Multi-objective prioritization of genes for high-throughput functional assays towards improved clinical variant classification

Yile Chen[1†], Shantanu Jain[2†], Daniel Zeiberg[2], Lilia M. Iakoucheva[3], Sean D. Mooney[1], Predrag Radivojac[2*], Vikas Pejaver[4*]

[1] *University of Washington, Seattle, WA,* [2] *Northeastern University, Boston, MA,* [3] *University of California San Diego, La Jolla, CA,* [4] *Icahn School of Medicine at Mount Sinai, New York, NY*

The accurate interpretation of genetic variants is essential for clinical actionability. However, a majority of variants remain of uncertain significance. Multiplexed assays of variant effects (MAVEs), can help provide functional evidence for variants of uncertain significance (VUS) at the scale of entire genes. Although the systematic prioritization of genes for such assays has been of great interest from the clinical perspective, existing strategies have rarely emphasized this motivation. Here, we propose three objectives for quantifying the importance of genes each satisfying a specific clinical goal: (1) Movability scores to prioritize genes with the most VUS moving to non-VUS categories, (2) Correction scores to prioritize genes with the most pathogenic and/or benign variants that could be reclassified, and (3) Uncertainty scores to prioritize genes with VUS for which variant pathogenicity predictors used in clinical classification exhibit the greatest uncertainty. We demonstrate that existing approaches are sub-optimal when considering these explicit clinical objectives. We also propose a combined weighted score that optimizes the three objectives simultaneously and finds optimal weights to improve over existing approaches. Our strategy generally results in better performance than existing knowledge-driven and data-driven strategies and yields gene sets that are clinically relevant. Our work has implications for systematic efforts that aim to iterate between predictor development, experimentation and translation to the clinic.

Keywords: Multiplexed Assays of Variant Effect; MAVE; clinical variant classification; variant pathogenicity prediction, gene prioritization.

1. Introduction

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have developed guidelines to standardize the practice of clinical variant classification and interpretation.[1] These guidelines group the disparate sources of information about a genetic variant into different lines of evidence, weigh them in terms of evidential strength, and provide rules to combine these differently weighted lines of evidence to assign a variant to one of five classes: pathogenic, likely pathogenic, benign, likely benign or a variant of uncertain significance (VUS). Despite the tremendous progress that the ACMG/AMP guidelines have brought about, a substantial number of variants, particu-

† These authors contributed equally

∗ To whom correspondence should be addressed (vikas.pejaver@mssm.edu)

larly missense, remain VUS due to the limited availability of evidence.[2] Furthermore, variants assigned to the remaining four classes are often reclassified due to initial misclassification.[3]

Among the evidential lines, functional evidence derived from *in vitro* assays holds the potential to address aforementioned challenges, as they are weighted highly in the ACMG/AMP guidelines. In particular, multiplexed assays of variant effects (MAVEs) can query the functional impact of all possible amino acid substitutions at every position in a protein within a single assay, allowing for the construction of a variant effect map for all missense variants for a gene.[4,5] However, only a limited number of genes have been assayed with the explicit intent of addressing the goal of clinical variant interpretation.

Historically, the selection of genes for MAVEs and functional characterization has been driven by study-specific motivations, including the study of sequence-structure-function relationships,[6] the characterization of biologically or medically important genes[7] and the development of new technology.[8] This is typically done on the basis of prior knowledge and expertise and is likely to recapitulate preferences for well-studied genes.[9] With the accumulation of large numbers of clinically interpreted variants in knowledgebases such as ClinVar,[10] it is now feasible to devise data-driven strategies to more directly address clinical objectives when prioritizing genes for MAVEs. To date, only one study has sought to systematically prioritize genes explicitly for clinical decision-making.[2] This study proposed a difficulty-adjusted impact score (DAIS) that accounted for the number of VUS in each gene, after adjusting for gene length, and up-weighted those that appeared in multiple patients and for which classifications were most likely to be impacted upon adding new functional evidence.

To the best of our knowledge, none of these strategies have incorporated computational predictors of variant pathogenicity. Variant pathogenicity predictors assign scores to each variant indicative of their pathogenicity based on different features such as sequence context, evolutionary history, protein structure and function, among others.[11] Recent work has suggested that at appropriate score thresholds, some predictors can provide strong evidence for both pathogenicity and benignity as per the ACMG/AMP guidelines.[12] This motivates an alternative strategy that uses computational variant pathogenicity predictors to guide the selection of genes for MAVEs such that when functional and predictive evidence are combined, they will be of sufficient strength to impact the overall clinical classifications of a large set of variants across different genes.

Here, we define three objectives for gene prioritization for MAVEs that improve clinical variant classification and operationalize these objectives through the use of variant pathogenicity predictors. We formalize the process of prioritizing genes for MAVEs solely from the perspective of clinical variant classification and define three objectives (two direct and one indirect) that are desirable in this context. The first two were devised to (1) move the most VUS towards more definitive classifications of pathogenicity or benignity, and (2) reassess and possibly correct existing classifications of the highest numbers of pathogenic and/or benign variants. The third objective emphasizes the use of MAVEs as a means to improve pathogenicity predictors themselves, which in turn, when combined with MAVE data can reclassify VUS. We then quantify to what extent the genes that have already been assayed in the literature or are registered to be assayed by MAVEs fulfill these objectives, along with other poten-

tial strategies that one could adopt. Finally, we present and evaluate alternative strategies to prioritize genes such that these objectives are optimized individually and when combined.

