MENDELSEEK: An algorithm that predicts Mendelian Genes and elucidates what makes them special

Authors: Hongyi Zhou[1], Jeffrey Skolnick[1]*

Affiliations:

[1]Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology; 950 Atlantic Drive, N.W., Atlanta, GA 30332, USA.

*Corresponding author. Email: skolnick@gatech.edu

Abstract

While individual Mendelian diseases (diseases caused by a single gene) are rare, their aggregate number is significant. Discovering which gene causes a Mendelian disease is crucial for accurate diagnosis and treatment. Despite decades of effort, the genetic cause driving over half of identified Mendelian diseases is 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 Ontology processes, and protein language models. In benchmarking on 16,946 human genes with 10-fold cross-validation, MENDELSEEK achieves an area under the receiver operating characteristic curve, AUC, and an area under precision-recall curve, AUPR, of 0.850 and 0.695 respectively, compared to the second best method that uses residue variation, ENTPRISE+ENTPRISE-X, with 0.781 and 0.604 scores, and the third best approach, REVEL, with 0.597 and 0.390 scores. Mendelian genes have significantly more protein-protein interactions than non-Mendelian genes and are evolutionarily ancient. Applying MENDELSEEK to 17,858 genes of the whole human genome, 1,024 novel Mendelian genes with a precision >0.7 are predicted. Thus, MENDELSEEK represents a major improvement over the state-of-the-art and provides valuable insights into the biochemical features that distinguish Mendelian from non-Mendelian genes.

Keywords: Mendelian disease, rare disease, genetic disease, genome sequencing, exome sequencing, genetic variation.

Introduction

80% of rare diseases are likely caused by a single gene, i.e., they are Mendelian; yet many of their causative genes have not been identified[1]. Understanding how Mendelian genes cause disease is also fundamental to the development of precision medicine - If one cannot understand their mode of action and how they cause a given disease, it is unlikely that one could fully understand how a collection of genes cause non-Mendelian diseases. Thus, many efforts have been made towards the discovery of the genes that cause these rare diseases[2, 3]. With next-generation sequencing (NGS) technologies, the discovery pace has significantly accelerated from around 170 to 240 Mendelian genes discovered per year[2]. However, exome sequencing itself does not identify which are the driver genes of Mendelian diseases, much less the diseases that they cause. For diseases with unknown causal genes, disease–gene relationships could be discovered by other means such as genome-wide association studies (GWAS)[4] and bioinformatics/computational approaches[5-13]. GWAS is applicable to germline and somatic variations but requires a sufficiently large number of individuals to achieve statistical significance. Thus, for a rare disease in which, by definition, few individuals are affected, GWAS is inapplicable. Another drawback of GWAS is that it only reveals which genes are associated with a given disease, not which are causal[14]. On the other hand, bioinformatics/computational genome variation annotation tools can prioritize disease-causing genes for a single patient[5, 6, 10-13]. Nevertheless, despite their promise, many bioinformatics approaches have not been tested/benchmarked on Mendelian genes, or at best, were tested on small protein sets[5-7].

In this contribution, we show that previous state-of-the-art methods tend to overpredict disease-causing variations and thus overprediction of disease-causing genes. On applying such approaches to rank patient variations, the true disease causing gene is usually ranked low because of many false positives. For example, in our earlier work, we tested ENTPRISE[10], SIFT[9], and

PolyPhen2-HDIV[7] on 10 patient samples. On average, ENTPRISE predicts ~100 disease causing genes, while SIFT, and PolyPhen2-HDIV predict ~400–500 disease-causing genes. Here, we develop a machine learning approach, MENDELSEEK, that predicts Mendelian genes by integrating the aggregate residue variation score with the gene's properties. At this point, the goal is not to predict which specific variations are causing the disease, but which genes are likely to cause Mendelian diseases/phenotypes. Thus, it could filter out the false predictions by variation-based methods. MENDELSEEK combines gene properties such as Reactome pathway information[15], protein language models[16], Gene Ontology (GO) processes[17] with the aggregate variation scores. The latter are assessed using ENTPRISE[10] for missense variations and ENTPRISE-X[11] for frame shift and stop codon variations. Such a combined approach is much more accurate than just using aggregate variation scores from various state-of-the-art methods including ENTPRISE[10] and ENTPRISE-X[11], REVEL[12], a meta method that combines 13 state-of-the-art methods: MutPred[18, 19], FATHMM[20], VEST[5], PolyPhen[7], SIFT[9], PROVEAN[21], MutationAssessor[22], MutationTaster[23], LRT[24], GERP[25], SiPhy[26], phyloP[27], and phastCons[28]. To assess performance, we benchmark MENDELSEEK with 10-fold cross-validation on 16,946 genes that includes 4,494 known Mendelian genes taken from the OMIM database, treated as true positives[29] and demonstrate that MENDELSEEK shows significant improvement over state-of-the-art approaches for discriminating a Mendelian from a non-Mendelian gene. Finally, using the input features of MENDELSEEK, we elucidate the biochemical properties that differentiate Mendelian from non-Mendelian genes.

Methods

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

Calculation of the aggregate gene variation score

As discussed in [5], 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. Ref. [5] also tested two other ways: Fisher's method[30] and Stouffer's Z-score[31]. Fisher's method requires a p-value and Stouffer's Z-score requires a Z-score; 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 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). 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)[32] for a given gene is averaged. For MAVERICK[6] 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.

Assignment of a gene to its pathways and GO processes

Pathways and their corresponding associated genes are obtained from the Reactome database[15]. The Reactome database contains 2,363 distinct pathways. Thus, its contribution to the assessment of whether a gene is Mendelian is described by a feature vector of 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/[17]. 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 component of the features. We tested GO molecular functions and cell components and found the best performing choice is the biological process.

The protein large language model (pLLM) of a gene is obtained from[16] (https://github.com/agemagician/ProtTrans). As was also found by the ligand virtual screening method ConPlex[33], the ProtBert embeddings model performs better than other alternatives. Here, we employ the 1,024-dimensional ProtBert model. Its 1,024-dimensional vector is a 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, 1,024 dimensions are from the pLLM.

