disorders, collectively referred to as tubulinopathies[57–59].

Gene-level enrichment analysis on PheMART's predictions. We aggregated phenotypic predictions at the gene level for enrichment analysis[30]. Specifically, for each target phenotype, we performed pathway enrichment analysis (false discovery rate (FDR) < 0.05, Benjamini–Hochberg) on genes harbouring missense variants predicted to be implicated in that phenotype. The results demonstrate that PheMART's predictions align with the biological mechanisms underlying phenotype development (Supplementary Fig. 5). For example, in developmental and epileptic encephalopathy, a neurological disorder, the implicated variants originate from 19 genes, which are enriched in synaptic function and neurodevelopmental pathways ($q = 5.5 \times 10^{-5}$, which represents the FDR-adjusted p value after multiple hypothesis testing correction; fold enrichment (FE) = 5.4). The involved genes include the *Chromodomain Helicase DNA Binding Protein 1* (*CHD1*) and *Chromodomain Helicase DNA Binding Protein 2* (*CHD2*), which act as chromatin remodellers and play roles in brain development. Also, voltage-gated potassium channels are involved, including *Potassium Voltage-Gated Channel Subfamily A Member 2* (*KCNA2*), *Potassium Voltage-Gated Channel Subfamily B Member 1* (*KCNB1*) and *Potassium Voltage-Gated Channel Subfamily Q Member 2* (*KCNQ2*), which regulate neuronal excitability and action potential generation. Similarly, in the metabolic disorder mitochondrial DNA depletion syndrome, the 11 associated genes, including *DNA Polymerase γ, Catalytic Subunit (POLG)* and *Ribonucleotide Reductase Regulatory TP53 Inducible Subunit M2B (RRM2B)*, show strong enrichment in energy metabolism and mitochondrial maintenance pathways, which is consistent with their roles in mitochondrial DNA synthesis and maintenance. For hereditary breast–ovarian cancer, associated variants span 10 genes, including *BRCA1 DNA Repair Associated (BRCA1)*, *Partner and Localizer of BRCA2 (PALB2)* and *RAD51 Paralog C/D (RAD51C/D)*, showing expected enrichment in telomere maintenance ($q = 4.0 \times 10^{-6}$, FE = 508.5) and homologous recombination-based DNA repair ($q = 1.50 \times 10^{-6}$, FE = 72.2), the defects of which lead to genomic instability and increased susceptibility to cancer. Furthermore, in the eye disorder retinitis pigmentosa, 29 associated genes show a high enrichment in visual phototransduction, including *Rhodopsin (RHO)*, *Phosphodiesterase 6A/B/C (PDE6A/B/C)* ($q = 2.5 \times 10^{-5}$, FE = 59.5), and retinal development pathways ($q = 8.4 \times 10^{-11}$, FE = 43.0), that is, *Crumbs Cell Polarity Complex Component 1 (CRB1)* and *Nuclear Receptor Subfamily 2 Group E Member 3 (NR2E3)*, which is consistent with the disruption of photoreceptor function and retinal degeneration observed in the disease. These functionally coherent enrichments further support the biological validity of PheMART's phenotypic predictions.

PheMART aids in genetic diagnosis of rare clinical disorders

By linking pathogenic missense variants to large-scale clinical phenotypes, PheMART provides greater clinical actionability than existing methods that solely classify variants as benign or pathogenic. For patients with genetic disorders, it not only aids in establishing clinical diagnoses but also pinpoints the causal variants (following the workflow in Fig. 4a).

PheMART pinpoints causal missense variants in patients with genetic disorders. A common scenario for an individual with a suspected genetic disorder is when a specific clinical diagnosis can be established through clinical review, enzymatic or metabolomics testing, imaging studies or tissue biopsy, but the causal genetic variant(s) underlying the disorder remain(s) unknown[60]. Pinpointing the disease-causal variant(s) can be critical for determining a patient's eligibility for clinical trials, potential therapeutic management and predicted disease prognosis, and can also enhance our understanding of the molecular mechanisms driving the disorders.

To evaluate PheMART's effectiveness in identifying disease-causing variants, we curated a set of 31 diagnosed patients with validated disease-causing missense variants from UDN[29]. These patients have extensive clinical profiles and complete sequencing data. Typically, disease-causing variants in such cases are validated through extensive experimental assays, model organism studies or by identifying additional genotype- and phenotype-matched individuals using services such as Matchmaker Exchange[61].

When ranking all missense variants by the proximities of their PheMART-predicted phenotypes to an individual's clinical diagnosis for each case, we found that the causal variant was the top-ranked variant in 58% of cases and within the top-3-ranked variants in 77% of cases (Fig. 4b). In comparison, AlphaMissense[10], which identifies phenotype-agnostic pathogenic variants based solely on sequencing data (comparison results are provided in Supplementary Section 1.12), and ClinPrior[62], which combines patients' sequencing data with phenotype information for variant prioritization, both fall short of PheMART's performance (Fig. 4b). For example, in a patient clinically diagnosed with early infantile epileptic encephalopathy-64, PheMART correctly prioritized the true causal variant Rho-related BTB Domain-Containing Protein 2 (RHOBTB2) (p.Arg483His) as the top candidate. Similarly, in a patient with cystic fibrosis, the true causal variant, Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) (p.Arg117His), was ranked first by PheMART among 51 prioritized variants (Fig. 4e). By considering pairs consisting of a diagnosis and a causal variant as positives, and those involving a diagnosis and a non-causal variant as negatives, PheMART achieves an auROC of 0.956, further highlighting PheMART's promise for clinical diagnosis.

PheMART facilitates diagnosing challenging clinical disorders. Missense variants are predominantly implicated in rare monogenic diseases that collectively impact millions of patients[3,4]. Identifying these conditions is particularly difficult because they often lack unique distinguishing features or standard laboratory tests that can provide definitive confirmation of suspected genetic abnormalities[61,63]. A staggering 70% of affected individuals actively seek a diagnosis, yet up to 50% of suspected conditions remain elusive[2,64]. An accurate and early diagnosis can result in better disease management, counselling options and identification of potential therapeutics, and avoid unnecessary treatments that may have severe side effects. In such cases, genetic information has been shown to provide valuable insights that can aid in pinpointing the underlying diseases[63,65].

Leveraging PheMART's phenotypic predictions, we proposed a workflow to nominate diagnoses for patients with rare diseases whose diagnosis was clinically challenging (Fig. 4a). It consists of three steps: variant prioritization, diagnosis nomination by PheMART and nomination refinement based on patients' reported symptoms. We validated its clinical promise by assembling a cohort comprising 83 patients from UDN[29] (demographic information is provided in Supplementary Fig. 3). Each patient underwent comprehensive sequencing and was provided with a set of variants prioritized by the UDN team through a combination of techniques encompassing allele frequency prevalence in control populations, pathogenicity prediction and consideration of the most probable inheritance pattern[61]. Diagnoses were generated on the basis of established knowledge of genotype–disease associations or recognized diagnostic guidelines[60]. Among the 83 patients, the UDN team assigned 82 unique diagnoses. For each patient, we derived potential diagnoses by integrating PheMART predictions with their reported symptoms. Prioritized variants were fed into PheMART to generate 4,179-dimensional phenotype scores, which were aggregated by taking the maximum across all variants to produce the initial diagnostic predictions. We further refined the diagnostic predictions through a two-step sequential process, enhancing alignment between

