TITLE

Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of diverse feature set

AUTHORS

David Stein[1,2], Çiğdem Sevim Bayrak[3], Yiming Wu[3], Peter D. Stenson[4], David N. Cooper[4], Avner Schlessinger[1,*], Yuval Itan[2,3,*]

AFFILIATIONS

[1]Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029

[2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029

[3]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[4]Institute of Medical Genetics, Cardiff University, Cardiff CF14 4XN, UK

*Corresponding Author

ABSTRACT

Gain-of-function (GOF) variants yield increased or novel protein function while loss-of-function (LOF) variants yield diminished protein function. GOF and LOF variants can result in markedly varying phenotypes even when occurring in the same gene. Experimental approaches for identifying GOF and LOF are slow and costly, and computational tools cannot accurately discriminate between GOF and LOF variants. We developed LoGoFunc, an ensemble machine learning method for predicting pathogenic GOF, pathogenic LOF and neutral variants, trained on 672 protein, gene, variant, and network annotations describing diverse biological characteristics. We analyzed features in terms of protein structure and function, splice disruption, and phenotypic associations, revealing previously unreported relationships between various biological phenomena and variant functional outcomes. For example, GOF and LOF variants demonstrate contrasting enrichments in protein structural and functional regions, and LOF variants are more likely to disrupt canonical splicing as indicated by splicing-related features employed by the model. Furthermore, by performing phenome-wide association studies (PheWAS), we identified strong associations between relevant phenotypes and high-confidence predicted GOF and LOF variants. LoGoFunc performs well on an independent test set of GOF and LOF variants, and we provide precomputed genome-wide GOF and LOF predictions for 71,322,505 missense variants.

MAIN

It is essential to investigate the functional consequences of genetic variations to understand the diverse array of human disease phenotypes. Genetic variations exert diverse functional effects on gene products and can impact a protein's stability and interactions with binding partners, its catalytic activity, and many other properties[1]. In particular, the functional consequences of genetic variations include two broad categories: gain-of-function (GOF) variants, characterized by enhanced or novel protein activity, and loss-of-function (LOF) variants which result in partial or complete knockdown of protein activity[2]. GOF and LOF variants are of particular interest because they can give rise to distinct phenotypes in the same gene via contrasting molecular mechanisms[3]. Thus, for example, GOF mutations in STAT1 cause Chronic mucocutaneous candidiasis (CMC) - a susceptibility to candida

infection of the skin, nails and mucous membranes[3]. By contrast, LOF variants in STAT1 result in Mendelian Susceptibility to Mycobacterial Disease (MSMD) - an immunodeficiency characterized by vulnerability to weakly virulent mycobacteria[3]. Given the established heterogeneity in phenotypic outcomes and their diverse modes of action, it is necessary to distinguish GOF and LOF variants to develop a greater understanding of the genetic mechanisms of human disease, estimate individual genetic disease risk, identify candidate drug targets, and construct effective treatment regimens.

To date, effective and practical methods for determining GOF and LOF variants are lacking. Experimental techniques are capable of accurately detecting GOF and LOF variants, but these methods are constrained by their significant cost and low throughput[4]. Rapid computational methods for assessing various aspects of variants such as pathogenicity or protein impact have been developed[5–9]. For example, CADD[5] leverages a range of functional annotations and conservation metrics to rank the relative deleteriousness of variants. PolyPhen-2[6] and SIFT[7] combine physical characteristics of proteins with evolutionary features such as sequence conservation to predict if a variant will impact protein structure or function. Other tools such as REVEL[8] and BayesDel[9] combine the outputs of other predictors to produce a meta-score indicating variant pathogenicity. Yet, none of these tools is tailored for GOF and LOF classification and, hence, they cannot accurately discriminate between GOF and LOF variants.

Here we present LoGoFunc, the first effective predictor of GOF, LOF, and neutral variants, and a characterization of missense GOF and LOF variants genome-wide. LoGoFunc is a machine learning model composed of an ensemble of LightGBM[10] classifiers trained on experimentally labeled GOF and LOF variants. We first assessed LoGoFunc's performance and demonstrated that LoGoFunc outperforms available computational tools in its prediction of GOF, LOF, and neutral variants. Next, we described the curation of 672 descriptors derived from structural and functional properties of proteins, pathogenicity scores, and measures of evolutionary constraint and conservation, to construct a predictor for variants across the genome, and we introduced highly impactful network-embedding features previously not used in popular variant assessment algorithms. In conjunction, we analyzed GOF and LOF variants for enrichments by particular features, highlighting structural and functional features of proteins as well as features related to disease mechanisms such as splice disruption. Finally,

we identified relationships between high confidence, predicted GOF and LOF variants and patient phenotypes. LoGoFunc's speed allowed us to provide precomputed GOF, LOF, and neutral predictions for 71,322,505 missense mutations in the human genome, which are freely available for rapid retrieval and analysis at https://itanlab.shinyapps.io/goflof/.

RESULTS

Training, architecture, and performance of LoGoFunc

LoGoFunc was trained on a dataset[11] of experimentally identified GOF and LOF variants, collected via a natural language processing (NLP) pipeline[11]. In brief, our NLP pipeline parses abstracts associated with high-confidence, disease-causing variants in the Human Gene Mutation Database (HGMD)[12] Professional version 2021.3, searching for terminology denoting GOF and LOF and related nomenclature (Figure 1a). In total, 1,492 GOF and 13,524 LOF mutations were collected and labeled, in addition to 20,000 putatively neutral variants randomly selected from gnomAD v2.1[13] exome and genome sequences (Figure 1b). Variants were annotated with the 672 variant-, gene-, protein-, and network-level features (Supplementary Table 2) which were collected and preprocessed (Supplementary Information) for use in the classifier before splitting the data into training and testing sets comprising 90% and 10% of the full dataset, respectively (Figure 1b). The dataset split was stratified such that the ratio of GOF to LOF to neutral variants reflected the ratio in the complete dataset. Finally, features were encoded, imputed, and scaled, fitting each transformation to the training set and transforming on both sets (Figure 1b).

Four machine learning algorithms, artificial neural networks (ANN), RandomForest[14], XGBoost[15], and LightGBM[10] were tested on the data. LightGBM[10] was selected as the best performer based on the 5-fold cross-validated, macro-averaged F1-score after hyperparameter tuning (Supplementary Figure 1). The final LoGoFunc model is composed of 15 LightGBM[10] classifiers, and learned signal discriminating GOF, LOF, and neutral variants. The model accepts an array of 805 encoded, imputed, and scaled features, and outputs three values corresponding to the predicted probability that the input variant results in a GOF, LOF, or neutral phenotype respectively (Figure 1c).

LoGoFunc achieved notable success in classifying GOF, LOF, and neutral variants with respective areas under the Receiver-operating characteristic (ROC-AUC) of .99, .99, and .99. Considering the class imbalance in the dataset, we also calculated the average precisions (AP), resulting in values .87, .98, and .99 calculated on the held-out testing data (Supplementary Figure 2a), as well as via the F1-score and Cohen's kappa[16]. LoGoFunc realized F1-scores of .78, .92, and .96 for GOF, LOF, and neutral variants and Cohen's kappa[16] scores of .77, .87, and .90 respectively (Supplementary Figure 2c). The set of all GOF and LOF variants has a high degree of concordance with the set of pathogenic variants. Thus, we assessed the predictor's ability to discriminate pathogenic and neutral variants generally (Supplementary Figure 3). LoGoFunc achieved a Cohen's kappa[16] of .75 and an F1-score of .91 on a set of 23,980 randomly selected putatively neutral single nucleotide variants (SNV) from gnomAD v2.1[13] and 91,599 high-confidence pathogenic mutations from HGMD[12] Professional version 2021.3 that had not been used in LoGoFunc's training. To assess the model's performance, we employed a cutoff of 0.5 to convert probabilities to class labels. We also calculated 95% confidence intervals for determining cutoffs for each class, as well as 95% confidence intervals for determining GOF, LOF, and neutral prediction cutoffs per gene (Supplementary Table 1).

