PLOS Computational Biology

RESEARCH ARTICLE

MENDELSEEK: An algorithm that predicts mendelian genes and elucidates what makes them special

Hongyi Zhou, Brice Edelman, Jeffrey Skolnick*

Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, United States of America

* skolnick@gatech.edu

OPEN ACCESS

Citation: Zhou H, Edelman B, Skolnick J (2026) MENDELSEEK: An algorithm that predicts mendelian genes and elucidates what makes them special. PLoS Comput Biol 22(2): e1013992. https://doi.org/10.1371/journal.pcbi.1013992

Editor: Jean Fan, Johns Hopkins University Whiting School of Engineering, UNITED STATES OF AMERICA

Received: September 18, 2025

Accepted: February 6, 2026

Published: February 17, 2026

Data availability statement: Scripts and necessary inputs for generating the results in this work are available at https://github.com/hzhou3ga/MENDELSEEK.

Funding: This research was supported in part by grant GM-118039 from the Division

Abstract

Although individual Mendelian diseases—those caused by a single gene—are rare, their collective disease burden is substantial. Identifying the causal gene for each condition is essential for accurate diagnosis and effective treatment. Yet, despite decades of research, the genetic basis of more than half of all known Mendelian diseases remains unresolved. To address this gap, we introduce **MENDELSEEK**, a machine learning framework that predicts Mendelian genes by integrating residue variation scores with pathway participation, Gene Ontology processes, and protein language model features. In benchmarking across 16,946 human genes with 10-fold cross-validation, MENDELSEEK achieved an AUC of 0.869 and an AUPR of 0.737—substantially outperforming the next best methods, ENTPRISE+ENTPRISE-X (AUC 0.781; AUPR 0.626), and REVEL (AUC 0.585; AUPR 0.401). When applied to the full set of 17,858 human genes, MENDELSEEK predicted 1,277 novel Mendelian gene candidates with precision greater than 0.7. Analysis further revealed that Mendelian genes engage in significantly more protein-protein interactions than non-Mendelian genes and are evolutionarily ancient. Together, these results highlight MENDELSEEK as a major advance over existing methods, offering new insights into the biochemical features that distinguish Mendelian from non-Mendelian genes.

Author summary

A patient with a rare, Mendelian disease can have hundreds of mutated genes. Identifying which gene causes the disease is crucial for accurate diagnosis and treatment and for understanding more complex diseases. However, despite decades of effort, the genetic causes of over half of identified Mendelian diseases are unknown. To address this, we describe MENDELSEEK, a machine learning approach that predicts Mendelian genes by integrating the gene's aggregate residue variation score with properties such as their involved pathways, Gene

of General Medical Sciences of the National Institutes of Health to JS. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. JS, HZ received a salary from the grant GM-118039 funded by NIH.

Competing interests: The authors have declared that no competing interests exist.

Ontology processes, and protein language models. We show that MENDELSEEK performs significantly better than state-of-the-art approaches such as DeepMind's AlphaMissense in distinguishing Mendelian from non-Mendelian genes. We also present significant findings that Mendelian genes have more protein-protein interactions than non-Mendelian genes and are evolutionarily ancient. In practice, the most relevant pathways and Gene Ontology processes of Mendelian genes are discovered through comprehensive U-test analysis. We further applied MENDELSEEK to whole human genome; 1,277 novel Mendelian genes with a precision >0.7 are predicted. This work not only helps understand what pathways and molecular processes cause a given Mendelian disease but by filtering hundreds of falsely identified genes by other methods, provides valuable guidance to geneticists.

Introduction

Roughly 80% of rare diseases are thought to arise from mutations in a single gene, i.e., they are Mendelian in nature; however, many of their causative genes remain unidentified [1]. Even when the genes are known, as documented in the OMIM database [2], the mechanisms by which they give rise to the observed disease phenotypes are still poorly understood [3]. Such understanding is fundamental to the further development of precision medicine. Indeed, without knowledge of how a single gene drives a Mendelian disorder, it is unlikely that we will fully comprehend how multiple genes interact to cause non-Mendelian diseases. Consequently, significant efforts have been devoted to identifying the genetic drivers of rare diseases [4,5]. With the advent of low-cost, high-throughput next-generation sequencing (NGS) technologies, the pace of discovery has markedly accelerated, with approximately 170–240 Mendelian disease genes identified each year [4]. Nevertheless, exome sequencing alone cannot pinpoint which genes are the true drivers of Mendelian diseases, let alone the specific diseases they cause. For disorders with unknown genetic origins, disease–gene relationships could be uncovered through genome-wide association studies (GWAS) [6] or bioinformatics and computational approaches [7–15]. GWAS can be applied to both germline and somatic variations, but it requires sufficiently large cohorts to achieve statistical power. Thus, for rare diseases—which by definition affect only a few individuals—GWAS is generally inapplicable. Moreover, GWAS reveals only disease-associated genes, not those that are truly causal [16]. In contrast, computational genome variation annotation tools can prioritize candidate causal genes at the level of an individual patient [7,8,12–15]. Despite their promise, however, many bioinformatics approaches have not been rigorously benchmarked on Mendelian genes, or at best, have been tested only on relatively small protein sets [7–9].

Accurately predicting which genes are likely to be Mendelian enables the identification of the key features that distinguish them from those genes causing polygenic

diseases. Recognizing such Mendelian genes also helps researchers and physicians prioritize those most likely to cause diseases or phenotypes. However, existing state-of-the-art methods often overpredict disease-causing variations, which in turn inflates the number of predicted disease-causing genes. This suggests that either the methods fail to capture the essential characteristics of disease-causing genes, or that machine learning approaches overfit the data, making the features non-transferable to new predictions. When applied to patient data, these methods frequently misrank the true disease-causing gene due to the abundance of false positives. For example, in our earlier work, we evaluated ENTPRISE [12], SIFT [11], and PolyPhen2-HDIV [9] on ten patient samples. On average, ENTPRISE predicted ~100 disease-causing genes per patient, while SIFT and PolyPhen2-HDIV predicted between 400 and 500 such genes.

To address this limitation, we developed MENDELSEEK, a machine learning framework that predicts Mendelian genes by integrating aggregate residue variation scores with intrinsic gene properties. The aim is not to identify the specific variations causing disease, but rather to determine which genes among the many mutated candidates likely underlie Mendelian diseases or phenotypes. By doing so, MENDELSEEK can filter out false positives produced by variation-based methods. MENDELSEEK accomplishes that by integrating multiple sources of gene-level information, including Reactome pathway data [17], Gene Ontology (GO) biological processes [18], and protein language model features [19], alongside aggregate variation scores. The aggregate variation scores are derived using ENTPRISE [12] for missense variations and ENTPRISE-X(13) for frameshift and stop codon variations. This combined approach substantially outperforms methods that rely solely on aggregate variation scores, such as ENTPRISE [12], ENTPRISE-X(13), and REVEL [14] —a meta-predictor that integrates 13 methods including MutPred [20,21], FATHMM [22], VEST [7], PolyPhen [9], SIFT [11], PROVEAN [23], MutationAssessor [24], MutationTaster [25], LRT [26], GERP [27], SiPhy [28], phyloP [29], phastCons [30], —as well as the more recently developed AlphaMissense [31].

To evaluate its performance, we benchmarked MENDELSEEK using 10-fold cross-validation on a dataset of 16,946 genes, including 4,823 known Mendelian genes from the OMIM database; all are treated as true positives [2]. MENDELSEEK demonstrated a significant improvement over state-of-the-art approaches in distinguishing Mendelian from non-Mendelian genes. Finally, by analyzing MENDELSEEK's input features, we elucidate the biochemical characteristics that differentiate Mendelian from non-Mendelian genes.

Methods

A flowchart of MENDELSEEK is given in Fig 1. The detailed steps are explained below.