https://doi.org/10.1038/s41551-026-01636-4

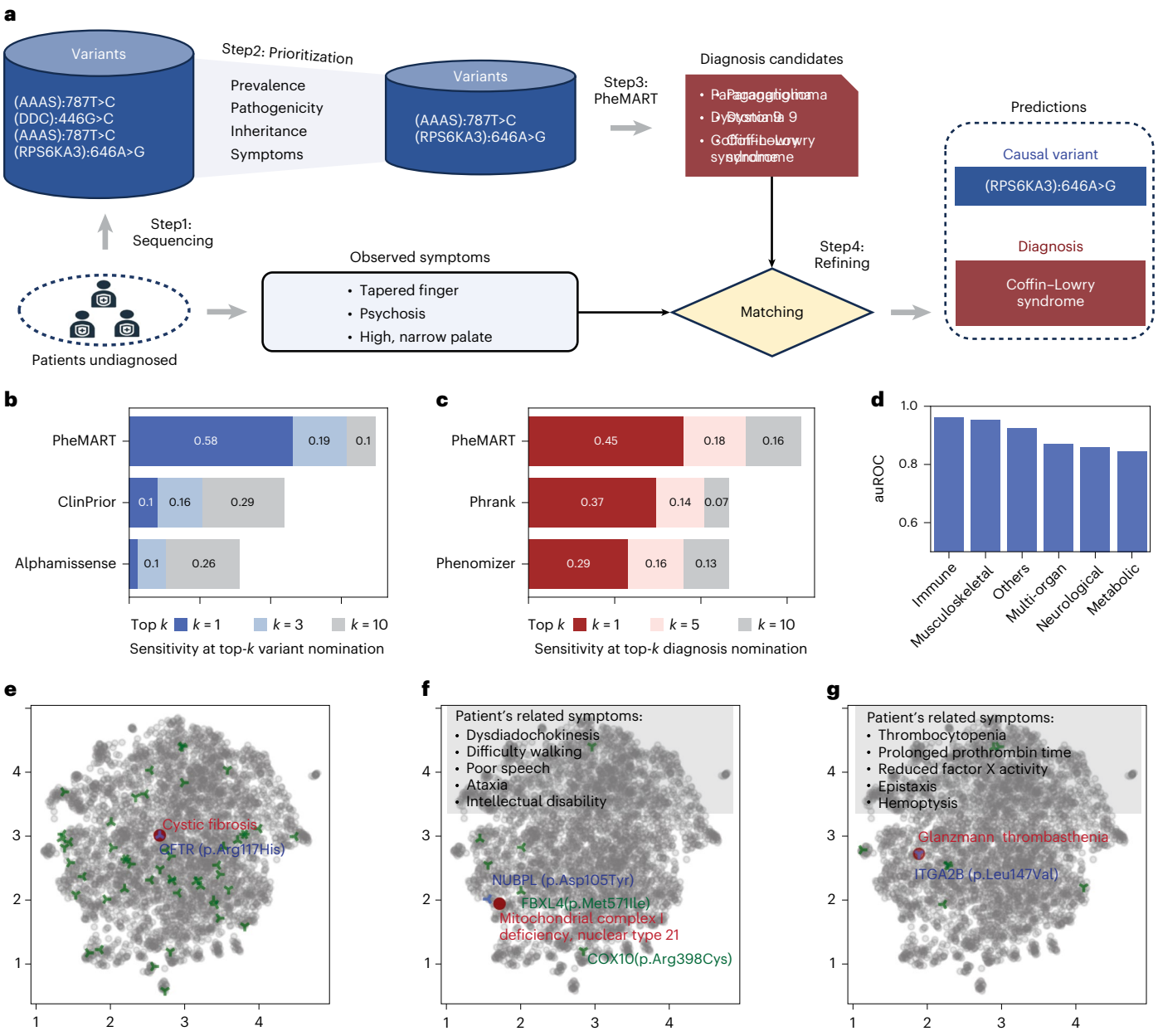

Fig. 4 | PheMART aids in diagnosing patients with rare genetic diseases. **a**, Workflow of PheMART to nominate diagnoses for patients with rare disorders. Following the guidelines[61], we obtained prioritized variants for each patient by combining sequence alignment, prevalence control, pathogenicity prediction, patient traits and considerations of inheritance patterns[61]. By projecting the prioritized variants and diseases into the metric embedding space, PheMART generates diagnosis candidates, which were filtered and re-weighted on the basis of patients' observed symptoms. **b**, PheMART nominates causal variants for patients with rare diseases. We curated a dataset of 31 patients from the UDN who have been diagnosed and have their causal missense variants validated. PheMART effectively pinpoints the causal variants, achieving a top-1 accuracy of 0.58 and a top-3 accuracy of 0.77. **c**, Performance of PheMART generating diagnosis nominations on a dataset consisting of 83 UDN patients[29]. PheMART demonstrates substantial improvements from existing methods. For 44.6% of the individuals, their clinical diagnoses are the top nominations by PheMART, 18.1% among the top 2–5, and 15.7% among the top 6–10. **d**, The auROC of PheMART across different disease categories in the diagnosis nomination. **e**, PheMART nominates the causal variant, CFTR(p.Arg117His), as the top candidate for a patient diagnosed with cystic fibrosis (observed variants are marked as triradiates in the embedding visualizations via UMAP[95]). **f**, For a patient with 7 variants prioritized by the UDN team[61], the given diagnosis, Mitochondrial complex I deficiency, nuclear type 21, is the top diagnosis nomination by PheMART. **g**, For an individual clinically undiagnosed, PheMART nominates a putative diagnosis that is supported by the observed symptoms.

PheMART predictions and the patients' reported symptoms (Fig. 4a). First, we required the diagnoses to have at least one main trait matched to patients' reported symptoms. To this end, we curated their main traits by integrating knowledge from Orphanet[66] and OMIM[67]. These symptoms are standardized as Human Phenotype Ontology (HPO) terms[68]. Then, we reweighed the prediction score of each candidate diagnosis based on the similarity between its set of main traits and the set of patients' reported symptoms. The similarity was obtained by calculating their best-match average[69] of the similarities in the CODER[25] embedding vectors.

PheMART identifies the correct clinical diagnoses for 44.6% of the patients with its top nomination out of the 4,179 disease candidates, substantially outperforming Phrank[70] and Phenomizer[71]. For example, PheMART nominates the given diagnosis, mitochondrial complex I deficiency, nuclear type 21, as the top candidate for a patient with 7 prioritized variants (Fig. 4f). For 18.1% of the patients, their diagnoses

rank among the 2nd to 5th nominations, while 15.7% of the diagnoses position within the 6th to 10th nominations (Fig. 4c). The predictive scores effectively distinguish true diagnoses from unrelated disorders across diverse phenotype groups, with particularly strong performance in 'musculoskeletal' and 'immune' groups, achieving auROC values of 0.961 and 0.953, respectively (Fig. 4d). However, it is important to highlight a discernible decline in the accuracy of the diagnosis nominations relative to phenotypic predictions based on ClinVar annotations, which can be attributed primarily to two factors. First, there are potential omissions of certain disease-relevant traits in patient reports or the disease–trait curations, which may compromise the thoroughness of the diagnosis refinement process. Second, some diagnoses nominated by PheMART may not yet be clinically manifested in patients, in which case PheMART's nominations could be instrumental in guiding future diagnostic endeavours or disease risk management. Given the promising results, PheMART could be used to generate diagnosis nominations for patients still undiagnosed and facilitate the diagnosing process. For example, on the basis of the 7 prioritized variants of a patient, PheMART nominates Glanzmann thrombasthenia as the top candidate diagnosis, which is supported by the patient's observed symptoms including prolonged prothrombin time, reduced factor X activity and epistaxis (Fig. 4g).

PheMART predictions as a community resource

We provided phenotypic predictions for 5.1 million amino acid alterations across 2,367 genes that were identified as 'likely pathogenic' by AlphaMissense[10]. Each mutation was associated with calibrated scores for 4,179 distinct phenotypes. Employing cut-off values that achieve 90% precision in 10-fold cross-validation, we identified a total of 1.44 million variant–phenotype pairs covering 1.1 million variants (all data downloadable as shown in Supplementary Fig. 8). Our analysis revealed that 28.9% of these pairs were involved in neurological disorders, 16.5% in metabolic conditions and 12.6% in musculoskeletal disorders (Fig. 5a). It is important to note that users can tailor the precision cut-offs to meet their specific requirements. In addition, we provided gene-level phenotypic predictions by aggregating the predictions across all missense variants within each gene (see the high-confidence predictions on the top 50 ClinVar genes in Fig. 5b). Also, 61.2% of the genes were found to be disproportionally involved in at least two phenotype categories, further underscoring the necessity of considering specific phenotypic impacts beyond merely categorizing predictions as pathogenic or benign. Of these 1.44 million variant–phenotype pairs, there are 5,125 unique gene–phenotype combinations involved, which are visualized by phenotype groups in Fig. 5c (4 phenotype groups) and Supplementary Fig. 7 (8 phenotype groups). Notably, 33.8% of these combinations have not been previously documented in ClinVar[7], indicating that these genes have not been associated with any closely related phenotypes (with a CODER cosine similarity above 0.7). These newly uncovered gene–phenotype combinations shed light on the gene's functional implications at a broader level. For example, while *ATPase Na^+/K^+ Transporting Subunit α 1 (ATP1A1)* and *ATP1A2*, which encode the alpha subunits of the Na^+/K^+ ATPase transporter, are known to be highly related to neurological disorders such as familial hemiplegic migraine[72], PheMART also reveals their potential involvements in dystonia (Fig. 5c), which is consistent with recent case reports[73]. On gene *Carbamoyl-Phosphate Synthetase 2, Aspartate Transcarbamylase and Dihydroorotase* (*CAD*), which encodes a trifunctional enzyme catalysing the initial steps of de novo pyrimidine biosynthesis, variants were predicted to cause congenital hyperammonemia type I (Fig. 5c), which aligns with the essential role of *CAD* in nucleotide metabolism and urea cycle[74]. The predictions also highlight functional correlations within gene families. For example, variants on *GATA Binding Protein 2 (GATA2)* are widely reported to cause primary lymphedema with myelodysplasia for which PheMART identifies enriched variants on gene *GATA3* (Fig. 5c), consistent with *GATA3*'s crucial function in T lymphoid cell development and immune regulation[75,76].

Discussion

In an individual's genome, there are typically over 9,000 missense variants that result in amino acid changes compared to a reference genome[77]. Accurately linking these variants to specific phenotypes is essential for making informed clinical decisions. Current tools for prioritizing genetic variants have successfully reduced a vast pool of potentially deleterious variants to a more manageable few hundred candidates. However, this number remains overwhelmingly large for practical clinical actions or experimental analyses. To address this problem, we introduced PheMART, an in silico tool designed to link missense variants to specific phenotypes. By mapping mutant protein sequences and 4,179 phenotypes into a unified metric space, PheMART provides comprehensive phenotypic characterizations for each variant, thereby enhancing the precision and practicality of variant prioritization in clinical and research settings.