Machine learning

We employed the Extreme Gradient Boosting (XGB) regression machine learning method[34]. XGB is optimized for memory usage and computational efficiency (e.g., for sparse matrix, parallel tree computing) 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 implemented in the Scikit-learn kit[35] was applied with the following

empirical parameters: n_estimators=1000, max_depth=6, and learning_rate=0.05. A known Mendelian gene as in the OMIM dataset[29] was set to an objective regression value of 1.0, otherwise, it is set to 0.0.

Evaluation of the relative importance of protein-protein interactions

The dataset of genes to be evaluated are combined with the unique genes from the STRING (with a cutoff score of 500) [36] and HIPPIE (with a cutoff score of 0.5) [37] databases to assess the relative importance of protein-protein interactions. Interestingly, we found 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 in the Results section.

Benchmarking protocol

The above analysis leads to a final set of 17,858 unique genes for evaluation; 4,494 genes overlap with the OMIM dataset[29] 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 sets are combined into a composite set for evaluation and as well as 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[38]. Here, ~25% of the total number of genes are true positives; thus, the total set is imbalanced. For all predictions, we convert the raw score to the predicted precision by:

$$predicted\ precision\ (S_0) = \frac{Number\ of\ true\ positives\ having\ raw\ score > S_0}{Number\ of\ genes\ having\ raw\ score > S_0} \quad (1).$$

For the unpaired t-test or Welch's t-test[39] on feature components for distinguishing between Mendelian genes and non-Mendelian genes, the t-value is calculated by:

$$t - value = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \quad (2),$$

where $\overline{x}_{1,2}$, $S_{1,2}$ $n_{1,2}$ are the average value, standard deviation and number of data points of known Mendelian and non-Mendelian genes, respectively.

Results

Comparison to other methods

We compared MENDELSEEK to other methods, most of which are based on variation scores, e.g. VEST[5]. 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, and MAVERICK. MENDELSEEK has an AUPR, AUC and precision for the top ranked 180 predictions of 0.695, 0.850 and 94.4%, respectively. The second-best method ENTPRISE+ENTPRISE-X has respective values of 0.604, 0.781 and 92.8%. MENDELSEEK, which includes ENTPRISE+ENTPRISE-X, performs better than ENTPRISE+ENTPRISE-X alone with a relative increase in its AUPR of 15%. 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 precision for the top 180 genes are 0.390, 0.597 and 68.3%, respectively. Some of the alternative methods even perform worse than random selection (enrichment factors < 1 or AUC < 0.5). Figure 2 shows the AUC and AUPR curves of the compared methods.

Table 2 and Figure 3 show the results on the 14,598 gene hard set that excluded the disease-causing training genes of ENTPRISE+ENTPRISE-X from the above consensus set. The best and second-best methods are again MENDELSEEK with an AUPR=0.424, AUC=0.784, and the precision of top 180 ranked genes of 71.1% and ENTPRISE+ENTPRISE-X with an AUPR=0.311, AUC=0.682, and the precision of the top 180 genes is 47.8%, respectively. The relative difference of the AUPR is ~36% (0.424 vs. 0.311). Although those AUPRs are much worse than those for the whole consensus set, they are still much better than the next best method, REVEL whose AUPR is 0.214. The enrichment factor of MENDELSEEK is even slightly better than that for the above "easier" set (4.40 vs. 3.56). Nevertheless, the maximal possible enrichment factor for the top 180 genes is 6.18 for the hard set and 3.77 for the easy set.

From Figure 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 Figure 2). MENDELSEEK's whole gene properties contribute more significantly for genes not seen in variation training. Thus, MENDELSEEK performed significantly better than all other methods for both the whole and hard sets in terms of AUPR, AUC, and top ranked 1% predictions.

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 and doing a 10-fold cross-validation for the 17,858 gene set. Table 3 shows the results. Without the ENTPRISE+ENTPRISE-X feature component, the AUPR decreases most from 0.694 to 0.598. Exclusion of the Reactome Pathway component results in the smallest decrease of AUPR to 0.684. The next smallest decrease is to 0.679, when the protein language model component is ignored.

Thus, the ENTPRISE+ENTPRISE-X score contributes the most to MENDELSEEK by increasing the AUPR from 0.598 to 0.694 (+16%), and the pathway component contributes the least, increasing the AUPR by ~1.5% from 0.684 to 0.694. The GO process component increases the AUPR by 6.9%, and the protein language model component increases the AUPR by ~2.2%.

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

Our results indicate that the ENTPRISE+ENTPRISE-X aggregate variation score has the largest contribution to MENDELSEEK. ENTPRISE+ENTPRISE-X scores are mainly determined by the protein's three-dimensional (3D) structure and structure-pathogenicity relationships learned from existing knowledge[10, 11]. The protein language model component embeds tokens describing protein amino acid sequences and encodes intrinsic properties of proteins also learned from existing knowledge[16]. Thus, we are not able to dissect them further. In contrast, the GO processes and Reactome pathways have biological meaning for each component of the feature vector. What are the GO processes and Reactome pathways that most distinguish Mendelian genes from non-Mendelian genes? We analyzed those components using the unpaired t-test between true positives and the unknown ones (the majority will be true negatives) in the 17,858 dataset (see Equation 2).

We first analyze the possible correlation of t-values of the GO processes and Reactome pathways with the maximal number of protein-protein interactions (PPI) of a given process or pathway. A gene's number of PPIs are computed from the combined set of the STRING (with a cutoff score of 500)[36] and HIPPIE (with a cutoff score of 0.5)[37] databases. The maximal number of PPIs for the top 20 t-values of GO processes and Reactome pathways are given in Table 4 and 5 (a full list can be found in Supplementary Material Table S1, S2). For the 11,920 GO

processes having a maximal number of PPIs, the Pearson's correlation between t-values and the corresponding maximal number of PPIs is 0.309 with a p-value of 0. For the 2,359 pathways having a maximal number of PPIs, the correlation coefficient is 0.283 with a p-value 9.8×10^{-45}. This means genes with GO processes or pathways having a higher number of protein-protein interactions are more likely to be Mendelian genes (higher t-values). Since they are reasonably correlated, inclusion of the number of protein-protein interactions does not improve performance.

