SYSTEMS BIOLOGY

Phenome-wide identification of therapeutic genetic targets, leveraging knowledge graphs, graph neural networks, and UK Biobank data

Lawrence Middleton[1], Ioannis Melas[1], Chirag Vasavda[2], Arwa Raies[1], Benedek Rozemberczki[3], Ryan S. Dhindsa[2,4,5], Justin S. Dhindsa[6], Blake Weido[7], Quanli Wang[2], Andrew R. Harper[1], Gavin Edwards[3], Slavé Petrovski[1,8], Dimitrios Vitsios[1]*

The ongoing expansion of human genomic datasets propels therapeutic target identification; however, extracting gene-disease associations from gene annotations remains challenging. Here, we introduce Mantis-ML 2.0, a framework integrating AstraZeneca's Biological Insights Knowledge Graph and numerous tabular datasets, to assess gene-disease probabilities throughout the phenome. We use graph neural networks, capturing the graph's holistic structure, and train them on hundreds of balanced datasets via a robust semi-supervised learning framework to provide gene-disease probabilities across the human exome. Mantis-ML 2.0 incorporates natural language processing to automate disease-relevant feature selection for thousands of diseases. The enhanced models demonstrate a 6.9% average classification power boost, achieving a median receiver operating characteristic (ROC) area under curve (AUC) score of 0.90 across 5220 diseases from Human Phenotype Ontology, OpenTargets, and Genomics England. Notably, Mantis-ML 2.0 prioritizes associations from an independent UK Biobank phenome-wide association study (PheWAS), providing a stronger form of triaging and mitigating against underpowered PheWAS associations. Results are exposed through an interactive web resource.

INTRODUCTION

Identifying and prioritizing genetic targets for treating a disease is a complex undertaking that involves carefully weighing various arguments and lines of evidence. One effective strategy is to focus on mechanisms with a clear genetic basis, as therapies targeting such mechanisms are more likely to succeed in clinical trials and regulatory processes (*1*, *2*). While identifying genetic targets is not trivial, genome-wide association studies, advances in next-generation sequencing, and precompetitive public-private collaborations have all substantially advanced our understanding of the genetics of biology and pathology. However, the flurry of data has already become too detailed to digest manually, with important discoveries effectively hidden in plain sight. The emergence of newer phenome-wide studies, which test for associations among tens of thousands of phenotypes, further necessitate a more high-throughput approach to parsing genome-phenome data. Here, we propose the need for sophisticated machine learning methods to assist in efficiently unifying disparate troves of research to find meaningful and actionable genetic correlates of disease.

In our first attempt to do so, we previously introduced an automated machine learning framework termed Mantis-ML that leveraged a few known disease-associated genes against publicly annotated genetic data to then predict candidate genes of interest (*3*). Mantis-ML models the underlying biology of a disease using a set of generic and disease-specific features (e.g., genic intolerance, tissue-specific expression, animal knockout models, and others) across known associated genes. By inferring the genetic, functional, and systems biology landscape of disease from these known genes, Mantis-ML then attempts to generalize these findings across all other genes, estimating the probability that each gene has biological relevance to the disease. With this approach, Mantis-ML prioritized several previously unidentified associations for a range of diverse diseases, including chronic kidney disease, amyotrophic lateral sclerosis, epilepsy, idiopathic pulmonary fibrosis, and spontaneous coronary artery dissection (*3*–*5*). Mantis-ML outperformed previous state-of-the-art methods (*6*, *7*) and additionally accurately predicted associations derived from other large-scale cohort studies (*3*).

Despite its utility, however, the first generation of Mantis-ML (Mantis-ML 1.0) relies on manually curated lists of disease-specific features and seed genes (i.e., well-established disease-associated genes). This curation process limits the scalability of Mantis-ML and its potential to unbiasedly uncover otherwise undiscovered genetic associations. Here, we report an evolution to Mantis-ML 2.0 in which we adopt and expand on several machine learning approaches to develop a fully streamlined and automated version of Mantis-ML, scaled across the phenome (5220 diseases; Fig. 1). Mantis-ML 2.0 now incorporates the AstraZeneca Biological Insights Knowledge Graph (BIKG) (*8*), a comprehensive network of known relationships among genes, proteins, diseases, and compounds, assembled across 55 different data sources. In addition, Mantis-ML 2.0 deploys natural language processing (NLP), eliminating the need to manually input the relevant features for a disease or phenotype. With this now automated and multidimensional foundation, we find that Mantis-ML 2.0 is ripe for discovering previously unidentified gene-disease associations that, when coupled with human genetic evidence, can serve as launchpads for future research and development programs. We report the phenome-wide Mantis-ML scores, as well as a number

[1]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK. [2]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Waltham, MA 02451, USA. [3]Biological Insights Knowledge Graph (BIKG), Research D&A, R&D IT, AstraZeneca, Cambridge, UK. [4]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [5]Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, TX 77030, USA. [6]Medical Scientist Training Program, Baylor College of Medicine, Houston, TX 77030, USA. [7]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [8]Department of Medicine, University of Melbourne, Austin Health, Melbourne, Victoria, Australia.
*Corresponding author. Email: dimitrios.vitsios@astrazeneca.com

of follow-up analyses, through the introduction of a publicly available web resource at: http://mantisml.public.cgr.astrazeneca.com.

RESULTS

Mantis-ML 2.0: Methodological improvements leveraging knowledge graphs and GNNs

To derive an automated and holistic view of gene-disease probabilities from genome-wide and phenome-wide studies, we redesigned and extended the machine learning framework underlying Mantis-ML 1.0. First, we incorporated AstraZeneca's BIKG containing 14 million nodes (representing genes, proteins, diseases, and compounds) and 136 million edges connecting these nodes, of which 8.7 million correspond to gene-gene interactions. Like other knowledge graphs, the BIKG summarizes data across multiple sources and depicts various relationships among biological entities; by integrating them together, these relationships may then be used to infer potentially previously unknown connections between genes

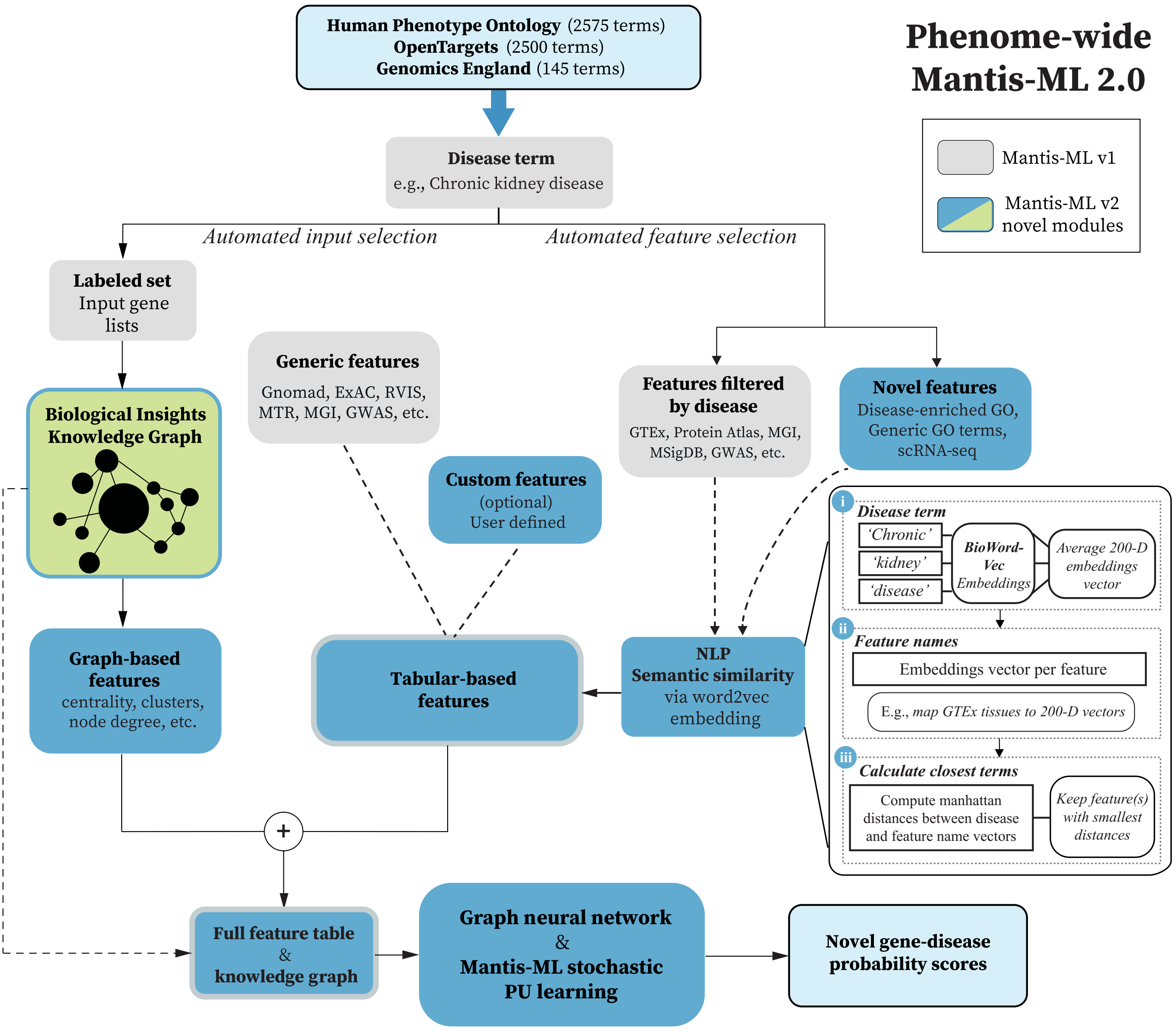
Human Phenotype Ontology (2575 terms)
OpenTargets (2500 terms)
Genomics England (145 terms)
Phenome-wide
Mantis-ML 2.0
Mantis-ML v1
Mantis-ML v2 novel modules
Disease term
e.g., Chronic kidney disease
Automated input selection
Automated feature selection
Labeled set
Input gene lists
Generic features
Gnomad, ExAC, RVIS, MTR, MGI, GWAS, etc.
Features filtered by disease
GTEx, Protein Atlas, MGI, MSigDB, GWAS, etc.
Novel features
Disease-enriched GO, Generic GO terms, scRNA-seq
Biological Insights Knowledge Graph
Custom features
(optional)
User defined
i
Disease term
'Chronic'
'kidney'
'disease'
BioWord-Vec Embeddings
Average 200-D embeddings vector
ii
Feature names
Embeddings vector per feature
E.g., map GTEx tissues to 200-D vectors
iii
Calculate closest terms
Compute manhattan distances between disease and feature name vectors
Keep feature(s) with smallest distances
Graph-based features
centrality, clusters, node degree, etc.
Tabular-based features
NLP
Semantic similarity
via word2vec embedding
+
Full feature table
&
knowledge graph
Graph neural network
&
Mantis-ML stochastic PU learning
Novel gene-disease probability scores

Fig. 1. Phenome-wide Mantis-ML deployment. Schematic of Mantis-ML workflow leveraging knowledge graphs, GNNs, and semantic similarity calculation. Each disease term from one of the resources (HPO, OT, and GEL) is chosen. From here, a set of relevant features is selected as outlined in the right branch, while the set of positively associated genes is identified. Mantis-ML then classifies each gene as associated or not based on the features identified in the left branch, which includes a comprehensive gene-gene network along with hundreds of tabular-based gene-level features. Calculation of semantic similarity between disease names and ontology terms for automated feature selection: Between any two terms I and j, the semantic similarity is calculated using the Euclidean distance between BioWordVec embeddings to produce a distance D_{ij}. The distance between two groups of words (e.g., "Abnormal circulating lipid concentration" and "Lipid metabolism") can then be calculated by averaging over all pairwise distances.

and diseases. To expose deeper relationships within the knowledge graph, we trained a graph neural network (GNN) (*9*, *10*) to propagate features to neighboring nodes such that the resulting feature vectors represent spatial proximity among nodes in the knowledge graph. We also now include additional single-cell transcriptomic data from a recent study (*11*).