PheMART stands out as an in silico approach in elucidating the phenotypic implications of missense variants, thus providing substantial advantages over existing methodologies. It eliminates reliance on population-wide data from GWAS for interpreting the phenotypic effects of missense variants, particularly in the context of rare phenotypes. PheMART surpasses traditional binary assessments of variant pathogenicity by offering nuanced and clinically relevant insights, aiding in patient diagnosis and the identification of causative variants. Although there have been computational efforts to predict the phenotypic consequences of genetic variants[15], these efforts have primarily focused on molecular-level impacts, which often do not directly translate to actionable clinical outcomes.

As an initial effort to computationally link missense variants to clinical phenotypes, PheMART revealed three critical insights. First, while pretrained large-scale PLMs are acknowledged for their use in predicting the pathogenicity of missense variants, PheMART highlights that relying solely on pretrained variant features is insufficient for effectively discerning subtle phenotypic distinctions among variants sharing the same wild-type sequence. Enhancing this approach by contrasting variant features with their wild-type counterparts substantially improves phenotypic differentiation. Second, incorporating existing knowledge to model phenotypic interrelations is pivotal. Given the limitation of variant data availablity for the majority of phenotypes, leveraging the semantic relationships among phenotypes during model training markedly boosts predictive accuracy. Lastly, devising effective learning strategies to connect variants and phenotypes is critical. Since only positive variant–phenotype pairs are reported, both negative sampling and knowledge-guided negative sampling (as in ref. 28) may inadvertently introduce false negatives, compromising model learning when using either a cross-entropy or standard contrastive learning objective[41]. By using a dynamic negative sampling strategy, CLIP-based contrastive training enhances the representation of variants and phenotypes by simultaneously contrasting them against multiple negatives. This approach ensures that the learned embeddings exhibit stronger generalization capabilities to unannotated variants.

Existing clinical workflows to derive patient diagnoses from sequencing data typically involve extensive collaboration among clinicians, geneticists and bioinformatics experts. While these multidisciplinary interactions are invaluable, they may not be readily accessible in resource-constrained settings, especially in developing or underdeveloped regions. PheMART's capability to leverage sequencing data for diagnostic purposes has the potential to greatly simplify the diagnostic process. It narrows down the potential disease spectrum and reduces the need for extensive genetic expertise among frontline clinicians. Furthermore, PheMART's ability to identify causal variants in patients with established or candidate diagnoses offers important clinical benefits. Pinpointing these variants enables more personalized

https://doi.org/10.1038/s41551-026-01636-4

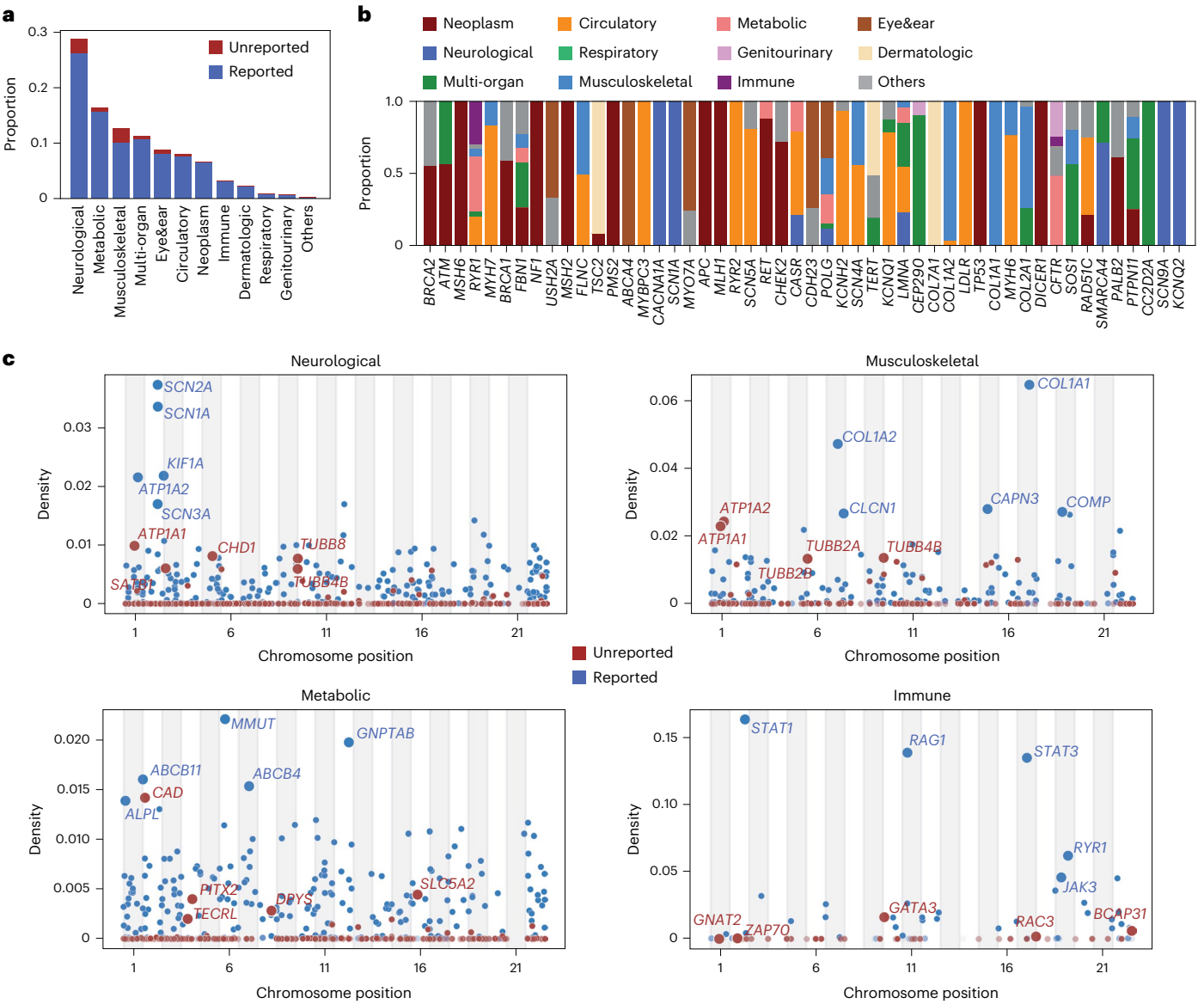

Fig. 5 | PheMART's phenotypic predictions as a community resource. **a**, Phenotype distributions of variants predicted with high confidence (HC), that is, employing a cut-off value that achieves a positive predictive value of 0.9 in 10-fold cross-validation. We provided phenotype distributions for the 5.1 million single-amino-acid alterations identified as 'likely pathogenic' by the AlphaMissense tool[10]. Among the 1.44 million HC variant–phenotype pairs, there are 5,125 gene–phenotype pairs involved, 45.8% of which have not been reported in ClinVar[7] (detailed distributions are provided in Supplementary Fig. 7). **b**, Examination of missense variants' HC phenotypic predictions in prominent ClinVar genes. We visualized the distributions of the associated phenotypes on the top 50 genes based on the number of HC variant–phenotype predictions. We observed that a notable portion, 61.2%, of these genes are implicated in more than two phenotype categories, and the variants are disproportionately pathogenic across phenotype categories. **c**, PheMART's HC phenotypic predictions aggregated at the gene level across four disease categories (blue for known gene–phenotype associations and red for discovered pairs unreported in ClinVar). Supplementary Fig. 7 provides the phenotypic predictions for the other phenotype categories. These newly uncovered gene–phenotype associations shed light on the gene's novel functional implications, for example, *ATP1A1* in dystonia[73] and *GATA3* in primary lymphedema with myelodysplasia[75,76].

and effective treatment strategies and interventions. In addition, this capability provides essential information for genetic counselling, aiding in informed decisions.

As a valuable resource for the scientific community, PheMART provides extensive phenotypic predictions, offering a unique perspective to delve into the molecular underpinnings of diseases and the functional roles of genes. By pinpointing gene regions with a high density of pathogenic variants, PheMART facilitates a deeper investigation into the molecular functions of these critical areas, such as their roles in protein interactions or catalytic activities, thereby illuminating the molecular basis of associated diseases. Furthermore, by projecting the phenotypes into the metric space, PheMART also reveals new relationships between phenotypes that are connected at the molecular level (Supplementary Fig. 6), potentially opening up opportunities for repurposing existing drugs for a variety of rare diseases that currently lack specific treatments.

