Proteome-scale prediction of molecular mechanisms underlying dominant genetic diseases

Mihaly Badonyi[1*], Joseph A Marsh[1]

[1] MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

*correspondence to mihaly.badonyi [at] ed.ac.uk

Abstract

Many dominant genetic disorders result from protein-altering mutations, acting primarily through dominant-negative (DN), gain-of-function (GOF), and loss-of-function (LOF) mechanisms. Deciphering the mechanisms by which dominant diseases exert their effects is often experimentally challenging and resource intensive, but is essential for developing appropriate therapeutic approaches. Diseases that arise via a LOF mechanism are more amenable to be treated by conventional gene therapy, whereas DN and GOF mechanisms may require gene editing or targeting by small molecules. Moreover, pathogenic missense mutations that act via DN and GOF mechanisms are more difficult to identify than those that act via LOF using nearly all currently available variant effect predictors. Here, we introduce a tripartite statistical model made up of support vector machine binary classifiers trained to predict whether human protein-coding genes are likely to be associated with DN, GOF, or LOF molecular disease mechanisms. We test the utility of the predictions by examining biologically and clinically meaningful properties known to be associated with the mechanisms. Our results strongly support that the models are able to generalise on unseen data and offer insight into the functional attributes of proteins associated with different mechanisms. We hope that our predictions will serve as a springboard for researchers studying novel variants and those of uncertain clinical significance, guiding variant interpretation strategies and experimental characterisation. Predictions for the human UniProt reference proteome are available at https://osf.io/z4dcp/.

Introduction

Genetic diseases are often caused by protein-altering mutations that act via different protein-level mechanisms. One such mechanism is loss of function (LOF), wherein the mutation completely ablates the function of the protein. While LOF-associated disease is often due to protein null mutations, in which little or no protein is produced due to premature termination codons and nonsense mediated decay, it can also be caused by protein-altering mutations, e.g., missense mutations that destabilise protein structure. Nearly all recessive disorders are associated with LOF. By contrast, although LOF mutations often cause dominant disease via haploinsufficiency when the lost function cannot be compensated by the remaining wild-type allele [1], many dominant disorders are caused by non-LOF mechanisms. These include gain of function (GOF), which is characterized by an altered or newly appeared function in the mutant protein, and the dominant-negative (DN) effect, whereby the mutant allele can directly or indirectly disrupt the function of the wild type [2]. Understanding the molecular mechanisms through which protein mutations exert their effects has great potential for improving diagnoses and developing adequate treatments of genetic disorders. Recent years have seen major improvements in experimental methods capable of probing a large number of variants at once. In particular, deep mutational scanning and its modifications hold huge promise for deciphering the effect of variants and, by extension, their mechanisms [3]. For now, relative to LOF, our understanding of the properties of variants that act via DN or GOF (collectively, non-LOF) mechanisms remains limited.

In many cases, the different dominant molecular mechanisms are not mutually exclusive in a gene. For example, a type of cardiomyopathy can be influenced by both titin haploinsufficiency (LOF) and truncated titin peptides that seem to “poison” wild-type complexes (DN) [4]. Although molecular mechanisms are indisputably variant-level phenomena, it has become evident that mutations in many genes are more likely to exhibit a specific mechanism [5]. This concept has shed light on certain structural and functional properties associated with non-LOF proteins at the systems level [5,6]. Moreover, these analyses have called attention to a limitation of current variant effect predictors (VEPs), as they struggle to accurately predict the pathogenicity of non-LOF variants [5,7,8]. This is a critical problem to address, as otherwise there is a strong possibility that we could miss variants on the account of difficulties in computationally predicting their effects. A gene-level predictor of molecular mechanisms could potentially be very useful for highlighting genes where VEPs could perform poorly and miss pathogenic variants.

Previously, we built a simple model to predict genes most likely to be associated with non-LOF mechanisms [6]. Predictions from this model were incomplete for the proteome due to limited availability of protein structural and functional features, in particular its reliance on quaternary structure information that comes from experimentally determined structures. Additionally, DN and GOF mechanisms were combined into a single non-LOF category because the genes associated with them tend to have similar attributes such as evolutionary constraints, functional niches, and protein complex topologies. To address these problems, here we train three support vector machine binary classifiers using surrogate structural and functional features complete for the proteome. Each

classifier is tasked with identifying one of the molecular mechanisms over another unique class (DN *vs* LOF and GOF *vs* LOF) or pooled classes (LOF *vs* non-LOF). This design maximizes the number of cases available for training and enables a flexible classification regime. Analysis of an unbiased set of disease-linked proteins with predicted mechanisms suggests that their properties, independent from model features, are consistent with known characteristics of the mechanisms, thus strongly supporting the utility of this approach.

To facilitate a community-wide effort in piecing together the human variant effect map, we make predictions for all human protein-coding genes. These predictions can pinpoint the likely mechanism of a variant in a dominant gene with no previous association to a molecular mechanism and offer valuable insight when assessing novel variants in a gene with no prior disease linkage. Additionally, they could help prioritise genes in clinical and population genetics data for laboratory studies and allow researchers to explore the functional, structural, and evolutionary properties inherent to the mechanisms.