Correlations of GO processes and Reactome pathways with the likelihood of disease causality

Although it is assumed that GO processes and pathways involved in Mendelian genes should more likely be disease driving than others, here we provide an assessment of the quantitative correlation between these two gene properties. A gene's likelihood of being disease driving is predicted by our previous work, LeMeDISCO[40, 41] and quantified by a value called a normalized p-value weighted frequency (F_{moa}). For each of the 3,608 diseases in our LeMeDISCO library, we used MEDICASCY[42] to derive the associated mode of action (MOA) proteins based on their p-value (which is then converted to a q-value, with a q-value cutoff of 0.05)[43]. To further prioritize them according to their likelihood of contributing to efficacy, the p-value weighted frequency of shared MOA proteins across all diseases is calculated. We define a p-value weighted frequency of an input MOA protein T as follows: If protein T is shared by a disease D and the p-value of T associated with D is P, then its weight defined by the min (1.0,-αlogP) is counted as T's frequency. In practice, we used data from ten cancer cell lines[44] to determine the coefficient α. There, we used this information to predict which are the driver proteins for a given cancer and then use this information to predict the most likely effective drugs. By varying the parameter α from 0.005 to 0.25 on these ten cell lines, α=0.025 gives the best drug

screening results[41]. The normalization is done by dividing the weighted frequency by the number of diseases considered.

As above, the maximal number of PPIs, for each GO process and pathway, the maximal normalized p-value weighted frequency defined as the maximal value of a gene involving the process or pathway (denoted as $maxF_{moa}$). For 11,358 processes with $maxF_{moa}$ values >0, the correlation coefficient between the t-value and $maxF_{moa}$ is 0.342 with a p-value of zero! For 2,335 pathways, the correlation coefficient is 0.406 with a p-value of zero. Thus, the t-values of both GO processes and pathways are highly correlated with $maxF_{moa}$, even higher than those of the number of PPIs. Thus the idea that Mendelian genes are effective single protein disease drivers is valid.

Correlations of GO processes and Reactome pathways with evolutional time

In Tables 4 and 5, we also present the top 20 GO processes and Reactome pathways ranked by their t-values along with their minimal LCA (Lowest Common Ancestor) evolutional time scale. LCA values range from 1 to 31, with 1 being the oldest, i.e., at the origin of life to first cellular organisms (i.e. Prokaryota) as determined in [45]. The minimal LCA is the minimal value of a gene's LCA having/involving the same GO process/pathway. For all 11,920 GO processes having minimal LCA values, the Pearson's correlation between the t-value and the corresponding minimal LCA is -0.197 with a p-value of 0. For the 2,354 pathways having minimal LCA values, the correlation coefficient is -0.173 with a p-value of 2.5×10^{-17}. Thus, genes likely to be Mendelian are the most ancient.

Since all of the above three properties (maximal PPIs, $maxF_{moa}$ and LCA) of GO processes and pathways are highly correlated with their t-values that characterize their ability of distinguishing Mendelian genes from non-Mendelian genes, and therefore to each other, as noted above, adding any of them to the feature vector does not improve MENDELSEEK's performance.

Again, this is because all these properties are already encoded in the GO processes, Reactome pathways, and the ENTPRISE+ENTPRISE-X aggregate score. In fact, the correlations between the ENTPRISE+ENTPRISE-X score and the LCA, number of PPIs and F_{moa} of the 17,858 genes are -0.338, 0.228 and 0.157, with corresponding p-values of zero, respectively. Direct correlations of Mendelian gene values (set to a value of 1.0 in regression training, and non-Mendelian gene set to 0.0) to the ENTPRISE+ENTPRISE-X score, LCA, number of PPIs and F_{moa} of 17,858 genes are 0.467, -0.183, 0.107 and 0.061, with p-values 0.0, 0.0, 0.0 and 3.38×10^{-16}, respectively.

The top 20 GO process and pathways by t-values ranked are the oldest with a minimal LCA of 1. The top three of these GO processes are related to sensory processes: *visual perception, sensory perception of sound, response to xenobiotic stimulus*. The GO process *visual perception* (GO:0007601) with the largest t-value of 16.6, has 29 phenotypes caused by 6 genes (BBS4, COL1A1, COL2A1, OAT, RDH11, RPE65) having a LCA of 1. While it may seem odd at first glance that visual perception has an LCA=1, these genes actually 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[46] 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[47] and involves 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[48]. COL2A1 is involved in 39 GO processes including *visual perception*, *sensory perception of sound, skeletal system development, collagen fibril organization, central nervous system development, cartilage development as well as others.* Furthermore, the current OMIM database[29] lists 15 phenotypes caused by COL2A1. OAT encodes the ornithine aminotransferase

enzyme, that is found in the 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[49]. RDH11, retinol dehydrogenase, in addition to being an essential gene in the eye, is involved in the cellular detoxification of aldehyde[17, 50], one of its 6 human GO processes. Finally, while RPE65 helps convert light into electrical signals that are sent to the brain, this protein is involved in 14 human GO processes including the *insulin receptor signaling pathway* (GO:0008286).

In humans, there are 29 phenotypes of these genes; many are eye diseases (see Table S3 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 caused by COL2A1. There are also non-eye diseases: Czech dysplasia, Achondrogenesis, 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 next three GO processes are replication related: *positive regulation of gene expression, positive regulation of transcription by RNA polymerase II, positive regulation of DNA-templated transcription*. These processes are essential for life.