2. Methods

2.1. *Data collection*

ClinVar variants. We extracted all missense variants in ClinVar (October 2021) and separated them by the category of clinical significance: Pathogenic (P), Likely Pathogenic (LP), Benign (B), Likely Benign (LB), variants of uncertain significance (VUS), and variants with conflicting interpretations of pathogenicity for each gene. The ClinVar data set contained 11,281 genes with 402,721 missense variants (Supp. Table 1).

gnomAD variants. VUS in ClinVar are likely to accumulate in a biased manner due to differences in the frequency with which different genes are tested. At the gene-level, variants in population-scale sequencing resources such as gnomAD accumulate in a less biased manner as all genes are likely to be uniformly sampled. To this end, we extracted missense variants from gnomAD (v2.1.1 GRCh38 dataset) as an additional set of variants that are not annotated as P, LP, B or LB.[13] Only variants with genotype quality (GQ) ≥ 20 and depth (DP) ≥ 10 were retained. We identified 17,988 genes that had 4,542,252 missense variants.

Genes with MAVEs. We extracted genes from three resources: MaveDB,[14] VariantEffect (`https://github.com/VariantEffect/MaveReferences`), and MaveRegistry,[15] to create a representative set of genes with functional data. The first two record and maintain information on which genes have been subject to MAVEs either by submission to the resource or by reviewing the literature. MaveRegistry hosts information on which genes are currently being assayed or are expected to be assayed in the near future. After accounting for overlaps between these resources, we were left with a set of 94 assayed genes.

2.2. *Data pre-processing*

We treated P, LP, and P/LP as a single pathogenic category; B, LB, and B/LB as a single benign category; VUS and conflicting interpretations of pathogenicity as the VUS category. Motivated by the clinical objectives that we define in Section 2.4, we only retained genes that had at least one VUS and at least one pathogenic or benign variant in the ClinVar data set, reducing our data set to 3,981 genes. Considering the increased difficulty in mapping variant effects for longer proteins, we removed genes that were longer than genes previously assayed by MAVEs. We also removed genes that were shorter than those previously assayed because these genes may have had too few known variants to justify prioritization for MAVEs. Only genes that appeared in both ClinVar and gnomAD were considered and variants that were recorded in both databases were removed from gnomAD data so as to avoid double-counting when scoring. The set of genes remaining after these pre-processing steps (3,829 genes with 321,619 VUS/P/B variants and 1,161,072 gnomAD variants) served as our starting gene set.

2.3. *Obtaining calibrated REVEL scores*

REVEL is a meta-predictor that combines scores from multiple pathogenicity predictors and has been shown to perform well for clinical variant interpretation.[11] For each variant in all

data sets, we extracted REVEL scores by mapping the chromosomal position and amino acid alteration to REVEL's prediction tables.[11] However, REVEL scores themselves are not calibrated for clinical use and our formulations for clinical objectives require that prediction scores best approximate the posterior probability of pathogenicity/benignity (Section 2.4). Therefore, we obtained a mapping of all possible REVEL scores to local posterior probability of pathogenicity and benignity from Pejaver *et al.*[12] We then recorded these local posterior probabilities for all variants in this study and used them in all analyses.

2.4. *Gene prioritization objectives: a clinical perspective*

From a clinical perspective, the overall goal of gene prioritization is to make definitive and accurate classifications for more variants appearing in patient populations, when combining new functional evidence and existing evidence. This includes: (1) assisting the movement of VUS to pathogenic and benign classes, (2) correcting for errors in current pathogenic and benign classifications and (3) improving predictors to assist clinical decision making. To operationalize these objectives we rely on pathogenicity predictions from REVEL for variants in ClinVar and gnomAD over a subset of ClinVar genes. While ClinVar variants are the most relevant to the clinical goal, we include gnomAD variants to account for biases in ClinVar VUS annotations that arise out of the preferential testing of some genes over others. We refer to this combined set of ClinVar VUS and gnomAD variants as the *unlabeled set* of variants.

Let $\mathcal{G}$ be a subset of ClinVar genes filtered based on constraints related to assay feasibility and other attributes of interest (Sections 2.1, 2.2). For a gene $g \in \mathcal{G}$, let $\mathcal{P}(g)$, $\mathcal{B}(g)$ be the set of variants in g annotated as P/LP and B/LB in ClinVar, respectively. Let $\mathcal{U}(g)$ be the *unlabeled* set of variants, i.e., the combined set of ClinVar VUS and gnomAD variants for gene g. For a variant v, let $\rho(v)$ be a variant's probability of pathogenicity, estimated by explicitly calibrating a predictor's pathogenicity scores on a set of pathogenic and benign variants, i.e., $\rho(v) = p(v \text{ is pathogenic}|\ \text{REVEL}(v))$ (Section 2.3). We then define three prioritization objectives, each serving different purposes in relation to our overall goal.