Methods

Software and databases

In this study, we rely on a number of properties derived from the AlphaFold-predicted structures of human proteins [9]. Solvent accessible surface area is calculated at residue level with AREAIMOL from the CCP4 programme suite [10]. FoldX version 5.0 [11] is run to determine the predicted Gibbs free energy of folding (ΔΔG) of missense substitutions by first calling the `RepairPDB` command followed by `BuildModel`. Residue-level SCRIBER scores [12] for the human proteome are extracted from the 9606_database.json file provided by the DescribePROT database [13]. Protein-level UniProt ProtNLM embeddings [14] are obtained via the UniProt FTP for the human proteome (2023_02 release). All 1024 embedding dimensions are extracted from the per-protein.h5 file using the `h5dump` command from the HDF5 command line tool. Clustal Omega version 1.2.4 [15] is used to create a distance matrix of protein sequences for homology filtering. Scores from VEPs ESM-1v [16], EVE [17], MetaRNN [18], and VARITY_R [19] are gathered using an in-house pipeline described in [20]. Machine learning methods are implemented in R version 4.3.0 [21] using the `tidymodels` framework (https://www.tidymodels.org/) with extension packages `themis`, `probably`, and `DALEX`. Statistical tests are performed with the `rstatix` package. We consider an alpha level of 0.05 to be statistically significant. Bootstrap confidence intervals are obtained from 1,000 resamples with the percentile method [22].

ClinVar and gnomAD mapping to UniProt reference proteome

Genomic coordinates of pathogenic and likely pathogenic (hereinafter “pathogenic”) missense variants are extracted from the ClinVar [23] variant calling file (VCF) (accessed on 2023-06-15) with `BCFtools` [24]. Putatively benign variants (hereinafter “benign”) with a “PASS” filter are extracted from gnomAD v2.1.1 [25] using the exomes all chromosomes VCF file. Mapping is performed with Ensembl VEP 107 [26] using the `--uniprot` and `--canonical` flags. For each variant, we select either the canonical transcript or the first UniProt isoform on condition that the mapped amino acid is identical to that in the sequence.

Data acquisition

Evidence for DN, GOF, or LOF molecular disease mechanism observed in a gene was collected using a combination of text-mining and semi-manual curation, as previously described [6]. We updated the gene set using newly added autosomal dominant genes from the Online Inheritance of Man (OMIM) database [27] (accessed via the API on 2023-05-25) and developmental disorder genes with monoallelic requirement from the Deciphering Developmental Disorders Genotype-to-Phenotype (DDG2P) database (https://www.deciphergenomics.org/ddd/ddgenes, accessed on 2023-05-25). The updated list contains 1,270 genes with one or more molecular mechanism class based upon available evidence. Of the genes, 874 are assigned to a unique class, 318 are assigned to two classes, and 78 genes intersect with all three classes. We provide the set of manually curated genes, their class labels and evidence lines along with their PubMed identifiers in **Table S1.**

Feature selection and engineering

We use gene and protein-level features previously described in [6], notably excluding subunit interface size, which is derived from experimentally determined structures of protein complexes, and the PANTHER functional classification of proteins [28], which is incomplete for the human proteome. We sought to rationally select and design surrogate features that make up for the loss of protein structural and functional information. To replace interface size, we calculate the median SCRIBER score of residues that have >5% relative surface accessible surface area (RSA) in the corresponding AlphaFold predicted structure. RSA is defined as the ratio of the maximum solvent accessible surface area determined from Gly-X-Gly tripeptides (where X denotes the amino acid) [29] and that of the residue in the context of the monomeric protein. To substitute protein functional classification, we use UniProt ProtNLM embeddings that capture information about Pfam domains and functional sequence features and transfer those to unannotated and manually unreviewed entries. In each model, to keep the ratio of primary cases to number of predictors high and avoid overfitting on a large number of uncorrelated features, we select the top 20 ProtNLM embeddings by Wilcoxon effect size between the relevant binary outcomes. In addition, we introduce a pair of features related to the likelihood of randomly drawing a missense substitution with the characteristics of a particular type of molecular mechanism. This heuristic metric is derived as the median of the ratio of ESM-1v score and the FoldX-computed ΔΔG of all missense mutations based on possible amino acid substitutions in the canonical sequence. We use both raw ΔΔG and its absolute value (|ΔΔG|), which has been shown to improve the performance of pathogenic variant prediction [30]. Finally, we complement existing population genetics constraint metrics with s_{het} [31], which relates the frequency of protein null mutations in a gene to the strength of selection against them, with higher values indicating stronger selection and lower tolerance to LOF mutations. We provide the list of all features used in the models and their detailed descriptions in **Table S2**.

Model design and data preprocessing

We opted for three separate binary classifiers rather than one multiclass predictor for several reasons. First, a multiclass approach would require seven classes to account for all mixed class combinations. The case numbers of the mixed classes are not sufficient for modelling, as they prohibit cross-validation and necessitate the use of fewer features to avoid overfitting [32]. Second, the prevalence of different mechanisms within the same gene is often unknown, making it difficult to weight them appropriately in a multiclass setting. Since many genes with a prevailing mechanism can have rare occurrences of other mechanisms, our current mixed-class groups are likely subject to a level of study bias, which would be unfavourably reflected in the model output. Third, binary models allow genes with mixed classes to contribute to more than one model, delegating the task of weighting mechanism propensities to the models. Thus, multiple binary models can essentially function like a multiclass predictor, without the data scarcity problems introduced by attempting a direct multiclass prediction [33].