Mantis-ML performance was measured by monitoring the area under the curve (AUC) of the classical receiver operating characteristic (ROC) curve. We implemented several advances to increase its performance over a number of iterations (Fig. 2A). To visualize how the change to Mantis-ML 1.0 affects its evolution, we set the Mantis-ML 1.0 default classifier to XGBoost (*12*). By first incorporating human tissue–specific data and NLP, we scaled the number of diseases from just a handful to thousands; we see a mild but nonsignificant reduction in the AUC from 0.84 to 0.81 ($P = 0.30$, two-sided Mann-Whitney U), suggesting that automating and standardizing Mantis-ML did not compromise prediction performance. The subsequent iteration—replacing InWeb (*13*) features with BIKG features—increases the median AUC relative to the first iteration (0.86 compared to 0.84; $P = 0.07$, one-sided Mann-Whitney U). With the addition of a graph convolutional network (GCN) in the subsequent iteration, this median AUC increased further to 0.88

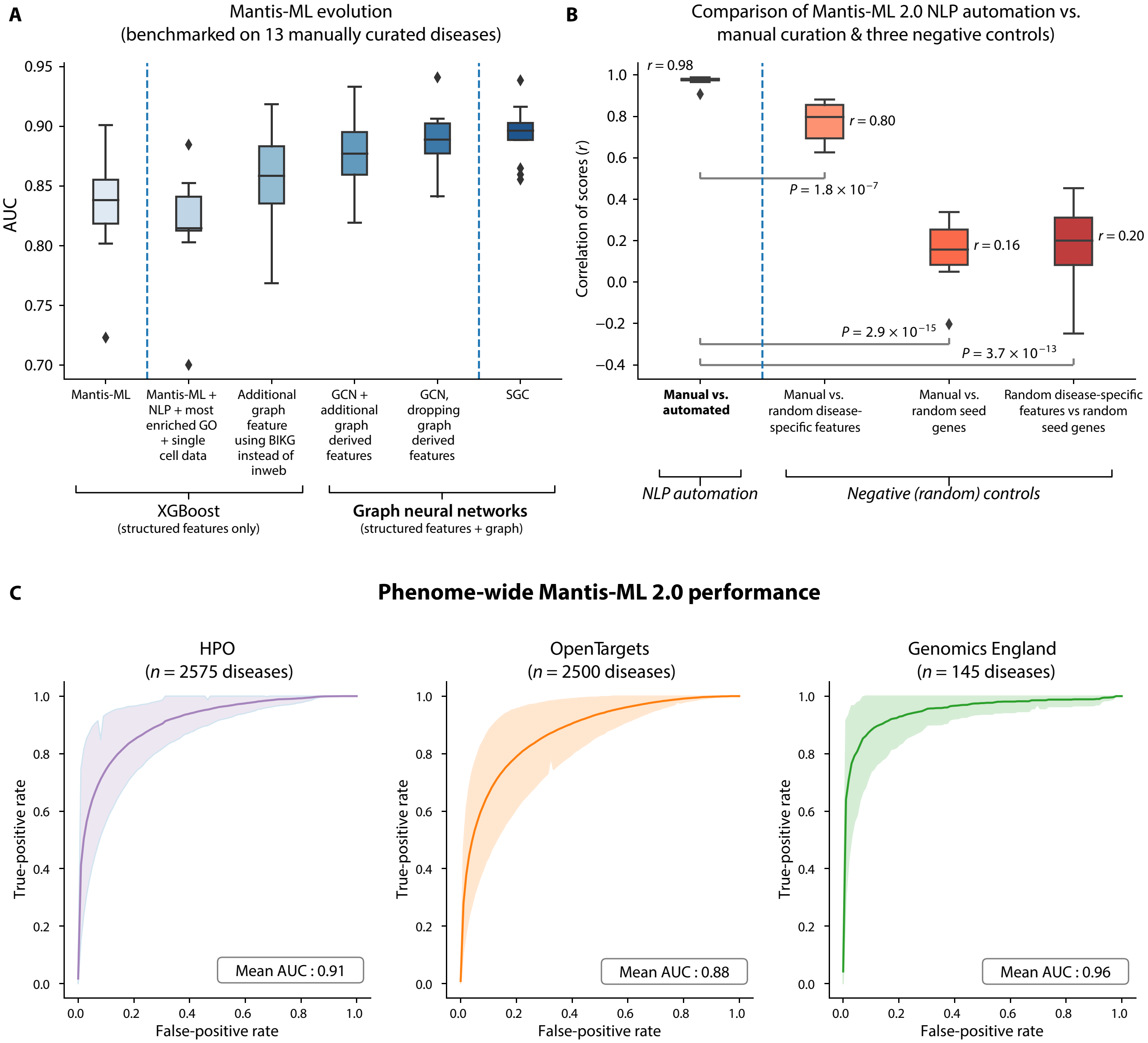

Fig. 2. Mantis-ML 2.0 performance improvements and phenome-wide Mantis-ML performance. (**A**) Iterations of Mantis-ML, successively introducing additional features and updating the modeling framework, using GNNs in the last three iterations, specifically GCN and SGC. (**B**) Validation of Mantis-ML results using automatically identified annotation terms and seed gene lists. (**C**) ROC curves for Mantis-ML predictions for each disease in each resource. Bands represent the lower 5th and upper 95th percentile in true-positive rate for a given false-positive rate (i.e., pointwise).

($P = 0.005$, one-sided Mann-Whitney U). We then tested whether stripping Mantis-ML of all graph-derived features, such as node degree and centrality (see Materials and Methods), and leaving only the graph structure could improve its predictive power. The AUC rose to 0.89, demonstrating that the GNN can capture the full signal just by leveraging the full structure of the graph. Thus, the graph-derived features were dropped from the final feature set. To then optimize Mantis-ML 2.0 to interface with phenome-wide data, we replaced the GCN with the lower-cost simple graph convolutions (SGCs) (*14*). The median AUC rose again to 0.90 ($P = 0.0005$, one-sided Mann-Whitney U compared to v1 AUC of 0.84), yielding a total 6.9% rise in predictive power overall from Mantis-ML 1.0. This improved AUC was accompanied by a 35-fold increase in processing speed (fig. S10).

With the ultimate goal of automating Mantis-ML 2.0, we explored the effect of introducing NLP approaches to identify and recognize features of a disease or phenotype. NLP was used to parse terms from MSigDB (*15*), MGI (*16*), and GTEX (*17*) related to the disease of interest. In addition, seed genes are now parsed automatically from the Human Phenotype Ontology (HPO) (*18*), Open Targets (OT) (*19*, *20*), and Genomics England (GEL) (*21*) resources, further streamlining the use of the next-generation Mantis-ML (Fig. 1). To evaluate the reliability of NLP automation, we first compared the outcomes of Mantis-ML with NLP methods versus manually curated inputs. We find that the scores from Mantis-ML are remarkably similar in both situations with a median Pearson's coefficient of 0.98 when tested over 13 heterogeneous diseases (Fig. 2B and table S1). In contrast, Mantis-ML performs substantially worse when fed random disease-specific features or random seed genes, suggesting that NLP automation does steer the model to incorporate valid disease-specific features. Mantis-ML also performs worse when fed random disease-specific features than with manually inputted features, but still exhibits a Pearson's coefficient of 0.80; this somewhat healthy correlation emphasizes that Mantis-ML is predominantly driven by genetic and physiologic variables, such as intolerance, lethality, and the BIKG graph. However, the improvement from 0.80 to 0.98 with NLP underscores that an automated Mantis-ML 2.0 performs as if manually executed with expert curation, paving the way for large-scale deployment of Mantis-ML across thousands of diseases.

Performance of Mantis-ML phenome-wide

Because Mantis-ML generates probability scores per gene, we tested its reliability by cross-validating scores for well-known disease-associated genes across three resources. We subsequently plotted these tests as ROC curves (*3*), thereby allowing us to measure how well Mantis-ML performs in individual datasets. The AUC provides a measure of how accurately Mantis-ML distinguishes between associated and nonassociated genes based on the original input seed genes. We deployed Mantis-ML to precalculate gene-disease probability scores for 5220 diseases catalogued among three resources: HPO ($n = 2575$), OT ($n = 2500$), and GEL ($n = 145$). Across all tested diseases in all three resources, the median AUC was 0.90 (Fig. 2C). Mantis-ML 2.0 performed best in GEL diseases with an AUC of 0.96.

Phenome-wide informed gene and disease networks

With Mantis-ML 2.0 in hand, we first set out to explore whether plotting gene-disease probability scores may reveal distinct patterns or clusters of diseases, thereby uncovering phenotypes with shared genetic or molecular fingerprints. The closer that two diseases are in such a plot, the greater the overlap in the genes associated between them. When scores are plotted for HPO, GEL, and OT, we observe meaningful clusters in each network, with HPO (Fig. 3A) and GEL exhibiting multiple disease clusters and OT exhibiting a distinct cluster for cancer physiology (figs. S11 and S12). Manual inspection of HPO disease scores reveals tight clusters of salient clinical traits (Fig. 3A), such as cerebellar with hepatic cysts, palpitations with arrhythmias, and intellectual disability with attention deficit hyperactivity disorder. By instead recategorizing these nonspecific or isolated clinical characteristics as features consistent with a molecular pathology, clinicians may be empowered to leverage Mantis-ML as a tool in diagnostic reasoning and treatment.

To identify what shared genetic architecture underlies each cluster, we subsequently regrouped diseases in a cluster based on their semantic similarities. For example, diseases in the "kidney" cluster contain five semantically similar terms, including "horseshoe kidney," "chronic kidney disease," "ectopic kidney," "enlarged kidney," and "renal tubular dysfunction." By prioritizing these subgroup features, we sought to glean additional insight into the pathophysiology that may relate the diseases to one another. In parallel, gene ontology (GO) analyses of these clusters also hint at aspects of shared physiology and may aid in discerning their genetic associations. In a liver HPO cluster, "GO_SMALL_MOLECULE_METABOLIC_PROCESS" ranks 8th, whereas "GO_REGULATION_OF_BODY_FLUID_LEVELS" ranks 16th in a hematologic disease OT cluster. "GO_RENAL_TUBULE_DEVELOPMENT" ranks 13th in a kidney disease GEL cluster (fig. S13 to S18).

Since genes that score similarly in their gene-disease probabilities share a phenotypic signature, we then considered whether we could calculate pathway enrichment analyses for a gene based on the other genes surrounding it. To do so, we performed an enrichment analysis for each gene in the exome with its 19 closest neighbors (for a total of 20 genes per analysis) in the transformed space of gene embedding distances. To test whether gene sets from Mantis-ML 2.0 can predict known pathways, we trialed analyses for the well-studied genes *PKD1*, *BRCA1*, and *APOB*. Kidney, breast carcinoma, and metabolic processes were enriched as expected (figs. S19 to S24), suggesting that Mantis-ML 2.0 disease probability scores can be exploited to generate meaningful gene networks. We report results for the rest of the exome for exploration in the Mantis-ML 2.0 web resource (http://mantisml.public.cgr.astrazeneca.com).

We also sought to extract the most contributing features during Mantis-ML training across the phenome. We focused on summarizing the feature importance scores across all diseases from HPO, reporting the top 30 features (Fig. 3B and table S4). Information derived from the BIKG graph (seed gene overlap) ranks as the most important feature indicating the value that can be derived from data-rich representations such as knowledge graphs. NLP-derived features, such as phenotype-specific animal knockout models, and GO generic terms introduced in this version of Mantis-ML also rank highly, demonstrating the importance of automation in feature selection by Mantis-ML.

Prioritizing less-studied genes in the knowledge graph

Akin to testing whether Mantis-ML 2.0 can infer genetic networks, we then explored whether Mantis-ML can be leveraged to identify promising yet lesser-studied genes of interest. We find that a gene's Mantis-ML 2.0 score partially correlates with how many BIKG nodes