The top four pathways of Mendelian genes are *Metabolism* (t-value=20.7), *Metabolic pathways* (t-value=20.1), *Pathways in cancer* (t-value=15.6), *Disease* (t-value=13.0). These pathways are crucial because they allow cells to efficiently capture and utilize energy from nutrients, enabling essential functions like growth, reproduction, maintaining structure, and when uncontrolled, they could result in cancers in some organisms. The top pathway of *Metabolism*

(REACT:R-HSA-1430728) is associated with 371 phenotypes caused by genes having a LCA of 1 (see Table S4 for 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)[36] and HIPPIE (with a cutoff score of 0.5)[37] 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 the interaction dataset. Equation 1 converts the raw regression score to the predicted precision score. With a predicted precision score cutoff of 0.7, we have 1,024 novel gene predictions (those not known to be true positives by MENDELSEEK, i.e. they are not in the OMIM database). These predictions of novel Mendelian genes are listed in Supplementary Material Table S5. We then asked ChatGPT (https://platform.openai.com, o3-mini model) if a given gene is disease causing based on known literature evidence. We also ask the same question for the 1,000 lowest ranked unknown genes. For the 1,024 top ranked genes with a predicted precision score >0.7, 175 genes have supporting evidence, whereas for the lowest ranked 1,000, only 23 genes are supported by the literature. The enrichment factor of the top ranked genes vs. lowest ranked genes is 7.43. For the randomly selected 1,000 genes, 79 have supporting evidence. This leads to MENDELSEEK's effective enrichment of 2.1 for discovering novel Mendelian genes. For ENTPRISE+ENTPRISE-X, 167 genes have supporting evidence which is close to MENDELSEEK's result and consistent with the benchmarking results in Table 1. In contrast, for REVEL, only 99 genes have supporting evidence; roughly half that of MENDELSEEK's 175 and slightly better than random selection. Again, these results are consistent with the benchmarking evaluation provided in Table 1 and 2.

Examples of the top predicted genes with literature evidence that are not known to the MENDELSEEK algorithm are: PLEC (plectin gene) having a predicted precision of 0.94. Mutations on this gene cause autosomal recessive epidermolysis bullosa simplex with muscular dystrophy[51]. FYN (precision=0.94) encoding a tyrosine kinase that its dysfunction together with Src causes a Reeler-like phenotype[52]. RIMS1 (precision=0.94) was documented as causing Cone-rod dystrophies (CORDs)[53]. PARK7 is a rare recessive form of early-onset Parkinson Disease (PD) caused by mutations in PARK7 (precision=0.94)[54]. Variants in SORL1 (precision=0.94) have been implicated in familial dementia[55]. Whether the other gene predictions are false positives or novel true positives are uncertain at this juncture. A full list of predicted genes can be found in Supplementary Material Table S5. Those with ChatGPT evidence of being disease causing are marked in Table S5 and those not marked, can serve as guidance for further bioinformatics/experimental validations/tests.

Difference between Mendelian genes and complex disease driving genes

The above analysis shows that Mendelian genes have high correlation with their likelihood of being disease drivers. Are they also drivers of complex diseases? Combining the known 4494 OMIM genes and the 1024 predicted Mendelian genes leads to 5518 putative Mendelian genes. For putative complex disease driving genes, applying a cutoff of 0.01 to the above p-value weighted frequency (this means approximately > 1% of the 3608 complex diseases have the gene as a driver) gives 7309 genes. The two sets of genes have 2242 overlaps. Thus, a subset of the Mendelian genes are also involved in complex diseases (see Table S6). The remaining 3277 putative Mendelian genes are not complex disease drivers, with 2674 of these found in OMIM. 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.

Discussions

We have shown that MENDELSEEK performs significantly better than other methods in discriminating Mendelian genes (those whose variations alone are sufficient to cause a disease) from non-Mendelian genes. MENDELSEEK's predictions show consistent performance on benchmarking tests and supporting literature evidence. The ablation study indicates that the ENTPRISE+ENTPRISE-X score which is associated with residue variation contributes the most (increases the AUPR by 16%) to its improved performance. These quite confident results reflects the low false positive rate of ENTPRISE[10] and ENTPRISE-X[11] in predicting disease causing variations. In practice, ENTPRISE has a 10.7% false positive rate compared to the 36.4% by PolyPhen-2-HVAR[7] on the 1,000 genome dataset[10]. On the VEST-indel test set, ENTPRISE-X has a false positive rate of 8.4% compared to 18.6% by VEST-indel[11, 56].

Second to the ENTPRISE+ENTPRISE-X score, the GO process component increases the AUPR by 6.9%. Unpaired t-test analysis on the individual GO processes that discriminate Mendelian genes from non-Mendelian genes shows that the most discriminating GO processes (large positive t-value) tend to involve the oldest genes (small LCA value), genes having a larger number of PPIs, and genes having a larger likelihood of being disease-driving. This tendency is also true for the most discriminating Reactome pathways.

As indicated by literature mining, ~20% of MENDELSEEK's new predictions have supporting literature evidence. The remaining novel predictions are potential high value targets for future experimental discovery/validation. Future work will involve the extension of MENDELSEEK to the prediction of the specific phenotypes (e.g. diseases) a gene causes and the pattern of genetic transmission (autosomal dominant or autosomal recessive). These studies assume that a specific phenotype is associated with specific GO processes/pathways and the

pattern of this association can be learned by artificial intelligence. Another issue that requires further investigation is when one gene causes multiple phenotypes. For example, since COL2A1 causes 15 different phenotypes according to the OMIM database, can one predict which of these phenotypes will actually occur in a given patient? Can one identify "protective" genes that prevent one/some phenotypes from happening? In general, there exist "genetic modifiers"[57] in Mendelian diseases that could affect the outcome of specific malfunctional genes. But how can one identify them and what is responsible for their behavior?

More generally, for non-Mendelian diseases, there are usually dozens to hundreds of mutated disease-causing genes, and which combination of genes/mutations actually cause the disease is usually not obvious[10]. In contrast, because of their unimodal character Mendelian genes have much to teach us about the interrelationship between a disease and the genetic variations that cause the disease. By developing tools that map such genes to their associated phenotypes, the resulting algorithms and insights are likely to have applications to non-Mendelian diseases as well.

Acknowledgments

This research was supported in part by grant, GM-118039, from the Division of General Medical Sciences of the National Institutes of Health. A gift from the Ovarian Cancer Institute is also gratefully acknowledged. We thank Jessica Forness for proofreading and polishing this manuscript and Bartosz Ilkowski for his computational support.