In each model, we prioritise the first event level of the binary outcome. For example, in the DN *vs* LOF model, we treat any gene with a DN class annotation as DN even if it has other class assignments;

this strategy is likewise applied to the GOF *vs* LOF and the LOF *vs* non-LOF models (where non-LOF represents pooled DN and GOF gene sets). We next remove features that have a Spearman correlation >0.9 with another, which eliminates the gnomAD oe_lof metric (observed over expected ratio for predicted LOF variants in transcript) in all three models due to its high correspondence with s_{het}. The features are normalised and missing values are imputed based upon all other features with five nearest neighbours [34]. We use a distance matrix of the protein sequences to create a non-redundant dataset with proteins sharing <50% sequence identity within each outcome, e.g., within DN or LOF genes, but not across, as it may be important to learn differences between homologues assigned to different classes. To improve the signal-to-noise ratio, when possible, instead of randomly removing one protein from a pair of proteins above the sequence identity cutoff, we remove the one that overlaps with another class. This procedure results in the following class proportions across the models (protein count, class percentage): DN (498, 52%) *vs* LOF (451, 48%), GOF (479, 51%) *vs* LOF (463, 49%) and LOF (610, 52%) *vs* non-LOF (559, 48%).

Initial model screen

To assess which statistical learning approach suits the data and the classification problem the best, we carried out a model screen. We chose five diverse modelling methods (tuned hyperparameters): lasso logistic regression (penalty), multilayer perceptron (penalty, hidden units, epochs), polynomial support vector machine (SVM) (cost, degree, scale factor, margin), and two tree-based methods, random forest (trees, mtry [minimum number of randomly sampled features at each split], min_n [minimum number of data points a node must contain]) and the LightGBM algorithm (trees, tree depth, mtry, min_n, loss reduction) [35]. Maintaining class ratios, we split the data into 75% training and 25% test sets and ran a 10-fold cross-validation, similarly divided into 75% analysis and 25% assessment sets. An efficient grid search was performed via the ANOVA race tuning method [36] on a random grid of size k = 10×(number of hyperparameters), which eliminated configurations unlikely to have been the best after 3 resamples using repeated measures ANOVA. We compared the models by both the area under the receiver operating characteristics (AUROC) curve of the ANOVA race-winning configuration measured on the cross-validation folds as well as the AUROC of the k random grid configurations of each model on the test set (**Figure 1**). By the former, the top-ranking models were LightGBM for DN *vs* LOF and LOF *vs* non-LOF models and polynomial SVM for the GOF *vs* LOF model. However, because the k random grid configurations of the polynomial SVM consistently gave the best performance on the test sets, suggesting it performs well on unseen data without hyperparameter tuning, it was chosen to reduce the risk of overfitting and to maintain a similar prediction profile across the three models.

Model building

The final DN *vs* LOF, GOF *vs* LOF, and LOF *vs* non-LOF models are built using an SVM with a polynomial kernel. The data are divided into 75% training and 25% test sets with 3-times repeated 10-fold cross-validation, which itself is split into 75% analysis and 25% assessment sets. We perform Bayesian hyperparameter optimisation on the tunable parameters of the model: cost, degree, scale factor, and margin. Initially, a random grid of size 10 is created for the Gaussian process model and a

maximum of 20 search iterations are performed, with an early stop after 10 cycles if no improvement is observed. The models are finalised based upon the parameters that achieve the highest AUROC. Exact values of the SVM hyperparameters used in the models are provided in **Table S2**.

Global and local model feature importance assessment

Global feature importance is estimated by loss in AUROC upon removal of the feature in 10 permutations. Local influence of features on the predicted outcome for a protein is estimated via computing the features' Shapley values [37] in 20 permutations.

Model evaluation

We compute threshold-agnostic performance measures for the models based upon the test sets, including AUROC and area under the precision recall curves (AUPRC) (**Figure 2**). The consistency of AUROCs measured on the test sets relative to the cross-validation folds suggest the models are not overfitted (DN *vs* LOF: $AUROC_{CV}$ = 0.718, $AUROC_{test}$ = 0.71; GOF *vs* LOF: $AUROC_{CV}$ = 0.73, $AUROC_{test}$ = 0.763; LOF *vs* non-LOF: $AUROC_{CV}$ = 0.739, $AUROC_{test}$ = 0.763). To assess the models in further analyses, we determine the probability thresholds at which the models reach 50% sensitivity on their respective test sets. At these thresholds, which we refer to as t_{50}, the models are able to correctly identify half of the positive cases and approximately 80% of the negative cases. Threshold-dependent performance measures at t_{50}, including accuracy, Matthews correlation coefficient (MCC) and the F1 score, are summarised in **Table 1**. Biologically interpretable evaluation is performed by examining the models' capacity to identify emergent properties known to be associated with dominant molecular disease mechanisms [5,6]. Specifically, we evaluate 3 properties:

1) The energetic impact of pathogenic missense mutations in the structure, approximated by the FoldX-predicted Gibbs free energy of folding (ΔΔG). Missense mutations that map to residues with a predicted local distance difference test (pLDDT) value [38] less than 70 are excluded. This is because residues with low pLDDT are less likely to be ordered and thus ΔΔG is less interpretable for them.
2) The degree to which pathogenic missense mutations spatially cluster in the structure, approximated by the extent of disease clustering (EDC) metric. EDC is calculated as previously described [5], however, alpha carbon atoms with a pLDDT<70 are excluded from the calculation and only proteins with at least 5 pathogenic variants after this procedure are used for the analysis. This is because pathogenic missense mutations are enriched in structured regions of proteins [39], hence AlphaFold structures with high disorder could exhibit pseudo-clustering.
3) The performance of pathogenic missense variant prediction in a binary classification task, using two unsupervised (ESM1-v and EVE) and two supervised (MetaRNN and VARITY_R) VEPs, measured by AUROC. Only proteins with at least 5 pathogenic and 5 benign variants, and all variants mutually shared across the VEPs (that is, have predicted values), are considered for the analysis to increase per-protein AUROC confidence.