Benchmark against variant assessment algorithms

Currently, there are no high-throughput, computational predictors for GOF and LOF variant classification[11]. We therefore compared LoGoFunc to seven established predictors of pathogenicity/deleteriousness: PolyPhen-2[6], MetaSVM[17], MutPred[18], SIFT[7], DEOGEN2[19], PrimateAI[20], and CADD[5]. To assess the methods' ability to discriminate GOF and LOF across tools, we selected the subset of 803 GOF, LOF, and neutral variants from the test set for which all predictors provided a score (Figure 2). Of these variants, 86 were GOF, 328 were LOF, and 389 were neutral. We tested each method's performance in separating GOF from neutral, LOF from neutral, both GOF and LOF combined from neutral, and GOF from LOF variants. LoGoFunc achieved an AP of .983 for GOF *vs*. neutral (Figure 2a) and .975 for LOF *vs*. neutral (Figure 2b). The next best tool, DEOGEN2[19], achieved AP values of .726 and .882 for GOF and LOF respectively (Figure 2a,b). We calculated the one-*vs*.-all AP for the neutral variants against the GOF and LOF variants. Again, LoGoFunc scored highest with an AP of .991, followed by DEOGEN2[19] with an AP of .903 (Figure 2c). We also compared the

model's ability to separate GOF from LOF variants. LoGoFunc achieved an AP of .9 followed again by DEOGEN2[19] with a score of .210 (Figure 2d). LoGoFunc outperformed the other models for each comparison, particularly for the separation of GOF and LOF variants from each other, indicating that it is the most effective method for identifying protein gain- and loss-of-function as a result of genetic variation. To estimate LoGoFunc's performance on genes with scarce variant functional annotation, we repeated the hyperparameter tuning protocol described above to construct a new LightGBM ensemble using training and testing sets that did not contain overlapping genes. We found that our method continued to outperform the general pathogenicity predictors (Supplementary Figure 4), indicating that LoGoFunc is likely to be the most effective computational method for assessing variant functional outcome irrespective of the functional annotations available for the gene's variants in the literature.

LoGoFunc leverages diverse biological signals for prediction

To interpret the LoGoFunc classifier's predictions, we leveraged the measure of feature importance native to the LightGBM[10] model, i.e. the frequency that a feature is used to split a node in the trees (Figure 3c). Because LoGoFunc is a voting ensemble composed of weighted LightGBM[10] models, we computed the weighted average of the number of splits for each feature across each sub-model. We observed that LoGoFunc learned from a diverse array of gene, protein, and variant level features including functional, conservation, structural, systems-based, and network information (Supplementary Table 3, Figure 3a).
For example, the top feature is the Mutation Significance Cutoff[21] (MSC) 95% confidence interval (CI) which aims to identify pathogenic variants with high recall by identifying the lowest biologically relevant CADD[5] cutoff for protein-coding genes. It is complemented by the rank score of VEST4[22] scores from dbNSFP[23] as well as PHRED-scaled and raw CADD[5] values. The most important variant-level annotations are GERP[24] conservation scores and the MOI-pred[25] mode of inheritance prediction of variants pathogenic for autosomal dominant (AD) and autosomal recessive (AR) disease. The model also learned from protein sequence and structural information, including top features amino acid sequence complexity (measured by Clark's distance[26]), relative and absolute solvent accessibility (RSA and ASA, respectively), and predicted binding sites in disordered regions (predicted by ANCHOR2[27]). Notably, protein-protein interaction (PPI) network features also had a significant impact on the model. We processed the STRING[28] protein-protein interaction (PPI) network using node2vec[29]

resulting in 64 tabular features summarizing the human protein interactome weighted by the probability of interaction between each pair of putatively interacting proteins. Several dimensions of the transformed PPI network appeared in the list of top features and, assessed cumulatively, the dimensions of the PPI have the strongest influence on the model as measured by node-splitting frequency (Figure 3a).

Based on our initial observations, we hypothesized that structural and functional features of proteins may help to stratify GOF, LOF, and neutral variants. We further examined protein features by calculating enrichments by class, determined via Fisher's exact test with continuous variables categorized according to their respective recommended cutoff values. Interestingly, GOF and LOF variants were significantly more likely to be found in Pfam or Interpro protein domains than neutral variants with log-odds (LO) of 1.026, 0.904, and -1.068 respectively (Figure 3b). Similarly, GOF and LOF were significantly enriched while neutral variants were depleted in post-translationally modified (PTM) residues and α-helices (Figure 3b). We found this trend to be reversed for molecular recognition features (MoRFs), regions of low sequence complexity, cleavage sites, disordered regions, and disordered binding residues (Figure 3b). When comparing all three classes, GOF and LOF demonstrated opposing trends of enrichment and depletion in post-translational modification recognition domains (Figure 3b). However, when comparing GOF and LOF variants only, the classes also demonstrated opposing trends of enrichment and depletion in cellular localization domains, weakly and moderately stabilizing residues, zinc fingers, Pfam or Interpro domains, α-helices, and transmembrane helices (Supplementary Figure 5). These results indicate potentially contrasting molecular mechanisms for pathogenic GOF and LOF, and for neutral variants at the protein level, thereby identifying future avenues for investigation of disease etiology. Moreover, these results support the presence of signal in our feature set and identify avenues by which the model may reach conclusions concerning the functional outcome of predicted variants.

Considering the stratification of GOF, LOF, and neutral variants by protein structural and functional features, we examined the localization of predicted variants on a three-dimensional structure of the CASR protein. CASR was selected because (i) it is a functionally important calcium-sensing receptor associated with neonatal severe hyperparathyroidism[30] and primary

hypercalciuria[31] among other disorders[30]; (ii) both GOF and LOF pathogenic variants have been reported in CASR[11]; and (iii) 3D structures[32] are available in the PDB[33] (Figure 3c). We highlighted six regions of the CASR structure grouped by protein family or domain – the ligand-binding domains 1 and 2 (LB1, LB2), the cysteine-rich region (CR), and three unstructured disordered regions (D1, D2, D3) (Figure 3c). Interestingly, results from this analysis indicated that the predicted GOF, LOF, and neutral variants are localized in distinct protein structural regions, as expected from our initial hypothesis. We observe that the model's GOF predictions are mildly enriched (p=1.124E-2, LO=0.185) in LB1 while neutral is depleted (p=1.433E-7, LO=-1.182). LB2 is significantly enriched for LOF (p=6.301E-25, LO=0.916) and depleted for GOF (p=2.954E-20, LO=-0.859) and neutral (p=4.031E-4, LO=-0.825). The CR region is also significantly enriched for LOF (p=2.110E-18, LO=1.123) and depleted for GOF (p=8.636E-18, LO=-1.184); furthermore, 29 pathogenic LOF variants have been reported in the LB2 and CR regions[11]. Conversely, D1 is significantly enriched for GOF (p=1.908E-48, LO=1.670) and neutral (p=6.825E-5, LO=0.872), and D2 is significantly enriched for GOF (p=1.014E-22, LO=1.828). Both D1 and D2 are depleted for LOF (p=6.912E-59,1.204E-23, LO=-1.890,-1.908), while D3 is enriched for neutral variants (p=3.203E-54, LO=3.853) and depleted for LOF (p=7.320E-24, LO=-2.444).

GOF, LOF, and neutral variant effects on splicing

Splice disrupting variants have been reported to constitute the second-largest class of known disease driving mutations, and have been demonstrated to yield both GOF and LOF phenotypes[34,35]. Given the importance of splice disruption as a general causal disease mechanism, as well as the high impact of various splicing relevant features to the model such as proximity to an exon-intron junction boundary, MMSplice[36] percent spliced-in (Ψ) predictions, and SpliceAI[37] predictions of cryptic splicing effects (Supplementary Table 3), we further investigated the distribution of these features among the classes. Interestingly, LOF variants clustered most closely to exon-intron junction boundaries followed by GOF and neutral variants (p-values 0.0009, 2.5193E-17, 6.3104E-129) (Figure 4a, 4d). Furthermore, LOF variants were significantly enriched for loss of cryptic splice acceptors and donors (p-values 8.5064E-5, 1.4044E-3), which are potentially important mechanisms of alternative splicing (Figure 4a). Notably, neutral variants were enriched for the gain of splice acceptors, consistent with their greater tendency to appear in proximity to the 3' end of introns than GOF

or LOF variants in the dataset (Figure 4a). The enrichment of neutral variants for the gain of cryptic splice sites (CSS) may be explained by the ability of canonical splice sites to suppress CSS activation[38]. Thus, these CSS gained via neutral mutations may not have a significant impact on transcript expression. After removing variants not predicted to impact splicing, LOF variants were predicted to lead to a greater decrease in Ψ than GOF or neutral variants based on estimates from the MMSplice[36] exon, donor, and acceptor predictors (Figure 4c). GOF variants lead to a greater decrease in Ψ than neutral variants based on the same models (Figure 4c). These results indicate that in particular LOF variants, but also GOF variants, may at times exert their pathogenic effect via disruption of canonical splicing patterns.

PheWAS corroborates LoGoFunc predictions.