Author contributions:

Conceptualization: JS, HZ
Methodology: HZ, JS
Investigation: HZ, JS
Funding acquisition: JS
Project administration: JS

Supervision: JS
Writing – original draft: HZ
Writing – review and editing: JS, HZ

Financial Interest

The authors declare that they have no financial interest in the outcome of this work.

Data and Software Availability

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

Clinical trial number: not applicable

References

1. Wright CF, FitzPatrick DR, Firth HV: **Paediatric genomics: diagnosing rare disease in children**. *Nat Rev Genet* 2018, **19**(5):253-268.
2. Seaby EG, Rehm HL, O'Donnell-Luria A: **Strategies to Uplift Novel Mendelian Gene Discovery for Improved Clinical Outcomes**. *Frontiers in Genetics* 2021, **12**.
3. Bamshad MJ, Nickerson DA, Chong JX: **Mendelian Gene Discovery: Fast and Furious with No End in Sight**. *Am J Hum Genet* 2019, **105**(3):448-455.
4. Manolio TA: **Genomewide Association Studies and Assessment of the Risk of Disease**. *The New England Journal of Medicine* 2010, **363**(2):166-176.
5. 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.
6. Danzi MC, Dohrn MF, Fazal S, Beijer D, Rebelo AP, Cintra V, Züchner S: **Deep structured learning for variant prioritization in Mendelian diseases**. *Nature Communications* 2023, **14**(1):4167.
7. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**(4):248-249.
8. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J *et al*: **Predicting the clinical impact of human mutation with deep neural networks**. *Nature Genetics* 2018, **50**(8):1161-1170.
9. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic Acids Res* 2003, **31**(13):3812-3814.

10. 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.

11. Zhou H, Gao M, Skolnick J: **ENTPRISE-X: Predicting disease-associated frameshift and nonsense mutations**. *PLoS One* 2018, **13**(5):e0196849.

12. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D *et al*: **REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants**. *Am J Hum Genet* 2016, **99**(4):877-885.

13. 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-d1154.

14. Pearson TA, Manolio TA: **How to interpret a genome-wide association study**. *Jama* 2008, **299**(11):1335-1344.

15. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R *et al*: **The reactome pathway knowledgebase**. *Nucleic Acids Res* 2020, **48**(D1):D498-d503.

16. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M *et al*: **ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning**. *IEEE Trans Pattern Anal Mach Intell* 2022, **44**(10):7112-7127.

17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**(1):25-29.

18. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H-J, Mort M, Cooper DN, Sebat J, Iakoucheva LM *et al*: **Inferring the molecular and phenotypic impact of amino acid variants with MutPred2**. *Nature Communications* 2020, **11**(1):5918.

19. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P: **Automated inference of molecular mechanisms of disease from amino acid substitutions**. *Bioinformatics* 2009, **25**(21):2744-2750.

20. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR: **Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models**. *Hum Mutat* 2013, **34**(1):57-65.

21. Choi Y, Chan AP: **PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels**. *Bioinformatics* 2015, **31**(16):2745-2747.

22. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics**. *Nucleic Acids Research* 2011, **39**(17):e118-e118.

23. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D: **MutationTaster evaluates disease-causing potential of sequence alterations**. *Nature Methods* 2010, **7**(8):575-576.

24. Chun S, Fay JC: **Identification of deleterious mutations within three human genomes**. *Genome Res* 2009, **19**(9):1553-1561.

25. 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.

26. 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.

27. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies**. *Genome Res* 2010, **20**(1):110-121.

28. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**(8):1034-1050.

29. **Online Mendelian Inheritance in Man, OMIM®** [https://omim.org/]

30. Fisher R, Bennett J, Yates F: **Statistical Methods, Experimental Design, And Scientific Inference: A Re-issue Of Statistical Methods For Research Worker.**; 1990.

31. Stouffer S, Suchman E, Devinney L, Star S, Williams R, Jr: **The American soldier: adjustment during army life.**, vol. 1; 1949.

32. 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 Medicine* 2020, **12**(1):103.

33. Singh R, Sledzieski S, Bryson B, Cowen L, Berger B: **Contrastive learning in protein language space predicts interactions between drugs and protein targets**. *Proceedings of the National Academy of Sciences* 2023, **120**(24):e2220778120.

34. 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: Association for Computing Machinery; 2016: 785–794.

35. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel BT, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss VD, Jake Vanderplas, Alexandre

Passos *et al*: **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research* 2011, **12**(85):2825-2830.

36. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S *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-d646.

37. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH: **HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks** *Nucleic Acids Research* 2017, **45**(D1):D408-D414.

38. Ozenne B, Subtil F, Maucort-Boulch D: **The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases**. *Journal of Clinical Epidemiology* 2015, **68**(8):855-859.

39. Welch BL: **THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED**. *Biometrika* 1947, **34**(1-2):28-35.

40. Courtney Astore, Zhou H, Ilkowski B, Forness J, Skolnick J: **LeMeDISCO is a computational method for large-scale prediction & molecular interpretation of disease comorbidity**. *Communications Biology* 2022, **5**:870.

41. Astore C, Zhou H, Jacob J, Skolnick J: **Prediction of severe adverse events, modes of action and drug treatments for COVID-19's complications**. *Scientific Reports* 2021, **11**(1):20864.

42. Zhou H, Cao H, Matyunina L, Shelby M, Cassels L, McDonald JF, Skolnick J: **MEDICASCY: A Machine Learning Approach for Predicting Small-Molecule Drug**

Side Effects, Indications, Efficacy, and Modes of Action. *Molecular Pharmaceutics* 2020, **17**(5):1558-1574.

43. John DS: **The positive false discovery rate: a Bayesian interpretation and the q-value**. *The Annals of Statistics* 2003, **31**(6):2013-2035.

44. **NCI-60 Human Tumor Cell Lines Screen: https://dtp.cancer.gov/discovery_development/nci-60/**.