Statistical overrepresentation test for molecular function

Functional enrichment of proteins that are subsets of predicted DN or GOF proteins is performed by comparing the protein lists against each other via PANTHER [40] using Fisher's exact test and the Gene Ontology release 2023-05-10. We consider functions with a false discovery rate <1%, a sample size (number of proteins) >50, and a fold-enrichment of >1.5. To avoid redundancy, we select a single term with the largest sample size for groups of similar terms.

Results and discussion

Global and local feature importance evaluation

In total, 21 interpretable features were used in the models, including properties derived from protein sequences, structures, networks, and gene mutational constraints (Badonyi and Marsh 2023). Additionally, 20 language model-based embeddings were also included, which are thought to represent protein function in their latent space [41]. As a measure of feature importance, we calculated the loss in AUROC relative to the full model (**Figure 3**). While we cannot draw conclusions on why a particular ProtNLM embedding is useful for a given model, their relatively high ranks in the GOF *vs* LOF model may reflect that genes whose dominant disease mutations act via GOF are functionally more dissimilar or diverse than DN or LOF genes. In terms of the interpretable features, s_{het}, a metric that relates the frequency of LOF mutations in a gene to the strength of selection against them, ranks 1st and 3rd in the DN *vs* LOF and LOF *vs* non-LOF models, respectively. This emphasizes the importance of eliminating length-bias from selective constraint metrics to improve their power [31]. Moreover, the number of paralogues a protein has in the genome seems similarly important for the latter two models, ranking 3rd and 4th, respectively, consistent with our previous results [6]. Other notable features are gnomAD metrics pLI, the probability of the gene being intolerant to loss of function, and pRec, the probability of the gene being associated with recessive disease, which rank consistently high. This is unsurprising given that genes with high pLI should be enriched in genuinely haploinsufficient genes and those with non-LOF disease mechanisms will tend to have higher pRec values, i.e. are more “recessive-like”. Interestingly, the median ESM-1v/|ΔΔG| metric ranks 4th in the DN *vs* LOF model, suggesting that DN genes may possess a missense variant repertoire that is biased for functionally important but structurally less damaging effects.

We next looked at the worst and best predicted genes from the training sets of each model and calculated Shapley values for the features. These values indicate the average contributions of different feature orderings, with positive values suggesting a tendency to predict a protein towards the primary outcome and vice versa. In the training set of the DN *vs* LOF model, the DNA-binding protein SATB2 has the lowest probability of belonging to the DN class (pDN), while the type I cytoskeletal keratin 14 (KRT14) has the highest (**Figure 4**). Shapely values for SATB2 suggest that primarily s_{het}, but also, for example, pLI and cotranslational assembly have contributed to its low pDN. Indeed, SATB2 has a particularly high s_{het} (0.652, which is 5.2 standard deviations above the proteome mean), a pLI of 1, and has been found to cotranslationally assemble in a recent study [42], consistent with a decreased likelihood of observing its role in disease through a DN mechanism [6]. In our source data, evidence for SATB2 dominant negativity came from a report that describes a frameshift mutation in the gene likely to escape nonsense-mediated decay, whose clinical phenotype is more severe than glass syndrome, attributed to SATB2 haploinsufficiency [43]. However, SATB2 is recognised by the ClinGen review panel to have sufficient evidence for haploinsufficiency [44]. Thus, it is possible that SATB2 either represents a false positive case or, due to regression to the mean, the model fails to identify it as its features are too consistent with the secondary outcome. In contrast, Shapely values for KRT14 indicate congruence with known characteristics of DN proteins, e.g., the

high number of paralogues (68 in the proteome, only 1 in the training set). Furthermore, keratin disorders are considered a classical group of DN diseases [45], reinforcing the prediction of the model.

Similar conclusions can be drawn for the other two models. AF4/FMR2 family member 4 (AFF4), implicated in CHOPS syndrome, is a GOF case in the source data with the lowest pGOF value among genes of the training set (**Figure 4**). Three heterozygous missense mutations in its gene were recently found to cause GOF due to decreased clearance of the protein by the ubuiquitin proteasomal system, leading to transcriptional overactivation [46]. Since transcription factors are frequently haploinsufficient [47], their gene-level features in our data may generally align more with a LOF mechanism, as supported by the Shapley values of pLI and s_{het} for AFF4, which could hinder the identification of similar edge-cases. Interestingly, the worst predicted positive case by the LOF *vs* non-LOF model is apolipoprotein E (APOE) (**Figure 4**), which does lack sufficient evidence for dosage sensitivity in ClinGen. On further review, evidence in our source data revealed that a heterozygous null mutation of APOE had been linked to protection against a type of amyloidosis observed in Alzheimer's disease, rather than being implicated in disease causation [48]. Hence, the model is able to correctly down-weight false positive cases.