We performed phenome-wide association study (PheWAS) analyses on ten high confidence predicted GOF and LOF missense variants which were either absent from or indicated as variants of uncertain significance (VUS) in ClinVar[39] (Supplementary Table 4). In brief, PheWAS evaluates the association between a genetic variant and a set of phenotypes. Although our analysis was insufficiently powered for phenome-wide significance, as expected due to the low frequency of our variants, we nevertheless uncovered meaningful associations between our variants and relevant phenotypes (Figure 5a). For example, we observed that the predicted LOF variant c.1648G>A (rs563131364) in the *SLC12A3* gene that encodes for a sodium-chloride co-transporter is strongly associated with increased risk for several phenotypes including severe chronic kidney disease (p=0.001, LO=1.723, ICD=N184), abnormal blood chemistry (p=0.003, LO=1.189, ICD=R7989), and retinal edema (p=0.006, LO=2.377, ICD=H3581), among several other conditions. Conversely, the mutation was found to be protective for pure hypercholesterolemia (p=.0396, LO=-1.512, ICD=E7800). The c.1648G>A variant has a CADD[5] PHRED score of 29.3 and an MSC[21] 95 CI of 13.89, indicating that the variant may be pathogenic taking into consideration the genic context of *SLC12A3* (Figure 5b). c.7471C>T (rs201746476) is a high confidence predicted GOF variant in the *PIEZO1* gene, which encodes for a mechanosensitive ion channel and has previously been linked to arrhythmia when overexpressed[40]. c.7471C>T was associated with risk for palpitations (p=0.009, LO=2.912, ICD=R002), abdominal pain (p=0.028, LO=1.842, ICD=R109), and viral hepatitis C (p=0.039,0.043, LO=2.438, 2.389, ICD=B182, B1920). Notably, GOF mutations in *PIEZO1* have been demonstrated to impair hepatic iron

metabolism[41], a mechanism of viral hepatitis C[42]. Similar to c.1648G>A, c.7471C>T had a CADD[5] PHRED score of 25.4, well over the MSC[21] 95 CI of 2.185 for the *PIEZO1* gene.

DISCUSSION

Describing the functional consequences of genetic variations is critical for the understanding of disease mechanisms. We have developed LoGoFunc, a rapid and accurate predictor of GOF, LOF, and neutral variants, and analyzed various features associated with these variant types. Four key findings emerge from this work.

First, LoGoFunc demonstrated strong performance on an independent test set of GOF, LOF, and neutral variants, and consistently outperformed generic pathogenicity predictors in the classification of GOF and LOF variants. Interestingly, the tools against which we compared LoGoFunc performed better on LOF variants than GOF. It has been demonstrated that several pathogenicity predictors such as CADD[5] and REVEL[8] tend to more often predict LOF variants as pathogenic or deleterious than GOF variants, whereas GOF variants are more often predicted to be benign[11]. This may be due in part to the underrepresentation of GOF variants in the training data used by these tools where applicable or may arise because GOF variants may be difficult to separate from neutral variants based on the features or methods employed by these tools. Importantly, these results suggest that LoGoFunc may be particularly useful for predicting GOF variants, and also for identifying false-negative predictions of other pathogenicity predictors.

Second, our analysis identified previously undocumented features that are associated with the functional outcomes of genetic variants. We assessed the importance of the features used to train LoGoFunc and found that the model learns from a diverse array of gene, protein, and variant level features including functional, conservation, structural, and network information (Figure 3a). For example, we processed the STRING[28] protein-protein interaction (PPI) network using node2vec[29] to summarize the human protein interactome. Whereas some models have included binary indications of the involvement of a protein in any protein interaction[43], to our knowledge, such PPI network features are rarely used in popular variant

assessment methods. Yet, many dimensions of the output are highly impactful for the LoGoFunc model, suggesting protein function at the pathway- and/or systems-level may have a significant bearing on variant pathogenicity and functional effect. Concordantly, PPI features are accompanied by several other protein sequence and structural features from which the model also learns, including top features sequence complexity and amino acid solvent accessibility. Genic context also has a substantial impact on the model's output as evidenced by the inclusion of several gene-level features such as s_{het}[44], the gene damage index[26], and the number of gene paralogs. Indeed, the most important feature for the model, the MSC[21] 95 CI, indicates that a gene-specific pathogenicity cutoff is valuable in the assessment of variant pathogenicity and functional outcome. Other important features such as the per variant predictions of pathogenicity for autosomal dominant or recessive disease, align with previous characterizations of GOF and LOF variants, supporting the biological plausibility of the predictions and lending credence to novel associations we identified between various features employed by the model and GOF, LOF, and neutral variants.

Third, given the stratification of GOF, LOF, and neutral variants by features derived from protein sequence, we demonstrate the plausible biological basis for LoGoFunc's predictions by highlighting their localization in a three-dimensional structure of CASR and suggesting a molecular mechanism for their disruptive effect. We found that predicted LOF variants were enriched in two structural domains, the cysteine-rich region (CR) and ligand-binding domain two (LB2). Notably, the CR and transmembrane regions of the CASR dimer are brought into close contact by the motion of the LB2 region when forming the active conformation[32]. Disruption of this contact by variants in the LB2 or CR regions suggests a mechanism for LOF, concordant with our predictions. Furthermore, LOF enrichment in the CR is consistent with the previously reported[11] enrichment of LOF variants involving cysteine residues which can act as stabilizers via the formation of disulfide bonds. We also noted that disordered regions D1 and D2 were enriched for GOF variants. The intracellular domain (ICD) of CASR, which includes the disordered regions, influences receptor trafficking, signaling, and desensitization rate among other functions[45]. Interestingly, variants in the ICD, particularly in D1 and D2 have been demonstrated to yield GOF phenotypes[31,45,46]. Specifically, the R990G (D2) variant results in an increased response of CASR to extracellular Ca^{2+}, possibly via its predicted modification of the α-helix between codons 984 and 997 resulting in modulated interaction with filamin-a which interacts with CASR in that region[31]. Similarly, variants in D1,

including S901del/fsX977 and S895_V1075del which result in shortened CASR intracellular C-tails, yield GOF phenotypes, possibly via the modulation of signal transduction or cell surface expression for which the ICD is partially responsible[45,46]. These results further support the biological validity of LoGoFunc's predictions, indicating their capture of information regarding local protein context.

Finally, we illustrate LoGoFunc's potential utility for characterizing variants of uncertain significance (VUS) and uncharacterized variants, an active challenge in genomics. We performed PheWAS on high-confidence, predicted GOF and LOF variants which were either marked as VUS in, or were absent from, ClinVar using patient records from the Mount Sinai BioMe Biobank. We discovered strong associations between the tested variants and relevant phenotypes. For example, the predicted LOF variant c.1648G>A (rs563131364) in *SLC12A3* was associated with severe chronic kidney disease and abnormal blood chemistry among other phenotypes. Notably, over 140 putative LOF *SLC12A3* variants have previously been identified in patients with Gitelman syndrome[47], a disorder characterized by impaired salt reabsorption in the kidneys. Analysis of specific features also suggested that c.1648G>A may be a pathogenic, loss-of-function variant. For example, the variant has a CADD[5] PHRED score of 29.3 and an MSC[21] 95 CI of 13.89, indicating that the variant may be pathogenic taking into consideration the genic context of *SLC12A3*. Similarly, the variant is predicted to manifest autosomal recessive inheritance, consistent with the inheritance pattern of Gitelman syndrome[48]. Together, these results provide preliminary evidence that LoGoFunc may provide utility in the assessment of VUS and uncharacterized variants in addition to providing predictions of functional effect.

In summary, we have developed LoGoFunc, an automated predictor of GOF, LOF, and neutral variants. Our model performed favorably compared to commonly used computational tools designed for the assessment of genetic variation and demonstrated strong predictive power across metrics on our test set of GOF, LOF, and neutral variants. We assessed the contribution of various features to the model's output and found that LoGoFunc learns from a diverse array of structural, functional, sequence-based, and systems-level information, indicating that these features have bearing on the functional outcome of genetic variants. Further, we demonstrated significant localization of GOF, LOF, and neutral variants in 3D structural and functional sites in proteins, and demonstrated LoGoFunc's ability to assess

previously uncharacterized variants. Our findings corroborated previously reported molecular mechanisms resulting in gain and loss of function and also suggest novel mechanisms that may shed light on disease etiology. We applied our method to 71,322,505 missense mutations in the human genome, and provide our predictions at https://itanlab.shinyapps.io/goflof/.