(a) **Movability.** We define movability as the 'movement' of a variant from a VUS annotation to a non-VUS (P, LP, B, LB) annotation when additional functional evidence is collected. This is similar to a previous definition[2] but allows for the incorporation of prediction outputs more explicitly towards the reduction of VUS annotations. To have maximal impact on the reclassification of VUS, we aim to prioritize genes that contain the highest expected number of movable variants, i.e., the expected number of pathogenic/benign variants among a gene's unlabeled variants. Since annotating new pathogenic variants and new benign ones have different benefits, we propose two movability scores for each gene: the *movability-to-P score* and the *movability-to-B score*, and calculate them as follows:

$$\text{Move}_P(g) = \sum_{v \in \mathcal{U}(g)} \rho(v) \quad \text{and} \quad \text{Move}_B(g) = \sum_{v \in \mathcal{U}(g)} 1 - \rho(v)$$

Optimizing this objective can also benefit the objective of improving predictors (see below), as it is expected to increase the number of P/LP and B/LB variants available for training.

(b) **Correction.** We define the 'correction' of a variant's clinical annotation as the update

of an existing P/LP classification to B/LB/VUS or of an existing B/LB classification to P/LP/VUS, when additional functional evidence is collected. To have maximal impact on pathogenic or benign variants that may be currently misclassified, we want to prioritize those genes that contain the highest expected number of variants whose clinical classification ought to be corrected, i.e., the expected number of pathogenic (benign) variants among a gene's variants annotated as benign (pathogenic). Again, since there are differences in importance between correcting misclassifications of pathogenic variants and benign ones, we propose two correction scores for each gene: the *correction-of-P score* and the *correction-of-B score*, and calculate them as follows:

$$\text{Correct}_P(g) = \sum_{v \in \mathcal{P}(g)} 1 - \rho(v) \quad \text{and} \quad \text{Correct}_B(g) = \sum_{v \in \mathcal{B}(g)} \rho(v)$$

(c) **Predictor improvement.** Though not obvious, increasing the number of VUS with more certain predictions towards benignity or pathogenicity has a significant role to play in moving more VUS to a non-VUS (P, LP, B, LB) annotation. If the improvement in the prediction of a VUS is large enough, it may directly provide an additional line of evidence that may be enough to push it to a non-VUS annotation. Furthermore, an improved prediction on variants from the same gene, might make the gene more likely to be assayed by an experimentalist motivated by the movability objective defined above. The new functional evidence thus obtained would help its movement to a non-VUS annotation.

In order to increase the number of VUS with more certain predictions, the predictors themselves ought to be improved. To that end, we intend to generate more functional evidence for unlabeled variants (VUS and gnomAD variants) with uncertain predictions and we prioritize genes with high average uncertainty over their unlabeled variant set. The new functional evidence accrued on these variants would help improve the predictors, either by incorporating it as a feature while training a pathogenicity predictor or via transfer learning from function to disease domain. Note that the improvement in the predictor thus obtained is not restricted to the assayed variants, but also to other variants due to the predictor's generalization capabilities. Inspired by the entropy-based uncertainty sampling approach in the active learning literature,[16] we prioritize genes for predictor improvement based on the average entropy of prediction on a gene's unlabeled variants. Intuitively, the criterion prioritizes genes having a higher fraction of unlabeled variants with calibrated pathogenicity score close to 0.5. Formally, we define the average entropy of a gene, adjusted for the number of unlabeled variants, as

$$\text{Entropy}_{\text{adj}}(g) = \sum_{v \in \mathcal{U}(g)} \frac{-\rho(v)\log_2 \rho(v) - (1-\rho(v))\log_2(1-\rho(v))}{|\mathcal{U}(g)|} \left(1 + \lambda \frac{\log_2 |\mathcal{U}(g)|}{\log_2 |\max_{h \in \mathcal{G}} \mathcal{U}(h)|}\right)$$

In this expression, the term $\left(1 + \lambda \frac{\log_2 |\mathcal{U}(g)|}{\log_2 |\max_{h \in \mathcal{G}} \mathcal{U}(h)|}\right)$, with $\lambda \in [0, 1]$, serves as an adjustment factor that prevents genes with very small number of unlabeled variants from being prioritized. The log scale gives genes with many unlabeled variants only a small advantage. The hyperparameter λ can be further used to moderate the advantage given to genes with a large number of unlabeled variants. In this work, we choose $\lambda = 1$.

2.5. *Gene prioritization strategies and their comparison*

There are several possible strategies to prioritize genes for high-throughput functional assays. We describe a diverse set of prioritization strategies below.

(1) **Knowledge- or expert-driven.** The set of 94 assayed genes described in Section 2.1 serve as an appropriate proxy for expert-driven gene prioritization. After applying the pre-processing steps described in Section 2.2, we were left with a set of 68 genes. This set is referred to as the *assayed* set. In addition, we simulated knowledge-driven selection in a simple manner by prioritizing genes in terms of the collective knowledge that we have about them. Here, we used publication counts as reported by PubMed (`https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz`) in July 2022. We refer to this gene set as the *highest publications* set.