Proteome-scale molecular mechanism prediction

From the test sets we derived probability threshold t_{50} for each model, which represents a classification performance where approximately half of the cases belonging to the primary outcome and 80% of the secondary outcome are correctly predicted (**Table 1**). This threshold constitutes a relatively good balance between sensitivity and specificity while being more stringent than what is commonly considered the optimal threshold, e.g., minimum distance from the [0,1] corner of the ROC curve (dashed vertical line in **Figure 5A**), and being more lenient than the maximum positive predictive value. To allow the specification of different thresholds, we provide estimates of common threshold metrics in **Table S2** for every 0.01 increase in class probability. In **Figure 5B**, we show t_{50} in the context of model probability distributions for unique and overlapping cases in the test sets. Intriguingly, both the DN *vs* LOF (pDN) and the GOF *vs* LOF (pGOF) models assign higher probabilities to mixed LOF/GOF and LOF/DN classes, respectively, relative to the unique secondary outcomes, despite the fact that the models are blind to the overlapping classification of these cases. This suggests that these genes may have intermediary characteristics or exhibit a higher false positive rate for the secondary class in the ground truth data.

We chose a distribution-agnostic approach to classify the human proteome into a single molecular mechanism class. By ranking the class probabilities rather than comparing their raw values, the approach is less sensitive to differences in the properties of the models' class probability distributions. First, class probabilities are computed and ranked for all proteins and then each protein is assigned to the mechanism with the highest rank relative to the proteome, on condition that its class probability is above the t_{50} value. Based upon this strategy, in the 2023_02 UniProt reference proteome, 6,058 proteins are predicted to be DN, 5,287 GOF, and 2,580 LOF, with 6,440 proteins lacking classification.

It is important to emphasize that the models do not predict disease involvement, i.e. the above proportions should not be misconstrued as implying a twofold prevalence of DN over LOF proteins contributing to disease. Only dominant disease genes were included in training, so the models were not designed to distinguish disease-associated from non-disease proteins. Instead, they can be interpreted as representing the most likely molecular mechanism by which mutations in a given gene could cause disease, if it was a dominant disease gene. The high number of predicted DN and GOF proteins may be ascribed to a higher number of paralogous proteins in these mechanism classes relative to LOF [6], which results in a larger fraction of the proteome aligning with their characteristics, regardless of whether or not they actually cause disease when mutated. When we consider only known dominant disease genes, the proportion of proteins predicted to be DN, GOF, and LOF are 27%, 28%, and 44%, respectively.

In practical scenarios, researchers are often concerned with anticipating a LOF or an alternative molecular mechanism when encountering missense variants in a gene. In such a situation, an increased pDN or pGOF score would suggest that the gene's characteristics deviate from those typically linked to LOF genes, a notion that would be further supported by a low pLOF score. Although the models do not excel at identifying positive cases at an apparent sensitivity level of 50%, they are much better at identifying negative cases. Therefore when the researcher is confident that the dominant missense variants are linked to a particular phenotype, the higher-ranking molecular mechanism is still more likely to be correct.

Importantly, however, relative to haploinsufficiency predictors introduced in previous studies [49–51] (comparisons with pLOF in **Figure S1**), our models are unique in that they were trained on strictly autosomal dominant genes and specifically designed to tell apart alternative molecular mechanisms from that of LOF. Thus, we expect them to be much more useful than haploinsufficiency predictors for the interpretation of missense variant effects, as they more accurately represent the properties of the underlying mechanisms. We provide the class probabilities, the rank-based classification results with and without t_{50} inclusion, and the labels for training set genes in **Table S3** for all human proteins (n = 20,365).

Biologically and clinically relevant validation of the models

We wished to test whether the models are able to recapitulate known properties of dominant molecular mechanisms [5]. These properties include the energetic impact of missense mutations in the protein structure, their degree of spatial clustering, and their current predictability by state-of-the-art VEPs (see **Methods**). None of these properties were explicitly used for training, thus making them simple and powerful means to assess the usefulness of the predictions. We created an unbiased analysis set where we removed genes used for training and those associated with autosomal recessive inheritance. We first looked at the predicted ΔΔG of pathogenic missense mutations across the different predicted classes in this dataset (**Figure 6A**). The results suggest mutations in DN and GOF proteins are significantly less damaging than in LOF proteins. This can be explained by the idea that destabilisation is one of the signature mechanisms of LOF mutations [5]. By contrast, GOF mutations should not be too damaging in order to alter a function, and most DN and many GOF genes

are in fact assembly-mediated [2]. Thus, the effect of their mutations should not preclude protein complex assembly, which makes them necessarily less damaging.

Next, we looked at how much pathogenic missense mutations cluster in the structures of proteins predicted to be DN or GOF. Our prior expectation is that both classes should exhibit higher degree of clustering than LOF, as measured by the EDC metric. We found this assumption to hold up, with both DN and GOF proteins having significantly higher clustering values than LOF proteins (**Figure 6B**). This observation agrees with the concept that LOF mutations are generally more sparsely distributed in the protein structure, but non-LOF mutations tend to be concentrated at protein interfaces and functional sites [5].

Lastly, we took advantage of an important bottleneck of contemporary VEPs, which is that they less well predict pathogenic missense mutations associated with DN and GOF genes [5]. In **Figure 6C**, we show a per-protein aggregated AUROC analysis of pathogenic *vs* benign missense variants, evaluated by two unsupervised (ESM1-v and EVE) and two supervised (MetaRNN and VARITY_R) VEPs, which we recently showed to be the top performing VEPs of their category [20]. Expectedly, missense mutations in proteins classified as LOF are much better predicted than those in DN or GOF proteins.