Despite its strong performance, PheMART has several limitations. It does not provide mechanistic insights into how variants perturb protein stability, folding, molecular interactions or enzymatic activity[78,79], or how such molecular changes propagate through cellular and physiological pathways to produce phenotypic outcomes. Because the phenotypic impact of a variant is intrinsically linked to the native function of its corresponding wild-type protein, incorporating experimentally derived functional annotations (such as structural stability

metrics, interaction perturbation profiles and domain-specific activity changes) could enhance both interpretability and predictive accuracy. Model performance is also sensitive to the choice of pretrained PLM (Supplementary Fig. 2) and declines markedly when training and evaluation are conducted on variants from non-overlapping domains (Supplementary Section 1.10 and Fig. 10), underscoring the importance of domain context and the need for models that are explicitly optimized to capture the fine-grained structural perturbations induced by single-amino-acid substitutions. In addition, PheMART's generalizability is limited to the 4,179 phenotypes with reported genetic associations and to genes with existing annotations. The current approach for linking variants across genes, which is based primarily on contrastive learning over protein–protein interactions and shared domain annotations, provides only a coarse approximation of cross-gene relationships and does not fully capture convergent pathogenic mechanisms that span multiple, functionally related genes (Supplementary Fig. 11). Finally, the framework is currently limited to missense variants, excluding other clinically important variant types such as insertions, deletions and frameshifts, which are also major contributors to genetic disorders.

Recent advancements in LLMs have demonstrated their remarkable capability in learning complex representations across various domains, particularly in understanding both natural languages and protein sequences. The findings of this study underscore the promise of leveraging LLMs to systematically characterize genetic variants and their phenotypic consequences. Building on PheMART's foundation, future work could integrate additional biomedical modalities, especially transcriptomic and functional data, to achieve a more mechanistically grounded framework for variant interpretation. Transcriptomic profiles, including gene expression outliers and allele-specific expression patterns[80–82], could provide critical context on regulatory consequences. In addition, high-throughput mutational scanning[9] provides direct evidence of altered protein activity, stability and interactions, which could further improve the accuracy of phenotypic effect predictions. Incorporating these data sources could enable a more mechanistically informed framework for variant interpretation. Moreover, PheMART currently predicts the phenotypic effects of each variant independently, without explicitly modelling their combinatorial influences on phenotypes. While aggregating predictions across multiple variants within an individual at the phenotype level provides a practical approximation, this approach does not fully capture fine-grained variant interactions. Developing models that explicitly integrate interactions at both the molecular and phenotypic levels would not only enhance predictive accuracy but also improve the clinical use of these predictions, enabling more precise diagnoses and personalized therapeutic strategies.

Methods

PheMART projects both variants and phenotypes into a low-dimensional metric embedding space, where the distance between them reflects clinical relevance. For each variant–phenotype pair, PheMART produces a score within the range [0,1], where a score of 1 indicates a strong likelihood that the variant contributes to the phenotype's development, and a score of 0 indicates no relevance. Variants are represented as inputs using two amino acid sequences: the altered amino acid sequence corresponding to the variant and the wild-type protein sequence. Phenotypes are represented by their associated text descriptions, which are transformed into LLM and EHR embeddings through medical embedding techniques[25,26]. We provide further details on data sources, preprocessing steps and the operational workflow of PheMART below.

Data sources and preprocessing

To train and validate PheMART, we used five primary data sources, along with pretrained PLMs[83]. We integrated summary-level EHR data from the Veteran Affairs (VA) and Mass General Brigham Biobank (MGBB), and UMLS database[20] to represent phenotypes. We used the ClinVar database[7] for model training and cross-validation. In addition, the UDN patient-level data[29] was used to showcase the clinical significance of PheMART in aiding the diagnosis of patients with rare diseases.

EHR data and UMLS database. The clinical characteristics of phenotypes were captured by combining the summary-level EHR data from the VA cohort containing 60,106 clinical concepts and 12.5 million patients, and the MGBB consisting of 84,995 clinical concepts and 60,000 patients[26]. A matrix factorization variant of the skip-gram algorithm was performed on MGBB and VA summary-level data separately to derive two sets of EHR-based embedding vectors, which were fused into one set of embeddings for a total of 61,009 EHR concepts via a block-wise overlapping noisy matrix integration (BONMI) algorithm[26]. These embeddings, denoted as BONMI-EHR, delineate the clinical relationships between phenotypes and other medical concepts including disease symptoms, medications, laboratory tests and medical procedures[28].

The UMLS database contains 4.27 million concepts, 15.48 million terms with several synonymous terms representing the same concept, and 87.89 million relations with 127 semantic types, 14 relation types and 923 relationship attributes[20]. These expert-curated relations were condensed via a deep language model, CODER[25,84], which is a BERT-based model further trained through contrastive learning based on the known synonymous terms and relationships in the UMLS.

ClinVar database. PheMART's labelled training data were sourced from the ClinVar database, with the phenotypic reports of missense variants up to 10 April 2023[7]. To ensure label reliability, reports with a review status of 'no assertion criteria provided' were excluded. In addition, we focused on genetic variants in this database that result in amino acid changes with a maximum length of 1,024 because the pretrained language models used have a maximum protein sequence length constraint[83]. Phenotypic reports on ClinVar were submitted by various organizations, hence the same phenotype can have different reported names. To efficiently harmonize the different wordings of synonyms, abbreviations and other non-standardized terms, we mapped these descriptions to UMLS CUIs[20]. The harmonization process comprises two main steps. First, we performed text matching to map 2,214 phenotypes out of the 11,023 unique phenotypes in ClinVar to their corresponding CUIs. Subsequently, we employed the LLM-based CODER embedding vectors[25,84] to enhance the matching process. This involves inputting each ClinVar phenotype text string into CODER to obtain an embedding vector and prioritizing the CUIs candidates on the basis of the embedding similarities. Finally, we manually inspected the matching between the ClinVar strings and the UMLS-provided CUI strings to ensure quality. The resulting dataset comprises 30,657 variant–phenotype pairs, encompassing 18,986 unique variants from 2,573 genes and 4,179 distinct phenotypes. This harmonization process ensures that the phenotypic data used in PheMART are comprehensive and standardized, allowing for reliable variant–phenotype association studies.

UDN. At the time of our validation study, the UDN dataset was composed of 12 clinical research centres across the USA where patients underwent an in-depth clinical evaluation, and clinical researchers leverage whole genome sequencing with other omics technologies and a wide array of diagnostic tools to investigate complex and elusive medical cases that have remained unresolved. Before applying to the UDN, most patients have experienced extensive testing by various clinicians. The UDN prioritizes accepting patients whose diagnoses are most likely to be established and whose disease is most likely to generate new knowledge about the underlying pathogenic mechanism[60]. Established diagnostic guidelines are used to create patient diagnoses. For diseases that do not have clear diagnostic criteria, diagnosis

involves synthesizing the available objective data and the judgement of the treating clinician.

Once admitted to the UDN, each patient underwent a thorough standardized phenotyping using HPO terms[68] alongside the necessary diagnostic tests. Samples from both affected and unaffected family members were collected for genomic sequencing. These data were then analysed in a collaborative, iterative process involving bioinformaticians, clinicians and genetic counsellors to identify the variants most likely responsible for the patient's conditions. This analysis involved four key steps: (1) alignment of the sequencing reads to a reference human genome, (2) identification of genetic variants from these aligned reads, (3) annotation of these variants, and (4) selection and filtering to pinpoint the variants most likely responsible for the patient's symptoms[10,61].

VEM

PheMART engineers features for variants and their corresponding wild-type sequences via a pretrained PLM [83,85]. Phenotypes were featurized by synthesizing knowledge from the LLM embedding vectors based on the feature textual descriptions and the BONMI-EHR embeddings. PheMART consists of three computational components, Variant Encoding Module (VEM), Phenotype Encoding Module (PEM) and Contrastive Variant-Phenotype Embedding (CVPE). The VEM learns low-dimensional variant encoding vectors that characterize a variant's phenotypic effects by contrasting it against the corresponding wild type. The PEM derives phenotype encoding vectors that integrate the phenotype knowledge from both the UMLS database and EHR data. Then, CVPE projects variant encoding vectors and phenotype encoding vectors onto a metric embedding space such that a variant is close to the phenotypes it is implicated in, allowing PheMART to elucidate the role of missense variants across all phenotypes

Input features. The variants and wild types were featurized as the input of VEM using a PLM, RGN2 (ref. 83), which was trained using ~250 million unlabelled natural amino acid sequences. PLM-derived features have been shown to facilitate various tasks, including protein structure prediction[83] and mutation effect inference[18,39]. The protein sequences of variants and wildtypes were transformed into the PLM $d^{\mathrm{v}} = 768$ dimensional embeddings as features. We note that the PLM was not further fine-tuned for variant–phenotype association learning.