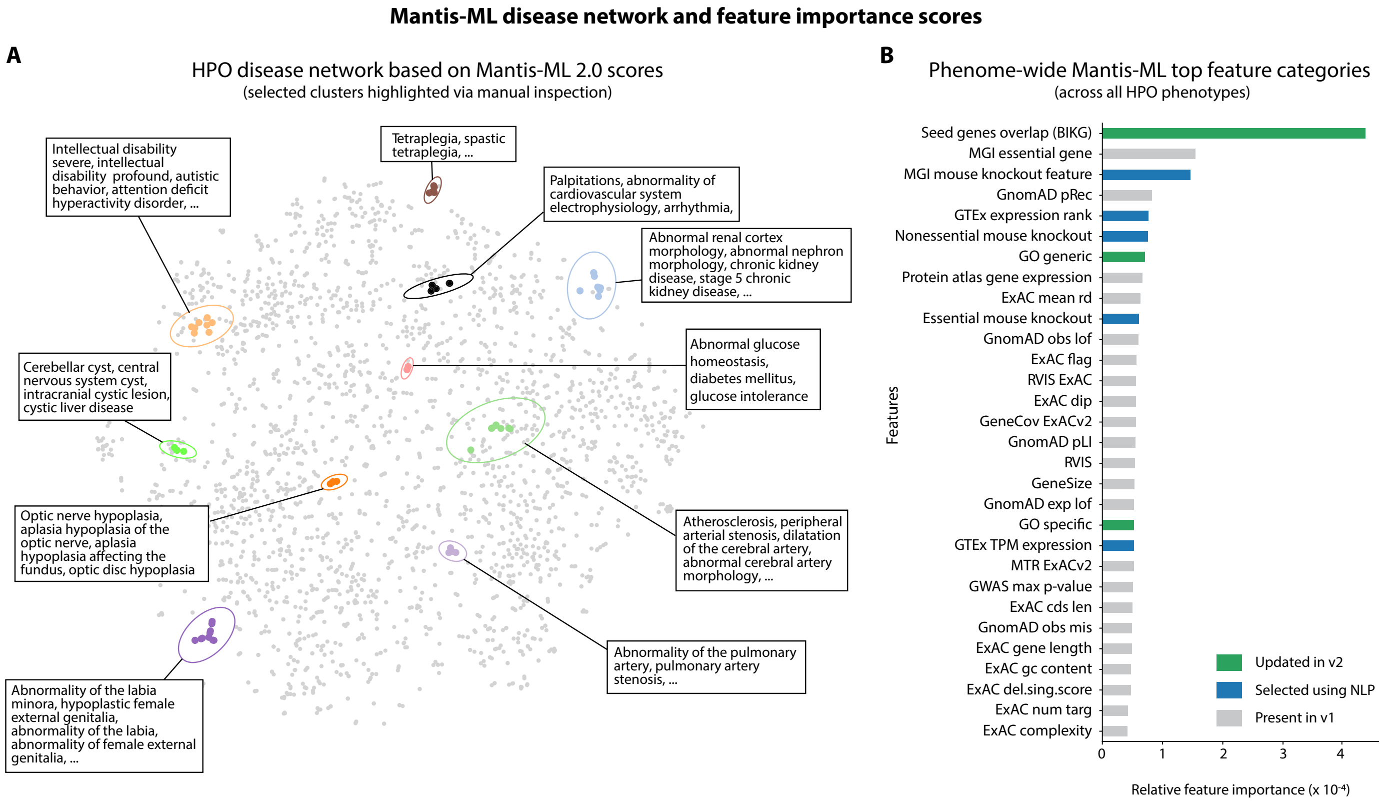

Fig. 3. Mantis-ML disease network and feature importance scores. (**A**) Disease similarity network of the 2575 diseases from the HPO dataset based on their Mantis-ML 2.0 rankings across all genes. Proximity of any two diseases in the network reflects a similar genetic fingerprint, showing that Mantis-ML probability scores over protein-coding genes between the two diseases are similar. Some example disease clusters, based on similar Mantis-ML signatures, are depicted in the t-SNE projection. (**B**) Top 30 features ranked by their feature importance for HPO diseases, aggregated over diseases. Summarization of features across all diseases was performed after harmonizing feature names (e.g., mapping tissue-specific GTEx scores to a single GTEx feature) so that feature importance scores are comparable between them. For each disease, feature importance scores were derived from an ensemble of models trained across all balanced datasets, eventually taking the mean value per feature from all models. Feature importance scores provide insight into which aspects of physiology are useful for discerning genetic associations. We see that the NLP selects features that are typically high ranking for the purposes of gene-disease predictions.

it intersects with, a proxy of how well it is annotated and studied (median correlation across all diseases = 0.69, median *P* value of all correlations $< 1 \times 10^{-308}$) (Fig. 4A); accordingly, genes with high Mantis-ML scores (>0.9) but fewer BIKG annotations may be ripe for inquiry. We first considered genes in a collated set of seven diabetes phenotypes ("diabetes insipidus," "diabetes mellitus," "insulin-resistant diabetes mellitus," "maternal diabetes," "neonatal insulin-dependent diabetes mellitus," "type II diabetes mellitus," and "type I diabetes mellitus") and observed that there is a correlation of 0.81 between Mantis-ML 2.0 scores and BIKG connectivity. Among genes with a Mantis-ML score > 0.9, the gene *BMP4* has curiously low BIKG connectivity. Similar analysis of a set of two atherosclerosis phenotypes ("atherosclerosis" and "coronary artery atherosclerosis" terms from HPO) suggests that the genes *APOC2* and *LIPA* are understudied despite being likely associated with cardiovascular disease. Notably, *BMP4* was recently reported to mediate insulin signaling (*22*), while loss of either *APOC2* (*23*) or *LIPA* (*24*) in rodents led to atherosclerosis and may be worthwhile avenues of research. Similar analyses between Mantis-ML and BIKG may uncover and prioritize previously unidentified candidate genes for future study. Furthermore, we observe two diseases (both in GEL) that exhibit negative correlations with BIKG node degree ("Laterality disorders and isomerism" and "Primary ciliary disorders"). We see that in these two cases, the fraction of input genes that were also included in the BIKG gene-gene graph was only 87.5 and 93.1%, respectively, placing them in the bottom 5th percentile of all GEL diseases by this measure. Moreover, in neither case was the seed gene overlap in the top 20 features learned by Mantis-ML 2.0. Both of these observations suggest that the effect of the graph in these two diseases may have a comparably smaller contribution than for other diseases and, instead, other tabular features may be dominating the predictions.

Validation of top Mantis-ML predictions with NLP-inferred gene-disease associations from literature

To identify more understudied gene candidates for diseases beyond diabetes and atherosclerosis, we automated our manual method above by integrating the existing machine learning tool AMELIE (Automatic Mendelian Literature Evaluation) (*25*). AMELIE crawls PubMed and scores the strength of queried gene-disease relationship based on the literature and other preexisting association scores. AMELIE also outputs a list of relevant PubMed articles for researchers to parse prior studies (Fig. 4D). By applying AMELIE to high-scoring Mantis-ML genes with low BIKG connectivity, AMELIE may accelerate research into understudied but promising genetic links to disease.

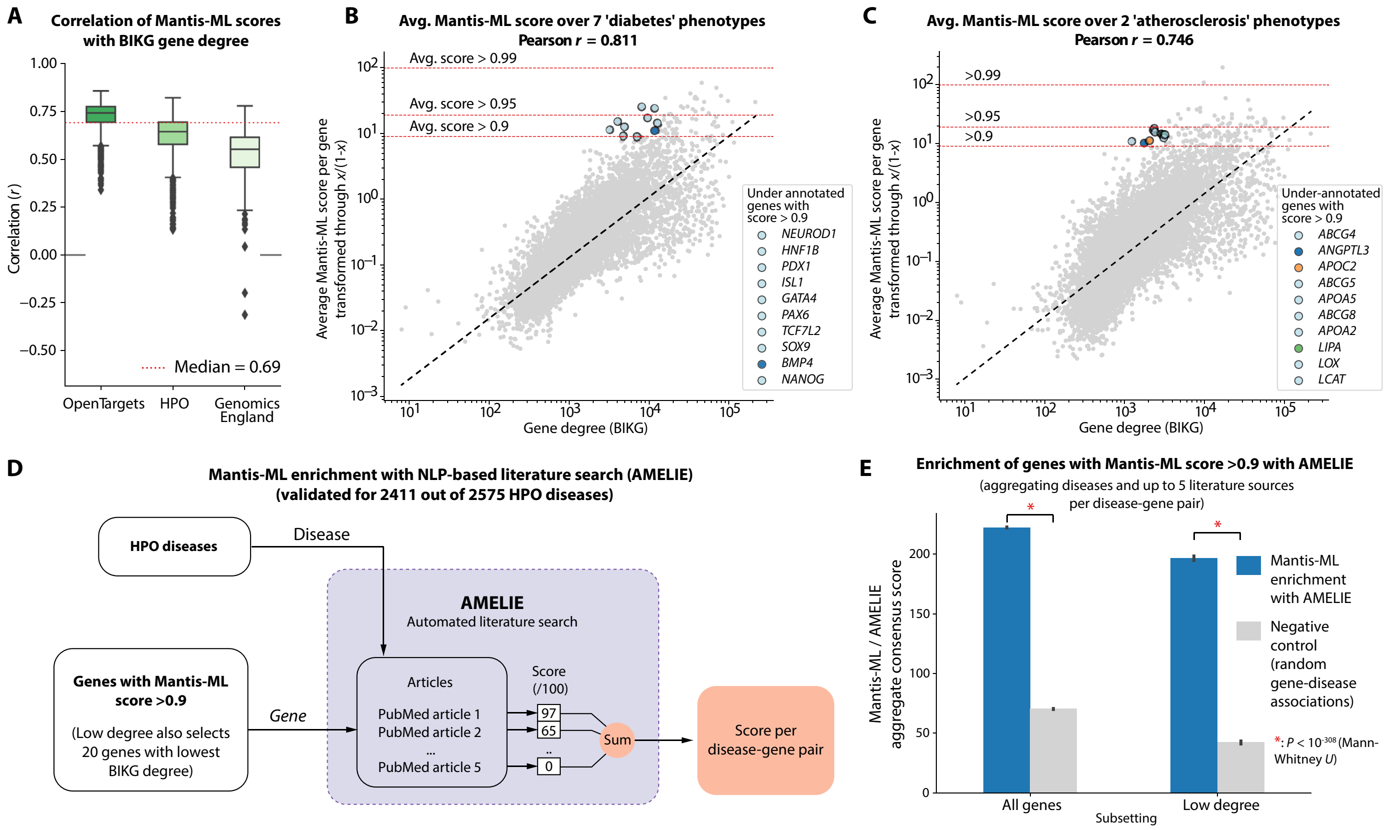

Fig. 4. Exploring Mantis-ML performance on under-annotated genes. We use the node degree in BIKG as a proxy for the number of annotations for each gene. (**A**) Boxplot of correlations between transformed scores [through log(x/(1 − x)] and log(BIKG degree). Both transformations are monotonic and correspond to the correlation parameters inferred from the subsequent scatter plots (each point in the box represents a unique disease). (**B**) Scatterplot of transformed scores against BIKG degree on log-log scales for a collection of seven diabetes phenotypes. (**C**) Scatterplot of transformed scores against BIKG degree for a collection of two atherosclerosis phenotypes. (**D**) Flowchart of enrichment process to validate Mantis-ML scores against the literature using an automated ML method (AMELIE). (**E**) Average enrichment of high-scoring Mantis-ML with AMELIE, focusing on all genes and also those 20 genes with the lowest node degree—capturing the enrichment with under-annotated genes.

Since AMELIE and Mantis-ML are two distinct approaches to scoring gene-disease associations, we first verified whether AMELIE favors a similar set of genes as Mantis-ML for a particular disease. We examined whether genes that score highly in Mantis-ML (>0.9 for a given disease) were enriched among those ranked highly by AMELIE compared to a random sample of genes output by Mantis-ML. The size of the random gene set was matched to the number of genes scoring >0.9 for a given disease. Compared to random gene sets, we find that there is an approximately threefold enrichment of highly scoring Mantis-ML genes among those favored by AMELIE ($P < 1 \times 10^{-308}$ subject to machine precision, one-sided t test), implying that Mantis-ML and AMELIE agree that certain genes are promising leads (Fig. 4E).

To shed light on more understudied candidates, we then applied AMELIE to high-scoring Mantis-ML 2.0 genes with low BIKG connectivity. We focused on the 20 genes with a Mantis-ML score > 0.9 but the lowest BIKG connectivity per HPO disease. We find that poorly connected genes are consistently favored by AMELIE compared to a random gene set as well, suggesting that AMELIE can help prioritize gene subsets for a broad set of phenotypes (Fig. 4E).

Phenome-wide validation of Mantis-ML predictions using exome-wide association studies

With the ever-increasing number of genetic datasets, we are identifying more gene- and variant-level relationships. However, more sequencing does not always lead to clearer answers. With large-scale genomic datasets, a major challenge is that some candidates fail to meet genome-wide significance after correcting for multiple hypothesis testing; among those that do not, it is not always clear which to prioritize for further experimental validation. In common variant studies, one of the biggest challenges that remain is the variant-to-function challenge of understanding the causal gene for a given significant locus. To test whether Mantis-ML can improve triaging among the highly ranked but not statistically significant genetic associations or among the many genes that could be driving a common variant loci signal, we cross-referenced genes with Mantis-ML predictions to associate with a trait against genes nominally significant ($P < 0.05$) for the trait in a recent phenome-wide association study (PheWAS) of 454,669 exomes from the UK Biobank (UKB) (*2*).

To compare gene sets between Mantis-ML and UKB, we first had to map the original disease terms in HPO, OT, and GEL to their most relevant ICD10 (International Classification of Diseases, 10th revision) codes in UKB. We matched disease terms to ICD10 codes by their semantic similarity as measured by BioWordVec (*26*), an established projection of millions of words mined from biomedical literature to a 200-dimensional Euclidean space (Fig. 5A). Ninety percent of disease terms (4688 of the original 5220) were validated, subject to filtering criteria designed to minimize the number of times semantic similarity was reassessed between disease terms and ICD10 codes (described further in Materials and Methods). We find that ICD10 chapter 17 ("Congenital malformations, deformations, and chromosomal abnormalities") represents many disease terms, comprising between 10 and 20% of each resource. In contrast, chapter 20 ("External causes of morbidity and mortality") matches to the fewest terms in each resource, as expected, since the indications included in the analysis must have some genetic basis. Our final compendium mapping disease terms across HPO, OT, and GEL to the most relevant ICD10 codes is available as fig. S25.