Importantly, the above trends are observed with the raw rank-based classification results without the t_{50} probability cutoff (**Figure S2**), demonstrating that they are not an artefact of a careful threshold selection. We conclude that the models effectively reproduce biologically and clinically relevant properties of dominant molecular mechanisms.

The functional landscape of predicted DN and GOF proteins

Finally, given that pDN and pGOF class probabilities are positively correlated by design (**Figure 7**), we decided to explore the molecular functions associated with proteins predicted exclusively as DN or GOF using statistical overrepresentation tests. It is well-established that DN effects are common in homomers [6,52], which are proteins that form complexes with copies of themselves. Correspondingly, we identified functions closely related to homomeric symmetry groups in exclusively DN proteins, including “organic cyclic compound binding”, “oxidoreductase activity”, and “hydrolase activity” [53]. Interestingly, a strongly enriched, large group of proteins possess “nucleic acid binding” function, potentially indicating a higher prevalence of DN effects among transcription factors and other DNA/RNA-binding proteins. This underscores the importance, and difficulty, of distinguishing between mutations that act through DN effects versus haploinsufficiency. In contrast, exclusively GOF proteins exhibit functions less directly associated with protein complexes but more susceptible to the impact of overactivation events. These functions include “molecular transducer activity”, “signaling receptor activity”, “kinase activity”, “enzyme regulator activity”, and “phosphotransferase activity”.

Altogether, these findings reinforce that the predictions are consistent with known and expected properties of non-LOF molecular mechanisms. Thus, they can be used to test hypotheses about the functional attributes of the proteins associated with the molecular mechanisms, which could facilitate the discovery of novel features for the next generation of VEPs.

Conclusion

We constructed three binary classifiers to predict DN, GOF, and LOF dominant molecular disease mechanisms by substituting structural and functional features of protein-coding genes with limited coverage. These models have a similar performance to our previous non-LOF *vs* LOF classifier, which was based upon such limited but interpretable features. We combined the models' output into a single class prediction using a proteome rank-based approach and examined the properties of proteins assigned to the different molecular mechanisms. Our analyses provide strong evidence that the models reproduce the properties linked to molecular mechanisms at systems level. Notably, our findings underscore the predictability of pathogenic mutations in DN and GOF proteins – a problem recognised to be more challenging than in LOF proteins. This observation was consistently upheld in our unbiased analysis set, further stressing the need for diversifying the predictive capabilities of future VEPs. We have made available the predictions for the human reference proteome, and we hope researchers will use the data to prioritise novel or challenging variants and test hypotheses about the functional roles fulfilled by these proteins. It is important to emphasize that the models do not predict the likelihood of disease involvement by a gene. Instead, a non-disease associated gene with a predicted mechanism implies that the protein's properties are most consistent with the given mechanism, which we expect will be particularly useful for researchers seeking to determine the molecular mechanisms underlying novel disease mutations.

Funding

This project was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101001169). MB was supported by the Biotechnology and Biological Sciences Research Council EASTBIO Doctoral Training Programme (BB/M010996/1). JAM is a Lister Institute Research Prize Fellow.