45. Lopes KP, 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.

46. Hernandez-Hernandez V, Pravincumar P, Diaz-Font A, May-Simera H, Jenkins D, Knight M, Beales PL: **Bardet-Biedl syndrome proteins control the cilia length through regulation of actin polymerization**. *Hum Mol Genet* 2013, **22**(19):3858-3868.

47. Małecki K, Fabiś-Strobin A, Sałacińska K, Kwas K, Stelmach W, Beczkowski J, Niedzielski K, Gach A: **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.

48. Glavey SV, Naba A, Manier S, Clauser K, Tahri S, Park J, Reagan MR, Moschetta M, Mishima Y, Gambella M *et al*: **Proteomic characterization of human multiple myeloma bone marrow extracellular matrix**. *Leukemia* 2017, **31**(11):2426-2434.

49. Urra M, Buezo J, Royo B, Cornejo A, López-Gómez P, Cerdán D, Esteban R, Martínez-Merino V, Gogorcena Y, Tavladoraki P *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):5581-5595.

50. Kasus-Jacobi A, Ou J, Bashmakov YK, Shelton JM, Richardson JA, Goldstein JL, Brown MS: **Characterization of mouse short-chain aldehyde reductase (SCALD), an enzyme regulated by sterol regulatory element-binding proteins**. *J Biol Chem* 2003, **278**(34):32380-32389.

51. Servidei S, Primiano G, Muto V, Cuccagna C, Bernardo D, Sauchelli D, Sancricca C, Lucchini M, Mirabella M, Tartaglia M: **PLEC gene mutations cause familial disto-proximal myopathy and long QT syndrome mimicking mitochondrial disease**. *Neuromuscular Disorders* 2017, **27**:S150-S151.

52. Kuo G, Arnaud L, Kronstad-O'Brien P, Cooper JA: **Absence of Fyn and Src causes a reeler-like phenotype**. *J Neurosci* 2005, **25**(37):8578-8586.

53. 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.

54. Xiao C, Markello T, Zein WM, Bishop R, Groden C, Gahl W, Toro C: **PARK7-Related Early Onset Parkinson Disease in the Setting of Complete Uniparental Isodisomy of Chromosome 1**. *Neurology Genetics*, **7**(4):e606.

55. Alvarez-Mora MI, Blanco-Palmero VA, Quesada-Espinosa JF, Arteche-Lopez AR, Llamas-Velasco S, Palma Milla C, Lezana Rosales JM, Gomez-Manjon I, Hernandez-Lain A, Jimenez Almonacid J *et al*: **Heterozygous and Homozygous Variants in SORL1 Gene in Alzheimer's Disease Patients: Clinical, Neuroimaging and Neuropathological Findings**. *Int J Mol Sci* 2022, **23**(8).

56. Douville C, Masica D, Stenson P, Cooper D, Gygax D, Kim R, Ryan M, Karchin R: **Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel)**. *Hum Mutat* 2016, **37**(1):28-35.

57. Rahit K, Tarailo-Graovac M: **Genetic Modifiers and Rare Mendelian Disease**. *Genes (Basel)* 2020, **11**(3).

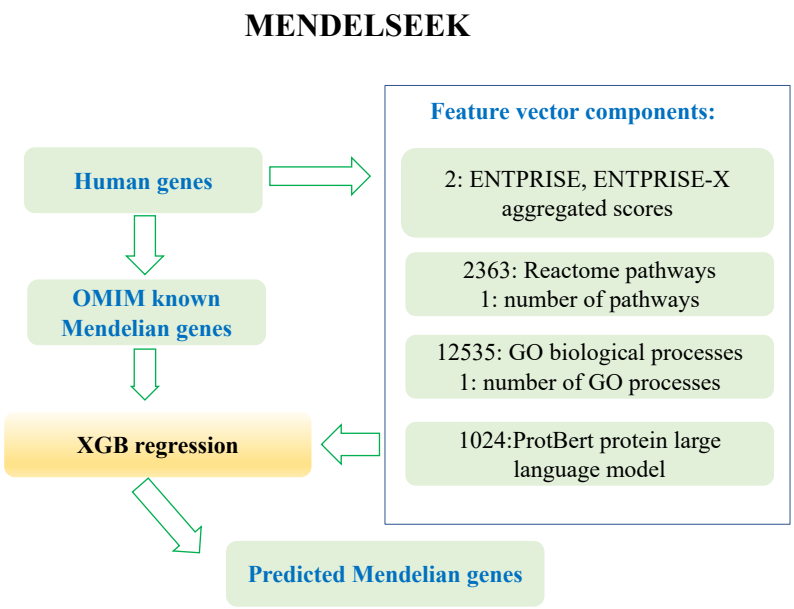

Figure 1: Flowchart of the MENDELSEEK algorithm.