(2) **Data-driven.** In this strategy, knowledgebases such as ClinVar are explicitly queried and genes are prioritized based on the numbers of variants of interest observed in them. For instance, genes with a high number of VUS are of particular interest because of the challenges in classifying such variants. We constructed a gene set ranked by the highest number of unlabeled variants (VUS and gnomAD). We refer to this gene set as the *highest unlabeled variants* set. Similarly, one may be interested in genes with the most number of VUS along with P/LP variants. We also constructed a gene set ranked by the highest total of VUS and P/LP. We refer to this gene set as the *highest non-benign variants* set.

Previous work introduced two sophisticated strategies to prioritize genes for MAVEs in addition to the number of ClinVar VUS in a gene.[2] The movability- and reappearance-weighted impact score (MARWIS) incorporated patient data from Invitae to define variants' movability and reappearance and give extra weight for reappearing and movable VUS. The other score, difficulty-adjusted impact score (DAIS) was a specialized version of MARWIS that was adjusted for protein length. DAIS was deemed to be better-performing in practice and a set of 100 genes with the highest DAIS was made available to the community. After applying the pre-processing steps in Section 2.2 to this set, 94 genes remained. We refer to this gene set as the *DAIS* set.

(3) **Single score optimization.** We constructed five gene sets by directly optimizing the five scores, derived to increase movability, correction and predictor improvement (Section 2.6). For each score, we picked the top-K genes to create a gene set of length K. We refer to the resulting five gene sets as the *highest movability-to-P*, the *highest movability-to-B*, *highest correction-of-P*, *highest correction-of-B* and the *highest uncertainty* sets. As these gene sets represent the best selection for their corresponding score, no other gene set can be better w.r.t. that score.

(4) **Multiple score optimization.** In order to obtain a single gene set that improves on all three objectives simultaneously, we implemented an approach to optimize a weighted combination of the five scores. The weights are learnt to incentivize improvement over the *assayed* gene set on all five scores (Section 2.6). The resultant gene set is referred to as the *combined score* set. This gene set makes tradeoffs between the five scores depending on how well the *assayed* gene set performs on each score.

(5) **Random selection.** To create a baseline gene set of length K, we sampled K genes

randomly from the starting gene set and refer to this gene set as the *random* set.

We evaluated these different strategies by computing their score distributions in terms of $Move_P(g)$, $Move_B(g)$, $Correct_P(g)$, $Correct_B(g)$, and $Entropy_{adj}(g)$. Then, we tested whether the single score optimization strategy was significantly better than all other strategies using one-sided Wilcoxon rank-sum tests. We also tested whether the multiple score optimization strategy was better than those that were used to generate the *assayed* and *DAIS* gene sets. To ensure a fair comparison, we only compared gene sets of the same length. Since the *assayed* and *DAIS* are extant gene sets of fixed length, they determined the length constraints on the remaining gene sets. For comparisons with the *assayed* set, K was set to 68, and for those with the *DAIS* set, K was set to 94.

2.6. *Multiple score optimization*

Let $\mathcal{G}$ be a starting set of genes available to be assayed. Let $\mathcal{A} \subseteq \mathcal{G}$ (e.g., *assayed* set) be an existing gene set of length K, determined to be suitable for assaying based on some criteria. We present an approach to create a novel gene set optimized to improve over $\mathcal{A}$, w.r.t. the five scores, derived to increase movability, correction and predictor improvement (Section 2.4). Let $\boldsymbol{w} = [w_i]_{i=1}^5$ be a weight vector with five non-negative entries such that $\sum_{i=1}^5 w_i = 1$. Let S_1, S_2, S_3, S_4 and S_5 be short-hands for $\text{Move}_P, \text{Move}_B, \text{Correct}_P, \text{Correct}_B$, and $\text{Entropy}_{\text{adj}}$, respectively. We define the combined weighted score as

$$\text{Combined}_{\boldsymbol{w}}(g) = \textstyle\sum_{i=1}^5 w_i \overline{S}_i(g)$$

where $\overline{S}(g)$ denotes a score $S(g)$ after z-score normalization on the entire gene set $\mathcal{G}$. The normalization ensures that the scores are on the same scale, which in turn allows us to define an optimization criteria that treats each score equally. It also allows the weights to be on the same scale, which makes it easier to find a good solution. In order to learn the optimal $\boldsymbol{w}$, we first create a sample, W, containing 10^5 candidate weights from Dirichlet$(1, 1, 1, 1, 1)$, a uniform distribution over the space of five dimensional probability vectors. For each candidate $\boldsymbol{w} \in W$, we sort the genes in $\mathcal{G}$ in the decreasing order of $\text{Combined}_{\boldsymbol{w}}(g)$. The top K genes are picked in a candidate gene set $\mathcal{O}_{\boldsymbol{w}}^K$. For a set of numbers X, let $\text{Median}(X)$ and $\text{Prctile}_{90}(X)$ denote the median and the 90^{th} percentile of those numbers. For $G \subseteq \mathcal{G}$, let $\overline{S}_i(G)$ denote the set containing the i^{th} normalized score evaluated on genes in G. If the median or the 90^{th} percentile of any normalized score on $\mathcal{O}_{\boldsymbol{w}}^K$ is less than that on $\mathcal{A}$, then discard $\boldsymbol{w}$, i.e., for any i, if $\text{Median}\big(\overline{S}_i(\mathcal{O}_{\boldsymbol{w}}^K)\big) < \text{Median}\big(\overline{S}_i(\mathcal{A})\big)$ or $\text{Prctile}_{90}\big(\overline{S}_i(\mathcal{O}_{\boldsymbol{w}}^K)\big) < \text{Prctile}_{90}\big(\overline{S}_i(\mathcal{A})\big)$, then discard $\boldsymbol{w}$. This ensures that each remaining weight leads to a gene set with higher median and 90^{th} percentile on each of the five score distributions compared to the $\mathcal{A}$. Let W_{good} be the set of remaining candidate weights. If $W_{\text{good}} \neq \emptyset$, a $\boldsymbol{w} \in W_{\text{good}}$ is guaranteed to give a better gene set than $\mathcal{A}$ on each of the five scores. In order to select an optimum weight from W_{good}, we define the following optimization criteria to find weights that lead to largest cumulative increase in the the normalized score medians compared to $\mathcal{A}$:

$$C(\boldsymbol{w}) = \textstyle\sum_{i=1}^5 \big[\text{Median}\big(\overline{S}_i(\mathcal{O}_{\boldsymbol{w}}^K)\big) - \text{Median}\big(\overline{S}_i(\mathcal{A})\big)\big].$$

The optimum weights are given by $\boldsymbol{w}_{\text{opt}} = \text{argmax}_{\boldsymbol{w} \in W_{\text{good}}} C(\boldsymbol{w})$. The corresponding gene set, $\mathcal{O}_{\boldsymbol{w}_{\text{opt}}}^K$ is the optimal gene set, referred to as the *combined score* set. Note that if a gene set

of a different size, $K_1 \neq K$, is needed, the top K_1 genes sorted based on Combined$_{\boldsymbol{w}_{\text{opt}}}(g)$ are selected. The resultant set is referred to as $\mathcal{O}^{K_1}_{\boldsymbol{w}_{\text{opt}}}$.

2.7. *Functional and phenotypic enrichment analyses*

To evaluate the biological and clinical relevance of the multiple score optimization strategy, we ranked all genes by their *combined score* and conducted a functional enrichment analysis on the top 100 genes using the *g:GOSt* function in the gProfiler web-server.[17] We used our starting gene set of 3,829 genes as the background set. Any Gene Ontology (GO) and Human Phenotype (HP) Ontology terms that were significantly enriched in the top 100 genes, after correcting for multiple hypothesis testing (P-value < 0.05) were recorded.

3. Results

3.1. *Multiple score optimization outperforms knowledge-driven and simple data-driven strategies*

We compared multiple gene sets (see Section 2.5), constructed through diverse prioritization strategies on the five scores, covering the three clinical objectives: movability, correction and