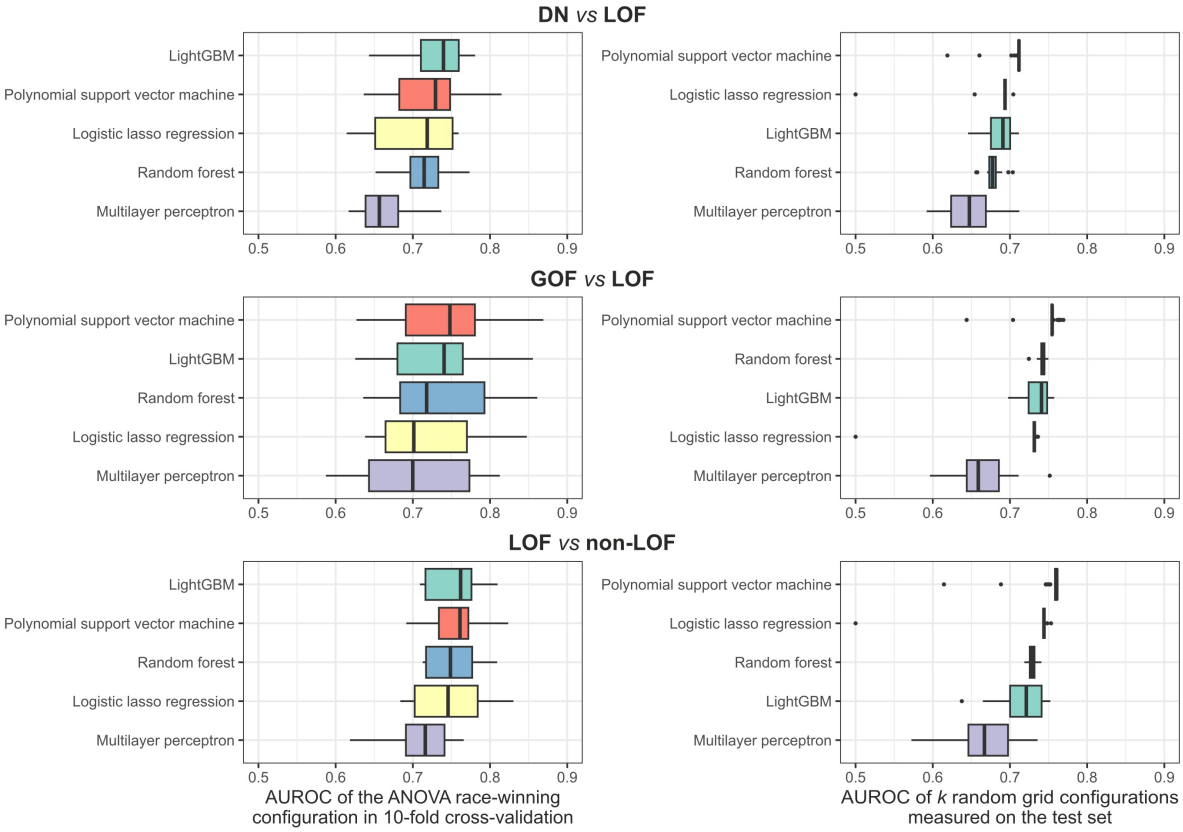
Acknowledgement

We thank Benjamin J. Livesey for generating the VEP scores and providing helpful comments on the manuscript.

Data availability

Data and code to build the models and run the analyses are available at https://osf.io/z4dcp/.

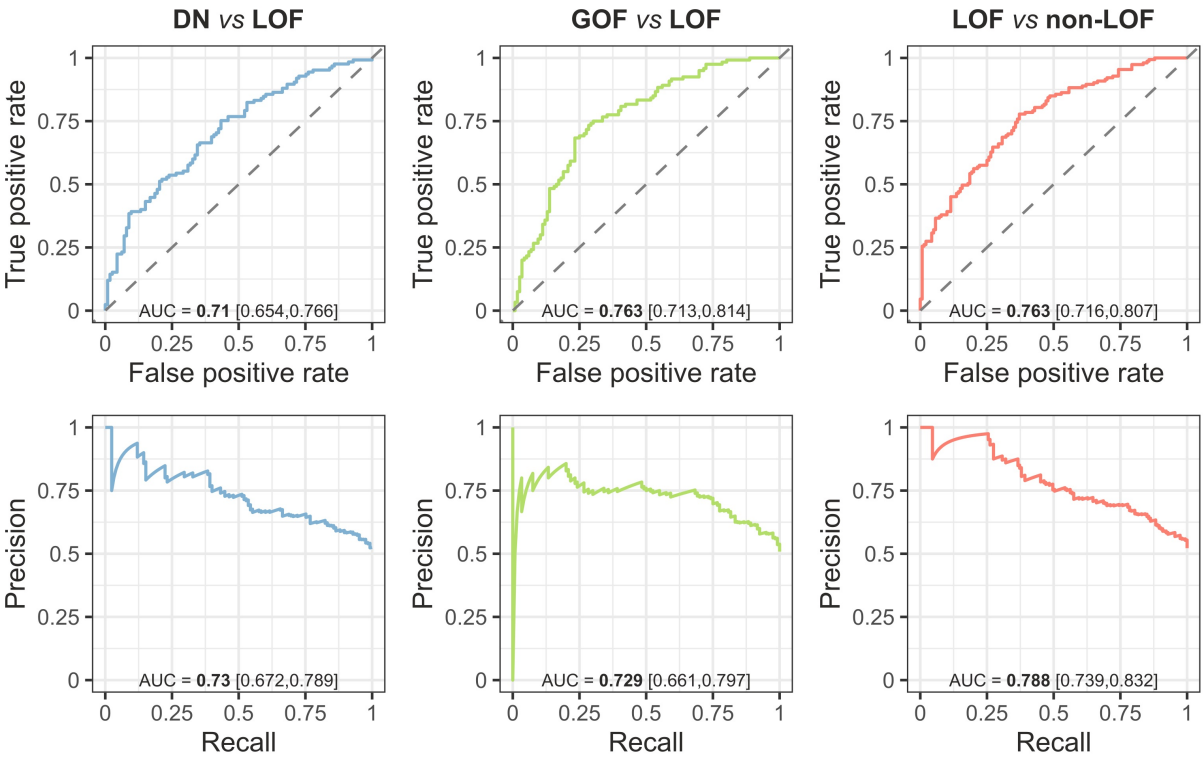
Figure 1



Results of the initial model screen.

Plots on the left show the AUROC of the best hyperparameter set of each model on the cross-validation folds. Plots on the right show the AUROC of *k* randomly generated parameter combinations as measured on the test sets, where *k* = 10×(number of hyperparameters). Boxes denote data within 25th and 75th percentiles and the middle line represents the median. Whiskers extend from the box to 1.5× the interquartile range.