METHODS

Dataset assembly

We used 11,370 variants corresponding to GOF and LOF variants labeled by Bayrak et al[11]. Additionally, we compiled 65,075 variants that were included in the HGMD[12] Professional version 2021.3 database and deposited in 2020 and 2021 and assigned labels using the same strategy that Bayrak et al[11] employed. From these variants, we first selected 32,911 disease-causing class (DM) variants. Then, we used the Spacy 3.0.6 natural language processing (NLP) library to search for GOF- and LOF-related phrases in associated publications for each DM variant. Using the phrase-based matching algorithm PhraseMatcher, we iteratively searched the paper titles and abstracts from all associated publications for the patterns “gain of function(s)”, “gain-of-function(s)”, “GOF”, “loss of function(s)”, “loss-of-function(s)”, and “LOF”. We converted the text to lowercase to allow for case sensitivity. When at least one of the publications indicated GOF or LOF, we labeled the corresponding variant accordingly. When there was a disagreement, i.e. a variant was found as GOF in one abstract and LOF in another abstract, the variant was excluded from the dataset. Up to five putatively neutral variants were randomly selected from all human genes for which they were available in the gnomAD v2.1[13] database. For each gene in the dataset containing labeled GOF or LOF variants, neutral variants were selected up to the number of GOF and LOF variants in the gene or all the available neutral variants, whichever was the lower. An additional 14,537 neutral variants were randomly selected from the remaining genes bringing the total neutral variants to 20,000.

Preprocessing of input data

Preprocessing steps were applied to prepare sample variants for prediction. Where dbNSFP[23] provided multiple annotations for different transcripts, the annotation pertinent to the canonical transcript was selected. The variants were split into training and test sets of 90% (n. LOF: 12,171, n. GOF: 1,343, n. Neutral: 18,000) and 10% (n. LOF: 1,353, n. GOF: 149, n. Neutral: 2,000) of the dataset respectively. A binary encoder was fit to the categorical features in the training set and used to encode the categorical features in the training and test sets. Missing allele frequencies were imputed with the median value from the training set and all other missing data were imputed using the average value of the respective columns from the training set. Scikit-learn's[50] MinMaxScaler was fit to the training data and used to scale training and test sets between 0 and 1. Values in the test set outside this range after scaling were clipped.

Model selection and assessment

We tested four algorithms on the dataset: RandomForest[14], LightGBM[10], XGBoost[15], and Neural Networks. For each algorithm, we performed hyperparameter tuning with the Optuna[51] optimization library for 500 rounds to maximize the F1-score averaging the one-*vs*-rest binary F1-score for each class (for hyperparameter search spaces see Supplementary Information). The F1-score is a function of the precision and recall, defined as follows, where *y* is the set of predicted sample, label pairs, and *y'* is the set of true sample, label pairs:

$$precision(y, y') = \frac{|y \cap y'|}{|y|}$$

$$recall(y, y') = \frac{|y \cap y'|}{|y'|}$$

$$F1(y, y') = 2 \times \frac{(precision(y, y') \times recall(y, y')}{(precision(y, y') + recall(y, y')}$$

To extend the F1-score to multiclass classification, we calculated the macro-averaged F1-score, defined as follows - *L* is the set of labels:

$$F1_{macro} = \frac{1}{|L|} \sum_{l \in L} y'_l \, F1(y_l, y'_l)$$

Each hyperparameter combination was evaluated with 5-fold stratified cross-validation and the combination yielding the highest average F1-score across folds was selected. The model with the best macro-averaged F1-score from the hyperparameter tuning was tested in a voting ensemble. The top 15 hyperparameter sets were used in a voting ensemble of 15 models. Model weights were tuned using the same hyperparameter tuning procedure as described above. The weights yielding the best macro-averaged F1-score were applied to the final model.

LoGoFunc's performance was assessed via ROC-AUC, average precision (AP), F1-score, and Cohen's kappa[16] calculated using scikit-learn version 1.0[50]. AP and Cohen's kappa[16] are defined as follows, where n is the n_{th} threshold, p_o is the observed agreement ratio, and p_e is the expected agreement when annotators assign labels randomly[50]:

$$AP = \sum_n (recall_n - recall_{n-1}) precision_n$$

$$Cohen's\ \ Kappa = \frac{p_o - p_e}{1 - p_e}$$

We estimated LoGoFunc's performance on scarcely annotated genes by splitting the labeled variants such that variants occurring in a given gene did not appear in both the training and testing sets. In particular, we randomly selected 1,200 genes represented in the labeled dataset and included all variants from these genes in the testing set. All other variants were included in the training set. The resulting training set was comprised of 17,920 neutral variants, 11,820 LOF variants and 1,308 GOF variants, whereas the testing set consisted of 2,080 neutral variants, 1,704 LOF variants and 184 GOF variants. Using these training and testing sets, we performed hyperparameter tuning and performance assessment, according to the process described above.

Prediction cutoffs

For each variant class, GOF, LOF, and neutral, we selected predictions from the training and testing sets for variants of that class. We applied the Kolmogorov-Smirnov test[51] for goodness of fit to predictions for these variants with each continuous distribution implemented in scipy[52] version 1.0.1. For each distribution, we first estimated the distribution parameters that best modeled the predictions using scipy[52], and then selected the parametrized distribution with

the highest p-value from the Kolmogorov-Smirnov test[51]. 95% confidence intervals were then calculated using the best fitting, parameterized distribution for predictions from each class respectively and clipped between zero and one where applicable. When five or more variants were available from a given class for a given gene, we repeated the above process to calculate gene-specific 95% confidence intervals. When fewer than five variants were available for a class in a gene, we defaulted to the 95% confidence intervals calculated for predictions from the entire dataset.

Model architecture

LoGoFunc is a voting ensemble of 15 LightGBM[10] models, with each of the 15 LightGBM[10] models assigned a weight from one to five describing its impact on the final prediction. The model receives as input an array of 805 numerical values and outputs three values between zero and one corresponding to the predicted probability that a variant results in GOF, LOF, or is neutral respectively. These outputs are the weighted average of the predictions of each LightGBM[10] model in the ensemble.

Enrichment analysis and feature importance

Feature enrichments were calculated via Fisher's exact test. Continuous features were categorized by assigning a cutoff according to the value recommended by the authors of the tool from which the features were derived. When no such cutoff was reported, a cutoff of 0.5 was selected for probabilistic features. The Benjamini-Hochberg correction[53] was applied at an alpha level of 0.05 to control for false positives as a result of multiple testing. Feature importance values were calculated as the weighted average of the frequency that a feature was used to split nodes in the trees. Weights correspond to the weights of the models in the voting ensemble.

Tool comparison

LoGoFunc was compared to other computational methods by assessing the area under the ROC curve and AP. All GOF (n. 86) and LOF (n. 328) variants from the test set for which all compared tools provided a prediction were collected. ROC-AUC and AP were calculated treating GOF as the positive class. To assess performance separating neutral variants from

GOF and LOF, we added all neutral (n. 389) variants from the test set for which each tool provided a prediction. ROC-AUC and AP were again calculated, this time with GOF and LOF variants as the positive classes respectively and neutral as the negative class. Finally, we calculated the one-*vs.*-all ROC-AUC and AP with GOF and LOF variants as the positive class and neutral variants as the negative class. Most of the compared tools provide predictions in which higher scores correspond to a greater likelihood that a given variant will be damaging. However, SIFT[7] outputs predictions between zero and one in which lower scores correspond to a greater likelihood of a damaging effect. Similarly, LoGoFunc outputs a value between zero and one where higher scores indicate a greater likelihood of neutrality. Thus, to ensure consistency between all compared tools when treating neutral as the negative class and GOF and LOF as the positive class, SIFT[7] and LoGoFunc neutral predictions were transformed by subtracting each prediction from one before assessing ROC-AUC and AP.

PheWAS of predicted GOF and LOF variants.

We analyzed 1,650 phenotypes for which there were at least 20 cases in the Mount Sinai BioMe BioBank. For each phenotype, controls were randomly sampled from non-cases to fix the ratio of cases to controls at 1:5. Overlapping individuals, i.e. those sharing phenotypes other than the phenotype of interest, were removed from the control set. We implemented principal component analysis (PCA) on 3,800 whole-exome sequencing samples in BioMe using independent variants before the PheWAS analysis and used the first five components to adjust for potential population stratification in cases and controls. We reduced linkage disequilibrium (LD) between markers by removing all markers with $r^2>0.2$ (window size 50, step size 5), as well as markers in known high LD regions. Furthermore, we retained variants with minor allele frequency (MAF) greater than 0.02 and genotyping rate greater than 95% across the dataset (excluding A/T, C/G mutations). PCA was conducted using Plink 1.9[54]. PheWAS was conducted using the R “PheWAS” package[55].