To gauge the agreement between Mantis-ML and UKB PheWAS, we evaluated the overlap between the highest-ranking Mantis-ML genes from each resource (top 5% in HPO, OT, and GE) and genes that achieve genome-wide significance ($P < 10^{-8}$) in the UKB PheWAS by Fisher's exact test. We tested overlap between genes that associate with a Mantis-ML trait and the 20 most semantically similar ICD10 codes because mapping traits to ICD10 codes is imperfect, and restricting overlap to only one ICD10 code may overlook more biologically relevant codes. We then compared how much more the gene sets derived from Mantis-ML overlap with the PheWAS of the relevant ICD10 codes. We find that genes derived from HPO and OT Mantis-ML significantly overlap with their relevant

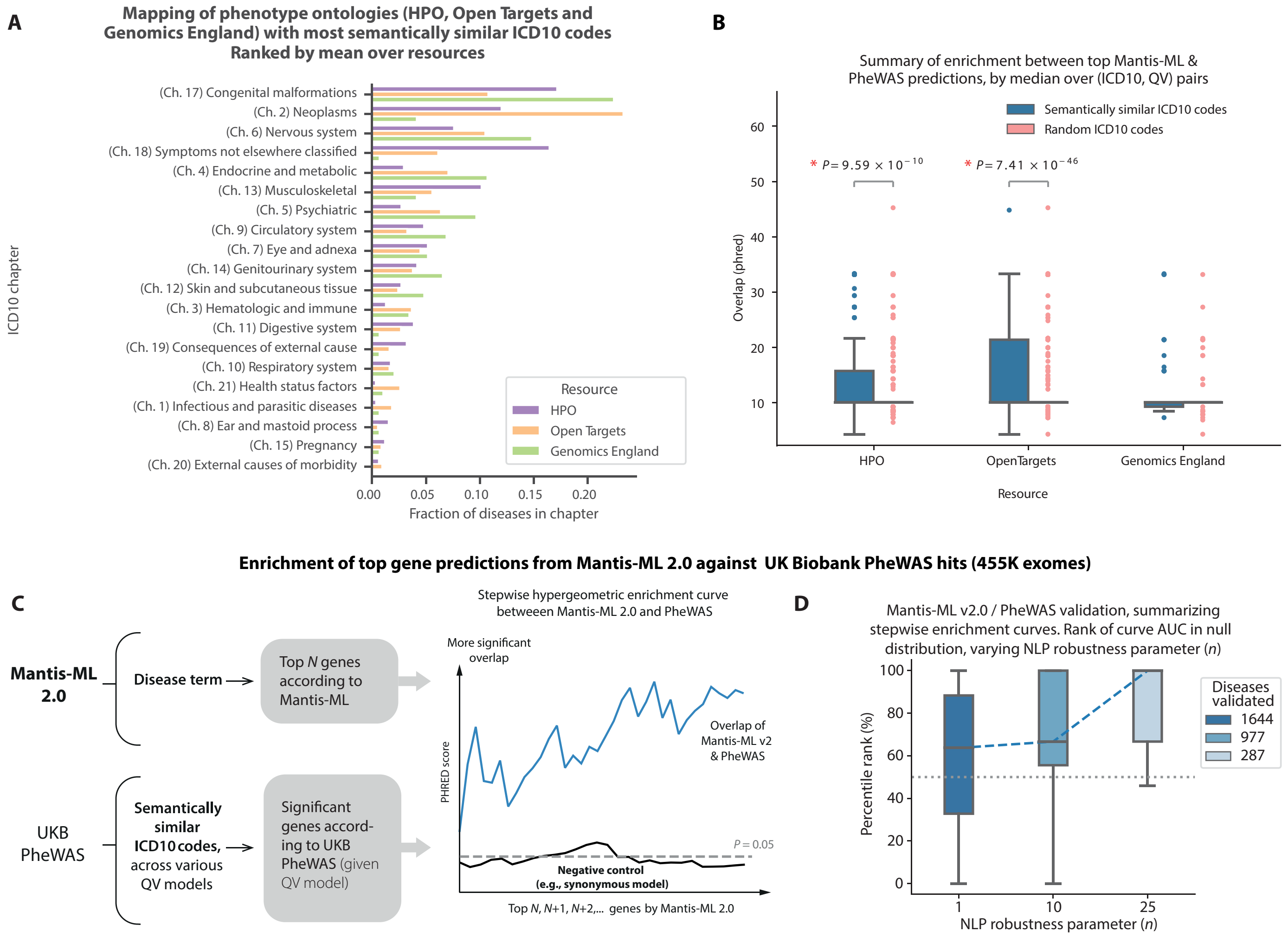

Fig. 5. Validating the utility of Mantis-ML predictions to triage significant hits derived from rare-variant collapsing analyses on semantically similar phenotypes. (**A**) Semantically similar ICD10 chapter to each disease from three resources considered (HPO, OT, and GEL). (**B**) PHRED scores measuring the overlap between high-ranking Mantis-ML genes and those that are significant for semantically similar PheWAS phenotypes. The overlap in each is quantified through Fisher's exact test. Overlaps were calculated between the (at most) 20 most semantically similar ICD10 codes to a given disease and the (at most) 11 genetic architectures (QV models). PHRED scores are broken down by resource, aggregating the overlap over similar ICD10 codes and QV models, with each data point in a box representing a unique disease. A negative control is provided (orange), which aggregates instead over a set of randomly chosen ICD10 codes, and asterisks indicate a significant difference between each pair of blue and orange boxes ($P < 0.05$, one-sided t test). (**C**) Schematic summarizing stepwise enrichment curves of diseases with semantically similar ICD10 codes in PheWAS. (**D**) Summaries of stepwise enrichment curve across a broad set of diseases. Each curve is collapsed to a single value (taking the AUC), and then the percentile of this value within a null distribution is evaluated.

PheWAS counterpart traits ($P = 9.6 \times 10^{-10}$ and $P = 7.4 \times 10^{-46}$ for HPO- and OT-derived genes, respectively) (Fig. 5B). In contrast, GEL-derived genes do not significantly overlap with the relevant UKB PheWAS traits ($P = 0.28$), although this is not unexpected since GEL emphasizes data from pediatric patients who are not enriched for in the UKB.

Stepwise hypergeometric enrichment with UKB PheWAS

With greater confidence that Mantis-ML can corroborate UKB human PheWAS, we explored whether Mantis-ML may guide how to interpret and prioritize more equivocal and not yet statistically significant highly ranked PheWAS gene results. Here, we expanded our enrichment analysis to measure the overlap between a trait's top 5% Mantis-ML candidates and all genes ranked by their PheWAS significance, rather than only those meeting genome-wide significance (Fig. 5C). We perform a series of enrichment analyses that measure the overlap between top Mantis-ML candidates with increasingly relaxed PheWAS thresholds and then estimate how likely the overlap is due only to chance (described further in Materials and Methods).

We again used pretrained word2vec embeddings to map the disease terms underlying Mantis-ML to relevant UKB ICD10 codes, identifying the *n* most semantically similar codes. This allowed for the introduction of *n* stepwise hypergeometric enrichment tests for each qualifying variant (QV) collapsing model. Data from the corresponding ICD10 codes and QV model were aggregated for each disease. Since both Mantis-ML and PheWAS contain thousands of phenotypes, many gene sets for unrelated phenotypes overlap by chance. To estimate this background overlap, we compared Mantis-ML outputs with PheWAS results obtained from randomly sampled ICD10 codes (Supplementary Methods). A gene that ranks highly in the background analysis might be associated stochastically with unrelated ICD10 codes, while a gene that ranks lower may be tied to semantically similar codes. This approach is limited by imprecise semantic maps between phenotypes, the absence of a genetic component for a disease, and/or the statistical power of patient cohorts.

As illustrated in Fig. 5D, more genes rank highly in the background analysis as the filtering requirements become more stringent, with the median percentile increasing from 64 to 100% as *n* increases. This may be because fewer ICD10 codes map to a disease term at higher values of *n*.

Our accompanying web resource includes enrichment curves to demonstrate validation results and a robust gene prioritization approach. It examines genes that show strong associations in both PheWAS and are also supported with high Mantis-ML probability scores. The resource enables users to explore validation results for the top-ranked Mantis-ML genes associated with different ICD10 codes. These codes are selected on the basis of their semantic similarity to the original disease from a pool of the 100 closest codes.

Benchmarking with other knowledge graph–based and pathogenicity tools

To determine how Mantis-ML 2.0 may aid researchers alongside other gene prioritization instruments, we compared Mantis-ML 2.0 to two recently published tools, PhenoApt (*27*) and Knowledge Graph Analytics Platform (KGAP) (*28*). As a trial, we measured how well each tool ranked genes for 14 different diseases/phenotypes by calculating how well the top 500 hits from each resulting gene set overlapped with top-ranked ($P < 0.05$) genes from a UKB PheWAS (*2*). We chose to test each tool in 14 diseases across various therapeutic areas: "abnormality of the immune system," "acute myeloid leukemia," "anxiety," "asthma," "cardiomyopathy," "chronic kidney disease," "congestive heart failure," "cystic liver disease," "dementia," "diabetes mellitus," "hypercholesterolemia," "parkinsonism," "pulmonary fibrosis," and "ventricular arrhythmia."

Mantis-ML outperformed both PhenoApt and KGAP, overlapping with more genes in the PheWAS validation set for 13 of the 14 diseases (Fig. 6A). Notably, KGAP did not yield any genes for three diseases: "chronic kidney disease," "abnormality of the immune system," and "cystic liver disease." In side-by-side comparisons, Mantis-ML 2.0 was enriched for more PheWAS hits than PhenoApt in 8 of 13 diseases and outmatched KGAP in 8 of 11 diseases (two-sided *t* test, $P < 0.05$) (Fig. 6A and table S2). When considering all UKB PheWAS hits in the aggregate, top-ranked Mantis-ML genes were also significantly more enriched for top-ranked PheWAS genes than PhenoApt (two-sided *t* test, $P = 1.5 \times 10^{-5}$) and KGAP (two-sided *t* test, $P = 7.3 \times 10^{-40}$) (Fig. 5B).

We also investigated which of the three tools better predicts the genes that will achieve greater significance as the size of a genetic cohort increases. Using the same set of 14 HPO phenotypes as above, we evaluated whether Mantis-ML, PhenoApt, or KGAP identified the more promising PheWAS hits ($P < 0.001$) from a UKB study that increased in sample size from 150,000 to 450,000 exome-sequenced participants (*29*). We compared each tool's scoring across a set of up to four biologically similar ICD10 codes per phenotype (table S3). Compared to a null model with a random gene set of the same size, Mantis-ML 2.0 was successfully enriched for genes that achieved higher significance as the UKB cohort size increased from 150,000 to 450,000 samples. PhenoApt performed similarly for 8 of the 10 top percentile thresholds but fared worse than Mantis-ML in 8 of 10 cases (Fig. 6C). In contrast, KGAP did not perform better than the null model. Overall, Mantis-ML 2.0 outperformed PhenoApt (two-sided *t* test, $P = 0.041$) and KGAP (two-sided *t* test, $P = 1.8 \times 10^{-7}$) across all 14 phenotypes and all examined top percentile thresholds (Fig. 6D). These results underscore the power of Mantis-ML 2.0 to identify promising biological candidate genes among the top-ranked genes that are not yet statistically unequivocal in large human genetic studies.

Last, in addition to the preceding benchmarks, we also examine whether an aggregate measure of gene pathogenicity derived by Mantis-ML's predictions correlates with existing intolerance metrics or pathogenicity tools (see Materials and Methods and fig. S37). Broken down by resource, we see that the highest measure of Mantis-ML–derived pathogenicity incurs a correlation of 0.43 ($P < 1 \times 10^{-308}$) with AlphaMissense (*30*), demonstrating that we are able to capture some degree of aggregate pathogenicity profile per gene using the derived scores.

Web resource overview

The web resource provides an interactive interface to query Mantis-ML probability scores computed phenome-wide as well as follow-up analyses (figs. S26 to S29). The web resource presents extrapolated disease and gene networks, suggesting additional candidate genes and phenotypes for future research. Mantis-ML can be explored from two perspectives—disease-centric or gene-centric.