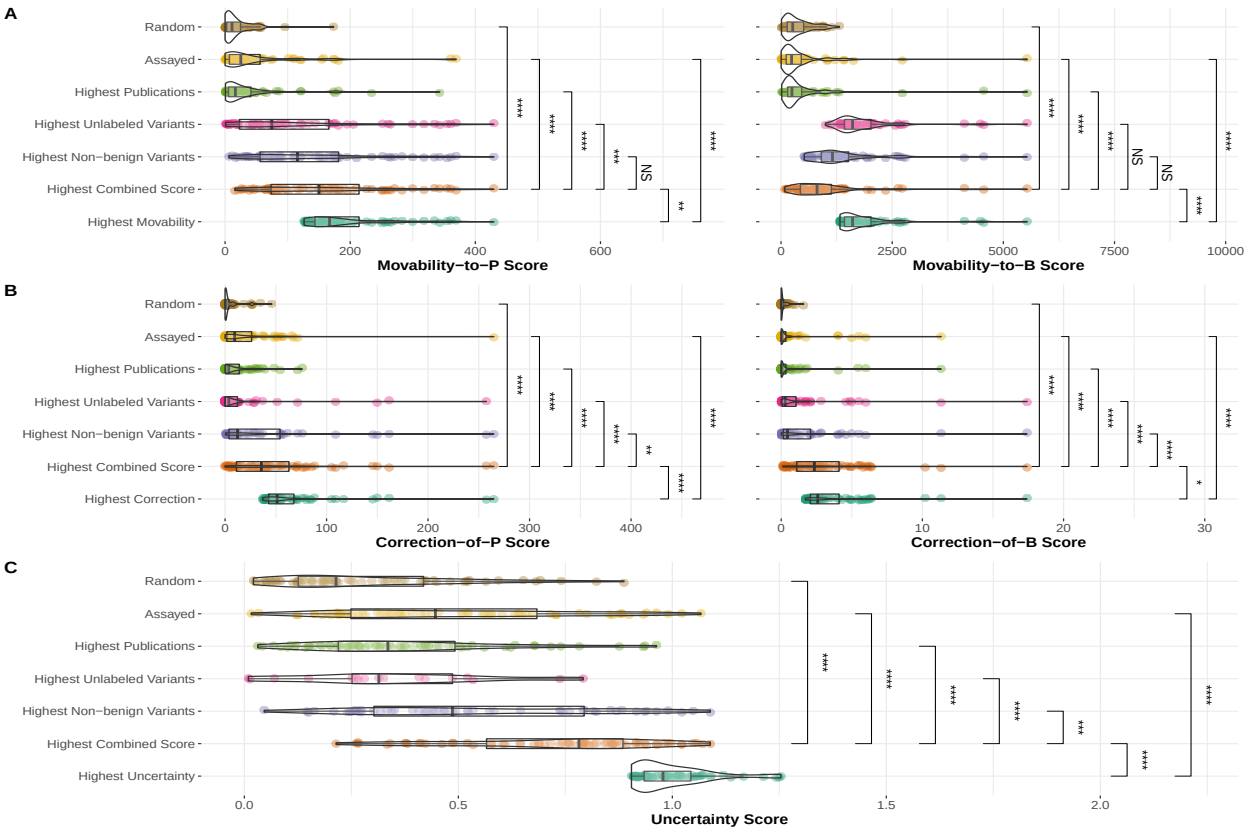

Fig. 1. **Score distributions 68-gene sets constructed based on seven prioritization strategies. A.** Score distribution of movability to pathogenic (left) and benign (right), **B.** Score distribution of correction of pathogenic (left) and benign (right) variants, **C.** Uncertainty score distribution.

predictor improvement (Figure 1). All the sets in this comparison had 68 genes, to be consistent with the *assayed* set. Unsurprisingly, for any given score, the *highest single score* gene set, being the best set for the score, outperformed all other gene sets. As expected, the *combined score* set performed better than the *assayed* gene set because it was explicitly constructed to improve over the *assayed* set. Overall, the *combined score* set performed better than all other gene sets except the respective highest single score sets. There were two exceptions to this. In the case of *movability-to-B* score, the *combined score* set did not perform better than the *highest unlabeled variants* and *highest non-benign variants* gene sets, suggesting that the number of unlabeled variants may be a strong determinant of *movability-to-B* due to the high prior probability of benignity in general. In particular, the scope of improvement in *movability-to-B* score over the *highest unlabeled variants* set is limited as can be observed in comparison to *highest movability-to-B* set, the best possible set for that score. Furthermore, among all comparisons of the *combined score* where it performs better, it does so with statistical significance, except in one case: comparison with *highest non-benign variants* set on *movability-to-P* score.

The *assayed* set performed slightly better than *random* on most scores. Moreover, its score distributions were far away from that of the corresponding *highest single score* set. This suggests that there is a huge scope of improvement on the set of genes currently being assayed, with respect to clinical objectives. On all score criteria, the performance of the *highest publication* set is quite similar to that of the *assayed* set. This is consistent with the previous observation that genes with fewer publications are less likely to be functionally tested.[9]

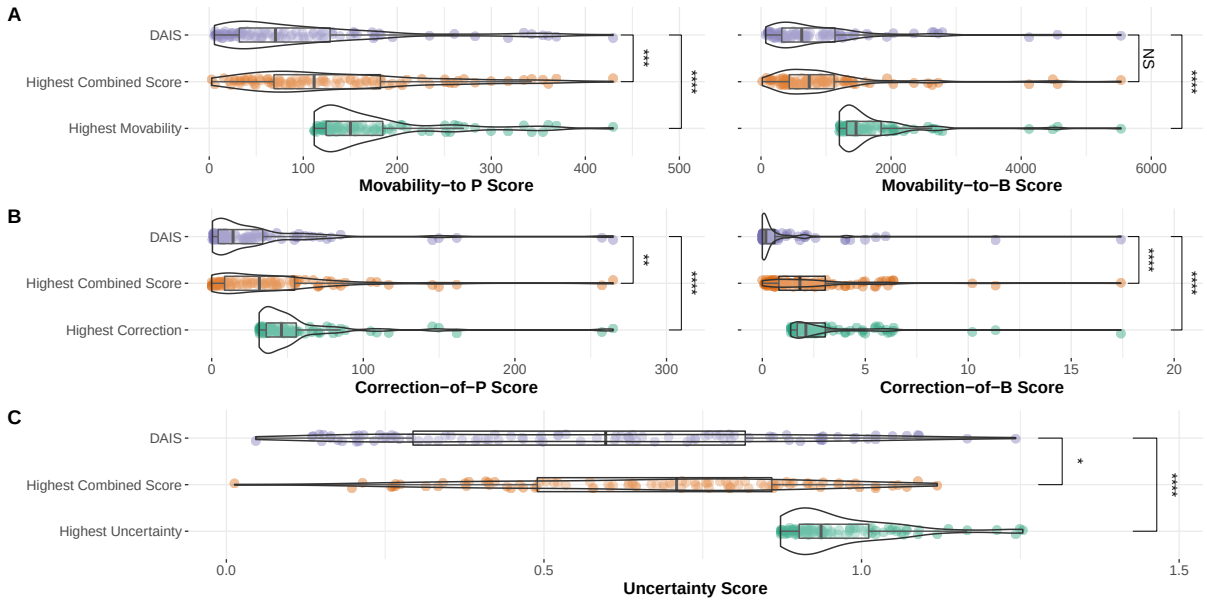

Fig. 2. **Score distributions for top 94 genes prioritized by our proposed strategies and by existing data-driven strategies. A.** Score distribution of movability to pathogenic (left) and benign (right), **B.** Score distribution of correction of pathogenic (left) and benign (right) variants, **C.** Uncertainty score distribution. DAIS, 94 genes out of the top 100 genes ranked by the difficulty-adjusted impact score.[2]