Calculation of the aggregate gene variation score

As discussed in [7], the simplest way of determining the aggregate variation score for a gene is to use the average score of all possible variations of a gene. Reference [7] also tested two other ways: Fisher's method [32] and Stouffer's Z-score [33]. Fisher's method requires a p-value and Stouffer's Z-score requires a Z-score; however, neither are readily available for many methods. Here, we adopt the average variation score for use in the comparison of different variation annotation methods to MENDELSEEK. In ENTPRISE's evaluation of the impact of missense variations, we assume that all wild-type residue positions are mutated to all other 19 amino acid types (in a real-life situation, this may not be realizable by a single nucleotide variation, but by combinations of such variations). Then, their variation scores are averaged to provide the aggregate score for a gene. Similarly, in ENTPRISE-X's evaluation of the impact of nonsense variations, all residue positions are assumed to be mutated to a stop codon, and the resulting scores averaged. For the other variation based methods, their score or rank score as provided by the dbNSFP database (v4.2a) [34] for a given gene is averaged. For MAVERICK [8] for which dbNSFP does not provide results, we obtained their published pre-computed whole human genome scores and averaged that score for a given gene. For AlphaMissense, the gene-level average predictions were computed by the authors by taking the mean pathogenicity over all possible missense variants in a transcript (i.e., to all

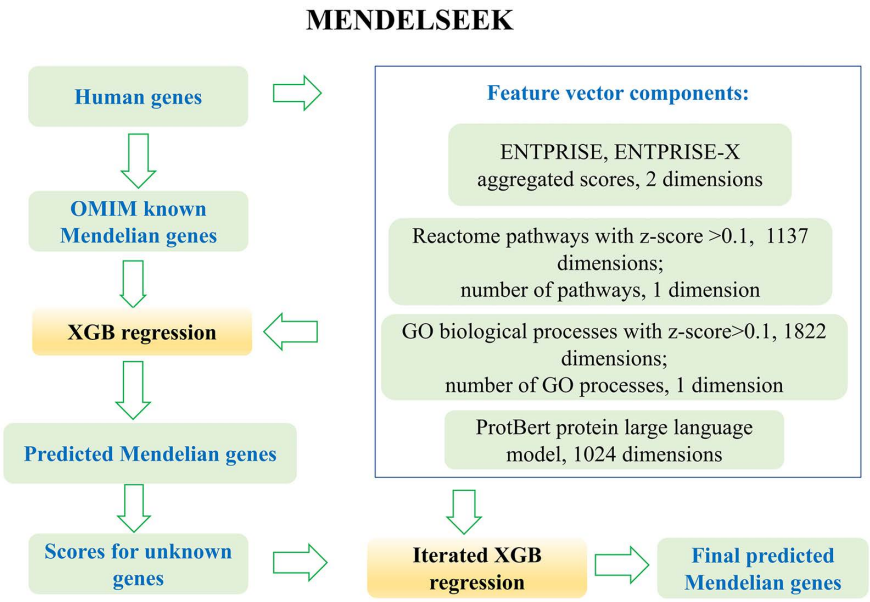

Fig 1. Flowchart of the MENDELSEEK algorithm.

https://doi.org/10.1371/journal.pcbi.1013992.g001

other 19 residue types, calculated the same way as ENTPRISE, see AlphaMissense_gene_hg19.tsv.gz which was downloaded from https://zenodo.org/records/8208688).

Assignment of a gene to its pathways and GO processes

Pathways and their corresponding associated genes are obtained from the 2,363 distinct pathways in the Reactome database [17]. Thus, its contribution to the assessment of whether a gene is Mendelian is described by a feature vector with 2,363 components. If a gene is present in a pathway, the component corresponding to this pathway is set to 1; otherwise, it is set to 0. We also include an additional component which is the total number of pathways in which a gene is involved.

The 12,535 unique human biological processes provided by the GO processes of genes are downloaded from http://geneontology.org/docs/download-ontology/ [18]. If a gene is involved in a GO process, the component corresponding to that process is set to 1; otherwise, it is set to 0. Again, the total number of processes a gene is involved with is added as an additional feature component. We tested GO molecular functions and cell components and found that the best performing choice is biological process.

The protein large language model (pLLM) embedding of a gene is obtained from [19] (https://github.com/agemagician/ProtTrans). As was also found by the ligand virtual screening method ConPlex [35], the ProtBert embeddings model performs better than alternatives. Here, we employ the 1,024-dimensional ProtBert model which is the vector output of the deep learning-based protein language model that embeds/represents a protein sequence. The exact meaning of each component depends on the embedding token dictionary that describes the amino acid sequence used in training the pLLM.

Concatenating all the above features leads to a 15,926-dimensional feature vector for each gene: 2 dimensions are from the ENTPRISE and ENTPRISE-X aggregate scores, 2,364 dimensions are from pathways, 12,536 dimensions are from GO biological processes, and 1,024 dimensions are from the pLLM.

Mann–Whitney U-test and feature reduction

To trace back the importance of each Reactome pathway and GO process, we employ the Mann–Whitney U-test on each component between Mendelian genes and unknown genes to calculate its z-score [36]. The component values of all genes are ranked according to their values with all tied values set to the average rank. For example, if three genes have a value of 1, they will be ranked 1, 2, 3 (which is ranked 1, or, 2, or 3 is random), their final assigned ranks will be (1 + 2 + 3)/3 = 2. Then, the z-score is calculated using the following equations:

$$U_1 = n_1 \times n_2 + \frac{n_1 \times (n_1 + 1)}{2} - T_1, \quad U_2 = n_1 \times n_2 + \frac{n_2 \times (n_2 + 1)}{2} - T_2 \tag{1}$$

$$U_{corr} = \sum_{i=1}^{k} \frac{t_i^3 - t_i}{12}, \qquad \sigma_{U_{corr}} = \sqrt{\frac{n_1 \times n_2}{n \times (n-1)}} \times \sqrt{\frac{n^3 - n}{12} - U_{corr}} \tag{2}$$

$$\mu_u = \frac{n_1 \times n_2}{2}, \qquad z_1 = \frac{U_1 - \mu_u}{\sigma_{U_{corr}}}, \quad z_2 = \frac{U_2 - \mu_u}{\sigma_{U_{corr}}} \tag{3}$$

where $n_{1,2}$ are the sample numbers (here gene numbers) of group 1 (here Mendelian) and 2 (unknown); $T_{1,2}$ are the sums of group 1, 2 ranks; $U_{1,2}$ are the U-values of group 1 and 2; U_{corr} is a correction value to the U-value for calculating its standard deviation, $\sigma_{U_{corr}}$; $k$is the number of tied ranks and t_i is the number of genes sharing rank i; μ_u is the expected U-value of both groups 1 and 2. $z_{1,2}$ are the z-scores of U-values from Mendelian and unknown genes.

In this work, we use z_1 to measure the importance of pathways and GO processes to Mendelian genes. To reduce the number of pathway features (a total of 2,364) and GO processes (a total of 12,536) and to avoid possible overfitting, we only keep those features having a z-score >0.1. The z-score cutoff of 0.1 is empirically determined by scanning a small number of values from 0 to 0.5, where we found that a value of 0.1 results in a very small reduction in performance compared to the full set of features, while providing a significant reduction in the number of features. Thus, the number of pathways is reduced from 2,364 to 1,057, and the number of GO processes is reduced from 12,536 to 1,672. The final total number of features is 3,755 (compared to full set's15,926 features).

Machine learning and iterated training

We employed the Extreme Gradient Boosting (XGB) regression machine learning method [37]. XGB is optimized for memory usage and computational efficiency (the sparse matrix caused by many of the pathway and GO process features having 0 value components) and is less likely to cause overfitting. Thus, it is well suited for the large dimensions of the feature vectors used in this work. In practice, the GradientBoostingRegressor was implemented in the Scikit-learn kit [38] with the following empirical parameters: n_estimators = 1000, max_depth = 6, and learning_rate = 0.05. A known Mendelian gene from the OMIM dataset [2] was set to an objective regression value of 1.0; otherwise, it is set to 0.0. To reduce the uncertainty of unknown genes that are treated as true negatives (thus, set to 0.0) in training, we utilize the predicted precision score (see Equation 4 below) for the unknown genes as the training values of unknown genes in a second or iterated round of training and prediction. The final predictions are provided by this iterated training model.

Evaluation of the relative importance of protein-protein interactions

The dataset of genes to be evaluated was combined with the unique genes from the STRING (with a cutoff score of 500) [39] and HIPPIE (with a cutoff score of 0.5) [40] databases to assess the relative importance of protein-protein interactions, PPIs. Interestingly, we find that including protein interactomes does not improve performance and thus is not included in the final version. We explain the reason for this lack of sensitivity to the further inclusion of PPIs in the Results section.