The VEM follows a two-stream encoder–decoder architecture. The encoder, denoted as $\mathbf{Enc}^{\mathrm{v}}(\cdot)$, first transforms the variant and wild-type input features, that is, $\mathbf{I}^{\mathrm{v}}$ and $\mathbf{I}^{\mathrm{wt}}$, respectively, separately using feedforward layers and then fuses them via a fusion module to obtain the variant encoding vector $\mathbf{h}^{\mathrm{v}}$. The decoder consists of two streams, a variant part $\mathbf{Dec}^{\mathrm{v}}_{\mathrm{v}}(\cdot)$ and a residual part $\mathbf{Dec}^{\mathrm{v}}_{\mathrm{r}}(\cdot)$. On the basis of the variant encoding vector $\mathbf{h}^{\mathrm{v}}$, the $\mathbf{Dec}^{\mathrm{v}}_{\mathrm{v}}(\cdot)$ learns to reconstruct the variant input feature $\mathbf{I}^{\mathrm{v}}$, while $\mathbf{Dec}^{\mathrm{v}}_{\mathrm{r}}(\cdot)$ reconstructs the residual feature between the variant and wild-type inputs, namely $\mathbf{I}^{\mathrm{v}} - \mathbf{I}^{\mathrm{wt}}$. By constructing the residuals, the variant encoding $\mathbf{h}^{\mathrm{v}}$ is encouraged to capture the residuals between the variants and the wild types. Specifically, the semisupervised training objective is

$$\begin{aligned}\mathcal{L}^{\mathrm{v}}_{\mathrm{con}} = \mathbb{E}_{i\in\mathbb{V}^{\mathrm{l}}\cup\mathbb{V}^{\mathrm{u}}}\Big(&||\mathbf{Dec}^{\mathrm{v}}_{\mathrm{v}}\left(\mathbf{Enc}^{\mathrm{v}}\left(\mathbf{I}^{\mathrm{v}}_i,\mathbf{I}^{\mathrm{wt}}_i\right)\right)-\mathbf{I}^{\mathrm{v}}_i||_2\\ &+||\mathbf{Dec}^{\mathrm{v}}_{\mathrm{r}}\left(\mathbf{Enc}^{\mathrm{v}}\left(\mathbf{I}^{\mathrm{v}}_i,\mathbf{I}^{\mathrm{wt}}_i\right)\right)-\left(\mathbf{I}^{\mathrm{v}}_i-\mathbf{I}^{\mathrm{wt}}_i\right)||_2\Big),\end{aligned} \tag{1}$$

where the sets $\mathbb{V}^{\mathrm{l}}$ and $\mathbb{V}^{\mathrm{u}}$ denote the indices of variants with specific phenotype annotations and the indices of large-scale VUS, respectively. We incorporated all the VUS in ClinVar[7] for the training to enhance the generalization of the variant encoding vectors.

Semisupervised residual learning. To elucidate the phenotypic consequences of missense variants from the wild-type protein sequence, the VEM learns to contrast the variants against the wild type in a phenotype-aware manner. Specifically, higher similarities are encouraged between the encoding vectors of variants implicated in the same phenotypes than those implicated in distinct phenotypes by optimizing the variant encoder with a contrastive learning objective:

$$\begin{aligned}\mathcal{L}_{\mathrm{vc}} = \mathbb{E}_{i\in\mathbb{V}^{\mathrm{l}}}\Bigg[w_i\Bigg(&\sum_{k\in\mathbb{P}^{\mathrm{v}}_i}\max\left(m-\mathrm{sim}\left(\mathbf{h}^{\mathrm{v}}_i,\mathbf{h}^{\mathrm{v}}_k\right),0\right)\\ &+\sum_{j\in\mathbb{N}^{\mathrm{v}}_i}\max\left(\mathrm{sim}\left(\mathbf{h}^{\mathrm{v}}_i,\mathbf{h}^{\mathrm{v}}_j\right)-n,0\right)\Bigg)\Bigg],\end{aligned} \tag{2}$$

where $\mathbf{h}^{\mathrm{v}}_i$, $\mathbf{h}^{\mathrm{v}}_j$ and $\mathbf{h}^{\mathrm{v}}_k$ denote the variant encoding vectors of variant i, j and k, respectively. The $\mathrm{sim}(\cdot,\cdot)$ term is the similarity function between two variant embedding vectors, and we used cosine similarity. $\mathbb{P}^{\mathrm{v}}_i$ denotes the positive set of i consisting of variants implicated in the same phenotypes as i. They are encouraged to have similarities higher than m. $\mathbb{N}^{\mathrm{v}}_i$ denotes the negative sets of i, which include both the wild type and the variants from the same wild type but are reported either as 'benign' or with other distinct phenotypes having a CODER similarity below 0.7 to any of the reported phenotypes of variant i. By minimizing $\mathcal{L}_{\mathrm{vc}}$, PheMART encourages their embedding vectors to have similarities smaller than threshold value n to variant i. w_i denotes the weight of variant i, defined as $w_i = \frac{1}{\log(1+30P^{\mathrm{wt}}_i)}$, where P^{wt}_i denotes the prevalence of the variant–phenotype annotations from the wild type of variant i. Through re-weighting, variants from wild types with smaller numbers of phenotypic annotations are highlighted during model training. Incorporating the encoder–decoder-based semisupervised training and the variant–wild-type contrastive learning to co-train the VEM would simultaneously enhance the phenotypic discriminability and generalization of the learned variant encodings.

Bridging variants from different wild types by leveraging previous knowledge. We exploited previous knowledge (PK) on PPIs, protein functional domains and gene pathways to connect variants across different wild types, further enhancing PheMART's generalization. PPI information has been shown to facilitate phenotypic inference clinically[62]. Disease-linked amino acid alterations were frequently observed to modify specific PPIs[86–88], on the basis of which we assumed variants involved in the same PPIs to have similar biological effects. In addition, variants from the same protein domains were also assumed to share similar biological effects. To this end, we compiled physical PPI data from several sources, including Interactome INSIDER[21] which specifies the amino acids as the PPI interfaces, HINT[22] which features high-quality manually curated PPIs, and HuRI[23] (details are provided in Supplementary Section 1.7). We collected the annotations of protein domains from UniProt[24]. Variants lying in the same protein domain or on wild types involved in the same PPIs were encouraged with higher similarities than those without these connections by optimizing the variant encoder with a variant PK loss:

$$\mathcal{L}_{\mathrm{PK}} = \mathbb{E}_{i\in\mathbb{V}^{\mathrm{PK}}}\left[q_i * \max\left(\mathrm{sim}\left(\mathbf{h}^{\mathrm{v}}_i,\mathbf{h}^{\mathrm{v}}_n\right)+c-\mathrm{sim}\left(\mathbf{h}^{\mathrm{v}}_i,\mathbf{h}^{\mathrm{v}}_p\right),0\right)\right], \tag{3}$$

where $\mathbf{h}^{\mathrm{v}}_i$ is the encoding vector of variant i, $\mathbb{V}^{\mathrm{PK}}$ denotes the set of variants connected by PPIs or functional domains; $\mathbf{h}^{\mathrm{v}}_{\mathrm{p}}$ and $\mathbf{h}^{\mathrm{v}}_{\mathrm{n}}$ respectively denote the encoding vector of a positive or negative variant, with or without such PK connections to variant i. To prevent false negatives, we further required that the wild types of negative variants do not have shared pathways, curated in PrimKG[30,89], with the wild type of variant i The q_i term denotes the quality score of the collected previous knowledge and controls their contributions. We assigned the highest importance to the variants located at the interfaces of PPIs involving the same partner, as identified in Interactome INSIDER. This is because mutations in a single wild type that disrupt interactions with various partners could lead to distinct phenotypes, and those lying in the interfaces with the same partner are more likely to yield similar phenotypic effects[86,90]. Subsequently, the

annotations of protein function domains and the high-quality PPIs in HINT, where at least two different publications reported the PPIs, were emphasized. Lastly, the PPIs reported in HuRI were leveraged (the detailed q_i values are provided in Supplementary Section 1.7). The similarity margin c ensures that the similarity between variants with PR connections is at least c greater than those without PR connections. For effective training, we performed hard negative sample mining by leveraging the features of wild type protein. Specifically, instead of random negative sampling for each variant, we prioritized the variants on wild types with higher RGN2 (ref. 83) similarities to its corresponding wild type as the negatives.