ACKNOWLEDGMENTS

We thank Bruce Gelb, Vikas Pejaver, Laura Huckins, Josh Milner, Nicole Zatorski, Keino Hutchinson, and Meltem Ece Kars for their thoughtful discussion and support for this project.

AUTHOR CONTRIBUTIONS

DS, CB, YI, and AS conceived of this project. CB and DS collected the labeled data for training and testing the model. YW carried out the PheWAS analysis. DS led the development of the model and the preparation of the manuscript. PDS and DNC provided data for training and testing. YI and AS oversaw this project and provided guidance. All authors contributed to the writing of the manuscript and approved the final manuscript.

FIGURE LEGENDS

Figure 1: LoGoFunc workflow and model architecture. **a**. Pipeline for the collection of labeled GOF and LOF variants. Related abstracts for high confidence pathogenic variants from the HGMD[12] were searched for nomenclature denoting gain or loss of function. **b**. Dataset preparation and preprocessing workflow. 1,492 GOF, 13,524 LOF, and 20,000 neutral variants were obtained from the GOF/LOF database[11], HGMD[33], and gnomAD[13] and annotated. Duplicate features were removed and annotations were selected for VEP[49] canonical transcripts during data cleaning. The cleaned data were split into training and test sets comprising 90% and 10% of the dataset respectively, stratified by variant label. A binary encoder was fit to the training set and used to encode the training and test sets. Missing allele frequencies in the training and test sets were imputed with the column median from the training set. All other missing data in the training and test sets were imputed with the column average from the training set. A scaler was fit to the training set and used to scale the training and test features between 0 and 1 with out of bounds values clipped. **c**. Model architecture and output. The encoded feature set consists of 805 variables for each sample. The model consists of an ensemble of fifteen LightGBM[10] classifiers. A probability is output for each class, GOF, LOF, and neutral. Created with BioRender.com.

Figure 2: **LoGoFunc evaluation.** PR curves comparing discriminatory power of computational tools on a set of variants from the test set for which predictions were available from all compared tools. **a**. GOF (n. 86) *vs*. neutral (n. 389). **b**. LOF (n. 328) *vs.* neutral (n. 389). **c**. GOF (n.86) and LOF (n.328) combined *vs*. neutral (n. 389). **d**. GOF (n. 86) *vs*. LOF (n. 328). Created with BioRender.com.

Figure 3: **Discriminating features of GOF, LOF, and neutral variants. a**. Features in the 90th percentile of importance as calculated by the weighted average of the frequency a feature was used to split nodes in the trees. **b**. Enrichments and depletions for protein structural and functional features used by the LoGoFunc model. GOF (blue), LOF (orange), and neutral (green) log odds are displayed for each feature. Significant enrichments and depletions are denoted by asterisks. Significance was calculated with Fisher's exact test, Benjamini-Hochberg corrected[53] to allow for multiple comparisons. **c**. 3D structure of the CASR protein colored by predicted GOF, LOF, and neutral residues. Residues for which missense variants are predicted to result only in GOF are in red, only LOF in blue, only neutral in white, GOF or LOF in purple, GOF, LOF, or neutral in brown, GOF or neutral in yellow, and LOF or neutral in orange. Distinct protein domains are denoted with brackets. Created with BioRender.com.

Figure 4: **Association between variant type and splice effects. a**. (Top) Density of GOF, LOF and neutral variants within 20 base-pairs of a splice junction. (Bottom) Proportion of GOF, LOF, and neutral variants predicted to yield a gain of splice acceptor or donor or a loss of splice acceptor or donor. **b**. Percentage of GOF, LOF, and Neutral variants in proximity (20 base-pairs) to acceptor and donor splice sites. **c**. MMSplice[36] sub-model alternate minus reference logit percent-spliced-in predictions. **d**. Box-and-whisker plot displaying distance to the nearest exon junction boundary in nucleotides by variant class. Boxes denote quartiles, while whiskers extend to the limits of the distribution with outliers removed when greater than 1.5 times the interquartile range from the low and high quartiles respectively. Created with BioRender.com.

Figure 5: **Relationship between variant type and phenotypes.** **a**. Associations between high-confidence, predicted GOF (c.7471C>T) and LOF (c.1648G>A) variants and phenotypes as determined by PheWAS[55] analysis of patients in the BioMe biobank. **b**. Distribution of CADD[5] PHRED scores in the dataset (green). CADD[5] PHRED scores and MSC[21] 95% CI cutoffs for c.7471C>T (solid and dashed blue lines) and c.1648G>A (solid and dashed red lines). Created with BioRender.com.

REFERENCES

1. Studer, R. A., Dessailly, B. H. & Orengo, C. A. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem. J.* **449**, 581–594 (2013).
2. Kagawa, R. *et al.* Alanine-scanning mutagenesis of human signal transducer and activator of transcription 1 to estimate loss- or gain-of-function variants. *J. Allergy Clin. Immunol.* **140**, 232–241 (2017).
3. Boisson-Dupuis, S. *et al.* Inborn errors of human STAT1: allelic heterogeneity governs the diversity of immunological and infectious phenotypes. *Curr. Opin. Immunol.* **24**, 364–378 (2012).
4. Gupta, K. & Varadarajan, R. Insights into protein structure, stability and function from saturation mutagenesis. *Curr. Opin. Struct. Biol.* **50**, 117–125 (2018).
5. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
6. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
7. Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452-457 (2012).
8. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
9. Feng, B.-J. PERCH: A Unified Framework for Disease Gene Prioritization. *Hum. Mutat.* **38**, 243–251 (2017).
10. Ke, G. *et al.* LightGBM: a highly efficient gradient boosting decision tree. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 3149–3157 (Curran Associates Inc., 2017).
11. Sevim Bayrak, C. *et al.* Identification of discriminative gene-level and protein-level features associated with pathogenic gain-of-function and loss-of-function variants. *Am. J. Hum. Genet.* **108**, 2301–2318 (2021).

12. Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD ®): optimizing its use in a clinical diagnostic or research setting. *Hum Genet* **139**, 1197-1207 (2020).
13. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
14. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
15. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 785–794 (2016) doi:10.1145/2939672.2939785.
16. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).
17. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).
18. Li, B. *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinforma. Oxf. Engl.* **25**, 2744–2750 (2009).
19. Raimondi, D. *et al.* DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.* **45**, W201–W206 (2017).
20. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
21. Itan, Y. *et al.* The mutation significance cutoff: gene-level thresholds for variant predictions. *Nat. Methods* **13**, 109–110 (2016).
22. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).
23. 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.* **12**, 103 (2020).
24. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).

25. Petrazzini, B. O. *et al.* Prediction of recessive inheritance for missense variants in human disease. 2021.10.25.21265472 (2021) doi:10.1101/2021.10.25.21265472.
26. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
27. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *ArXiv160700653 Cs Stat* (2016).
28. Hendy, G. N., Guarnieri, V. & Canaff, L. Calcium-sensing receptor and associated diseases. *Prog. Mol. Biol. Transl. Sci.* **89**, 31–95 (2009).
29. Vezzoli, G. *et al.* R990G polymorphism of calcium-sensing receptor does produce a gain-of-function and predispose to primary hypercalciuria. *Kidney Int.* **71**, 1155–1162 (2007).
30. Ling, S. *et al.* Structural mechanism of cooperative activation of the human calcium-sensing receptor by Ca2+ ions and L-tryptophan. *Cell Res.* **31**, 383–394 (2021).
31. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
32. Mesman, R. L. S. *et al.* Alternative mRNA splicing can attenuate the pathogenicity of presumed loss-of-function variants in BRCA2. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **22**, 1355–1365 (2020).
33. Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr. Protoc. Bioinforma.* **Chapter 1**, Unit1.13 (2012).
34. Cheng, J. *et al.* MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).
35. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).
36. Kapustin, Y. *et al.* Cryptic splice sites and split genes. *Nucleic Acids Res.* **39**, 5837–5844 (2011).
37. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

38. Jiang, F. *et al.* The mechanosensitive Piezo1 channel mediates heart mechano-chemo transduction. *Nat. Commun.* **12**, 869 (2021).
39. Andolfo, I. *et al.* Gain-of-function mutations in PIEZO1 directly impair hepatic iron metabolism via the inhibition of the BMP/SMADs pathway. *Am. J. Hematol.* **95**, 188–197 (2020).
40. Zou, D.-M. & Sun, W.-L. Relationship between Hepatitis C Virus Infection and Iron Overload. *Chin. Med. J. (Engl.)* **130**, 866–871 (2017).
41. Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* **11**, 5918 (2020).
42. Cassa, C. A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).
43. Itan, Y. *et al.* The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13615–13620 (2015).
44. Obermannova, B. *et al.* Novel calcium-sensing receptor cytoplasmic tail deletion mutation causing autosomal dominant hypocalcemia: molecular and clinical study. *Eur. J. Endocrinol.* **174**, K1–K11 (2016).
45. Lienhardt, A. *et al.* A large homozygous or heterozygous in-frame deletion within the calcium-sensing receptor's carboxylterminal cytoplasmic tail that causes autosomal dominant hypocalcemia. *J. Clin. Endocrinol. Metab.* **85**, 1695–1702 (2000).
46. Knoers, N. V. & Levtchenko, E. N. Gitelman syndrome. *Orphanet J. Rare Dis.* **3**, 22 (2008).
47. Riveira-Munoz, E. *et al.* Transcriptional and functional analyses of SLC12A3 mutations: new clues for the pathogenesis of Gitelman syndrome. *J. Am. Soc. Nephrol. JASN* **18**, 1271–1283 (2007).
48. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
49. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *ArXiv12010490 Cs* (2018).
50. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *ArXiv190710902 Cs Stat* (2019).