In the disease-centric view, users can search for a specific disease to identify genes with the greatest probability scores for biological relevance. Mantis-ML provides raw scores, GO enrichments for the top-ranking genes, and a comparison with PheWAS results. The

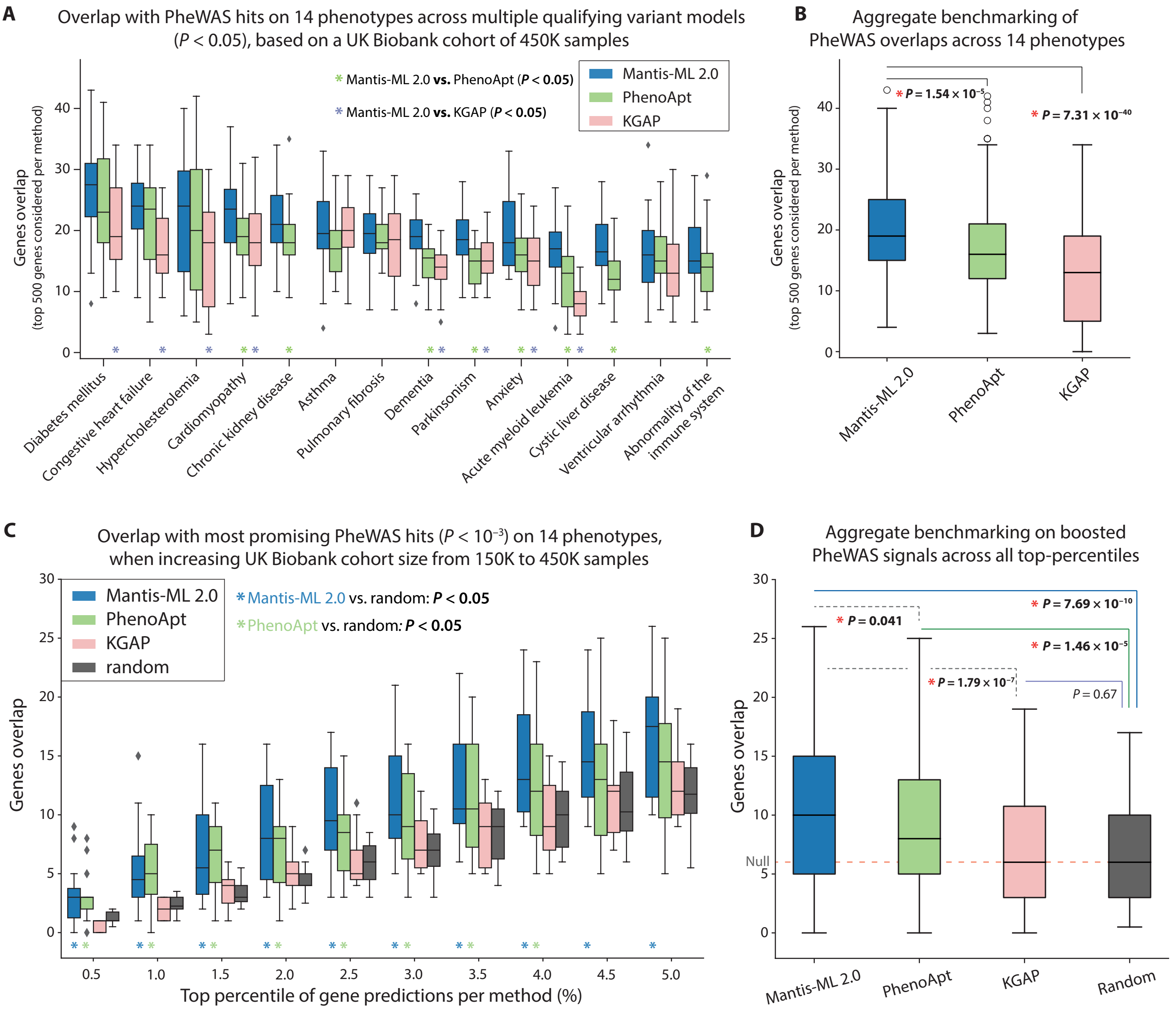

Fig. 6. Benchmarking with other methods and validation using results from rare-variant collapsing analyses on 450,000 whole exomes from UKB. (**A**) Overlap of top 500 gene predictions per method with nominally significant PheWAS hits ($P < 0.05$) across 14 HPO diseases and 10 QV models. (**B**) Aggregate benchmarking of gene predictions on 14 HPO diseases. (**C**) Overlap of top percentile predictions per method against most promising PheWAS hits ($P < 10^{-3}$) when increasing the UKB cohort size from 150,000 to 450,000 whole exomes. (**D**) Aggregate benchmarking of gene predictions per method on the boosted PheWAS signals.

gene-centric view provides a comprehensive profile of diseases associated with a gene, helping determine the extent of phenotypes the gene has a high probability of biological relevance. Users can also access information on the feature set used for gene classification and the initial list of positively associated seed genes adopted to construct the corresponding probability scores.

DISCUSSION

The phenome-wide application of Mantis-ML offers a tangible machine learning framework to organize and absorb the growing array of genetic data, representing a powerful tool for prioritizing genetic targets for further experimental research. Whereas Mantis-ML 1.0 relied on manually inputting features of a phenotype or disease, Mantis-ML 2.0 now uses NLP reasoning across three sets of ontologies to enable automated selection of disease-specific features and integrates a far more comprehensive knowledge graph capturing gene-gene interactions. Our results demonstrate a high degree of predictability across diseases within each ontology. Compared to other contemporary tools, Mantis-ML is a major step forward in triaging among the top-ranking signals from genetic studies without clear-cut statistically significant findings.

Since Mantis-ML 2.0 is a semi-supervised learning framework, the selection of seed genes plays an important role in determining the gene-disease prediction scores across the whole exome. We have used three diverse phenotype resources for selecting seed genes

(HPO, OT, and GEL), capturing a broad spectrum of diseases and phenotypes. Some of the disease terms Mantis-ML has been trained on may not directly correspond to clinical disease terms but rather refer to more generic phenotypes, some of which may still be relevant to clinical phenotypes. Thus, the translational utility and further application of the Mantis-ML predictions are influenced by the clinical relevance of the seed genes provided as input. To this end, we accompany this work with an open-source release of the Mantis-ML 2.0 software package so that researchers can train additional models using custom sets of seed genes that are directly related to their diseases and/or subphenotypes of interest.

Mantis-ML 2.0 was rigorously validated by comparing high-scoring genes against specific ICD10 codes in an independent UKB PheWAS. Compared to randomly sampled ICD10 codes from the UKB PheWAS, Mantis-ML agreed more with semantically similar codes. Such strong concordance with large-scale human genetic studies supports the applied utility of Mantis-ML to triaging the top-ranked gene-disease associations to identify the signals with the greatest biological plausibility in an entirely human-unbiased manner for further experimental study.

Limitations to validating Mantis-ML 2.0 scores stem from noise introduced at different stages of the pipeline, particularly when NLP mapping between the three resources (HPO, OT, and GEL) and those in PheWAS; it is unlikely that all high-ranking Mantis-ML genes were well matched to semantically similar PheWAS terms. Mantis-ML and PheWAS studies are also only as reliable as the underlying data. Although human genetic studies are uninfluenced by prior literature, they are limited by the statistical power affordable to a given genetic study. Mantis-ML probabilities, on the other hand, are the result of a powerful synthesis of large volumes of existing public knowledge beyond what a human could achieve. However, the theoretical constraints may come in finding truly unprecedented human biological discoveries to which there is no/little prior literature. Together, however, they are highly complementary approaches that enable better informed and unbiased decisions.

Furthermore, a number of nontrivial methodological extensions exist, going beyond the phenome-wide deployment of Mantis-ML described here. First, we have focused on gene-gene subgraphs of the entire BIKG; however, a full incorporation of all the different entities (e.g., disease-gene and disease-disease links) may better capture the underlying disease biology and allow information sharing between the different node types. A full treatment of this is outside the scope of this article; however, it may be feasible to refine the scores beyond their current predictions using edge prediction methodology, such as generic node embeddings (*31*) or learning link heuristics using GNNs (*32*). Second, it may be possible to additionally exploit NLP techniques to expand the input gene lists used by Mantis-ML, thereby enhancing score robustness, with gene lists from semantically similar diseases aggregated (taking, e.g., the union) to ensure a larger number of positive examples in training. Third, it could be instructive to include additional datasets, leveraging results from any number of preexisting knowledge graph databases to further enhance predictions (*33*, *34*). Last, it may also be possible to integrate PheWAS results directly in training, taking their association indicator as another gene feature and predicting scores across the exome using a form of cross-validation, ensuring that final scores are “out of fold.” This is left as further research for the future versions of Mantis-ML.

In addition to our findings and methods detailed here, we have developed an accompanying interactive web resource for researchers to explore Mantis-ML 2.0 in detail. This resource facilitates gene prioritization, detailing top-scoring Mantis-ML genes that are highly ranked signals in UKB cohort studies. We have also created interactive networks of diseases and genes derived from Mantis-ML; these clusters represent similarities between diseases or genes based on their reciprocal genetic or phenotypic components and may serve as reservoirs for research into further genetic interactions. As a bridge for the community to Mantis-ML, we hope that this resource will provide insights into the relationships between genes, diseases, and their shared genetic components to advance treatment strategies.

MATERIALS AND METHODS

Gene prioritization using stochastic semi-supervised learning

Mantis-ML leverages the following types of information to prioritize genes related to a disease of interest and potentially uncover previously unidentified gene-disease associations: (i) known gene-disease associations in the form of a seed gene list; (ii) known disease mechanisms such as tissue, pathways, and processes in the form of free text provided by the user; (iii) extensive gene annotation such as genic intolerance and tissue expression; and (iv) gene connectivity in the form of a knowledge graph.

On the basis of the data above, the problem of gene prioritization can be formulated as a semi-supervised classification problem, where the structured gene features correspond to the features/predictors matrix (genes as rows, features as columns) and the seed genes correspond to the positive labeled samples (genes) of the response vector. For negatively labeled genes, we randomly sample genes not present in the seed list. The sampling is repeated as many times as needed to guarantee full coverage of the exome, and the whole process is repeated for 10 stochastic iterations to provide robustness in the results. While knowledge graphs themselves are capable of capturing heterogeneous data types, we focus on a subset of the knowledge graph capturing connections between entities of gene type only. The working hypothesis is that genes that are highly connected according to the knowledge graph may have a higher likelihood of affecting the same biological pathways—and so incur similar disease associations. We incorporate this connectivity through GNNs, exploring a sophisticated (but computationally expensive) model as well as a simpler but computationally cheaper model—GCNs and SGCs, respectively. In addition, we note that the semi-supervised learning is inherited directly from the original Mantis-ML method (i.e., repeatedly sampling random partitions and then aggregating predicted scores). Mantis-ML 2.0 refines on this by converting the traditional feature classifier to a graph-structured node classifier (fig. S33).

Newly implemented gene features in Mantis-ML 2.0

Here, we discuss the newly implemented gene features in Mantis-ML 2.0: (i) graph-derived features, (ii) generic GO (*35*, *36*) signature, (iii) disease-specific GO signature, (iv) single-cell transcriptomics data as gene features, and (v) user specified—custom gene features.

Graph-derived features

Network science is the application of graph theory to understand complex systems (*37*). The subfields of network biology (*38*) and network medicine (*39*) aim to better understand the behavior of molecular networks and the role networks play in human disease. By applying the ideas of graph theory to biomedical knowledge graphs and generating graph-based features, it can capture important information about a node's roles within the network. For example, its popularity, influence, and communities, which correspond to how influential a gene is or identifying disease modules.

Therefore, the incorporation of the following graph-derived features was explored in Mantis-ML 2.0: (i) Leiden cluster centrality (*40*) for communities, (ii) core number (*41*) for identifying cohesive subgroups with relatively strong links, (iii) node degree (in, out, and total degree) for highlighting highly connected node hubs and specifically out degree for how outwardly interactive the node is, (iv) Louvain cluster centrality (*42*) also for communities, (v) Katz cluster centrality (*43*) for the influence of a node, (vi) and PageRank (*44*) for node importance. Despite the seemingly more flexible nature of introducing more graph-derived features, we ultimately did not use any except the seed gene overlap (defined as the percentage overlap between one-hop neighbors and the seed gene input list) as the rest of graph-derived features did not contribute to an increased AUC (Fig. 2).

Generic GO signature

To derive a generic GO signature across the exome, we first built a GO similarity network based on the genes each GO term/set includes. Specifically, for each gene, we created a binary vector where each element corresponds to the presence or absence of a specific GO term. We calculated gene similarity based on the Jaccard index for each pair of genes (represented as binary vectors). To define the presence of an edge between a pair of genes, we assessed whether their Jaccard similarity score fell in the top 50th percentile of the whole distribution, eventually translating similarity scores into binary values, i.e., presence (upper 50th percentile, represented by value of 1) or absence (lower 50th percentile, represented by value of 0) of an edge. We then clustered the GO similarity network using k-means ($n = 20$) and selected from each cluster the GO term with the maximum number of genes. With this approach, we aim to maximize both variance (sampling one term from each cluster) and coverage across the exome (sampling the largest GO set from each cluster). Eventually, we managed to capture ~10,000 genes among the 20 derived clusters, which represent more than 50% of the entire exome, and using only 20 instead of ~12,500 GO terms.

The set of 20 GO terms that we have extracted are the following: cell adhesion, cell cycle, chemical synaptic transmission, immune system process, innate immune response, ion transport, lipid metabolic process, mRNA processing, neutrophil degranulation, oxidation-reduction process, phosphorylation, positive regulation of guanosine triphosphatase (GTPase) activity, posttranslational protein modification, protein dephosphorylation, protein transport, proteolysis, regulation of ion transmembrane transport, regulation of transcription by RNA polymerase II, signal transduction, and translation. Within Mantis-ML, the GO signature of each gene is then calculated as its membership (Boolean values) to the respective GO terms, yielding 20 additional binary gene-level features.

Disease-specific GO signature

Here, we use the seed genes to identify GO terms they are significantly enriched in. Enrichment is calculated using Fisher's exact test. The top *N* significantly enriched terms are extracted where *N* is defined by the user ($N = 10$ was used throughout this paper). Then, the disease-specific GO signature of each gene is calculated, defined as its membership (Boolean) to the respective highly enriched GO terms.

Single-cell transcriptomics

Here, we use single-cell transcriptomics data as gene features (*11*). The single-cell transcriptomics is summarized as the average expression value of each gene across the available cell populations. Each cell population yields a new gene feature in the features matrix. Specifically, the datasets were pulled from publicly available single-cell data resources that have already been gone through extensive quality checks, including the Human Protein Atlas, the UCSC cell browser, and Azimuth. Using Seurat (*45*), we first normalized read counts using the NormalizeData function and then computed the average expression of each gene in each cell type using the AverageExpression function. These datasets are derived from putatively healthy tissues, specifically blood, brain, breast, colon, esophagus, heart, kidney, liver, lung, lymph, bone marrow, muscle, ovary, pancreas, prostate, skin, spleen, stomach, and testis (table S5). Thus, this comprehensive list includes most of the relevant cell types for human disease. Identification of the most suitable tissue for the disease of interest is carried out using NLP. More details about how NLP is used are included in the section "NLP for automated selection of disease-specific features."