Benchmarking protocol

The above analysis yielded a final set of 17,858 unique genes for evaluation. Of these, 4,823 genes overlap with the OMIM dataset [2] and are considered to be Mendelian genes. We randomly partition the genes into 10 sets and use 9 sets for training and 1 set for testing. Then, the predictions for the 10 testing sets are combined into a composite set for evaluation and novel Mendelian gene prediction (see Results). In practice, we choose cutoff independent metrics because ranking rather than scoring is often used in practical applications. In addition, for many of the other methods that we considered, there is no appropriate cutoff information available. Commonly used cutoff independent metrics are the area under receiver operating characteristic curve (AUC) and the area under precision-recall curve (AUPR). AUPR is better than AUC for measuring the ability of a method to rank true positives at the top when the dataset is unbalanced and true positives are in the minority class [41]. Here, ~27% of the total number of genes are known true positives; thus, the total set is unbalanced. For all predictions, we convert the raw score to the predicted precision by:

$$\textit{predicted precision}\ (S_0) = \frac{\textit{Number of true positives having raw score} > S_0}{\textit{Number of genes having raw score} > S_0} \quad (4)$$

Results

Comparison to other methods

We compared MENDELSEEK including a non-iterated training version MENDELSEEK (no-iteration) to other methods, most of which are based on variation scores, e.g., VEST [7]. Table 1 shows the results on the 16,946-consensus gene set for the following evaluated methods besides the ENTPRISE+ENTPRISE-X score: SIFT, PolyPhen2-HDIV, PolyPhen2-HVAR, VEST4, REVEL, PrimateAI, CADD, MAVERICK, and AlphaMissense. MENDELSEEK has an AUPR, AUC and enrichment factor for the top ranked 180 gene predictions (~top 1%) of 0.737, 0.869 and 3.28, respectively. MENDELSEEK (no-iteration) is only slightly worse than MENDELSEEK for AUC, AUPR, but slightly better for top 180 enrichment. The second-best method ENTPRISE+ENTPRISE-X has respective values of 0.626, 0.781 and 3.39. MENDELSEEK, which includes ENTPRISE+ENTPRISE-X, performs better than ENTPRISE+ENTPRISE-X alone, with a relative increase in its AUPR of 17.7%. All three measures of these two approaches perform much better than the third best method, REVEL, which is a meta-approach and whose AUPR, AUC and enrichment factor for the top 180 genes

Table 1. Comparison of the performance of different methods on the consensus 16,946 gene set.

Method	Enrichment factor of top ranked 180 genes[a]	AUC	AUPR
MENDELSEEK	3.28	0.869	0.737
MENDELSEEK (no-iteration)	3.47	0.845	0.712
ENTPRISE+ENTPRISE-X	3.39	0.781	0.626
REVEL	2.53	0.585	0.401
MAVERICK	1.56	0.597	0.354
AlphaMissense	1.11	0.576	0.324
VEST4	1.50	0.527	0.310
CADD	1.05	0.529	0.292
PolyPhen2-HVAR	0.73	0.473	0.260
PrimateAI	0.87	0.469	0.256
PolyPhen2-HDIV	0.72	0.474	0.250
SIFT	0.84	0.467	0.245

[a]The maximal possible enrichment factor of the top ranked 180 genes is 3.55.

https://doi.org/10.1371/journal.pcbi.1013992.t001

are 0.401, 0.585 and 2.53, respectively. The performance of the AlphaMissense method is surprising since it utilizes the most state-of-the-art artificial intelligence (AI) approach [31]; yet, it seems to have significantly overpredicted Mendelian or disease causing genes. Indeed, its AUPR 0.324 is behind the 0.354 result of MAVERICK and less than half of MENDELSEEK's. Some of the alternative methods even perform worse than random selection (enrichment factors < 1 or AUC < 0.5). Fig 2 shows the AUC and AUPR curves of the compared methods. MENDELSEEK and ENTPRISE+ENTPRISE-X are well separated from the other approaches.

Table 2 and Fig 3 show the results on the 14,598 hard gene set that excluded the disease-causing training genes of ENTPRISE+ENTPRISE-X from the above consensus set, i.e., in this set the disease causing genes used in ENTPRISE+ENTPRISE-X training are excluded from evaluation to avoid possible bias to ENTPRISE+ENTPRISE-X. Note that the 14,598 genes may still contain training genes used within the other compared methods. Unfortunately, this information is unavailable. The best and second-best methods are again MENDELSEEK with an AUPR = 0.489, AUC = 0.811, and enrichment factor of the top 180 ranked genes of 4.31 and ENTPRISE+ENTPRISE-X with an AUPR = 0.334, AUC = 0.683, and an enrichment factor of the top 180 genes of 2.88, respectively. The relative difference of the AUPR is ~ 46% (0.489 vs. 0.334). MENDELSEEK (no-iteration) again is only slightly worse than MENDELSEEK for enrichment factor AUC, AUPR. Although those AUPRs are considerably lower than those for the whole consensus set, they remain substantially

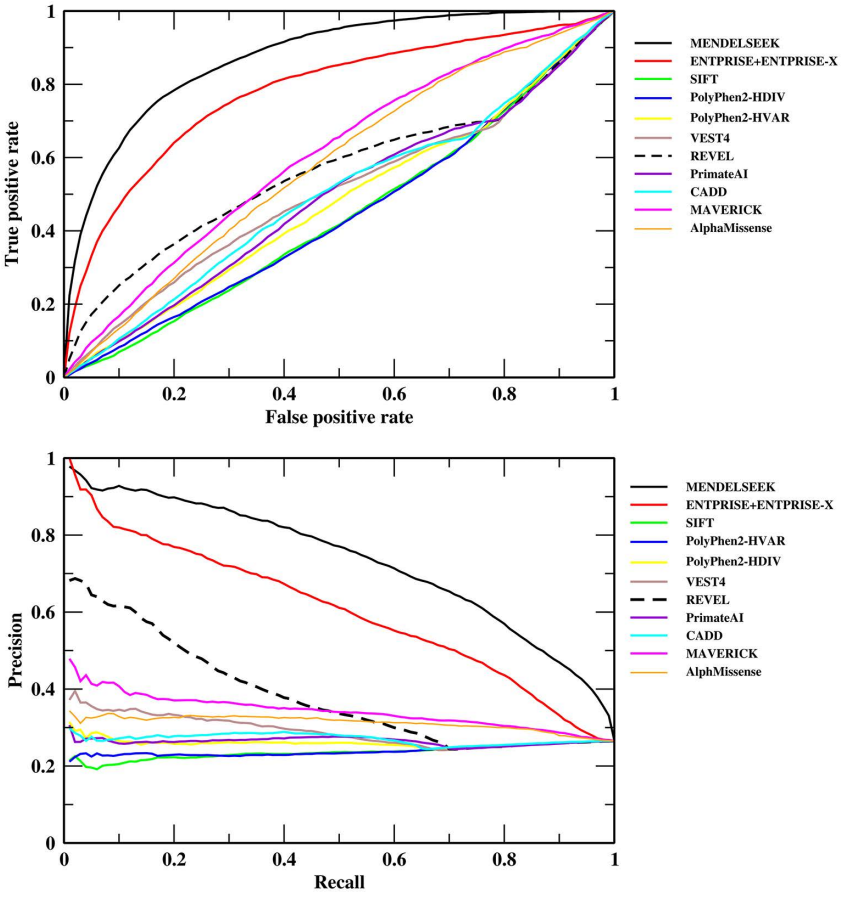

Fig 2. Classification performance of MENDELSEEK on the consensus 16,946 set in comparison to other methods. Receiver Operating Characteristic (upper Fig) and precision-recall curve (lower Fig).

https://doi.org/10.1371/journal.pcbi.1013992.g002

Table 2. Comparison of the performance of different methods on the 14,598-member hard set.

Method	Enrichment factor of top ranked 180 genes [a]	AUC	AUPR
MENDELSEEK	4.31	0.811	0.489
MENDELSEEK (no-iteration)	4.28	0.780	0.456
ENTPRISE+ENTPRISE-X	2.88	0.683	0.334
REVEL	2.03	0.559	0.225
AlphaMissense	1.49	0.582	0.219
MAVERICK	1.62	0.582	0.216
VEST4	1.68	0.537	0.203
CADD	1.17	0.540	0.192
PrimateAI	1.33	0.525	0.187
PolyPhen2-HVAR	1.20	0.503	0.174
PolyPhen2-HDIV	0.98	0.476	0.164
SIFT	0.89	0.468	0.158

[a]The maximal possible enrichment factor of the top ranked 180 genes is 5.70.