51. Massey, F. J. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **46**, 68–78 (1951).
52. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
53. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
54. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
55. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinforma. Oxf. Engl.* **30**, 2375–2376 (2014).

FIGURES

Figure 1

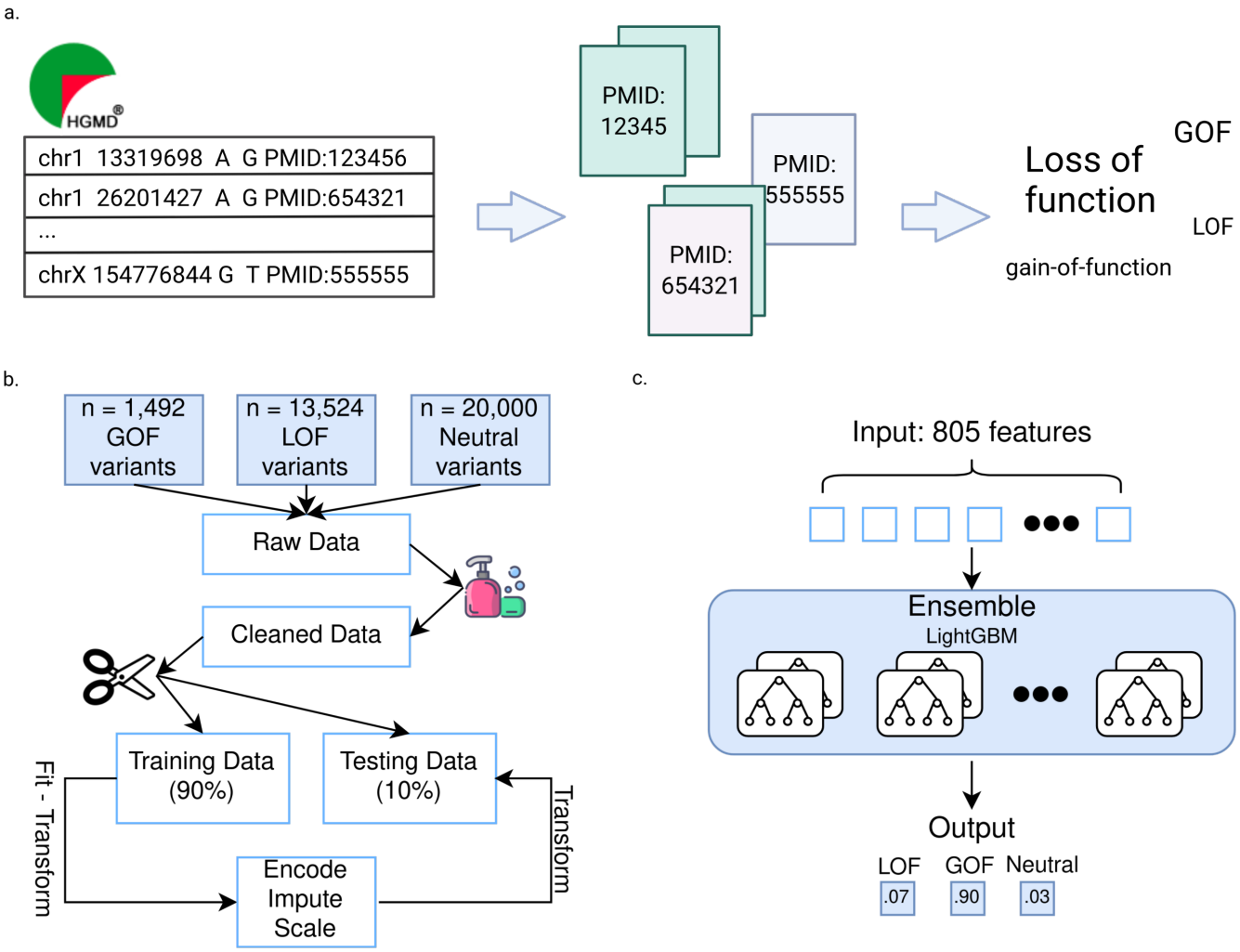

Figure 1: **LoGoFunc workflow and model architecture**. **a**. Pipeline for the collection of labeled GOF and LOF variants. Related abstracts for high confidence pathogenic variants from the HGMD[12] were searched for nomenclature denoting gain or loss of function. **b**. Dataset preparation and preprocessing workflow. 1,492 GOF, 13,524 LOF, and 20,000 neutral variants were obtained from the GOF/LOF database[11], HGMD[33], and gnomAD[13] and annotated. Duplicate features were removed and annotations were selected for VEP[49] canonical transcripts during data cleaning. The cleaned data were split into training and test sets comprising 90% and 10% of the dataset respectively, stratified by variant label. A binary encoder was fit to the training set and used to encode the training and test sets. Missing allele frequencies in the training and test sets were imputed with the column median from the training set. All other missing data in the training and test sets were imputed with the column average from the training set. A scaler was fit to the training set and used to scale the training and test features between 0 and 1 with out of bounds values clipped. **c**. Model architecture and output. The encoded feature set consists of 805 variables for each sample. The model consists of an ensemble of fifteen LightGBM[10] classifiers. A probability is output for each class, GOF, LOF, and neutral. Created with BioRender.com.