PEM

Phenotypes were featurized to incorporate existing knowledge from two sources: (1) LLM-derived CODER embedding vectors, which condense the curated medical knowledge graph in UMLS[20,25], and (2) EHR-derived BONMI-EHR embeddings, which capture the clinical relationships among phenotypes and other medical concepts, including disease symptoms, medications and so on, based on their observed concurrence[26]. For phenotypes that were not observed in the BONMI-EHR embedding, we trained a mapping network for imputation based on their corresponding LLM embeddings (details are provided in Supplementary Section 1.3). For each phenotype, CODER produces a 768-dimensional vector $\mathbf{I}^{\mathrm{p}} \in \mathbb{R}^{d_I}$, and BONMI produces a 300-dimensional vector $\mathbf{F}^{\mathrm{p}} \in \mathbb{R}^{d_F}$. PEM fuses them into one using a two-stream encoder–decoder architecture. The encoder $\mathbf{Enc}^{\mathrm{p}}(\cdot)$ first transforms the LLM-derived features and EHR-derived features separately using feedforward layers, and then fuses them via a fusion module to obtain the low-dimensional phenotype encoding vector $\mathbf{h}^{\mathrm{p}}$. The decoder consists of two streams, denoted as $\mathbf{Dec}_{\mathrm{I}}^{\mathrm{p}}(\cdot)$ and $\mathbf{Dec}_{\mathrm{F}}^{\mathrm{p}}(\cdot)$. On the basis of $\mathbf{h}^{\mathrm{p}}$, the $\mathbf{Dec}_{\mathrm{I}}^{\mathrm{p}}(\cdot)$ was trained to reconstruct the LLM-derived feature $\mathbf{I}^{\mathrm{p}}$, while $\mathbf{Dec}_{\mathrm{F}}^{\mathrm{p}}(\cdot)$ was trained to reconstruct the EHR-derived feature $\mathbf{F}^{\mathrm{p}}$ (the details of model architecture are provided in Supplementary Section 1.3). Formally, we obtained the training objective as:

$$\begin{aligned}\mathcal{L}_{\mathrm{con}}^{\mathrm{p}} = \mathbb{E}_{p\in\mathbb{P}^{\mathrm{l}}\cup\mathbb{P}^{\mathrm{u}}}\Big(&\left\|\mathbf{Dec}_{\mathrm{I}}^{\mathrm{p}}\left(\mathbf{Enc}^{\mathrm{p}}\left(\mathbf{I}^{\mathrm{p}},\mathbf{F}^{\mathrm{p}}\right)\right)-\mathbf{I}^{\mathrm{p}}\right\|_2 \\ &+\left\|\mathbf{Dec}_{F}^{\mathrm{p}}\left(\left(\mathbf{Enc}^{\mathrm{p}}\left(\mathbf{I}^{\mathrm{p}},\mathbf{F}^{\mathrm{p}}\right)\right)-\mathbf{F}^{\mathrm{p}}\right\|_2\right),\end{aligned} \tag{4}$$

where $\mathbb{P}^{\mathrm{l}}$ and $\mathbb{P}^{\mathrm{u}}$ denote the sets of annotated phenotypes and unannotated phenotypes, respectively. This dual decoding from $\mathbf{h}^{\mathrm{p}}$ encourages it to harness the strengths of both CODER and BONMI, combining semantic nuances captured by CODER with clinical relationships elucidated in EHR data.

CVPE

The variant–phenotype embedding involves jointly projecting variants and phenotypes onto a metric embedding space. In this space, variants were positioned close to the phenotypes they are implicated in. Variants and phenotypes were embedded by leveraging a contrastive learning strategy, CLIP[35]. CLIP training does not require explicit negative sampling of either phenotypes or variants, making it a good fit for variant–phenotype pairwise relationship learning. This is because the variant–phenotype pairs currently reported are non-exclusive, and random negative sampling of variant–phenotype pairs would inevitably include false negatives and lead to deteriorated performances. Specifically, we further transformed the encoding vectors, $\mathbf{h}^{\mathrm{v}}$ and $\mathbf{h}^{\mathrm{p}}$, with one-layer linear mapping to obtain the final embedding vectors $\mathbf{e}^{\mathrm{v}}$ and $\mathbf{e}^{\mathrm{p}}$, respectively. During that process, the reported variant–phenotype pairs were encouraged to have higher similarities than those unreported by minimizing the variant-phenotype co-map loss:

$$\begin{aligned}&\mathcal{L}_{\mathrm{cmap}} = -\mathbb{E}_{(i,j)\in\mathbb{A}} \\ &\left[w_i\left(\log\frac{\exp\left(\mathrm{sim}\left(\mathbf{e}_i^{\mathrm{v}},\mathbf{e}_j^{\mathrm{p}}\right)/\tau\right)}{\sum_{k\in\mathbb{P}^{\mathrm{l}}}\exp\left(\mathrm{sim}\left(\mathbf{e}_i^{\mathrm{v}},\mathbf{e}_k^{\mathrm{p}}\right)/\tau\right)}+\log\frac{\exp\left(\mathrm{sim}\left(\mathbf{e}_i^{\mathrm{v}},\mathbf{e}_j^{\mathrm{p}}\right)/\tau\right)}{\sum_{k\in\mathbb{V}^{\mathrm{l}}}\exp\left(\mathrm{sim}\left(\mathbf{e}_k^{\mathrm{v}},\mathbf{e}_j^{\mathrm{p}}\right)/\tau\right)}\right)\right],\end{aligned} \tag{5}$$

where the pair of variant i and phenotype j is sampled from the annotated variant–phenotype set $\mathbb{A}$, and τ is the temperate that controls the similarity matching. For variant i, we maximized its relative similarity to phenotype j over the other phenotypes. For each phenotype j, we maximized its relative similarity to variant i over all the labeled variants, namely $\mathbb{V}^{\mathrm{l}}$. w_i denotes the weight of variant i, as in equation 1, highlighting the contributions of variants from wild types with fewer phenotypic reports.

Training procedures

PheMART was trained in three steps to enhance its generalization to interpret the large-scale VUS. First, we pretrained the variant–phenotype encoders and decoders in an unsupervised manner by leveraging the large-scale ClinVar variants, irrespective of their annotations, and optimized the model using the objective $\mathcal{L}_{\mathrm{con}}^{\mathrm{v}}$ in equation 1 and $\mathcal{L}_{\mathrm{con}}^{\mathrm{p}}$ in equation 4. In this step, we guided the encoder–decoder to contrast variant sequences and their corresponding wild-type sequences, and to learn variant representations that capture the residual information between variants and wild types. Meanwhile, for phenotype, we guided the phenotype encoder to synthesize the phenotype knowledge from the two sources. Then, we refined the variant representation to differentiate the varied phenotypic effects of variants on the same wild type. Variants with the same phenotypic effects were encouraged to have similar representations and to be dissimilar to the variants that are benign or implicated in different phenotypes. To this end, we additionally incorporated the variant contrastive learning objective $\mathcal{L}_{\mathrm{vc}}$ in equation 2 and PPI-derived $\mathcal{L}_{\mathrm{PK}}$ in equation 3 to train the model. Finally, we linked variants to phenotypes in the metric space by integrating the objectives described above to fine-tune PheMART:

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{cmap}} + \alpha\left(\mathcal{L}_{\mathrm{con}}^{\mathrm{v}} + \mathcal{L}_{\mathrm{con}}^{\mathrm{p}}\right) + \beta\mathcal{L}_{\mathrm{vc}} + \gamma\mathcal{L}_{\mathrm{PK}}, \tag{6}$$

where α controls the involvement of the construction losses $\mathcal{L}_{\mathrm{con}}^{\mathrm{v}}$ and $\mathcal{L}_{\mathrm{con}}^{\mathrm{p}}$, and β and γ determine the involvements of the variant–wild-type contrastive learning and the previous knowledge of PPIs, respectively. We selected the values of α β and γ via grid search and set $\alpha = 0.6$, $\beta = 10.0$ and $\gamma = 5.0$ in the joint training stage.

Phenotype-aware prediction calibrations

To improve the reliability of the predicted scores across phenotypes, we performed phenotype-aware calibrations of the prediction scores. Specifically, we obtained calibrated predictions by training a linear logistic rescaling function that transforms the output logits and phenotype embeddings to a probability score:

$$\hat{p} = \sigma\left(a_0 \times s + [w_1, w_2, \ldots, w_d]\,[e_1, e_2, \ldots, e_d]^T + b_0\right), \tag{7}$$

where s denotes PheMART's raw output score for a variant–phenotype pair, σ is the sigmoid$(\cdot)$ function, a_0 and b_0 are scalar learnable parameters, and $[w_1, w_2, \ldots, w_d]$ is a d-dimensional learnable vector. All aforementioned learnable parameters are shared across all variant–phenotype pairs. The $[e_1, e_2, \ldots, e_d]$ term denotes the d-dimensional embedding vectors of the phenotype. To prevent overfitting in the calibration, we set $d = 2$ and transformed the 80-dimensional phenotype embedding into 2 dimensions via t-distributed stochastic neighbor embedding (t-SNE). The phenotype-aware calibrations were performed during both model validation and final training using a balanced set of annotations consisting of 24,000 variant–phenotype pairs, with 2,000 pairs in each phenotype category. We applied resampling to annotated pairs within disease categories that had fewer than 2,000 annotations. Since the variant–phenotype relationships documented in ClinVar are incomplete, we performed semantic-guided negative sampling to prevent the inclusion of false-negative variant–phenotype pairs during calibration. Specifically, for each phenotype in the annotated variant–phenotype pairs, we selected negative

variants by ensuring that they were not associated with any phenotype exhibiting a cosine similarity above 0.7 in the CODER embedding[25] to the given phenotype.

Evaluation metrics

We assessed the performance of phenotypic predictions using three key metrics: auROC, MRR and sensitivity@k.