https://doi.org/10.1371/journal.pcbi.1013992.t002

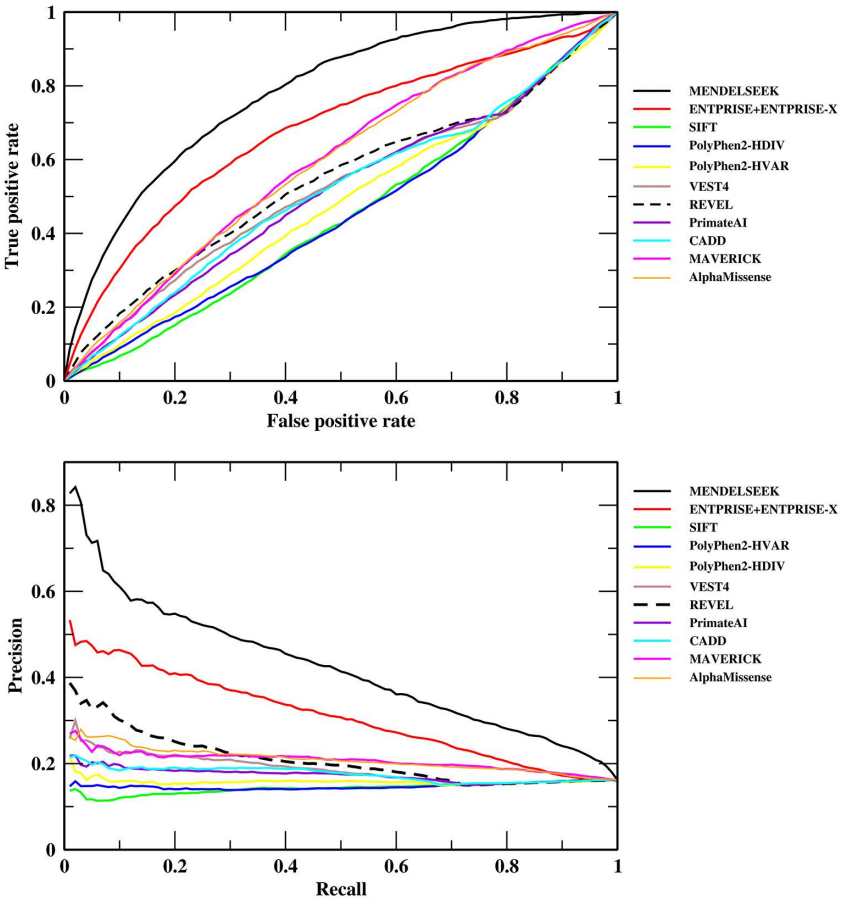

Fig 3. Classification performance of MENDELSEEK on the consensus 14,598 hard set in comparison to other methods. Receiver Operating Characteristic (upper Fig) and precision-recall curve (lower Fig).

https://doi.org/10.1371/journal.pcbi.1013992.g003

higher than the next best method REVEL, whose AUPR is 0.225. The enrichment factor of MENDELSEEK is even slightly better than that for the above "easier" set (4.31 vs. 3.28). Nevertheless, the maximal possible enrichment factor for the top 180 genes is 5.70 for the hard set and 3.55 for the easy set. The AlphaMissense's ranked fourth AUPR 0.219 is behind the 0.225 value provided by REVEL (ranked third).

From Fig 3, we see that MENDELSEEK and ENTPRISE+ENTPRISE-X are again well separated from the other methods, and the gap between MENDELSEEK and ENTPRISE+ENTPRISE-X becomes relatively larger compared to that of the full set (see Fig 2). MENDELSEEK's whole gene properties contribute more significantly for genes not seen in training for ENTPRISE+ENTPRISE-X. Thus, MENDELSEEK performed significantly better than all other methods for both the whole and hard sets in terms of AUPR, AUC.

To further assess MENDELSEEK in comparison to other methods, we employ a temporal validation strategy and constructed a validation set that consists of the current OMIM 2025 version's excessive 1,600 genes compared to the 2015 version. We randomly selected 4,800 unknown genes from the 2025 version (roughly close to the current Mendelian/unknown gene ratio). We then use the 2015 OMIM version to label the whole set of 17,858 genes and performed the exact same training/predictions. The consensus 6,046 gene set (1,584 positives, 4,462 unknown) of this 6,400 gene validation set with other compared methods are evaluated and the results are provided in Table 3. Although training on the 2015 OMIM version is to MENDELSEEK's disadvantage (some other methods, e.g., MAVERICK, AlphaMissense, CADD have been developed or updated recently), MENDELSEEK (2015 OMIM) still performs better than other methods across the three measures (enrichment factor, AUC, AUPR). ENTPRISE (2016) and ENTPRISE-X (2018) were developed close to 2015 and in terms of AUC and AUPR are again the second best. AlphaMissense comes in third; perhaps its recent training and deep learning helped it gain an advantage. We also provide the U test's z-score of each method to distinguish Mendelian genes from non-Mendelian genes in this set. They are highly correlated with AUPR values. The maximal possible enrichment factor is ~3.8; note that MENDELSEEK has an enrichment factor of 2.2 that is close to the literature validation of 1.9.

Ablation investigation.

To tease out the relative contribution of each component of MENDELSEEK to its performance, an ablation study was performed by removing one component at a time in training and doing a 10-fold cross-validation for the 17,858 gene set. Table 4 shows the results. Without the ENTPRISE+ENTPRISE-X feature component, the AUPR decreases most from 0.739 to 0.646. Exclusion of the Reactome Pathway component results in the smallest decrease of AUPR to 0.731. The

Table 3. Comparison of the performance of different methods on the 6046 validation set.

Method	Enrichment factor of top ranked 60 genes [a]	AUC	AUPR	U test z-score
MENDELSEEK (2015 OMIM)	2.16	0.648	0.386	17.5
ENTPRISE+ENTPRISE-X	1.84	0.600	0.353	11.6
AlphaMissense	1.34	0.597	0.328	10.3
MAVERICK	1.78	0.591	0.323	10.2
REVEL	1.91	0.566	0.321	4.7
VEST4	1.59	0.559	0.312	3.9
CADD	1.53	0.567	0.304	3.4
PrimateAI	1.21	0.549	0.290	2.7
PolyPhen2-HVAR	1.34	0.514	0.267	-1.9
PolyPhen2-HDIV	0.83	0.483	0.248	-5.9
SIFT	0.76	0.473	0.237	-6.8

[a]The maximal possible enrichment factor of the top ranked 60 genes is 3.82.

https://doi.org/10.1371/journal.pcbi.1013992.t003

Table 4. Ablation results on the 17,858 gene set.

Method	Enrichment factor of top ranked 180 genes [a]	AUC	AUPR
MENDELSEEK	3.41	0.876	0.739
ENTPRISE and ENTPRISE-X removed	3.27	0.825	0.646
Protein language model removed	3.44	0.854	0.712
Iterated training removed	3.62	0.853	0.713
GO processes removed	3.39	0.870	0.729
Reactome pathways removed	3.46	0.872	0.731

[a]The maximal possible enrichment factor of the top ranked 180 genes is 3.70.

https://doi.org/10.1371/journal.pcbi.1013992.t004

next smallest decrease is to 0.729, when the GO process component is ignored. Removing protein language model features results in AUPR reduction to 0.712. Without iterated training, AUPR decreases to 0.713. Thus, the ENTPRISE+ENTPRISE-X score contributes the most to MENDELSEEK by increasing the AUPR from 0.646 to 0.739 (+14%). Iterated training increases the AUPR by ~3.6%, and the protein language model component increases the AUPR by ~3.8%. The pathway and GO process components contribute the least, which when included increases the AUPR by ~1% from 0.729 and 0.731 to 0.739. This small increase is due to their correlations with the ENTPRISE+ENTPRISE-X score and to each other (see below). The ~ 1% improvement, though small, indicates that there is no overfitting. If both features are removed, the AUPR drops from 0.739 to 0.718 (a 3% reduction).

Iterated training doubles the computation cost and increases performance by ~3.6%. Can one get better performance by careful selection of negatives without iteration? To test this, we apply Valsci [42], our latest literature mining tool, to search for possible false negatives among the 12,919 unknown genes. Valsci is an open-source, self-hostable automated literature-review system that combines retrieval-augmented generation with bibliometric scoring to verify claims against the Semantic Scholar corpus and related scholarly sources. For each query, it retrieves and ranks relevant papers (augmented using citation counts, author h-index, and journal venue). It then creates a structured report evaluating each claim with a reasoning analysis and an ordinal 1–5 rating (from "No Evidence" to "Highly Supported") and traceable and true literature citations. In this study, we ran Valsci with an OpenAI-compatible LLM (gpt-5-mini) as the artificial intelligence backend. We asked Valsci if a given gene is likely associated with at least one Mendelian disease based on known literature evidence. Excluding those genes with rating score >0, 10,843 genes are retained for the confident negative set in training. With the same 10-fold cross-validation but only including these confident genes in training and testing as negatives, we obtained an AUPR of 0.737, compared to 0.742 of the original version without iteration. With iteration, for the 15,666 gene set the AUPR increases to 0.765 for the likely confident 10,843 negative and 4,823 positive genes. This indicates that including only confident negatives may have missed some true negatives and does not improve overall performance.

Correlations of GO processes and Reactome pathways with the number of protein-protein interactions

The above results indicate that the ENTPRISE+ENTPRISE-X aggregate variation score has the largest contribution to MENDELSEEK's improved performance. ENTPRISE+ENTPRISE-X scores are mainly determined by the protein's three-dimensional (3D) structure and structure-pathogenicity relationships learned from existing knowledge [12,13]. The protein language model component embeds tokens describing protein amino acid sequences and encodes intrinsic properties of proteins learned from existing knowledge [19]. As such, we are unable to dissect them further. In contrast, the GO processes and Reactome pathways have biological meaning for each component that they contribute to the feature vector. What, then, are the GO processes and Reactome pathways that most distinguish Mendelian genes from non-Mendelian genes? We analyzed those components using the Mann–Whitney U-test between true positives and the unknown ones (the majority will be true negatives) in the 17,858 dataset (see Equations 1–3) [36].