Figure 2

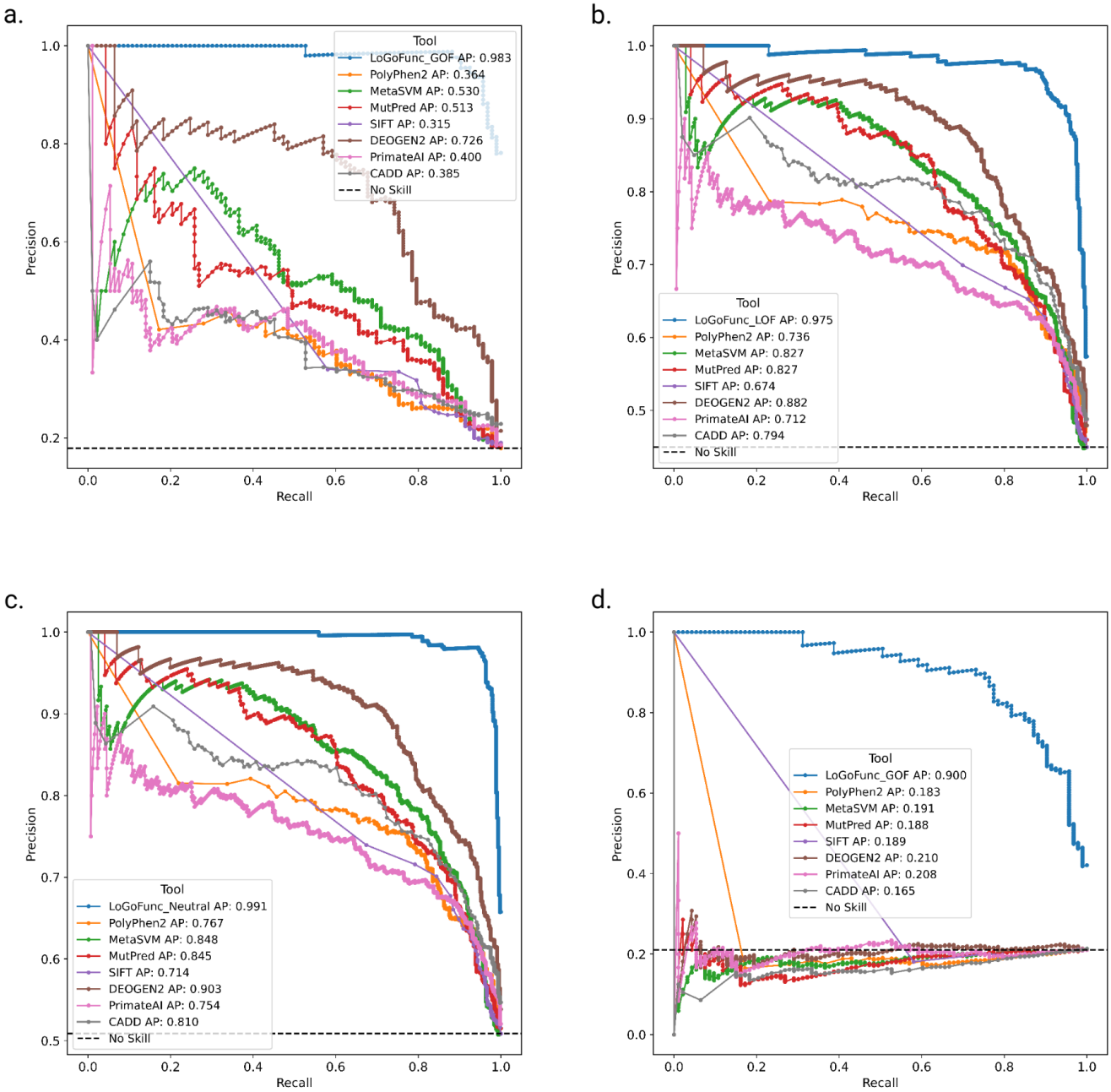

Figure 2: **LoGoFunc evaluation.** PR curves comparing discriminatory power of computational tools on a set of variants from the test set for which predictions were available from all compared tools. **a**. GOF (n. 86) *vs*. neutral (n. 389). **b**. LOF (n. 328) *vs*. neutral (n. 389). **c**. GOF (n.86) and LOF (n.328) combined *vs*. neutral (n. 389). **d**. GOF (n. 86) *vs*. LOF (n. 328). Created with BioRender.com.

Figure 3

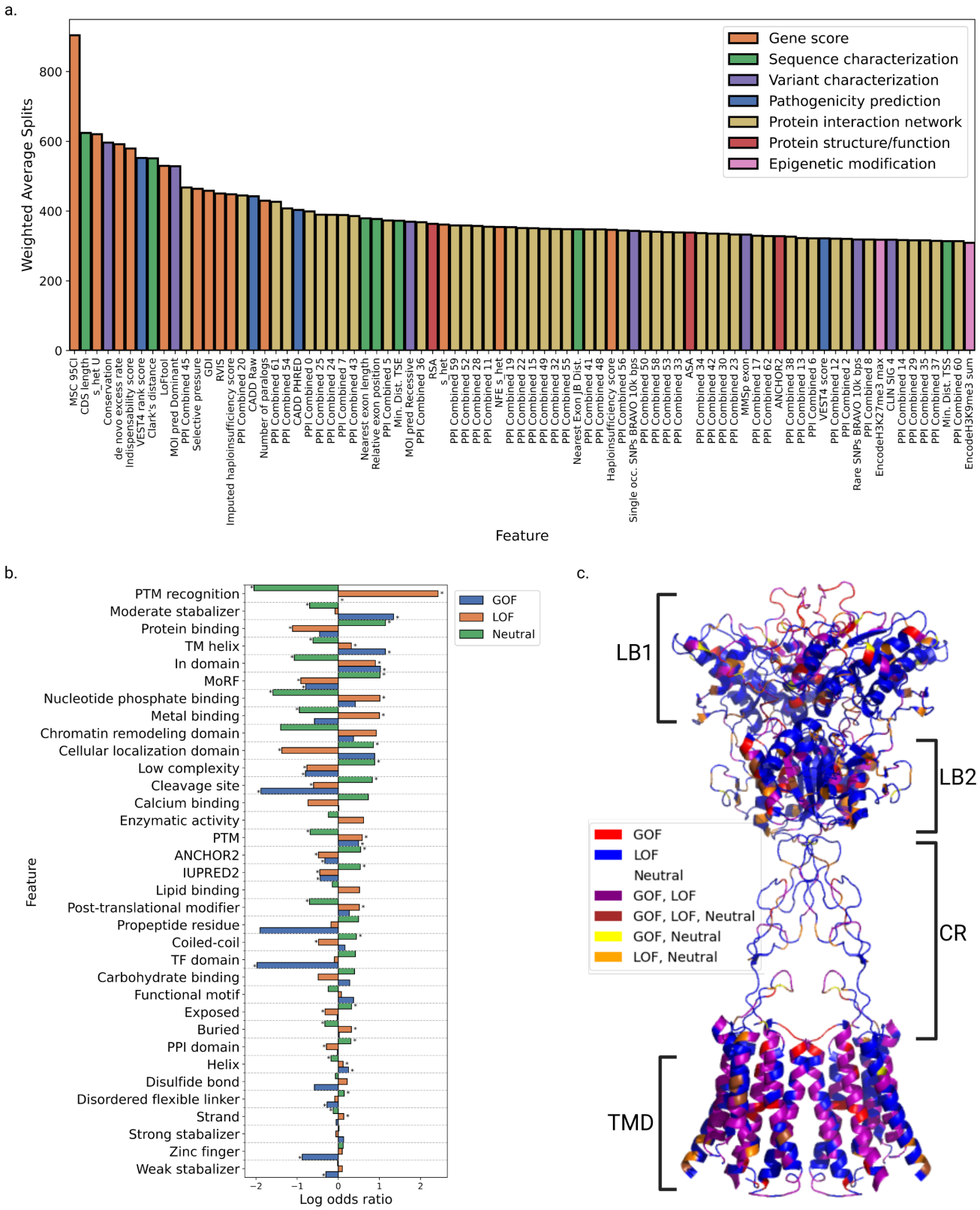
a.
Weighted Average Splits
0
200
400
600
800
Gene score
Sequence characterization
Variant characterization
Pathogenicity prediction
Protein interaction network
Protein structure/function
Epigenetic modification
MSC 95CI
CDS length
s_het U
Conservation
de novo excess rate
Indispensability score
VEST4 rank score
Clark's distance
LoFtool
MOI pred Dominant
PPI Combined 45
Selective pressure
GDI
RVIS
Imputed haploinsufficiency score
PPI Combined 20
CADD Raw
Number of paralogs
PPI Combined 61
PPI Combined 54
CADD PHRED
PPI Combined 0
PPI Combined 25
PPI Combined 24
PPI Combined 7
PPI Combined 43
Nearest exon length
Relative exon position
PPI Combined 5
Min. Dist. TSE
MOI pred Recessive
PPI Combined 36
RSA
s_het
PPI Combined 59
PPI Combined 52
PPI Combined 28
PPI Combined 11
NFE s_het
PPI Combined 19
PPI Combined 22
PPI Combined 15
PPI Combined 49
PPI Combined 32
PPI Combined 52
Nearest Exon JB Dist.
PPI Combined 41
PPI Combined 48
Haploinsufficiency score
PPI Combined 56
Single occ. SNPs BRAVO 10k bps
PPI Combined 50
PPI Combined 58
PPI Combined 53
PPI Combined 33
ASA
PPI Combined 34
PPI Combined 42
PPI Combined 30
PPI Combined 23
MMSp exon
PPI Combined 17
PPI Combined 62
ANCHOR2
PPI Combined 38
PPI Combined 13
PPI Combined 6
VEST4 score
PPI Combined 12
PPI Combined 2
Rare SNPs BRAVO 10k bps
PPI Combined 8
EncodeH3K27me3 max
CLIN SIG 4
PPI Combined 14
PPI Combined 29
PPI Combined 35
PPI Combined 37
Min. Dist. TSS
PPI Combined 60
EncodeH3K9me3 sum
Feature
b.
GOF
LOF
Neutral
PTM recognition
Moderate stabalizer
Protein binding
TM helix
In domain
MoRF
Nucleotide phosphate binding
Metal binding
Chromatin remodeling domain
Cellular localization domain
Low complexity
Cleavage site
Calcium binding
Enzymatic activity
PTM
ANCHOR2
IUPRED2
Lipid binding
Post-translational modifier
Propeptide residue
Coiled-coil
TF domain
Carbohydrate binding
Functional motif
Exposed
Buried
PPI domain
Helix
Disulfide bond
Disordered flexible linker
Strand
Strong stabalizer
Zinc finger
Weak stabalizer
Feature
−2
−1
0
1
2
Log odds ratio
c.
LB1
LB2
CR
TMD
GOF
LOF
Neutral
GOF, LOF
GOF, LOF, Neutral
GOF, Neutral
LOF, Neutral

Figure 3: **Discriminating features of GOF, LOF, and neutral variants. a**. Features in the 90th percentile of importance as calculated by the weighted average of the frequency a feature was used to split nodes in the trees. **b**. Enrichments and depletions for protein structural and functional features used by the LoGoFunc model. GOF (blue), LOF (orange), and neutral (green) log odds are displayed for each feature. Significant enrichments and depletions are denoted by asterisks. Significance was calculated with Fisher's exact test, Benjamini-Hochberg corrected[53] to allow for multiple comparisons. **c**. 3D structure of the CASR protein colored by predicted GOF, LOF, and neutral residues. Residues for which missense variants are predicted to result only in GOF are in red, only LOF in blue, only neutral in white, GOF or LOF in purple, GOF, LOF, or neutral in brown, GOF or neutral in yellow, and LOF or neutral in orange. Distinct protein domains are denoted with brackets. Created with BioRender.com.