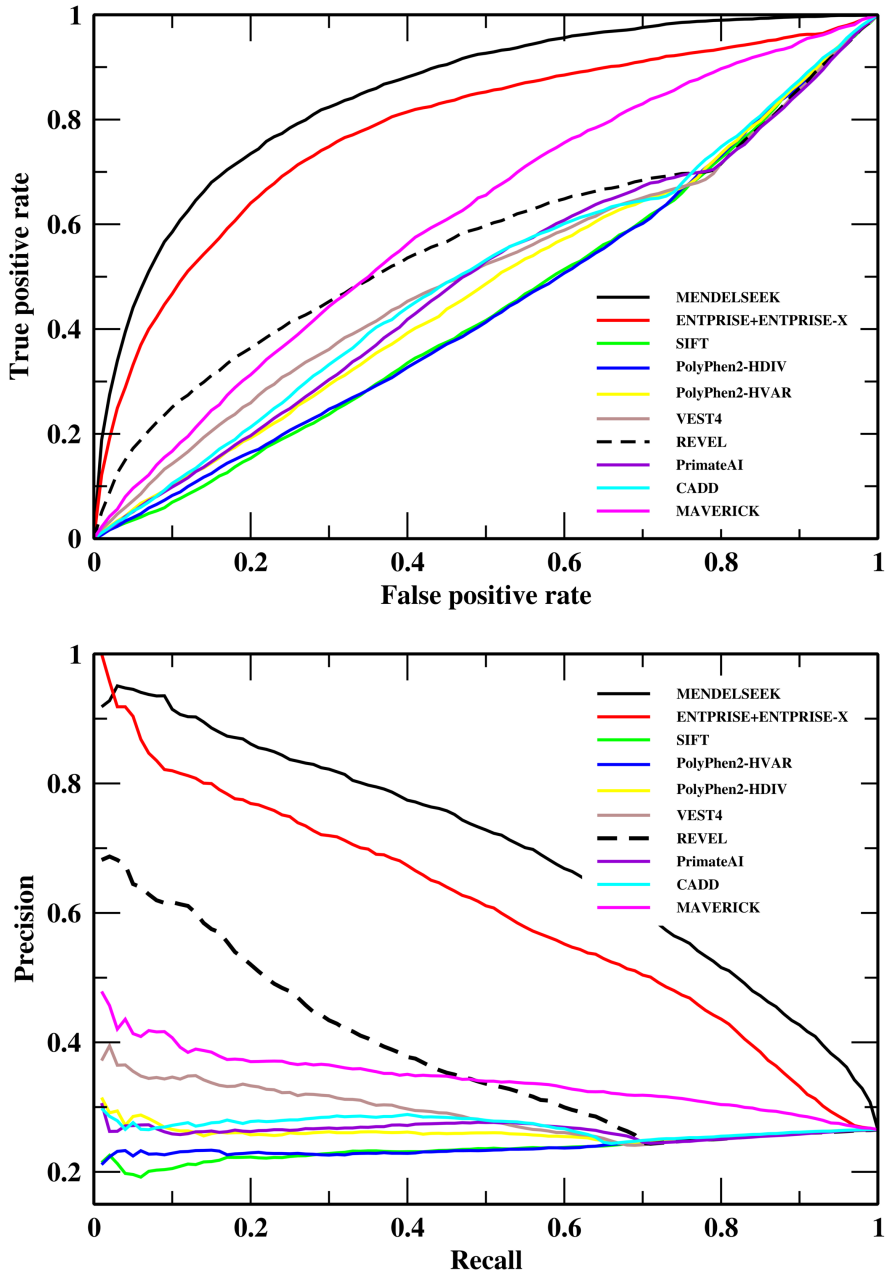

Figure 2: Classification performance of MENDELSEEK on the consensus 16,946 set in comparison to other methods. Receiver Operating Characteristic (upper figure) and precision-recall curve (lower figure).

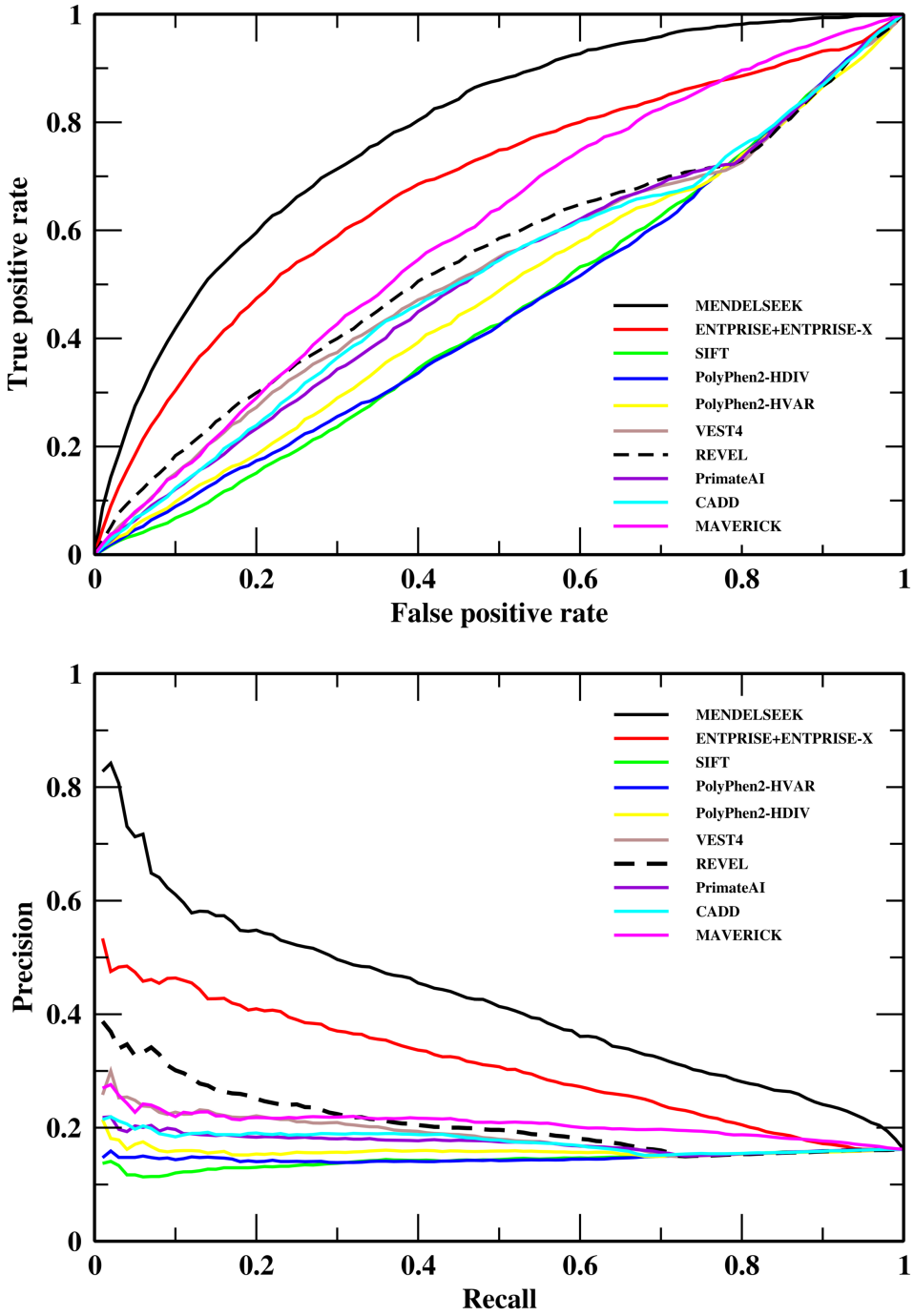

Figure 3: Classification performance of MENDELSEEK on the consensus 14,598 hard set in comparison to other methods. Receiver Operating Characteristic (upper figure) and precision-recall curve (lower figure).