Custom gene features

In addition to the gene features integrated in Mantis-ML 2.0, we allow the user to provide their own custom gene features in csv format. Rows in the csv correspond to genes, while columns correspond to the custom features. Missing values are by default imputed with zeros, but the user may select to impute using the median or the mean.

Biological Insights Knowledge Graph

The previously published BIKG (*46*) aims to model the fundamental interactions within biological systems by combining data from 55 public, licensed, and internal AstraZeneca data sources into a unified knowledge graph that can then be used for drug development tasks (*46*–*48*). The full BIKG includes 14 million entities covering 25 node types such as gene targets (genes and proteins), pathways, biological processes, diseases, and compounds. The entities are linked together by over 146 million edges that describe many biological relationships such as protein-protein interaction, drug-drug interaction, and gene-disease association, as well as relationships extracted from the scientific literature.

The knowledge graph we use in Mantis-ML 2.0 is a subset of BIKG. The chosen subgraph is composed of gene targets (genes and proteins) and the relationships between them. In total, the specific knowledge graph we use contains 8.7 million edges between 17,197 genes. Graph-derived features are then calculated on the BIKG subgraph to capture the structural information.

In addition, for 13 heterogenous diseases, we perform a direct comparison between BIKG-derived scores and InWeb-derived scores. We see, first, that InWeb requires more iterations (30 rather than 10) for the stochasticity robustness in the scores to be comparable with those derived from BIKG (fig. S35). With this configuration, we do observe lower AUC for InWeb; however, despite the differences in graphs, we do see that diseases can share up to 40% overlap in the top 10 to 10,000 genes (fig. S36). We note, however, that InWeb is comparably smaller—with only 600,000 edges, compared to BIKG's 8.7 million edges.

Modeling gene prioritization using GNNs

GNNs are used to effectively leverage the connectivity (such as gene-to-gene) in the BIKG network within a semi-supervised learning framework for the prioritization of gene-disease associations. GNNs in this context take as an input a graph, a node features matrix, and a response vector. Their predictive power comes from using the graph to propagate the node features and response variable to neighboring nodes.

We consider two classes of GNN models, principally, GCNs and SGCs (*14*). The former uses deep learning to capture nonlinear dependencies in graph-structured data, while the latter reduces to a linear model capturing network features with improved computational efficiency (though with a potential cost of lower model expressivity). Both may be used to perform node classification, which is later exploited within Mantis-ML to perform semi-supervised learning—classifying genes as either "associated" or "not associated" to a given disease based on a set of positive examples and an assembled list of gene-level features (fig. S31 to S33).

In Mantis-ML, as the graph, we use the BIKG network, and as node features, we use a wide range of gene features such as graph-derived features, GO, biological processes, gene-disease associations from genome-wide association study, Online Mendelian Inheritance in Man (OMIM), HPO, and gene expression across different tissues. As a response vector, we use a binary variable denoting whether the corresponding gene is known to be associated with the disease of interest (positive labeled genes) or not (negative labeled genes).

Since the number of negative labeled genes is much greater than the number of positive labeled genes, the negative labeled genes are subsampled to a ratio 3:2 (three negative labeled genes for every two positive labeled genes), yielding a more balanced dataset better suited for binary classification. This subsampling happens multiple times, generating an equal number of predictions for each gene. Gene predictions are finally averaged across samplings to yield a single gene prediction (fig. S33). The subsampling is implemented in the GCN framework by setting the value of the corresponding sample weights to 0 (for genes left out) or 1 (for genes included in the sampling). This is preferred over removing the genes from the response vector as this would lead to gaps in the graph, hindering performance. For SGC, model parameters are learned on a subset of the graph used only in training, and predictions are made on a graph that includes nodes in both the training and test data.

The GCN implementation we use is the one in Stellargraph. We have experimented with various configurations in terms of the number of hidden layers, number of filters, dropout ratio, learning rate, and number of epochs. The configuration we decided on for the bulk of the runs here uses two hidden layers, 16 filters in each hidden layer, dropout ratio of 0.5, 200 epochs, and a learning rate of 0.01. ReLU activation was used in the hidden layers, and sigmoid activation was used in the output layer (*49*). The Adam (*50*) optimizer was used and binary cross entropy as a loss function.

Fine-tuning SGCs

SGCs provide a computationally efficient alternative to GCNs, with the caveat that they only offer a simpler linear dependence between features and responses. Nonetheless, we explore them as a potential modeling alternative after a degree of parameter and architecture tuning. At their core, SGCs in the classification framework comprise logistic regression classifiers using features matrices X (rows represent independent samples, columns represent different features) premultiplied by a matrix related to the graph adjacency matrix. Denoting the latter matrix as $\widetilde{A}$, we see that the classifier acts on features by

$$Y \sim \text{LogisticRegression}(\widetilde{A}^K X)$$

where K is a power of the modified adjacency matrix $\widetilde{A}$. The power controls the size of the neighborhood influencing predictions at each node—the higher the power, the larger the neighborhood. Such an approach is designed to mimic GCNs but without the nonlinearities between layers.

We identify three possible ways in which the methodology can be modified and tune each on a subset of the diseases processed by Mantis-ML 2.0. These are the following:

1. The matrix power, K
2. The regularization parameter used by the logistic regression
3. The adjacency matrix $\widetilde{A}$

For tuning the matrix power K, we consider a random subset of 20 diseases from each resource and compare the AUC predictive performance of each. We see, overall, that there is a comparatively small effect on the median AUC of changing the matrix power. That being said, we identify the optimal value to be $K = 2$ (figs. S1 to S3).

Logistic regression has relatively few tuning parameters compared to deep learning methods; however, it is possible to tune the regularization parameter. The regularization parameter controls the penalty of the norm of the weights in the linear classifier. Tuning this on 10 of the previous 20 diseases (the reduced number required for computational expediency), we see that across the resources the optimal value of the regularization parameter is C= 1 (fig. S2).

Last, we explore tuning the adjusted adjacency matrix $\widetilde{A}$. The original SGC framework sought to average information from neighboring nodes. Instead, it is possible to sum information. Another dimension explored is that the adjacency matrix from BIKG has entries >1, as it is possible to have multiple edges between the same two nodes (representing different types of gene-gene relationships). We therefore also explore thresholding the adjacency matrix to take values in {0,1}, proposing a potentially more robust set of relations between the genes. We assess these four different combinations—mean/sum neighbors and thresholding/no thresholding of the adjacency matrix—and observe that no configuration outperforms the default as originally proposed (*14*).

Disease and gene networks from phenome-wide Mantis-ML scores

Disease and gene networks were generated using Mantis-ML scores across diseases in a given resource. Disease networks were generated by projecting the 18,626 gene association scores per disease into a two-dimensional (2D) vector using t-distributed Stochastic Neighbor Embedding (t-SNE). Before the nonlinear transformation, the values were first rank-transformed. Such a transformation is invariant on whether it uses the raw scores or the normalized scores, as the latter is just an affine transformation of the former (transformed to ensure mean zero and unit variance). Gene networks were generated by taking the m-dimensional vectors of disease associations for each gene and projecting them into 2D again using t-SNE. Here, m depends on the resource, but taking OT, for example, it would be 2500 long, each entry corresponding to a different disease in the resource. As for the previous network, values were rank transformed before the nonlinear transformation. For gene networks, perplexity of 10 was used. For disease networks, perplexity of 20 was used except for GEL, which was 10. These values were subsequently validated in the case of the gene networks through inspection of neighborhoods around small gene clusters (figs. S4 to S9).

NLP for automated selection of disease-specific features

Mantis-ML receives as user input a list of terms (free-text) that capture certain aspects of the disease biology, e.g., relevant tissue, biological/signaling pathways, and biological processes. These free-text terms are then matched against the resources integrated in Mantis-ML such as MSigDB (*15*), MGI (*16*), and GTEX (*17*) to extract their associated genes. In Mantis-ML v1, this matching was implemented using regular expressions; however, this ignored the semantic similarity between terms, e.g., kidney and renal have similar meaning, although they would not be matched using regular expressions. To this end, in Mantis-ML 2.0, we use NLP for the identification of relevant annotation terms.

Robustness analysis of ICD10 chapter assignment by semantic similarity

To examine the breakdown of diseases considered in each resource, each disease was assigned an ICD10 chapter based on semantic similarity. Such a technique was also used when exploring the breakdown of PheWAS overlaps by ICD10 chapter. To assign each disease an ICD10 chapter, the semantic similarity was used between ICD10 codes and each disease term across all resources. On this basis, the ICD10 chapter was extracted from the ICD10 code and a majority vote was taken within the closest 5, 10, and 20 codes to each disease.

We see that the assigned chapters are highly robust to this choice of tuning parameter and the relative counts of each chapter do not vary substantially between these values of n (fig. S25). The agreement between chapter assignments at each value of n can be measured by examining the fraction of diseases that share exactly the same chapter assignment across each value. However, in some cases, the majority vote results in a tie between two or more chapters. Under the most stringent policy of requiring all chapter assignments to match across all diseases and for diseases with tied chapters to also be tied with the same values, the fraction of diseases in agreement is 67.5%. Under a less stringent policy of requiring only that there is at least one common chapter among the set of ties across each value of n, this figure rises to 80.2%.

Semantic similarity between disease names and annotation terms

To extract the most relevant annotation terms for the disease of interest, we use the BioWordVec (*26*) word2vec embeddings. BioWordVec embeddings is a mapping/projection of millions of words mined from biomedical literature to a 200-dimensional Euclidean space. The projection was created in such a manner that words with similar meanings are mapped closer together in the Euclidean space (fig. S34).

For the disease of interest, each word separately is embedded using the BioWordVec embeddings into the Euclidean space. Then, its pairwise distance is calculated against all available annotation terms in MSigDB, MGI, and GTEX and also embedded using the BioWordVec embeddings. The top 2 annotation terms from MGI and GTEX and the top 20 terms from MSigDB with smaller distance to the disease name are extracted and used in the disease-specific features.

Phenome wide Mantis-ML deployment across >5000 diseases

Mantis-ML is deployed on HPO (n = 2575), OT (n = 2500), and GEL (n = 145).

Automated feature selection summary

We summarize deployment as follows: Starting from the diseases in one of the three resources, we calculate its semantic similarity to the available gene sets in MSigDB (*15*), MGI (*16*), and GTEx (*17*) by leveraging the BioWordVec (*26*) word2vec embeddings. Semantic similarity between terms of varying length can be estimated by calculating an average of the pairwise distances between individual word embeddings (Fig. 2B). The best-matching gene sets from each resource will be used as annotation terms in the disease-specific features. After compiling the disease-specific features and input gene lists, Mantis-ML is deployed to all available diseases/phenotypes, including on average (median over diseases) 138 features of 16,497 (0.8%).

Selection of diseases from each resource

The input gene lists are extracted from either HPO, OT, or GEL. For HPO, we are using a local installation, while for OT we are using the OT Application Programming Interface (API). The top (N) most associated genes are used for each disease based on a series of thresholds as follows. For each disease, gene-disease associations are extracted either from a local installation (HPO and GEL) or using the OT API. For HPO and GEL, we only consider diseases with a minimum of 30 associated genes. All the gene-disease associations provided in the two resources are taken into account. For OT, we only consider diseases with a minimum of 100 associated genes. Gene-disease associations are sorted on the basis of their overall association score, and only associations with a score greater or equal to 0.2 are taken into account. If more than 500 gene-disease associations are provided, then the top 500 are considered.

Validation with UKB PheWAS results

UKB PheWAS disease-gene associations

Mantis-ML disease-gene associations are compared with a rare-variant collapsing analysis performed on UKB data, phenome-wide. The UKB PheWAS consists of 454,669 sequenced individuals and can be found at https://azphewas.com. Sequencing data have been aggregated and collapsed on the gene level with variants split into the following categories: protein truncating variants (ptv), rare damaging variants (raredmg), protein-truncating rare damaging variants (ptvraredmg), ultra-rare variants (UR), synonymous variants (negative control), etc. In addition to the genotype data, the clinical presentation of each individual is captured using the ICD10 codes. There are ~7000 ICD10 codes numbering one or more patients in UKB. The association between each gene to the available ICD10 codes is quantified using Fisher's exact test, with P values capturing the degree of association between genotypes and phenotypes.

Stepwise hypergeometric test for external validation

Validation made routine use of a stepwise hypergeometric test between Mantis-ML and genes associated under a PheWAS rare-variant collapsing analysis. The test itself proceeds by going down the ranking of Mantis-ML predictions, one by one, and calculating the overlap of the top N, $N + 1$, $N + 2$, …, etc. genes of the ranking against the statistically significant genes in the external resource. The overlap is quantified via Fisher's exact test, where at each iteration the total number of genes is held constant and only the number of genes to include in the first list increases incrementally. We extract the P value and calculate the PHRED score. PHRED scores define the stepwise enrichment curve showing the overlap of the Mantis-ML gene ranking with the PheWAS across the different positions in the ranking.