MRR. MRR evaluates the ranking quality of predicted phenotypes for variant annotation. In each annotation, we obtained the rank of its associated phenotypes within all the 4,179 phenotypes based on the models' prediction scores. The reciprocal rank is the inverse of this position, and the MRR is the average of these reciprocal ranks across all variant annotations. A higher MRR suggests that the reported phenotypes are ranked more highly in the prediction scores. Mathematically, MRR is defined as:

$$\mathrm{MRR} = \frac{1}{|\mathbb{A}_e|}\sum_{i=1}^{|\mathbb{A}_e|}\frac{1}{\mathrm{rank}_i} \tag{8}$$

where rank_i is the rank of the associated phenotype in the ith variant–phenotype annotation and $|\mathbb{A}_e|$ denotes the total number of annotated variant–phenotype pairs in the evaluation set $\mathbb{A}_e$.

Ethics

The study protocol was approved by the MGB Human Research Committee (IRB00010756). No patient contact occurred during this study which relied on secondary use of data, thus allowing for a waiver of informed consent as detailed by 45 CFR 46.116. The methods were performed in accordance with relevant guidelines and regulations, and approved by the VA Central Institutional Review Board (IRB). They were supported by the Million Veteran Program, VA Central IRB 10-02, and approved under VA Central IRB protocol 18–38. This publication does not represent the views of the Department of Veteran Affairs or the US Government. The UDN study is approved by the US NIH IRB (protocol 15HG0130). All patients accepted to UDN provide written informed consent to share their data across the UDN as part of a network-wide informed consent process.

Reporting summary

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

Data availability

Processed datasets are provided in Zenodo at https://zenodo.org/records/17402388 (ref. 91). The high-confidence phenotypic predictions by PheMART are available in Zenodo at https://zenodo.org/records/17402574 (ref. 92) and are also visualized by genes, phenotypes and protein domains on https://shiny.parse-health.org/PheMART/. In accordance with VA policy, the VA-derived EHR embedding vectors used in this study are available upon request. For access to UDN data, please refer to the UDN official data availability guidelines at https://undiagnosed.hms.harvard.edu/research/data-availability/. The Human Gene Mutation Database used for external validation cannot be publicly shared due to licensing restrictions, as it was obtained under a paid institutional license from QIAGEN, which prohibits redistribution. Source data are provided with this paper.

Code availability

The source codes are available in GitHub at https://github.com/celehs/PheMART (ref. 93).

References

1. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science* **373**, 1464–1468 (2021).
2. Chong, J. X. et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).
3. Venugopal, A. et al. Monogenic diseases in India. *Mutat. Res. Rev. Mutat. Res.* **776**, 23–31 (2018).
4. Konishi, C. T. & Long, C. Progress and challenges in CRISPR-mediated therapeutic genome editing for monogenic diseases. *J. Biomed. Res.* **35**, 148–162 (2021).
5. Johannesen, K. M. et al. Genotype-phenotype correlations in *SCN8A*-related disorders reveal prognostic and therapeutic implications. *Brain* **145**, 2991–3009 (2022).
6. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
7. Landrum, M. J. et al. ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
8. Uffelmann, E. et al. Genome-wide association studies. *Nat. Rev. Methods Primers* **1**, 59 (2021).
9. Starita, L. M. et al. Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).
10. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
11. Wu, Y. et al. Improved pathogenicity prediction for rare human missense variants. *Am. J. Hum. Genet.* **108**, 1891–1906 (2021); erratum **108**, 2389 (2021).
12. Slatkin, M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).
13. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
14. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
15. Pejaver, V. et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* **11**, 5918 (2020).
16. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
17. 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).
18. 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).
19. Chesmore, K., Bartlett, J. & Williams, S. M. The ubiquity of pleiotropy in human disease. *Hum. Genet.* **137**, 39–44 (2018).
20. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
21. Meyer, M. J. et al. Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods* **15**, 107–114 (2018).
22. Das, J. & Yu, H. HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 92 (2012).
23. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
24. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* **53**, D609–D617 (2025).
25. Yuan, Z. et al. CODER: knowledge-infused cross-lingual medical term embedding for term normalization. *J. Biomed. Inform.* **126**, 103983 (2022).
26. Zhou, D., Cai, T. & Lu, J. Multi-source learning via completion of block-wise overlapping noisy matrices. *J. Mach. Learn. Res.* **24**, 1–43 (2023).
27. Hong, C. et al. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *npj Digit. Med.* **4**, 151 (2021).

28. Wen, J. et al. Multimodal representation learning for predicting molecule–disease relations. *Bioinformatics* **39**, btad085 (2023).
29. Ramoni, R. B. et al. The Undiagnosed Diseases Network: accelerating discovery about health and disease. *Am. J. Hum. Genet.* **100**, 185–192 (2017).
30. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
31. Stenson, P. D. et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136**, 665–677 (2017).
32. Ryu, J. Y., Kim, H. U. & Lee, S. Y. Deep learning improves prediction of drug–drug and drug–food interactions. *Proc. Natl Acad. Sci. USA* **115**, E4304–E4311 (2018).
33. Kingma, D. P., Rezende, D. J., Mohamed, S. & Welling, M. Semi-supervised learning with deep generative models. *Adv. Neural Inf. Process. Syst.* **27**, 3581–3589 (2014).
34. Dehghan, A., Abbasi, K., Razzaghi, P., Banadkuki, H. & Gharaghani, S. CCL-DTI: contributing the contrastive loss in drug–target interaction prediction. *BMC Bioinformatics* **25**, 48 (2024).
35. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning (PMLR 2021)* (eds Meila, M. & Zhang, T.) 8748–8763 (PMLR, 2021)
36. Miller, D. T. et al. ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **24**, 1407–1414 (2022).
37. Marchant, R. G. et al. Genome and RNA sequencing boost neuromuscular diagnoses to 62% from 34% with exome sequencing alone. *Ann. Clin. Transl. Neurol.* **11**, 1250–1266 (2024).
38. Koutsofti, C. et al. Massive parallel DNA sequencing of patients with inherited cardiomyopathies in Cyprus and suggestion of digenic or oligogenic inheritance. *Genes* **15**, 319 (2024).
39. 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).
40. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
41. Hadsell, R., Chopra, S. & LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR' 06)* (eds Fitzgibbon, A. et al.) 1735–1742 (IEEE, 2006).
42. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proc. 34th International Conference on Machine Learning (PMLR 2017)* (eds Precup, D. & Teh, Y. W.) 1321–1330 (PMLR, 2017)
43. George, A. L. Jr. Inherited disorders of voltage-gated sodium channels. *J. Clin. Invest.* **115**, 1990–1999 (2005).
44. Kullmann, D. M. & Waxman, S. G. Neurological channelopathies: new insights into disease mechanisms and ion channel function. *J. Physiol.* **588**, 1823–1827 (2010).
45. Bers, D. M. Cardiac excitation–contraction coupling. *Nature* **415**, 198–205 (2002).
46. Opalińska, M. & Jańska, H. AAA proteases: guardians of mitochondrial function and homeostasis. *Cells* **7**, 163 (2018).
47. Cloonan, S. M. & Choi, A. M. K. Mitochondria in lung disease. *J. Clin. Invest.* **126**, 809–820 (2016).
48. Klaassen, S. et al. Mutations in sarcomere protein genes in left ventricular noncompaction. *Circulation* **117**, 2893–2901 (2008).
49. Holm, H. et al. A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320 (2011).
50. D'Cruz, A. A., Babon, J. J., Norton, R. S., Nicola, N. A. & Nicholson, S. E. Structure and function of the SPRY/B30.2 domain proteins involved in innate immunity. *Protein Sci.* **22**, 1–10 (2013).
51. Van Bergen, N. J. et al. CDKL5 deficiency disorder: molecular insights and mechanisms of pathogenicity to fast-track therapeutic development. *Biochem. Soc. Trans.* **50**, 1207–1224 (2022).
52. Yan, G.-X. & Antzelevitch, C. Cellular basis for the Brugada syndrome and other mechanisms of arrhythmogenesis associated with ST-segment elevation. *Circulation* **100**, 1660–1666 (1999).
53. Jones, S. An overview of the basic helix-loop-helix proteins. *Genome Biol.* **5**, 226 (2004).
54. Duverger, O. & Morasso, M. I. Role of homeobox genes in the patterning, specification, and differentiation of ectodermal appendages in mammals. *J. Cell. Physiol.* **216**, 337–346 (2008).
55. Lewis, D. L. et al. Ectopic gene expression and homeotic transformations in arthropods using recombinant sindbis viruses. *Curr. Biol.* **9**, 1279–1287 (1999).
56. Hwang, J. L. et al. *FOXP3* mutations causing early-onset insulin-requiring diabetes but without other features of immune dysregulation, polyendocrinopathy, enteropathy, x-linked syndrome. *Pediatr. Diabetes* **19**, 388–392 (2018).
57. Brock, S. et al. Tubulinopathies continued: refining the phenotypic spectrum associated with variants in *TUBG1*. *Eur. J. Hum. Genet.* **26**, 1132–1142 (2018).
58. Cai, S., Li, J., Wu, Y. & Jiang, Y. De novo mutations of *TUBB2A* cause infantile-onset epilepsy and developmental delay. *J. Hum. Genet.* **65**, 601–608 (2020).
59. Watanabe, K. et al. Identification of two novel de novo *TUBB* variants in cases with brain malformations: case reports and literature review. *J. Hum. Genet.* **66**, 1193–1197 (2021).
60. Splinter, K. et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. *N. Engl. J. Med.* **379**, 2131–2139 (2018).
61. Kobren, S. N. et al. Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases. *Genet. Med.* **23**, 1075–1085 (2021).
62. Schlüter, A. et al. ClinPrior: an algorithm for diagnosis and novel gene discovery by network-based prioritization. *Genome Med.* **15**, 68 (2023).
63. Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med.* **14**, 23 (2022).
64. Gahl, W. A. et al. The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet. Med.* **14**, 51–59 (2012).
65. Zhu, X. et al. Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet. Med.* **17**, 774–781 (2015).
66. Weinreich, S. S., Mangon, R., Sikkens, J. J., En Teeuw, M. E. & Cornel, M. C. Orphanet: a European database for rare diseases [in Dutch with English abstract]. *Ned. Tijdschr. Geneeskd.* **152**, 518–519 (2008).
67. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
68. Robinson, P. N. et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).
69. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274–1281 (2007).
70. Jagadeesh, K. A. et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. *Genet. Med.* **21**, 464–470 (2019).