We first analyze the possible correlation of the z-scores of the GO processes and Reactome pathways with the maximal number of protein-protein interactions (PPI) of a given process or pathway. A given gene's number of PPIs is computed from the combined set of the STRING (with a cutoff score of 500) [39] and HIPPIE (with a cutoff score of 0.5) [40] databases. The maximal numbers of PPIs for the top 20 z-scores of GO processes and Reactome pathways are given in Tables 5 and 6 respectively (with the full list found in S1 Table and S2 Table). For the 12,287 GO processes having a maximal number of PPIs (>0), the Pearson's correlation between z-scores and the corresponding maximal number of PPIs is 0.421 with a p-value of 0. For the 2,362 pathways having a maximal number of PPIs, the correlation coefficient is 0.382 with a p-value 0. These results mean that genes whose GO processes or pathways have a higher number of protein-protein interactions are more likely to be Mendelian (higher z-scores). Since they are reasonably well correlated, inclusion of the number of protein-protein interactions does not improve performance.

Correlations of GO processes and Reactome pathways with evolutional time

In Tables 5 and 6, we also present the top 20 GO processes and Reactome pathways ranked by their z-scores along with their minimal Lowest Common Ancestor (LCA) evolutional time scale. LCA values range from 1 to 31, with 1 being the oldest, i.e., at the origin of life to 31 for the first cellular organisms (i.e., Prokaryota) as determined in [43]. The minimal LCA is the minimal value of a gene's LCA having/involving the same GO process/pathway. For all 12,220 GO processes having minimal LCA values, the Pearson's correlation between the z-score and the corresponding minimal LCA is -0.215 with a p-value of 0. For the 2,355 pathways having minimal LCA values, the correlation coefficient is -0.153 with a p-value of 8.3×10^{-14}. Thus, genes likely to be Mendelian are the most ancient. This is intuitively reasonable as ancient genes

Table 5. Top 20 GO processes that distinguish Mendelian genes.

z-score	Minimal LCA [a]	Maximal PPIs	GO process ID	Name
3.92	1	2458	GO:0045944	positive regulation of transcription by RNA polymerase II
3.09	1	1018	GO:0045893	positive regulation of DNA-templated transcription
2.96	1	2133	GO:0010628	positive regulation of gene expression
2.85	1	370	GO:0007601	visual perception
2.27	1	1196	GO:0000122	negative regulation of transcription by RNA polymerase II
2.23	1	960	GO:0009410	response to xenobiotic stimulus
2.17	1	1018	GO:0043066	negative regulation of apoptotic process
1.82	1	2133	GO:0008285	negative regulation of cell population proliferation
1.76	1	332	GO:0007605	sensory perception of sound
1.74	1	1196	GO:0008284	positive regulation of cell population proliferation
1.68	1	2133	GO:0006468	protein phosphorylation
1.58	1	473	GO:0007420	brain development
1.58	1	1018	GO:0001701	in utero embryonic development
1.55	1	2133	GO:0010629	negative regulation of gene expression
1.54	1	895	GO:0007507	heart development
1.53	1	2458	GO:0006357	regulation of transcription by RNA polymerase II
1.45	1	287	GO:0060271	cilium assembly
1.42	1	1108	GO:0007165	signal transduction
1.35	1	332	GO:0001501	skeletal system development
1.15	1	680	GO:0001666	response to hypoxia

[a] Minimal LCA (Lowest Common Ancestor) is the minimal value of a gene's LCA as determined in [43] having the same GO process. Gene LCA values are from 1 to 31 with 1 being the oldest.

https://doi.org/10.1371/journal.pcbi.1013992.t005

Table 6. Top 20 Reactome pathways that distinguish Mendelian genes.

z-score	Minimal LCA[a] Ls LCA LCA [a]	Maximal PPIs	Pathway ID	Name
10.4	1	2458	REACT:R-HSA-1430728	Metabolism
8.14	1	387	KEGG:hsa01100	Metabolic pathways
5.72	1	2133	REACT:R-HSA-162582	Signal Transduction
5.43	1	2133	REACT:R-HSA-168256	Immune System
4.20	1	2133	REACT:R-HSA-168249	Innate Immune System
4.16	1	2133	REACT:R-HSA-1643685	Disease
3.88	1	2133	REACT:R-HSA-392499	Metabolism of proteins
3.78	1	1196	REACT:R-HSA-1266738	Developmental Biology
3.69	1	1108	KEGG:hsa05200	Pathways in cancer
3.30	1	666	REACT:R-HSA-382551	Transmembrane transport of small molecules
3.20	1	1196	REACT:R-HSA-1280215	Cytokine Signaling in Immune system
2.91	1	2458	REACT:R-HSA-556833	Metabolism of lipids and lipoproteins
2.89	1	2133	REACT:R-HSA-109582	Hemostasis
2.68	1	1196	REACT:R-HSA-422475	Axon guidance
2.37	1	1108	KEGG:hsa04151	PI3K-Akt signaling pathway
2.34	1	1196	REACT:R-HSA-449147	Signaling by Interleukins
2.29	1	895	REACT:R-HSA-1474244	Extracellular matrix organization
2.18	1	1640	REACT:R-HSA-597592	Post-translational protein modification
2.14	1	666	REACT:R-HSA-71291	Metabolism of amino acids and derivatives
2.12	1	1196	REACT:R-HSA-166520	Signaling by NGF

[a]Minimal LCA (Lowest Common Ancestor) is the minimal value of a gene's LCA as determined in [43] having the same pathway. Gene LCA values are from 1 to 31 with 1 being the oldest.

https://doi.org/10.1371/journal.pcbi.1013992.t006

and their associated functions are likely to be essential for life. As such, their disruption should have a major phenotypical effect on the organism.

Since maximal protein–protein interactions (PPIs) and minimal lowest common ancestors (LCA) of Gene Ontology (GO) processes and pathways are highly correlated with their z-scores—which characterize their ability to distinguish Mendelian genes from non-Mendelian genes—they are also strongly correlated with each other. As a result, adding any of these features to the vector does not improve MENDELSEEK's performance; these properties are already (but implicitly) encoded in the GO processes, Reactome pathways, and the ENTPRISE+ENTPRISE-X aggregate scores. Indeed, the correlations between the ENTPRISE+ENTPRISE-X score and the LCA or number of PPIs across 17,858 genes are -0.338 and 0.228, respectively, with p-values effectively equal to 0. Direct correlations of Mendelian gene values (set to 1.0 for regression training, and 0.0 for non-Mendelian/unknown genes) with the ENTPRISE+ENTPRISE-X score, LCA, and number of PPIs are 0.477, -0.124, and 0.107, respectively, all with p-values of 0.0. Defining a gene's maximal z_{path} or $z_{go\ proc}$ as the maximal z-scores of all pathways or GO processes in which it is involved, we find correlations between the ENTPRISE+ENTPRISE-X score and maximal z_{path} and $z_{go\ proc}$ of 0.405 and 0.291. Direct correlations of Mendelian gene values with maximal z_{path} and $z_{go\ proc}$ are 0.221 and 0.201, respectively, compared to 0.477 for the ENTPRISE+ENTPRISE-X score. There is also a significant correlation of 0.096 (p-value = 7.8×10^{-38}) between z_{path} and $z_{go\ proc}$. These findings explain why ENTPRISE+ENTPRISE-X contributes most strongly to MENDELSEEK's accuracy. When either pathway or GO process features are removed, MENDELSEEK's AUPR decreases by only ~1%. If both features are removed, AUPR

drops from 0.739 to 0.718 (a 3% reduction). For comparison, AlphaMissense mean score's correlations with Mendelian gene value, maximal z_{path} and $z_{go\ proc}$ are 0.130, 0.239, and 0.239, respectively, compared to 0.477, 0.405, and 0.291 for the ENTPRISE+ENTPRISE-X score. This difference demonstrates that the current approach more effectively captures the essential features underlying MENDELSEEK's superior performance.

Analysis of the top GO processes and pathways

The top 20 GO process and pathways ranked by z-scores are the oldest with a minimal LCA of 1. The top three of these GO processes are related to gene expression: *positive regulation of transcription by RNA polymerase II, positive regulation of DNA-templated transcription,* and *positive regulation of gene expression.* These processes are essential for life. When these processes malfunction, variations/mutations in genes will happen and cause diseases or even death.