Stepwise hypergeometric tests for validation with UKB PheWAS

Mantis-ML association scores were validated using data derived from an independent large-scale cohort study of ~455,000 exomes contained in the UKB. In general, there is no clear mapping between the phenotypes measured in UKB organized in the ICD10 ontology, with disease terms in HPO, OT, or GEL. Hence, we identify the n most similar ICD10 codes to each disease (using BioWordVec embeddings as before) and calculate the overlap with significant associations in PheWAS using a stepwise hypergeometric test (*3*).

The test relies on, first, a list of genes ranked by PheWAS significance for a given ICD10 code and QV model and, second, a subset of high-ranking Mantis-ML genes. The former is defined on the basis of one of the n most semantically similar ICD10 codes, thereby eliminating the need to manually curate which ICD10 codes pertain to which diseases in HPO, OT, and GEL. For a given ICD10 code, the overlap with all genetic architectures—i.e., across 10 QV models—is calculated. The significance threshold for PheWAS was set to 0.05.

Summarizing stepwise enrichment curves across multiple ICD10 codes and QV models

Stepwise hypergeometric curves were calculated between high-ranking Mantis-ML genes and ranked lists of significant genes according to PheWAS. This was performed and aggregated over multiple QV models. Specifically, for a user-defined disease, the n most semantically similar ICD10 codes are identified. For each of these codes, stepwise enrichment curves are calculated between the ranked list of genes by Mantis-ML score and those genes that are significant under PheWAS. This is repeated for each of the 11 genetic architectures (including a synonymous model), resulting in $11n$ enrichment curves for each disease term. To summarize the result of this procedure, each of these curves is reduced to a single number, using either the AUC or taking the maximum of the curve (excluding curves that involved the synonymous QV model), and then aggregated into a single number for a given disease (taking, e.g., the median). This statistic represents the "average" overlap of high-ranking Mantis-ML genes and PheWAS significant genes over a number of semantically similar ICD10 codes and across a range of genetic architectures. This was repeated with n varied in a grid of [1, 10, 25].

For the purposes of comparison, we calculate average measures of overlap with randomly selected ICD10 codes, rather than semantically similar ICD10 codes. In particular, we are able to estimate the rank of the summary statistic derived using semantic similarity within those generated by randomly sampling ICD10 codes, termed "null statistics." Details of the precise sampling procedure for generating null statistics are described in detail in Supplementary Methods. Null statistics were calculated by computing stepwise enrichment curves with randomly sampled ICD10 codes and across all QV models (excluding the synonymous model). Mimicking the way the original method relied on aggregating over the "n ICD10 codes, null statistics were calculated by sampling sets of n (nonoverlapping) ICD10 codes and aggregating over these in a similar fashion. All QV models were included with the exception of the synonymous model, and each curve was summarized using either the AUC or its maximum. In some instances, a filtering procedure was performed to both the original statistics and the null statistics.

Disease filtering criterion when validating with UKB PheWAS

Different filtering criteria were used depending on the statistical test, i.e., whether it was Fisher's exact test or the stepwise hypergeometric test. In the case of the former, we impose a weak genetic basis for selecting ICD10 codes initially, in that the 100 most semantically similar ICD10 codes are first identified of all the ICD10 codes with at least one significant gene at 0.05. Following this, the 20 most similar ICD10 codes that have at least one significant gene are identified. For a given QV model and ICD10 code, a gene was considered significant if its P value under the collapsing analysis was less than 10^{-8} . In some cases, this will mean that none of the original 100 ICD10 codes are sufficient, in which case we exclude the disease from validation. In addition, for a given disease-ICD10-QV triplet to be included in the final analysis, the overlap (quantified by Fisher's exact test) P value must satisfy a relaxed threshold of $P < 0.5$ to eliminate any clearly nonsignificant enrichments that may occur in either the positive and null sets due to noise in ontology mapping and/or lack of statistical power from PheWAS. Such a criterion was included on the basis of the observation that many of the overlaps referring to unrelated phenotypes (in both the analysis and negative control) were either close to or equal to 1. Hence, the threshold is applied equally in the main analysis and negative control (provided by the null distribution). After all this filtering, the total number of diseases reduced from 5220 to 4688 (90.0%).

Benchmark against machine learning tools

We compared the performance of Mantis-ML 2.0 against two recently published tools: (i) PhenoApt (*27*), which uses a directional graph of genes, diseases, and phenotypes and applies graph embedding to prioritize candidate genes per disease, and (ii) KGAP (*28*), which uses a graph database to prioritize drug targets for associated diseases. We compared the performance of the three tools in ranking genes for 14 diseases by calculating the overlap of the top 500 ranked genes per disease from each method against significant ($P < 0.05$) genes from PheWAS on ~455,000 exomes contained in the UKB (*2*). We used PhenoApt web application to retrieve genes for each disease. As for KGAP, we installed and queried the database for each disease.

Benchmark against gene-level pathogenicity tools

As an additional validation, we compare an approximate measure of pathogenicity derived using phenome-wide Mantis-ML 2.0 scores with AlphaMissense's (*30*) mean pathogenicity per gene (fig. S37). To construct the Mantis-ML gene-level score, we average the scores for a given gene across all diseases in each resource, while for AlphaMissense we consider the mean pathogenicity per gene as its gene-level representation. We observe a relatively high degree of correlation between the transformed scores from Mantis-ML and AlphaMissense (Pearson $r = 0.43$, 0.36, and 0.38 for HPO, OT, and GEL, respectively; $P < 1 \times 10^{-308}$ across all three comparisons). We also compared the gene-level Mantis-ML scores with another established gene-level intolerance metric: Residual Intolerance to Variation Score (RVIS) (*51*). We observe that across all three resources, the correlation of pathogenicity is higher with AlphaMissense than it is with RVIS (fig. S37D). Correlations were obtained after log-transforming AlphaMissense and Mantis-ML's pathogenicity scores. In total, 92.3% of AlphaMissense Entrez transcript IDs could be mapped to HUGO Gene Nomenclature Committee (HGNC) using gProfiler (*52*). Of these, 90.0% shared a gene name with Mantis-ML, leaving 15,536 genes included ultimately in the analysis.

Code availability

We expose the code used to generate the phenome-wide Mantis-ML 2.0 scores on a public Zenodo repository: https://zenodo.org/records/10465793, as well as on a public GitHub repository: https://github.com/astrazeneca-cgr-publications/mantis-ml-release-2.0. The repositories include the InWeb gene-gene interaction graph as the default knowledge graph to be used during learning. However, any other knowledge graph may also be provided as input to Mantis-ML 2.0. The "Custom Features" option (Fig. 1) is also available should a user wish to use additional structured gene-level features in their own analyses.

Supplementary Materials

This PDF file includes:

Supplementary Methods
Figs. S1 to S37
Tables S1 to S5

REFERENCES AND NOTES

1. R. M. Plenge, E. M. Scolnick, D. Altshuler, Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
2. Q. Wang, R. S. Dhindsa, K. Carss, A. R. Harper, A. Nag, I. Tachmazidou, D. Vitsios, S. V. V. Deevi, A. Mackay, D. Muthas, M. Hühn, S. Monkley, H. Olsson, B. R. Angermann, R. Artzi, C. Barrett, M. Belvisi, M. Bohlooly-Y, O. Burren, L. Buvall, B. Challis, S. Cameron-Christie, S. Cohen, A. Davis, R. F. Danielson, B. Dougherty, B. Georgi, Z. Ghazoui, P. B. L. Hansen, F. Hu, M. Jeznach, X. Jiang, C. Kumar, Z. Lai, G. Lassi, S. H. Lewis, B. Linghu, K. Lythgow, P. Maccallum, C. Martins, A. Matakidou, E. Michaëlsson, S. Moosmang, S. O'Dell, Y. Ohne, J. Okae, A. O'Neill, D. S. Paul, A. Reznichenko, M. A. Snowden, A. Walentinsson, J. Zeron, M. N. Pangalos, S. Wasilewski, K. R. Smith, R. March, A. Platt, C. Haefliger, S. Petrovski, Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
3. D. Vitsios, S. Petrovski, Mantis-ml: Disease-agnostic gene prioritization from high-throughput genomic screens by stochastic semi-supervised learning. *Am. J. Hum. Genet.* **106**, 659–678 (2020).
4. R. S. Dhindsa, J. Mattsson, A. Nag, Q. Wang, L. V. Wain, R. Allen, E. M. Wigmore, K. Ibanez, D. Vitsios, S. V. V. Deevi, S. Wasilewski, M. Karlsson, G. Lassi, H. Olsson, D. Muthas, S. Monkley, A. Mackay, L. Murray, S. Young, C. Haefliger, T. M. Maher, M. G. Belvisi, G. Jenkins, P. L. Molyneaux, A. Platt, S. Petrovski, Identification of a missense variant in SPDL1 associated with idiopathic pulmonary fibrosis. *Commun. Biol.* **4**, 392 (2021).
5. K. J. Carss, A. A. Baranowska, J. Armisen, T. R. Webb, S. E. Hamby, D. Premawardhana, A. Al-Hussaini, A. Wood, Q. Wang, S. V. V. Deevi, D. Vitsios, S. H. Lewis, D. Kotecha, N. Bouatia-Naji, S. Hesselson, S. E. Iismaa, I. Tarr, L. McGrath-Cadell, D. W. Muller, S. L. Dunwoodie, D. Fatkin, R. M. Graham, E. Giannoulatou, N. J. Samani, S. Petrovski, C. Haefliger, D. Adlam, Spontaneous coronary artery dissection: Insights on rare genetic variation from genome sequencing. *Circ. Genom. Precis. Med.* **13**, e003030 (2020).
6. H. Yang, P. N. Robinson, K. Wang, Phenolyzer: Phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12**, 841–843 (2015).
7. J. Chen, E. E. Bardes, B. J. Aronow, A. G. Jegga, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, 305–311 (2009).
8. D. Geleta, A. Nikolov, G. Edwards, A. Gogleva, R. Jackson, E. Jansson, A. Lamov, S. Nilsson, M. Pettersson, V. Poroshin, B. Rozemberczki, T. Scrivener, M. Ughetto, E. Papa, Biological Insights Knowledge Graph: An integrated knowledge graph to support drug development. bioRxiv 2021.10.28.466262 [Preprint] (2021). https://doi.org/10.1101/2021.10.28.466262.
9. T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in *5th International Conference on Learning Representations* (*ICLR*), *Conference Track Proceedings,* Toulon, France (2017).
10. X. M. Zhang, L. Liang, L. Liu, M. J. Tang, Graph neural networks and their current applications in bioinformatics. *Front. Genet.* **12**, 690049 (2021).
11. R. S. Dhindsa, B. Weido, J. S. Dhindsa, A. J. Shetty, C. Sands, S. Petrovski, D. Vitsios, A. W. Zoghbi, Genome-wide prediction of dominant and recessive neurodevelopmental disorder risk genes. bioRxiv 2022.11.21.517436 [Preprint] (2022). https://doi.org/10.1101/2022.11.21.517436.
12. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery,* New York, NY, USA, 785–794 (2016).
13. T. Li, R. Wernersson, R. B. Hansen, H. Horn, J. Mercer, G. Slodkowicz, C. T. Workman, O. Rigina, K. Rapacki, H. H. Stærfeldt, S. Brunak, T. S. Jensen, K. Lage, A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64 (2016).
14. F. Wu, T. Zhang, A. H. de Souza, C. Fifty, T. Yu, K. Q. Weinberger, Simplifying graph convolutional networks, in *36th International Conference on Machine Learning* (*ICML*), Long Beach, CA, USA, 6861–6871 (2019).
15. A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J. P. Mesirov, Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
16. J. T. Eppig, Mouse genome informatics (MGI) resource: Genetic, genomic, and biological knowledgebase for the laboratory mouse. *ILAR J.* **58**, 17–41 (2017).
17. F. Aguet, A. N. Barbeira, R. Bonazzola, A. Brown, S. E. Castel, B. Jo, S. Kasela, S. Kim-Hellmuth, Y. Liang, M. Oliva, E. D. Flynn, P. Parsana, L. Fresard, E. R. Gamazon, A. R. Hamel, Y. He, F. Hormozdiari, P. Mohammadi, M. Muñoz-Aguirre, Y. S. Park, A. Saha, A. V. Segrè, B. J. Strober, X. Wen, V. Wucher, K. G. Ardlie, A. Battle, C. D. Brown, N. Cox, S. Das, E. T. Dermitzakis, B. E. Engelhardt, D. Garrido-Martín, N. R. Gay, G. A. Getz, R. Guigó, R. E. Handsaker, P. J. Hoffman, H. K. Im, S. Kashin, A. Kwong, T. Lappalainen, X. Li, D. G. MacArthur, S. B. Montgomery, J. M. Rouhana, M. Stephens, B. E. Stranger, E. Todres, A. Viñuela, G. Wang, Y. Zou, S. Anand, S. Gabriel, A. Graubert, K. Hadley, K. H. Huang, S. R. Meier, J. L. Nedzel, D. T. Nguyen, B. Balliu, D. F. Conrad, D. J. Cotter, O. M. de Goede, J. Einson, E. Eskin, T. Y. Eulalio, N. M. Ferraro, M. J. Gloudemans, L. Hou, M. Kellis, X. Li, S. Mangul, D. C. Nachun, A. B. Nobel, Y. Park, A. S. Rao, F. Reverter, C. Sabatti, A. D. Skol, N. A. Teran, F. Wright, P. G. Ferreira, G. Li, M. Melé, E. Yeger-Lotem, M. E. Barcus, D. Bradbury, T. Krubit, J. A. McLean, L. Qi, K. Robinson, N. V. Roche, A. M. Smith, L. Sobin, D. E. Tabor, A. Undale, J. Bridge, L. E. Brigham, B. A. Foster, B. M. Gillard, R. Hasz, M. Hunter, C. Johns, M. Johnson, E. Karasik, G. Kopen, W. F. Leinweber, A. McDonald, M. T. Moser, K. Myer, K. D. Ramsey, B. Roe, S. Shad, J. A. Thomas, G. Walters, M. Washington, J. Wheeler, S. D. Jewell, D. C. Rohrer, D. R. Valley, D. A. Davis, D. C. Mash, P. A. Branton, L. Sobin, L. K. Barker, H. M. Gardiner, M. Mosavel, L. A. Siminoff, P. Flicek, M. Haeussler, T. Juettemann, W. J. Kent, C. M. Lee, C. C. Powell, K. R. Rosenbloom, M. Ruffier, D. Sheppard, K. Taylor, S. J. Trevanion, D. R. Zerbino, N. S. Abell, J. Akey, L. Chen, K. Demanelis, J. A. Doherty, A. P. Feinberg, K. D. Hansen, P. F. Hickey, L. Hou, F. Jasmine, L. Jiang, R. Kaul, M. G. Kibriya, J. B. Li, Q. Li, S. Lin, S. E. Linder, B. L. Pierce, L. F. Rizzardi, K. S. Smith, M. Snyder, J. Stamatoyannopoulos, H. Tang, M. Wang, P. A. Branton, L. J. Carithers, P. Guan, S. E. Koester, A. R. Little, H. M. Moore, C. R. Nierras, A. K. Rao, J. B. Vaught, S. Volpi, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
18. S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, T. J. Callahan, C. G. Chute, J. L. Est, P. D. Galer, S. Ganesan, M. Griese, M. Haimel, J. Pazmandi, M. Hanauer, N. L. Harris, M. J. Hartnett, M. Hastreiter, F. Hauck, Y. He, T. Jeske, H. Kearney, G. Kindle, C. Klein, K. Knoflach, R. Krause, D. Lagorce, J. A. McMurry, J. A. Miller, M. C. Munoz-Torres, R. L. Peters, C. K. Rapp, A. M. Rath, S. A. Rind, A. Z. Rosenberg, M. M. Segal, M. G. Seidel, D. Smedley, T. Talmy, Y. Thomas, S. A. Wiafe, J. Xian, Z. Yüksel, I. Helbig, C. J. Mungall, M. A. Haendel, P. N. Robinson, The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
19. D. Carvalho-Silva, A. Pierleoni, M. Pignatelli, C. K. Ong, L. Fumis, N. Karamanis, M. Carmona, A. Faulconbridge, A. Hercules, E. McAuley, A. Miranda, G. Peat, M. Spitzer, J. Barrett, D. G. Hulcoop, E. Papa, G. Koscielny, I. Dunham, Open Targets Platform: New developments and updates two years on. *Nucleic Acids Res.* **47**, D1056–D1065 (2019).
20. D. Ochoa, A. Hercules, M. Carmona, D. Suveges, A. Gonzalez-Uriarte, C. Malangone, A. Miranda, L. Fumis, D. Carvalho-Silva, M. Spitzer, J. Baker, J. Ferrer, A. Raies, O. Razuvayevskaya, A. Faulconbridge, E. Petsalaki, P. Mutowo, S. MacHlitt-Northen, G. Peat, E. McAuley, C. K. Ong, E. Mountjoy, M. Ghoussaini, A. Pierleoni, E. Papa,