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

Method	Top 180 precision	Enrichment factor of top ranked 180 genes[a]	AUC	AUPR
MENDELSEEK	94.4%	3.56	0.850	0.695
ENTPRISE+ENTPRISE-X	92.8%	3.50	0.781	0.604
REVEL	68.3%	2.58	0.597	0.390
MAVERICK	43.9%	1.65	0.611	0.345
VEST4	40.0%	1.51	0.535	0.298
PolyPhen2-HVAR	31.1%	1.17	0.514	0.269
PrimateAI	28.3%	1.07	0.523	0.271
CADD	28.3%	1.07	0.543	0.281
SIFT	22.2%	0.84	0.476	0.245
PolyPhen2-HDIV	20.0%	0.75	0.481	0.248

[a] The maximal possible enrichment factor at the top ranked 180 genes is 3.77.

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

Method	Top 180 precision	Enrichment factor of top ranked 180 genes[a]	AUC	AUPR
MENDELSEEK	71.1%	4.40	0.784	0.424
ENTPRISE+ENTPRISE-X	47.8%	2.95	0.682	0.311
REVEL	33.3%	2.06	0.571	0.214
MAVERICK	27.8%	1.72	0.600	0.208
VEST4	27.8%	1.72	0.551	0.193
PolyPhen2-HVAR	21.1%	1.31	0.516	0.164
PrimateAI	21.1%	1.31	0.538	0.178
CADD	19.4%	1.20	0.554	0.183
PolyPhen2-HDIV	16.1%	1.00	0.488	0.154
SIFT	13.3%	0.82	0.482	0.149

[a] The maximal possible enrichment factor at the top ranked 180 genes is 6.18.

Table 3. Ablation results on the 17,858 gene set

Method	Precision of the top ranked 180 genes	AUC	AUPR
MENDELSEEK	94.4%	0.858	0.694
With ENTPRISE and ENTPRISE-X removed	87.2%	0.812	0.598
With protein language model removed	90.0%	0.850	0.679
With GO processes removed	90.6%	0.825	0.649
With Reactome pathways removed	92.8%	0.853	0.684

Table 4. Top 20 GO processes that distinguish Mendelian genes

t-value	Minimal LCA[a]	Maximal # of PPIs	ID	Name
16.6302	1	1839	GO:0007601	Visual perception
12.9975	1	1843	GO:0007605	Sensory perception of sound
12.3455	1	2052	GO:0009410	Response to xenobiotic stimulus
12.0455	1	2288	GO:0010628	Positive regulation of gene expression
10.7547	1	6164	GO:0045944	Positive regulation of transcription by RNA polymerase II
10.6129	1	6136	GO:0045893	Positive regulation of DNA-templated transcription
10.3748	1	164	GO:0001501	Skeletal system development
9.74437	1	1810	GO:0007507	Heart development
9.6254	1	1771	GO:0007368	Determination of left/right symmetry
9.40371	1	230	GO:0001701	In utero embryonic development
9.27579	1	7425	GO:0060041	Retina development in camera-type eye
9.02412	1	2225	GO:0010467	Gene expression
8.94095	1	7545	GO:0060271	Cilium assembly
8.92409	1	5502	GO:0043066	Negative regulation of apoptotic
8.83786	1	1893	GO:0008203	Cholesterol metabolic process
8.77946	1	355	GO:0001947	Heart looping
8.69407	1	1722	GO:0007224	Smoothened signaling pathway
8.68066	1	2289	GO:0010629	Negative regulation of gene expression
8.67107	1	5962	GO:0045494	Photoreceptor cell maintenance
8.53973	1	1793	GO:0007420	Brain development

[a] Minimal LCA (Lowest Common Ancestor) is the minimal value of a gene's LCA as determined in [45] having the same GO process. Gene LCA values are from 1 to 31 with 1 the oldest.

Table 5. Top 20 Reactome pathways that distinguish Mendelian genes

t-value	Minimal LCA[a]	Maximal # of PPIs	ID	Name
20.6922	1	667	REACT:R-HSA-1430728	Metabolism
20.1039	1	90	KEGG:hsa01100	Metabolic pathways
15.5678	1	284	KEGG:hsa05200	Pathways in cancer
13.0431	1	781	REACT:R-HSA-1643685	Disease
11.8945	1	747	REACT:R-HSA-162582	Signal transduction
11.7972	1	614	REACT:R-HSA-1266738	Developmental biology
11.7168	1	1421	REACT:R-HSA-382551	Transmembrane transport of small molecules
11.4752	1	851	REACT:R-HSA-168256	Immune system
11.2803	1	149	KEGG:hsa04151	PI3K-Akt signaling pathway
11.2081	1	679	REACT:R-HSA-1474244	Extracellular matrix organization
10.5965	1	847	REACT:R-HSA-168249	Innate immune system
10.4678	1	288	KEGG:hsa05205	Proteoglycans in cancer
10.4416	1	85	KEGG:hsa00970	Aminoacyl-tRNA
10.3852	1	616	REACT:R-HSA-1280215	Cytokine signaling in immune system
10.3044	1	1528	REACT:R-HSA-422475	Axon guidance
10.1506	1	1396	REACT:R-HSA-379724	tRNA Aminoacylation
10.1181	1	91	KEGG:hsa01200	Carbon metabolism
9.95835	1	508	REACT:R-HSA-109582	Hemostasis
9.94678	1	2100	REACT:R-HSA-71387	Metabolism of carbohydrates
9.92088	1	441	KEGG:hsa M00359	Aminoacyl-tRNA

[a] Minimal LCA (Lowest Common Ancestor) is the minimal value of a gene's LCA as determined in [45] having the same pathway. Gene LCA values are from 1 to 31 with 1 the oldest.