Figure 4

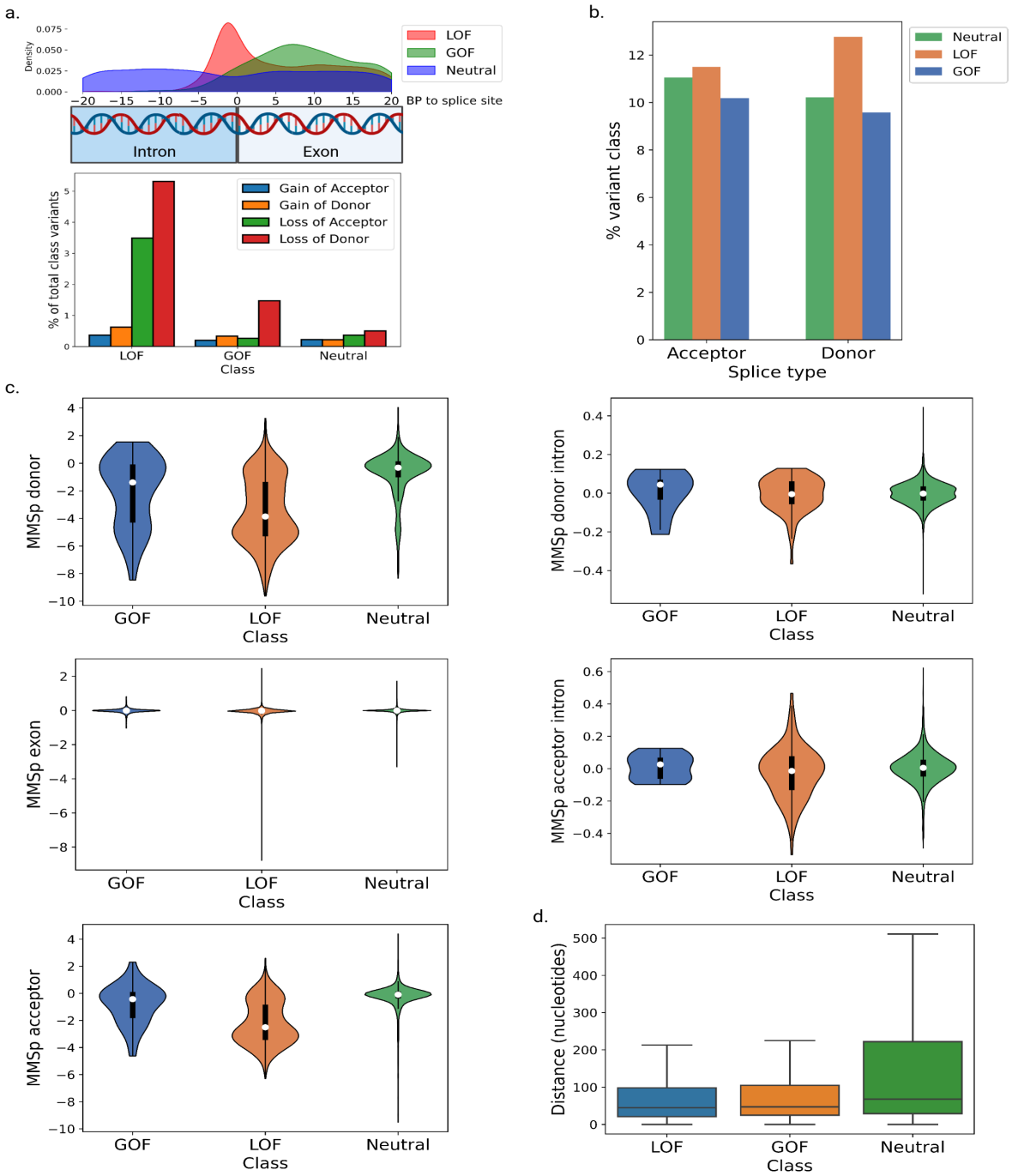

Figure 4: **Association between variant type and splice effects. a**. (Top) Density of GOF, LOF and neutral variants within 20 base-pairs of a splice junction. (Bottom) Proportion of GOF, LOF, and neutral variants predicted to yield a gain of splice acceptor or donor or a loss of splice acceptor or donor. **b**. Percentage of GOF, LOF, and Neutral variants in proximity (20 base-pairs) to acceptor and donor splice sites. **c**. MMSplice[36] sub-model alternate minus reference logit percent-spliced-in predictions. **d**. Box-and-whisker plot displaying distance to the nearest exon junction boundary in nucleotides by variant class. Boxes denote quartiles, while whiskers extend to the limits of the

distribution with outliers removed when greater than 1.5 times the interquartile range from the low and high quartiles respectively. Created with BioRender.com.

Figure 5

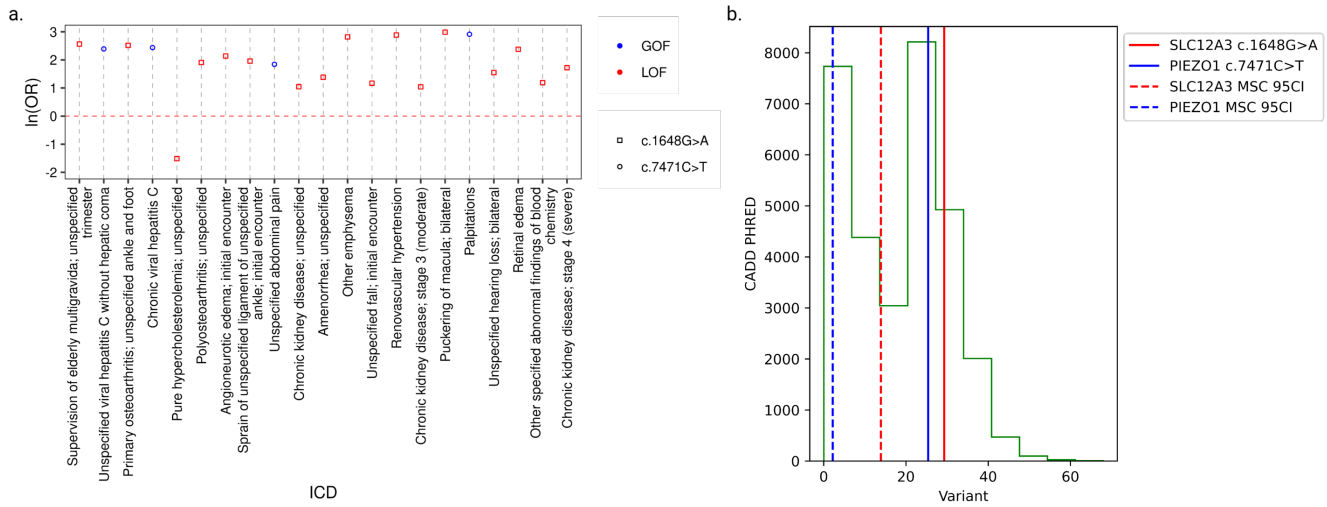

Figure 5: **Relationship between variant type and phenotypes. a**. Associations between high-confidence, predicted GOF (c.7471C>T) and LOF (c.1648G>A) variants and phenotypes as determined by PheWAS[55] analysis of patients in the BioMe biobank. **b**. Distribution of CADD[5] PHRED scores in the dataset (green). CADD[5] PHRED scores and MSC[21] 95% CI cutoffs for c.7471C>T (solid and dashed blue lines) and c.1648G>A (solid and dashed red lines). Created with BioRender.com.