71. Köhler, S. et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* **85**, 457–464 (2009).
72. Li, Y. et al. Functional correlation of *ATP1A2* mutations with phenotypic spectrum: from pure hemiplegic migraine to its variant forms. *J. Headache Pain* **22**, 92 (2021).
73. Wojciechowska, K., Pikulicka, A., Drgas, O., Żarnowska, I. & Brudkowska, Z. Heterozygous de novo mutation in the *ATP1A2* gene in a patient with alternating hemiplegia of childhood. *Pediatr. Pol.* **98**, 258–263 (2023).
74. Ng, B. G. et al. Biallelic mutations in CAD, impair de novo pyrimidine biosynthesis and decrease glycosylation precursors. *Hum. Mol. Genet.* **24**, 3050–3057 (2015).
75. Yang, H. et al. Noncoding genetic variation in *GATA3* increases acute lymphoblastic leukemia risk through local and global changes in chromatin conformation. *Nat. Genet.* **54**, 170–179 (2022).
76. Abunimye, D. A., Okafor, I. M., Okorowo, H. & Obeagu, E. I. The role of GATA family transcriptional factors in haematological malignancies: a review. *Medicine* **103**, e37487 (2024); retraction **103**, e38232 (2024).
77. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
78. Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R. & Alexov, E. Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.* **425**, 3919–3936 (2013).
79. Backwell, L. & Marsh, J. A. Diverse molecular mechanisms underlying pathogenic protein mutations: beyond the loss-of-function paradigm. *Annu. Rev. Genom. Hum. Genet.* **23**, 475–498 (2022).
80. Frésard, L. et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat. Med.* **25**, 911–919 (2019).
81. Lee, H. et al. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genet. Med.* **22**, 490–499 (2020).
82. Postel, M. D., Culver, J. O., Ricker, C. & Craig, D. W. Transcriptome analysis provides critical answers to the "variants of uncertain significance" conundrum. *Hum. Mutat.* **43**, 1590–1608 (2022).
83. Chowdhury, R. et al. Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* **40**, 1617–1623 (2022).
84. Zeng, S., Yuan, Z. & Yu, S. Automatic biomedical term clustering by learning fine-grained term representations. In *Proc. 21st Workshop on Biomedical Language Processing* (eds Demner-Fushman, D. et al.) 91–96 (Association for Computational Linguistics, 2022)
85. Putkowski, S. The National Organization for Rare Disorders (NORD): providing advocacy for people with rare disorders. *NASN Sch. Nurse* **25**, 38–41 (2010).
86. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
87. Ng, P. K.-S. et al. Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* **33**, 450–462 (2018).
88. Cheng, F. et al. Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nat. Genet.* **53**, 342–353 (2021).
89. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Sci. Data* **10**, 67 (2023).
90. Wang, X. et al. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159–164 (2012).
91. Jun, W. et al. Phenotypic prediction of missense variants (high-confidence predictions). *Zenodo* https://doi.org/10.5281/zenodo.17402573 (2025).
92. Jun, W. et al. Phenotypic prediction of missense variants (dataset). *Zenodo* https://doi.org/10.5281/zenodo.17402387 (2025).
93. Jun, W. et al. PheMART. *GitHub* https://github.com/celehs/PheMART (2025).
94. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
95. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. *J. Open Source Softw.* **3**, 861 (2018).

Acknowledgements

Part of this research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration, and was supported by award no. MVP000 (K.C.). Research reported in this manuscript was supported by the US National Institutes of Health (NIH) under award nos. U01HG007530 (S.N.K. and I.S.K.), P30 AR072577 (K.P.L.), R01 CA297832 (H.W.) and R01 GM152814-01 (J.S.L.). This work was also supported by the US National Science Foundation (NSF) under award no. IIS-2127918 (H.W.) and NSF CAREER Award IIS-2340125 (H.W.), and by an Amazon Faculty Research Award, and Microsoft AI and Society Fellowship (H.W.). The content is solely the authors' responsibility and does not necessarily represent the official views of the NIH.

Author contributions

J.W., T.C., S.Z., J.D., J.S.L., A.C.P., M. Zitnik, I.S.K., H.W., M. Zhu, S.C. and F.L. contributed to the conceptualization of the study. Writing was carried out by J.W., T.C., S.Z., A.C.P., Y.C., S.N.K., M. Zitnik, C.-L.B. and H.G.Z. Data curation was performed by J.W., S.Z., J.D., S.N.K., Y.C., S.C. and I.S.K. Funding was provided by T.C., I.S.K., K.P.L. and K.C.

Competing interests

K.P.L. was a past one-time consultant for the University of California, Berkeley. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-026-01636-4.

Correspondence and requests for materials should be addressed to Tianxi Cai.

Peer review information *Nature Biomedical Engineering* thanks Xiaoming Liu, Jianyi Yang and Matthew Jensen for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Tianxi Cai

Last updated by author(s): Oct 15, 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 *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 *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	The data collection pipeline is described in the manuscript, including variant-phenotype annotations from ClinVar, protein-protein interactions from INSIDER, HINT and HuRI, protein domains from UniProt, pathway information from Gene Ontology, clinical embedding vectors derived from the MGB and VA EHR cohort, and patient-level data from the undiagnosed disease network.
Data analysis	We built the model using Python 3.8 and Tensorflow 2.9.2. Codes for pre-processing the raw data and for computing the phenotypic predictions for missense variants is available on our GitHub repository (https://github.com/celehs/PheMART)

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 ClinVar annotations (https://www.ncbi.nlm.nih.gov/clinvar/), protein-protein interactions (http://interactomeinsider.yulab.org/; https://hint.yulab.org/),

protein domain annotations (https://www.uniprot.org/) and gene pathway information (https://github.com/mims-harvard/PrimeKG) are publicly available. The patient-level data from Undiagnosed Disease Network can be accessed by following the UDN official guidelines at https://undiagnosed.hms.harvard.edu/research/data-availability/. The EHR-derived clinical embedding vectors are available upon request. The high-confidence phenotypic predictions of the proposed method are available at
https://doi.org/ 10.5281/zenodo.17402573. The other related data are provided at https://doi.org/10.5281/zenodo.17402387.

The source codes are available at https://github.com/celehs/PheMART.

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	Not applicable. Sex and gender information were not included in the analyzed data.
Reporting on race, ethnicity, or other socially relevant groupings	For the 83 individuals in UDN, 67 are White, 9 are Asian, and 7 are black.
Population characteristics	The study utilized de-identified data from two sources: (1) summary-level co-occurrence matrices derived from 12.5 million U.S. veterans' EHRs data from the Veterans Affairs healthcare system and 60,000 patients from Mass General Brigham Biobank, and (2) genomic and clinical data from 83 participants enrolled in the Undiagnosed Diseases Network who consented to data sharing under UDN protocols.
Recruitment	Not applicable. No new participants were recruited for this study. All data were obtained from existing, approved repositories (VA EHR and UDN) under appropriate data-use agreements and de-identification protocols.
Ethics oversight	The study was conducted in accordance with ethical standards and data-use regulations. The VA EHR data were de-identified and accessed under institutional review and data-use approval. The UDN data were collected under informed consent and ethical approval by the UDN consortium.

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	No formal statistical method was used to predetermine sample size. 89 patients from UDN were used for validations.
Data exclusions	1. Clinvar phenotype filtering: the non-specific phenotypes, e.g., TP53-related disorders, were removed during the annotation curation.
Replication	We used 10-fold cross-validations to assess the stability of model predictions. For external validations, we replicated the model training and evaluation 10 times.
Randomization	We split the ClinVar data into training and test sets to measure model performance. Additionally, the external HGMD and UDN data were used for further validations.
Blinding	Blinding was not applicable because the study exclusively used pre-existing datasets where no new data collection or subjective labeling was performed. All data for model training or evaluation were defined prior to analysis and were not influenced by the investigators, eliminating potential observer or experimenter bias.

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	N/A.
Novel plant genotypes	N/A.
Authentication	N/A.