The fourth ranked GO process *visual perception* (GO:0007601) with a z-score of 2.85, has 29 phenotypes caused by 6 genes (BBS4, COL1A1, COL2A1, OAT, RDH11, RPE65) having a LCA of 1. While it may, at first glance, seem odd that visual perception has an LCA = 1 (ancient organisms did not have eyes but they could perceive light [44]), these genes also engage in other essential, nonvisual processes. For example, the BBS4 gene (Bardet-Biedl syndrome 4) is a protein-coding gene that plays a role in the development and function of cilia [45] and involves 49 human GO processes including *gene expression* (GO: 0010467). The COL1A1 gene produces a component of type I collagen that strengthens and supports many tissues in the body [46]; it is involved in 5 human GO processes including *skeletal system development* (GO:0001501). The COL2A1 gene encodes the alpha-1 chain of type II collagen which is essential for the structure and function of cartilage [47]. COL2A1 is involved in 39 GO processes including *visual perception* [48,49], *sensory perception of sound* [50], *skeletal system development [51,52], central nervous system development* [53], as well as other important biological functions. Furthermore, the current OMIM database [2] lists 15 phenotypes caused by mutations in COL2A1. OAT encodes the ornithine aminotransferase enzyme, that is found in mitochondria where it helps break down ornithine. Ornithine is involved in the urea cycle and in maintaining the balance of amino acids in the body [54]. For RDH11, retinol dehydrogenase, in addition to being an essential gene in the eye, another of its 6 human GO processes involves the cellular detoxification of aldehyde [18,55]. Finally, while RPE65 helps convert light into electrical signals that are sent to the brain, this protein is also involved in 14 human GO processes including the *insulin receptor signaling pathway* (GO:0008286).

In humans, there are 29 phenotypes associated with these genes; many are eye diseases (see S3 Table for the full list) including Retinitis pigmentosa, Leber congenital amaurosis caused by RPE65, Gyrate atrophy of choroid and retina with or without ornithinemia by OAT; Retinal dystrophy caused by RDH11, and Vitreoretinopathy with phalangeal epiphyseal dysplasia (the latter is not eye related) caused by COL2A1. There are also completely non-eye related diseases: Czech dysplasia, Chondrogenesis, type II or hypochondrogenesis, Spondyloperipheral dysplasia by COL2A1; Osteogenesis imperfecta, type I, Caffey disease, Ehlers-Danlos syndrome that are caused by COL1A1. Bardet-Biedl syndrome which is caused by BBS4 affects vision, body weight, genital abnormalities and kidney functions.

The top four pathways of Mendelian genes are *Metabolism* (z-score = 10.4), *Metabolic pathways* (z-score = 8.1), *Signal Transduction* (z-score = 5.7), *Immune System* (z-score = 5.4). These generic pathways are crucial because they allow cells to efficiently capture and utilize energy from nutrients, enabling essential functions such as growth, reproduction, maintaining structure, and when uncontrolled, they could result in cancers in some organisms. The top pathway, *Metabolism* (REACT:R-HSA-1430728) is associated with 371 phenotypes caused by genes having a LCA of 1 (see S4 Table for the full list).

Literature evidence that substantiates the predictions of novel Mendelian genes

The dataset of 17,858 unique genes obtained by combining the genes from the STRING (with a cutoff score of 500) [39] and HIPPIE (with a cutoff score of 0.5) [40] databases are also used for novel Mendelian gene prediction (these genes are absent in the OMIM database). We restricted our attention to predictions of genes within this interaction dataset that have

known protein-protein interactions, as we have shown that genes having a higher number of protein-protein interactions are more likely to be Mendelian. Equation 4 converts the raw regression score to the predicted precision score (see S5 Table). With a predicted precision score cutoff of 0.7 (corresponding to raw score 0.79), we have 1,277 novel gene predictions (those that are not in the OMIM database). These predictions of novel Mendelian genes are listed in S6 Table.

How can we validate these predictions? To do so, we again employ Valsci for validation [42]. We asked Valsci if a given gene is likely associated with at least one Mendelian disease based on known literature evidence. We also asked the same question for ~1,000 randomly chosen known and unknown Mendelian genes, respectively. For the 991 known Mendelian genes, Valsci finds 509 genes with rating score ≥ 3, whereas for the 997 unknown genes Valsci has only 44 genes with rating score ≥ 3. This leads to Valsci's precision/recall of 0.92/0.51, respectively, assuming the unknown ones are true negatives. The high precision of Valsci means MENDELSEEK has a low false positive rate (0.04) and its validated predictions are highly accurate.

Valsci returns supported literature evidence with a rating score ≥ 3 for 108 genes of the 1,277 novel Mendelian gene predictions. This leads to an effective enrichment factor of 1.9 compared to the 44/997 random unknown set. If we check around the top 1% of the 13,035 unknown genes, or the top 100 predicted novel genes, we get 10 genes with supportive evidence whereas random expected 100*44/997 = 4.41 genes, this results in an enrichment of 2.3. If we check the top 50 [20] novel predictions, we get an enrichment factor of 2.7(4.5). This means higher ranked Mendelian genes are more likely to be recalled. Since the predicted unknown genes have not been curated by the OMIM database, their literature evidence is rare, we cannot expect the recall rate (108/1277 = 0.08) to be comparable to those of the known Mendelian Genes, 0.51. The 108 genes with Valsci score ≥ 3 are highly accurate based on Valsci's precision of 0.92 for this purpose.

Examples of the top predicted genes with literature evidence that are not known to the MENDELSEEK algorithm are: ITGB1 (precision = 0.90) encoding integrin beta 1, is involved in 61 pathways, and 10 of them are within the top 20 z-scores in Table 6 including Signal Transduction (z-score = 5.72) and the Immune System (z-score = 5.43). Its dysfunction causes kidney/renal diseases [56]. ND6 has a predicted precision of 0.90 and is involved in 9 pathways including the top two Metabolism (z-score = 10.4) and Metabolic pathways (z-score = 8.14). Mutations in this gene cause mitochondrial disease [57] and Leber's hereditary optic neuropathy (LHON) [58]. RIMS1 (precision = 0.90) was documented as causing Cone-rod dystrophies (CORDs) [59]. It is involved in 21 GO processes including visual perception (z-score = 2.85). Variants in SORL1 (precision = 0.86) have been implicated in familial dementia [60] and is involved in protein Metabolism (z-score = 3.88).

Whether the other gene predictions with no Valsci supportive evidence are false positives or novel true positives is uncertain at this juncture. A full list of predicted genes can be found in S6 Table. Those with Valsci evidence (with scores of 3 or above indicating that at least some evidence in support has been found in the existing literature) for being disease causing are indicated in S6 Table and those that are not (rating score < 3) are marked "NONE". These can serve as guidance for further bioinformatics/experimental validations/tests. A detailed report of Valsci's results is included in the Supplemental Material.

The above validated 108 genes have a small recall (8%). To further validate our predicted genes, we compare the 1,277 genes with 4,069 genes from the DECIPHER database [61] after excluding those already in the OMIM database. The 1,277 predicted genes have 465 overlaps with this list, resulting a p-value of 0.0. Combined with the 108 Valsci validated genes (some overlaps with this 465 set), we get 530/1277 = 41.5% validated genes. This significantly increases the number of validated genes. We also marked these genes as "DECIPHER" in S6 Table.

Difference between Mendelian genes and complex disease driving genes

Are Mendelian genes also drivers of complex diseases? Combining the known 4,823 OMIM genes in our test set and the 1,277 predicted Mendelian genes leads to 6,101 putative Mendelian genes. For putative complex disease driving genes, we have previously derived a set of 7,311 genes from 3,608 complex diseases having the gene as a driver [62]. The two

sets of genes have 2,532 overlaps. Thus, a subset of the Mendelian genes are also involved in complex diseases (see S7 Table). The remaining 3,569 putative Mendelian genes are not complex disease drivers, with 2,834 of these found in OMIM and another 735 predicted. Thus, roughly 60% of this set of Mendelian genes appear to be drivers of a single disease, with the remaining ~40% being possible drivers of complex diseases as well. This is not surprising in that the malfunction of Mendelian genes is associated with the disruption of key biochemical processes.

To characterize differences between genes of that are only associated with Mendelian diseases, those genes that are drivers of both Mendelian and complex disease, and those that only drive complex diseases, we analyzed their number of protein-protein interactions, number of involved pathways, number of involved GO processes, protein RNA expression levels [63], and LCA values. The results are compiled in S8 Table. Genes that drive both Mendelian and complex disease have a greater number of protein-protein interactions, involve more pathways and GO processes compared to those that are only Mendelian drivers or that drive only complex diseases. The latter have a much smaller number of PPIs, pathways and GO processes. Although the absolute differences between genes driving only Mendelian disease and driving both Mendelian and complex diseases are small, their U test z-scores are significant (>1.65, p-value<0.05, see S8 Table). Furthermore, RNA expression levels have no significant difference (U test z-score = -0.71) between Mendelian only genes and Mendelian/complex disease drivers. Both expression levels are more than doubled that of complex disease only drivers. These results are in agreement with earlier findings that genes associated with both Mendelian and complex diseases tend to present higher functional relevance in the protein network and higher expression levels than genes associated only with complex disorders [64].