3.2. *Multiple score optimization outperforms existing clinically motivated prioritization strategies*

We next compared our single and multiple score optimization strategies to a previously proposed strategy that explicitly aimed to improve clinical variant classification, DAIS[2] (Figure 2). Since the DAIS set comprised of 94 genes, we considered the top 94 genes with the highest single and combined scores. The single and multiple score optimization strategies yielded statistically significant improvements over DAIS in all situations, with one exception. When considering the *movability-to-B* score, the *combined score* set showed improvement over DAIS, although not significantly, similar to our observations in Section 3.1.

3.3. *Multiple score optimization yields clinically relevant genes*

We characterized the properties of the highest-scoring genes in the *combined score* set and investigated to what extent our strategy aligned with biomedical interests. Among the top 20 genes, six genes were in our *assayed* gene set, and 12 genes were also prioritized by DAIS, albeit with differences in ranking (Table 1). All identified genes generally have a large number of variants recorded in ClinVar and gnomAD, with the exception of *SCN10A*, which has no

Table 1. Missense variant counts and scores for the top 20 genes from the *combined score* gene set. Similar counts and scores are available for all genes in this study here: `https://igvfgenecard.shinyapps.io/GeneCardApp/` Genes in bold were also present in the *assayed* set. The Movability and Correction scores are rounded to the closest integer. The Combined score is given as the weighted sum of the five scores after z-score normalization. The weights for *movability-to-P*, *movability-to-B*, *correction-of-P*, *correction-of-B*, and *uncertainty* were 0.143, 0.160, 0.380, 0.310, and 0.006, respectively.

		DAIS	ClinVar					Movability		Correction		Entropy	
Rank	Gene	rank	P/LP	B/LB	VUS	gnomAD	Total	to P	to B	of P	of B	adjusted	Combined
1	*TSC2*	32	80	185	2178	273	2716	318	2035	29	17	0.8	13.3
2	***BRCA1***	10	120	206	2817	160	3303	181	2727	71	11	0.5	10.5
3	***LDLR***	40	635	62	564	176	1437	155	547	265	4	0.9	10.1
4	*FBN1*	39	873	17	1338	536	2764	335	1451	257	2	0.9	9.9
5	***BRCA2***	9	57	236	5453	325	6071	173	5533	37	6	0.3	7.5
6	*IDS*	1055	120	57	49	125	351	32	134	39	10	0.7	7.0
7	*MYH7*	2	271	17	1284	297	1869	355	1129	150	2	1.1	6.7
8	*SCN1A*	66	452	39	670	361	1522	283	683	146	3	1.0	6.6
9	*NF1*	11	232	19	2750	224	3225	261	2632	162	0	0.5	6.4
10	***MSH2***	4	73	26	1757	123	1979	369	1409	28	6	1.0	5.9
11	*COL4A5*	1839	414	87	66	372	939	72	347	80	6	0.7	5.6
12	*SCN8A*	468	122	44	346	250	762	125	438	43	6	0.9	5.3
13	***SCN5A***	63	83	33	1058	386	1560	361	998	23	5	1.0	5.3
14	*MLH1*	8	122	33	1103	80	1338	175	957	62	4	0.8	5.0
15	*SCN10A*	391	0	55	381	831	1267	226	930	0	6	0.8	4.8
16	*FLNA*	211	32	85	560	493	1170	150	858	16	6	0.8	4.7
17	*CACNA1S*	323	12	44	393	777	1226	251	858	3	6	0.9	4.7
18	*FBN2*	155	33	58	708	1005	1804	300	1331	13	5	0.9	4.6
19	***TP53***	1	143	76	717	27	963	176	525	54	4	1.0	4.5
20	*ABCA4*	130	235	17	582	845	1679	252	1110	109	0	0.8	4.3

variants classified as pathogenic or likely pathogenic. In addition, our *combined score* also prioritized important genes that may have been overlooked previously. For example, IDS, which has more than 200 IDS variants were found in Hunter syndrome patients[18] was ranked

6^{th}. *COL4A5*, with over 400 variants thata cause Alport syndrome, was (ranked 11^{th}). Many sodium voltage-gated channels (*SCN*)-related genes were also ranked within the top 20, and mutations in these genes can lead to channel defects and cause channelopathies.[19] Since the objective of improving predictors may not necessarily yield genes that are clinically relevant, we systematically explored the functional and phenotypic characteristics associated with the *combined score* set. We conducted an enrichment analysis on the top 100 genes ranked by their combined score and reported significantly enriched GO terms and the 40 most significant HP terms (Supp. Figure 1A). This top-100 gene set was enriched in many biological processes such as neuronal action, membrane depolarization, and molecular functions such as multiple channel activities and transmembrane transporter activity. From the phenotypic perspective, enriched high-level HP terms included abnormalities of different organ systems such as skin, gastrointestinal tract, nervous system, among others (Supp. Figure 1B). More specific HP terms included cardiovascular related disease, limitation of mobility, and stroke, among others.