M. Pignatelli, G. Koscielny, M. Karim, J. Schwartzentruber, D. G. Hulcoop, I. Dunham, E. M. McDonagh, Open Targets Platform: Supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021).

21. M. Caulfield, J. Davies, M. Dennys, L. Elbahy, T. Fowler, S. Hill, T. Hubbard, L. Jostins, N. Maltby, J. Mahon-Pearson, G. Mcvean, K. Nevin-Ridley, M. Parker, V. Parry, A. Rendon, L. Riley, C. Turnbull, K. Woods, S. Mckee, A. Moffatt, J. Mccarroll, *The 100,000 Genomes Project Protocol* (The Genomics England Protocol, 2017).
22. R. K. Baboota, M. Blüher, U. Smith, Emerging role of bone morphogenetic protein 4 in metabolic disorders. *Diabetes* **70**, 303–312 (2021).
23. M. Gao, C. Yang, X. Wang, M. Guo, L. Yang, S. Gao, X. Zhang, G. Ruan, X. Li, W. Tian, G. Lu, X. Dong, S. Ma, W. Li, Y. Wang, H. Zhu, J. He, H. Yang, G. Liu, X. Xian, ApoC2 deficiency elicits severe hypertriglyceridemia and spontaneous atherosclerosis: A rodent model rescued from neonatal death. *Metabolism* **109**, 154296 (2020).
24. G. E. Morris, P. S. Braund, J. S. Moore, N. J. Samani, V. Codd, T. R. Webb, Coronary artery disease-associated *LIPA* coding variant rs1051338 reduces lysosomal acid lipase levels and activity in lysosomes. *Arterioscler. Thromb. Vasc. Biol.* **37**, 1050–1057 (2017).
25. J. Birgmeier, M. Haeussler, C. A. Deisseroth, E. H. Steinberg, K. A. Jagadeesh, A. J. Ratner, H. Guturu, A. M. Wenger, M. E. Diekhans, P. D. Stenson, D. N. Cooper, C. Ré, A. H. Beggs, J. A. Bernstein, G. Bejerano, AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci. Transl. Med.* **12**, eaau9113 (2020).
26. Y. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu, BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data* **6**, 52 (2019).
27. Z. Chen, Y. Zheng, Y. Yang, Y. Huang, S. Zhao, H. Zhao, C. Yu, X. Dong, Y. Zhang, L. Wang, Z. Zhao, S. Wang, Y. Yang, Y. Ming, J. Su, G. Qiu, Z. Wu, T. J. Zhang, N. Wu, PhenoApt leverages clinical expertise to prioritize candidate genes via machine learning. *Am. J. Hum. Genet.* **109**, 270–281 (2022).
28. J. J. Yang, C. R. Gessner, J. L. Duerksen, D. Biber, J. L. Binder, M. Ozturk, B. Foote, R. McEntire, K. Stirling, Y. Ding, D. J. Wild, Knowledge graph analytics platform with LINCS and IDG for Parkinson's disease target illumination. *BMC Bioinformatics* **23**, 37 (2022).
29. Q. Wang, R. S. Dhindsa, K. Carss, A. Harper, A. Nag, I. Tachmazidou, D. Vitsios, S. V. V. Deevi, A. Mackay, D. Muthas, M. Hühn, S. Monkley, H. Olsson, S. Wasilewski, K. R. Smith, R. March, A. Platt, C. Haefliger, S. Petrovski, B. R. Angermann, R. Artzi, C. Barrett, M. Belvisi, M. Y. Bohlooly, O. Burren, L. Buvall, B. Challis, S. Cameron-Christie, S. Cohen, A. Davis, R. F. Danielson, B. Dougherty, B. Georgi, Z. Ghazoui, P. B. L. Hansen, F. Hu, M. Jeznach, C. Kumar, Z. Lai, G. Lassi, S. H. Lewis, B. Linghu, K. Lythgow, P. Maccallum, C. Martins, A. Matakidou, E. Michaëlsson, S. Moosmang, S. O'Dell, Y. Ohne, A. O'Neill, D. S. Paul, A. Reznichenko, M. Snowden, A. Walentinsson, J. Zeron, Surveying the contribution of rare variants to the genetic architecture of human disease through exome sequencing of 177,882 UK Biobank participants. biorXiv 2020.12.13.422582 [Preprint] (2020). https://doi.org/10.1101/2020.12.13.422582.
30. J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė, T. Applebaum, A. Pritzel, L. H. Wong, M. Zielinski, T. Sargeant, R. G. Schneider, A. W. Senior, J. Jumper, D. Hassabis, P. Kohli, Ž. Avsec, Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
31. W. L. Hamilton, R. Ying, J. Leskovec, Inductive representation learning on large graphs, in *Advances in Neural Information Processing Systems,* Long Beach, CA, USA (2017).
32. M. Zhang, Y. Chen, Link prediction based on graph neural networks in *Advances in Neural Information Processing Systems* (2018) vols. 2018 December.
33. P. Chandak, K. Huang, M. Zitnik, Building a knowledge graph to enable precision medicine. *Sci. Data* **10**, 67 (2023).
34. F. Feng, F. Tang, Y. Gao, D. Zhu, T. Li, S. Yang, Y. Yao, Y. Huang, J. Liu, GenomicKB: A knowledge graph for the human genome. *Nucleic Acids Res.* **51**, D950–D956 (2023).
35. Gene Ontology Consortium, The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
36. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
37. National Research Council, Division on Engineering, Physical Sciences, Board on Army Science, & Committee on Network Science for Future Army Applications, *Network Science* (National Academies Press, 2006).
38. A. L. Barabási, Z. N. Oltvai, Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
39. A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
40. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
41. S. B. Seidman, Network structure and minimum degree. *Soc. Netw.* **5**, 269–287 (1983).
42. V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech. Theory E* **2008**, P10008 (2008).
43. L. Katz, A new status index derived from sociometric analysis. *Psychometrika* **18**, 39–43 (1953).
44. L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web. *World Wide Web Internet Web Info Syst.* **54**, 3283 (1998).
45. Y. Hao, T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda, R. Satija, Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2023).
46. D. Geleta, A. Nikolov, G. Edwards, A. Gogleva, R. Jackson, E. Jansson, A. Lamov, S. Nilsson, M. Pettersson, V. Poroshin, Biological Insights Knowledge Graph: An integrated knowledge graph to support drug development. biorXiv 2021.10.28.466262 [Preprint] (2021). https://doi.org/10.1101/2021.10.28.466262.
47. B. Rozemberczki, A. Gogleva, S. Nilsson, G. Edwards, A. Nikolov, E. Papa, MOOMIN: Deep molecular omics network for anti-cancer drug combination therapy, in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management.* Association for Computing Machinery, New York, NY, USA, 3472–3483 (2022).
48. G. Edwards, S. Nilsson, B. Rozemberczki, E. Papa, Explainable biomedical recommendations via reinforcement learning reasoning on knowledge graphs. arXiv:2111.10625 [cs.LG] (2021).
49. A. Apicella, F. Donnarumma, F. Isgrò, R. Prevete, A survey on modern trainable activation functions. arXiv:2005.00817 [cs.LG] (2021).
50. D. P. Kingma, J. L. Ba, Adam: A method for stochastic optimization, in *3rd International Conference on Learning Representations* (*ICLR*), *Conference Track Proceedings*, San Diego, CA, USA (2015).
51. S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, D. B. Goldstein, Genic intolerance to functional variation and the interpretation of personal genomes. *PLOS Genet.* **9**, e1003709 (2013).
52. L. Kolberg, U. Raudvere, I. Kuzmin, P. Adler, J. Vilo, H. Peterson, G:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res.* **51**, W207–W212 (2023).

Acknowledgments

Funding: We received no funding in support of this research. **Author contributions:** Conception and design of the study: D.V. and S.P. Developed the method and generated the phenome-wide scores: L.M., I.M., B.R., and D.V. Contributed to methods and analytical strategies: L.M., I.M., B.R., R.S.D., J.S.D., B.W., Q.W., A.R.H., G.E., S.P., and D.V. Performed validation analyses: L.M., I.M., and A.R. Wrote the paper: L.M., C.V., A.R., R.S.D., G.E., S.P., and D.V. All authors provided input and revisions for the final paper. **Competing interests:** L.M., C.V., B.R., A.R., R.S.D., Q.W., A.R.H., G.E., S.P., and D.V. are employees of AstraZeneca. B.R., A.R., R.S.D., Q.W., A.R.H., G.E., S.P., and D.V. are shareholders of AstraZeneca. L.M.'s work was funded by the AstraZeneca postdoctorate program. I.M. declares no competing interests. **Data and materials availability:** The phenome-wide Mantis-ML 2.0 generated gene predictions diseases are available in a public web resource: http://mantisml.public.cgr.astrazeneca.com. The Mantis-ML 2.0 method is publicly available at https://zenodo.org/records/10465793 and https://github.com/astrazeneca-cgr-publications/mantis-ml-release-2.0, and can be applied to any disease or phenotype using any public or proprietary knowledge graph. All other data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 8 June 2023
Accepted 4 April 2024
Published 8 May 2024
10.1126/sciadv.adj1424