The mean LCA value of Mendelian/complex disease drivers is greater than that of Mendelian only drivers and their U test z-score of difference is 2.11 indicating it is not a statistical fluctuation. This suggests that Mendelian disease/complex disease drivers emerged later than the Mendelian only drivers. Since their PPIs, pathway, GO process numbers are close, we hypothesize that the Mendelian/complex disease drivers evolved from Mendelian only disease drivers by gaining additional PPIs, pathways & GO processes without changing their expression levels. The large differences of numbers of PPIs, pathways & GO processes, expression level, and the larger LCA value (emerged later) of complex disease only drivers compared to those of the other two types genes indicate that they emerged later likely and evolved independently.

Discussion

In this contribution, we demonstrated that MENDELSEEK significantly outperforms other approaches including the state-of-the-art AlphaMissense [31], in distinguishing Mendelian genes—those whose variations alone are sufficient to cause disease—from non-Mendelian genes. MENDELSEEK's predictions are consistently supported by benchmarking tests and corroborating literature. Ablation analysis shows that the ENTPRISE+ENTPRISE-X score, reflecting residue variation, contributes the most to performance, improving AUPR by 14%. These robust results reflect the low false-positive rates of ENTPRISE [12] and ENTPRISE-X(13) in predicting disease-causing variants. For example, ENTPRISE exhibits a 10.7% false-positive rate compared to 36.4% for PolyPhen-2-HVAR [9] on the 1,000 Genomes dataset [12], while ENTPRISE-X shows an 8.4% false-positive rate compared to 18.6% for VEST-indel [13,65] on the 1,000 Genomes dataset.

Mann–Whitney U-test analysis of individual GO processes that discriminate Mendelian from non-Mendelian genes reveals that the most discriminating processes (those with large positive z-scores) are typically associated with the oldest genes (lowest LCA values) and genes with a higher number of PPIs. A similar trend is observed for discriminative Reactome pathways.

Literature mining indicates that approximately 8% (108/1277, 530/1277 = 41.5% after including DECIPHER overlaps) of MENDELSEEK's novel Mendelian gene predictions have existing literature support, while the remaining candidates represent high-value targets for experimental validation. Future directions include extending MENDELSEEK to predict not only whether a gene is Mendelian, but also its associated phenotypes or symptoms and the mode of inheritance (autosomal dominant or autosomal recessive). This approach assumes that specific phenotypes are linked to particular GO processes

or pathways and that artificial intelligence can learn these patterns. A further challenge arises when a single gene gives rise to multiple phenotypes. For instance, COL2A1 is associated with 15 phenotypes in the OMIM database; thus, an important question is whether one can predict which phenotypes manifest themselves in a given patient and whether "protective" genes can prevent certain outcomes. More broadly, genetic modifiers [66] play critical roles in Mendelian disease phenotypes, but identifying them and understanding their mechanisms remain unresolved challenges. A major limitation of these methods for predicting the specific phenotype(s) of a gene is that there are not enough gene samples to train/test machine learning methods for a given phenotype. A potential solution which will be explored in the future involves decomposing each phenotype into its symptom set, and then for each symptom, there will be enough genes to learn from. A given disease is then the aggregation of a set of symptoms.

For non-Mendelian diseases, where dozens to hundreds of gene variants may contribute, the specific causal combinations are often unclear [12]. By contrast, the unimodal nature of Mendelian genes provides valuable insights into the link between genotype and phenotype. Developing tools that map Mendelian genes to their phenotypes will not only advance understanding of these rare disorders but also yield algorithms and principles applicable to complex, non-Mendelian diseases.

Supporting information

S1 Table. List of GO process z-scores of Mann–Whitney U tests and their minimal LCA, maximal number of PPIs.
(XLSX)

S2 Table. List of Reactome pathway z-socres of Mann–Whitney U tests and their minimal LCA, maximal number of PPIs.
(XLSX)

S3 Table. Phenotypes caused by genes with LCA = 1 and GO process visual perception.
(XLSX)

S4 Table. Phenotypes caused by genes with LCA = 1 and pathway Metabolism.
(XLSX)

S5 Table. Predicted precision as function raw score.
(XLSX)

S6 Table. Top predicted Mendelian genes with predicted precision (normalized score) > 0.7.
(XLSX)

S7 Table. Mendelian genes that are also complex disease drivers.
(XLSX)

S8 Table. Properties of different disease drivers.
(XLSX)

S1 Data. Valsci detailed report for the 1277 predicted Mendelian genes (see S6 Table).
(ZIP)

Acknowledgments

A gift from the Ovarian Cancer Institute is gratefully acknowledged. We thank Jessica Forness for proofreading and polishing this manuscript and Bartosz Ilkowski for his computational support.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors have not used generative AI and AI-assisted technologies in the writing process.

Financial interest

The authors declare that there are no competing interests

Availability of data and materials

Scripts and necessary inputs for generating the results in this work are available at https://github.com/hzhou3ga/MENDELSEEK.

Author contributions

Conceptualization: Hongyi Zhou, Jeffrey Skolnick.

Data curation: Hongyi Zhou.

Formal analysis: Hongyi Zhou, Brice Edelman.

Funding acquisition: Jeffrey Skolnick.

Investigation: Hongyi Zhou, Brice Edelman.

Methodology: Hongyi Zhou, Brice Edelman.

Project administration: Jeffrey Skolnick.

Resources: Jeffrey Skolnick.

Software: Hongyi Zhou, Brice Edelman.

Supervision: Jeffrey Skolnick.

Validation: Hongyi Zhou, Brice Edelman.

Visualization: Hongyi Zhou.

Writing – original draft: Hongyi Zhou.

Writing – review & editing: Hongyi Zhou, Brice Edelman, Jeffrey Skolnick.

References

1. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018;19(5):253–68. https://doi.org/10.1038/nrg.2017.116 PMID: 29398702
2. McKusick-Nathans Institute of Genetic Medicine JHU. Online Mendelian Inheritance in Man. OMIM.
3. Condò I. Rare monogenic diseases: molecular pathophysiology and novel therapies. Int J Mol Sci. 2022;23(12).
4. Seaby EG, Rehm HL, O'Donnell-Luria A. Strategies to uplift novel mendelian gene discovery for improved clinical outcomes. Frontiers in Genetics. 2021;12.
5. Bamshad MJ, Nickerson DA, Chong JX. Mendelian gene discovery: fast and furious with no end in sight. Am J Hum Genet. 2019;105(3):448–55. https://doi.org/10.1016/j.ajhg.2019.07.011 PMID: 31491408
6. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010;363(2):166–76. https://doi.org/10.1056/NEJMra0905980 PMID: 20647212
7. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3(Suppl 3):S3. https://doi.org/10.1186/1471-2164-14-S3-S3 PMID: 23819870
8. Danzi MC, Dohrn MF, Fazal S, Beijer D, Rebelo AP, Cintra V, et al. Deep structured learning for variant prioritization in Mendelian diseases. Nat Commun. 2023;14(1):4167. https://doi.org/10.1038/s41467-023-39306-7 PMID: 37443090

9. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512
10. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50(8):1161–70. https://doi.org/10.1038/s41588-018-0167-z PMID: 30038395
11. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4. https://doi.org/10.1093/nar/gkg509 PMID: 12824425
12. Zhou H, Gao M, Skolnick J. ENTPRISE: An algorithm for predicting human disease-associated amino acid substitutions from sequence entropy and predicted protein structures. PLoS One. 2016;11(3):e0150965. https://doi.org/10.1371/journal.pone.0150965 PMID: 26982818
13. Zhou H, Gao M, Skolnick J. ENTPRISE-X: Predicting disease-associated frameshift and nonsense mutations. PLoS One. 2018;13(5):e0196849. https://doi.org/10.1371/journal.pone.0196849 PMID: 29723276
14. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877–85. https://doi.org/10.1016/j.ajhg.2016.08.016 PMID: 27666373
15. Schubach M, Maass T, Nazaretyan L, Röner S, Kircher M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. Nucleic Acids Res. 2024;52(D1):D1143–54.
16. Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA. 2008;299(11):1335–44. https://doi.org/10.1001/jama.299.11.1335 PMID: 18349094
17. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48(D1):D498–503. https://doi.org/10.1093/nar/gkz1031 PMID: 31691815
18. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9. https://doi.org/10.1038/75556 PMID: 10802651
19. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell. 2022;44(10):7112–27. https://doi.org/10.1109/TPAMI.2021.3095381 PMID: 34232869
20. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H-J, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun. 2020;11(1):5918. https://doi.org/10.1038/s41467-020-19669-x PMID: 33219223
21. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744–50.
22. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57–65. https://doi.org/10.1002/humu.22225 PMID: 23033316
23. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31(16):2745–7. https://doi.org/10.1093/bioinformatics/btv195 PMID: 25851949
24. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118. https://doi.org/10.1093/nar/gkr407 PMID: 21727090
25. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575–6. https://doi.org/10.1038/nmeth0810-575 PMID: 20676075
26. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553–61. https://doi.org/10.1101/gr.092619.109 PMID: 19602639
27. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025. https://doi.org/10.1371/journal.pcbi.1001025 PMID: 21152010
28. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):i54-62. https://doi.org/10.1093/bioinformatics/btp190 PMID: 19478016
29. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110–21. https://doi.org/10.1101/gr.097857.109 PMID: 19858363
30. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034–50. https://doi.org/10.1101/gr.3715005 PMID: 16024819
31. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492. https://doi.org/10.1126/science.adg7492 PMID: 37733863
32. Fisher R, Bennett J, Yates F. Statistical methods, experimental design, and scientific inference: A re-issue of Statistical methods for research worker. 1990.
33. Stouffer S, Suchman E, Devinney L, Star S, Williams R, Williams RJr. The American soldier: adjustment during army life. 1949.
34. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12(1):103. https://doi.org/10.1186/s13073-020-00803-9 PMID: 33261662
35. Singh R, Sledzieski S, Bryson B, Cowen L, Berger B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. Proc Natl Acad Sci U S A. 2023;120(24):e2220778120. https://doi.org/10.1073/pnas.2220778120 PMID: 37289807