4. Discussion

Genetic and genomic testing are now routinely used in healthcare systems to provide diagnoses and infer lifetime risk for disease symptoms, particularly in the identification of hereditary susceptibility to cancer, metabolic conditions, intellectual and physical developmental disorders, among others. The classification of genetic variants detected in a patient's gene panel or genome is a key step in this context. In this regard, our study presented three objectives that explicitly captured the goal of improving clinical classification of variants and derived five scores to operationalize them. We derived an optimal gene set for each score and also derived a *combined score* gene set by optimizing a weighted combination of the five scores to explicitly improve over the existing *assayed* set.

As expected, all single score optimization strategies, led to the best performance on the corresponding score. More importantly, evaluating the existing approaches relative to the single score optimization, demonstrated a considerable performance gap, suggesting a significant scope of improvement on each objective. Even though our *combined score* gene set was obtained by optimizing directly over the three objectives relative to the *assayed* set, its observed improvement over the *assayed* and DAIS gene sets on all scores is not entirely obvious due to the inherent trade-offs between the objectives (movability vs. predictor improvement). This is a further testament to the scope of simultaneous improvement on all objectives along with an approach that demonstrably does so.

DAIS, a more sophisticated strategy, presented higher scores in general but did not outperform our approach. Unlike DAIS, our approach does not use any proprietary patient data, but despite this, one-third of our genes overlapped with the DAIS set. Our approach can be potentially complementary to DAIS, since we accounted for conflicting variants, incorporated non-VUS and less biased gnomAD variants and focused on correction and predictor improvement as objectives. Another strength of our strategy is its interpretability. The *movability* scores and *correction* scores are interpreted as the expected number of pathogenic or benign variants, and the *uncertainty* score as predictive uncertainty. In addition, our approach for multiple score optimization could be easily extended to incorporate other scores such as DAIS,

if appropriate data were available, or could directly optimize the combined score to improve over both the assayed and DAIS sets.

Though our movability objective quantifies the expected number unlabeled variants in a gene that are pathogenic (or benign), it is possible that after running a given assay the number of variants moved to the P/LP (B/LB) categories as per the ACMG/AMP guidelines might differ. This might happen either because the assay might not capture the functional mechanism that leads to the disease, or the strength of the new evidence combined with existing evidence might not be enough to move the variant. Without functional assay outcomes, this is difficult to discern and is a limitation of our study. In future, when additional information on an assay's relevance to specific diseases is available, refined criteria that take that information into account might better quantify the movement. Similarly, if all existing evidence for a variant is accessible, the criteria may be refined to take it into account, as done so by Kuang et al.[2] Our study is currently limited in this regard, as ClinVar does not detail which specific lines of evidence were used to classify a variant. Similar considerations apply to the correction scores as well.

In conclusion, we defined three objectives in terms of improving clinical classification by using variant pathogenicity predictors. Our final *combined* scores provided a list of prioritized genes for MAVEs but this list will keep updating with iterated future work between prediction and experimentation. All data sets, analysis scripts, and supplementary results for this study can be accessed here: `https://github.com/strongbeamsprout/Gene-Prioritization`.

5. Acknowledgments

This work was supported by National Institutes of Health grant U01 HG012022.

References

1. S. Richards et al. *Genet. Med.*, 17(5):405–424, 2015.
2. D. Kuang et al. . *Bioinformatics*, 36(22-23):5448–5455, 2020.
3. S. M. Harrison and H. L. Rehm. *Genome Med.*, 11(1):72, 2019.
4. D. M. Fowler and S. Fields. *Nat. Methods*, 11(8):801–807, 2014.
5. J. Weile and F. P. Roth. *Hum. Genet.*, 137(9):665–678, 2018.
6. P. A. Romero et al. *Proc. Natl. Acad. Sci. U.S.A*, 112(23):7159–7164, 2015.
7. X. Jia et al. *Am. J. Hum. Genet.*, 108(1):163–175, 2021.
8. K. A. Matreyek et al. *Nat. Genet.*, 50(6):874–882, 2018.
9. T. Stoeger et al. *PLoS Biol.*, 16(9):e2006643, 2018.
10. M. J. Landrum et al. *Nucleic Acids Res.*, 46(D1):D1062–D1067, 2018.
11. N. M. Ioannidis et al. *Am. J. Hum. Genet.*, 99(4):877–885, 2016.
12. V. Pejaver et al. *bioRxiv*, 10.1101/2022.03.17.484479, 2022.
13. K. J. Karczewski et al. *Nature*, 581(7809):434–443, 2020.
14. D. Esposito et al. *Genome Biol.*, 20(1):223, 2019.
15. D. Kuang et al. *Bioinformatics*, 37(19):3382–3383, 2021.
16. B. Settles. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012.
17. U. Raudvere et al. . *Nucleic Acids Res.*, 47(W1):W191–W198, 2019.
18. V. Ricci et al. *Am. J. Med. Genet. A*, 120A(1):84–87, 2003.
19. M. de Lera Ruiz and R. L. Kraus. *J. Med. Chem.*, 58(18):7093–7118, 2015.