36. McKnight PE, Najab J. Mann-Whitney U Test. The Corsini Encyclopedia of Psychology. Wiley; 2010. 1.
37. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA. 2016. 785–94.
38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Grisel O, Blondel M, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–30.
39. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):D638–46. https://doi.org/10.1093/nar/gkac1000 PMID: 36370105
40. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. Nucleic Acids Res. 2017;45(D1):D408–14. https://doi.org/10.1093/nar/gkw985 PMID: 27794551
41. Ozenne B, Subtil F, Maucort-Boulch D. The precision--recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol. 2015;68(8):855–9. https://doi.org/10.1016/j.jclinepi.2015.02.010 PMID: 25881487
42. Edelman B, Skolnick J. Valsci: an open-source, self-hostable literature review utility for automated large-batch scientific claim verification using large language models. BMC Bioinformatics. 2025;26(1):140. https://doi.org/10.1186/s12859-025-06159-4 PMID: 40437377
43. Lopes K de P, Campos-Laborie FJ, Vialle RA, Ortega JM, De Las Rivas J. Evolutionary hallmarks of the human proteome: chasing the age and coregulation of protein-coding genes. BMC Genomics. 2016;17(Suppl 8):725. https://doi.org/10.1186/s12864-016-3062-y PMID: 27801289
44. Williams DL. Light and the evolution of vision. Eye (Lond). 2016;30(2):173–8. https://doi.org/10.1038/eye.2015.220 PMID: 26541087
45. Hernandez-Hernandez V, Pravincumar P, Diaz-Font A, May-Simera H, Jenkins D, Knight M, et al. Bardet-Biedl syndrome proteins control the cilia length through regulation of actin polymerization. Hum Mol Genet. 2013;22(19):3858–68. https://doi.org/10.1093/hmg/ddt241 PMID: 23716571
46. Małecki K, Fabiś-Strobin A, Sałacińska K, Kwas K, Stelmach W, Beczkowski J, et al. Clinical significance of polymorphisms of genes encoding collagen (COL1A1, COL5A1) and their correlation with joint laxity and recurrent patellar dislocation in adolescents. Sci Rep. 2023;13(1):22300. https://doi.org/10.1038/s41598-023-49378-6 PMID: 38102224
47. Glavey SV, Naba A, Manier S, Clauser K, Tahri S, Park J, et al. Proteomic characterization of human multiple myeloma bone marrow extracellular matrix. Leukemia. 2017;31(11):2426–34. https://doi.org/10.1038/leu.2017.102 PMID: 28344315
48. Okazaki S, Meguro A, Ideta R, Takeuchi M, Yonemoto J, Teshigawara T, et al. Common variants in the COL2A1 gene are associated with lattice degeneration of the retina in a Japanese population. Mol Vis. 2019;25:843–50. PMID: 31908402
49. Metlapally R, Li Y-J, Tran-Viet K-N, Abbott D, Czaja GR, Malecaze F, et al. COL1A1 and COL2A1 genes and myopia susceptibility: evidence of association and suggestive linkage to the COL2A1 locus. Invest Ophthalmol Vis Sci. 2009;50(9):4080–6. https://doi.org/10.1167/iovs.08-3346 PMID: 19387081
50. Liu Z, Bai X, Wan P, Mo F, Chen G, Zhang J, et al. Targeted Deletion of Loxl3 by Col2a1-Cre Leads to Progressive Hearing Loss. Front Cell Dev Biol. 2021;9:683495. https://doi.org/10.3389/fcell.2021.683495 PMID: 34150778
51. Markova T, Kenis V, Melchenko E, Osipova D, Nagornova T, Orlova A. Clinical and genetic characteristics of COL2A1-associated skeletal dysplasias in 60 Russian patients: part I. Genes. 2022;13(1).
52. Zhang B, Wang C, Zhang Y, Jiang Y, Qin Y, Pang D, et al. A CRISPR-engineered swine model of COL2A1 deficiency recapitulates altered early skeletal developmental defects in humans. Bone. 2020;137:115450. https://doi.org/10.1016/j.bone.2020.115450 PMID: 32450343
53. Hwang DW, Kim KT, Lee SH, Kim JY, Kim DH. Association of COL2A1 gene polymorphism with degenerative lumbar scoliosis. Clin Orthop Surg. 2014;6(4):379–84. https://doi.org/10.4055/cios.2014.6.4.379 PMID: 25436060
54. Urra M, Buezo J, Royo B, Cornejo A, López-Gómez P, Cerdán D, et al. The importance of the urea cycle and its relationships to polyamine metabolism during ammonium stress in Medicago truncatula. J Exp Bot. 2022;73(16):5581–95. https://doi.org/10.1093/jxb/erac235 PMID: 35608836
55. Kasus-Jacobi A, Ou J, Bashmakov YK, Shelton JM, Richardson JA, Goldstein JL, et al. Characterization of mouse short-chain aldehyde reductase (SCALD), an enzyme regulated by sterol regulatory element-binding proteins. J Biol Chem. 2003;278(34):32380–9. https://doi.org/10.1074/jbc.M304969200 PMID: 12807874
56. Wu W, Kitamura S, Truong DM, Rieg T, Vallon V, Sakurai H, et al. Beta1-integrin is required for kidney collecting duct morphogenesis and maintenance of renal function. Am J Physiol Renal Physiol. 2009;297(1):F210-7. https://doi.org/10.1152/ajprenal.90260.2008 PMID: 19439520
57. Chen D, Zhao Q, Xiong J, Lou X, Han Q, Wei X, et al. Systematic analysis of a mitochondrial disease-causing ND6 mutation in mitochondrial deficiency. Mol Genet Genomic Med. 2020;8(5):e1199. https://doi.org/10.1002/mgg3.1199 PMID: 32162843
58. Wallace DC, Singh G, Lott MT, Hodge JA, Schurr TG, Lezza AM, et al. Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. Science. 1988;242(4884):1427–30. https://doi.org/10.1126/science.3201231 PMID: 3201231
59. Gill JS, Georgiou M, Kalitzeos A, Moore AT, Michaelides M. Progressive cone and cone-rod dystrophies: clinical features, molecular genetics and prospects for therapy. British Journal of Ophthalmology. 2019;103(5):711.
60. Alvarez-Mora MI, Blanco-Palmero VA, Quesada-Espinosa JF, Arteche-Lopez AR, Llamas-Velasco S, Palma Milla C. Heterozygous and Homozygous Variants in SORL1 Gene in alzheimer's disease patients: clinical, neuroimaging and neuropathological findings. Int J Mol Sci. 2022;23(8).
61. Foreman J, Perrett D, Mazaika E, Hunt SE, Ware JS, Firth HV. DECIPHER: Improving genetic diagnosis through dynamic integration of genomic and clinical data. Annu Rev Genomics Hum Genet. 2023;24:151–76. https://doi.org/10.1146/annurev-genom-102822-100509 PMID: 37285546

62. Zhou H, Edelman B, Skolnick J. A mode of action protein based approach that characterizes the relationships among most major diseases. Sci Rep. 2025;15(1):9668. https://doi.org/10.1038/s41598-025-93377-8 PMID: 40113859

63. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419. https://doi.org/10.1126/science.1260419 PMID: 25613900

64. Spataro N, Rodríguez JA, Navarro A, Bosch E. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. Hum Mol Genet. 2017;26(3):489–500. https://doi.org/10.1093/hmg/ddw405 PMID: 28053046

65. Douville C, Masica DL, Stenson PD, Cooper DN, Gygax DM, Kim R, et al. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). Hum Mutat. 2016;37(1):28–35. https://doi.org/10.1002/humu.22911 PMID: 26442818

66. Rahit K, Tarailo-Graovac M. Genetic modifiers and rare mendelian disease. Genes (Basel). 2020;11(3).