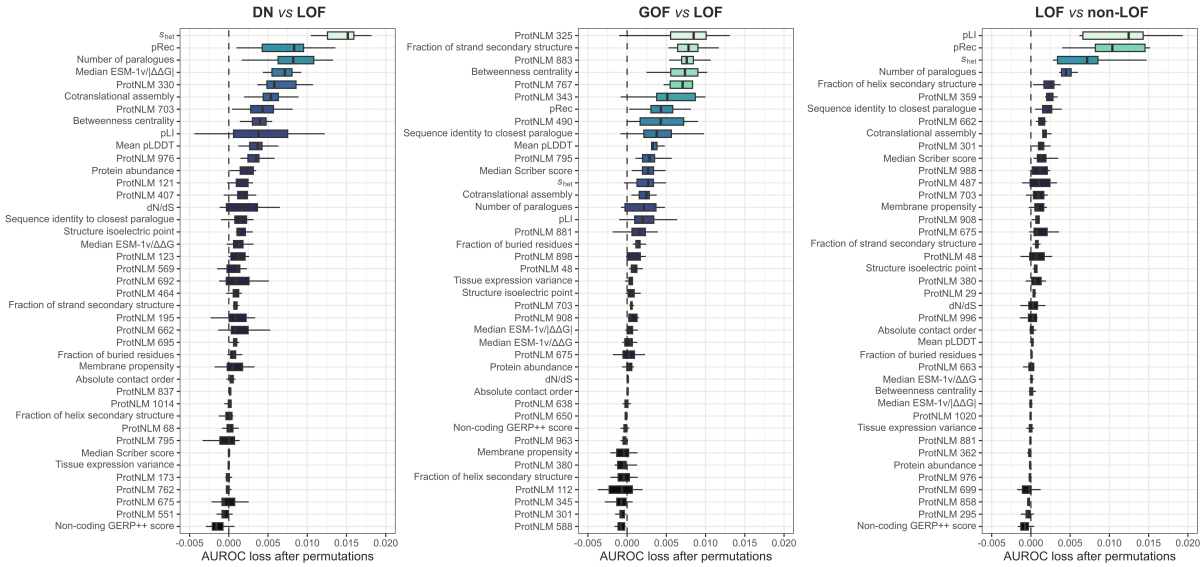
Figure 2



ROC and PR curves of the models measured on the test sets.

Numbers in bold denote point estimates and those in brackets represent 90% bootstrap confidence intervals.

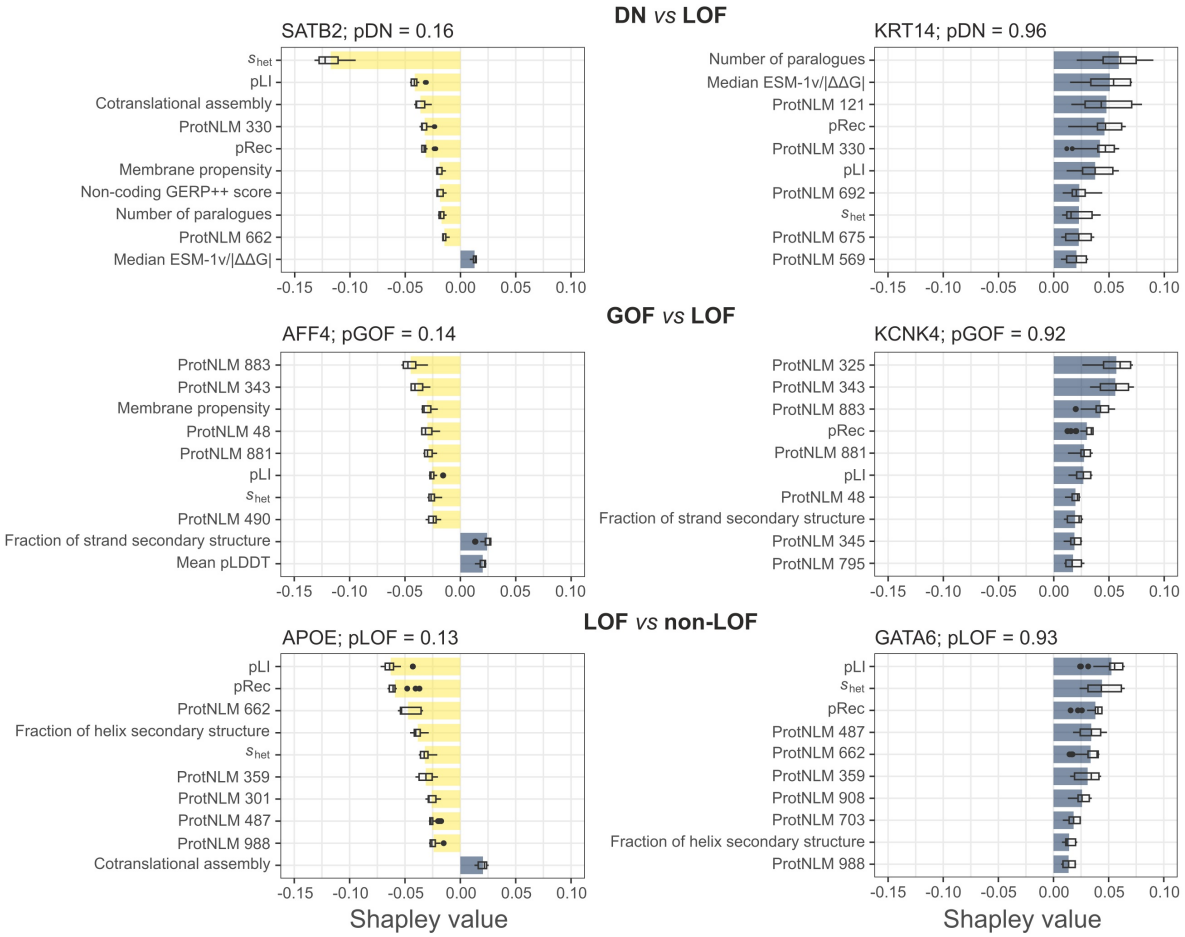
Figure 3



Feature importance of the models.

Feature importance is estimated by loss in AUROC upon removal of the feature from the model in 10 permutations. Boxes denote data within 25th and 75th percentiles and the middle line represents the median. Whiskers extend from the box to 1.5× the interquartile range. Dashed line is at AUROC loss = 0.

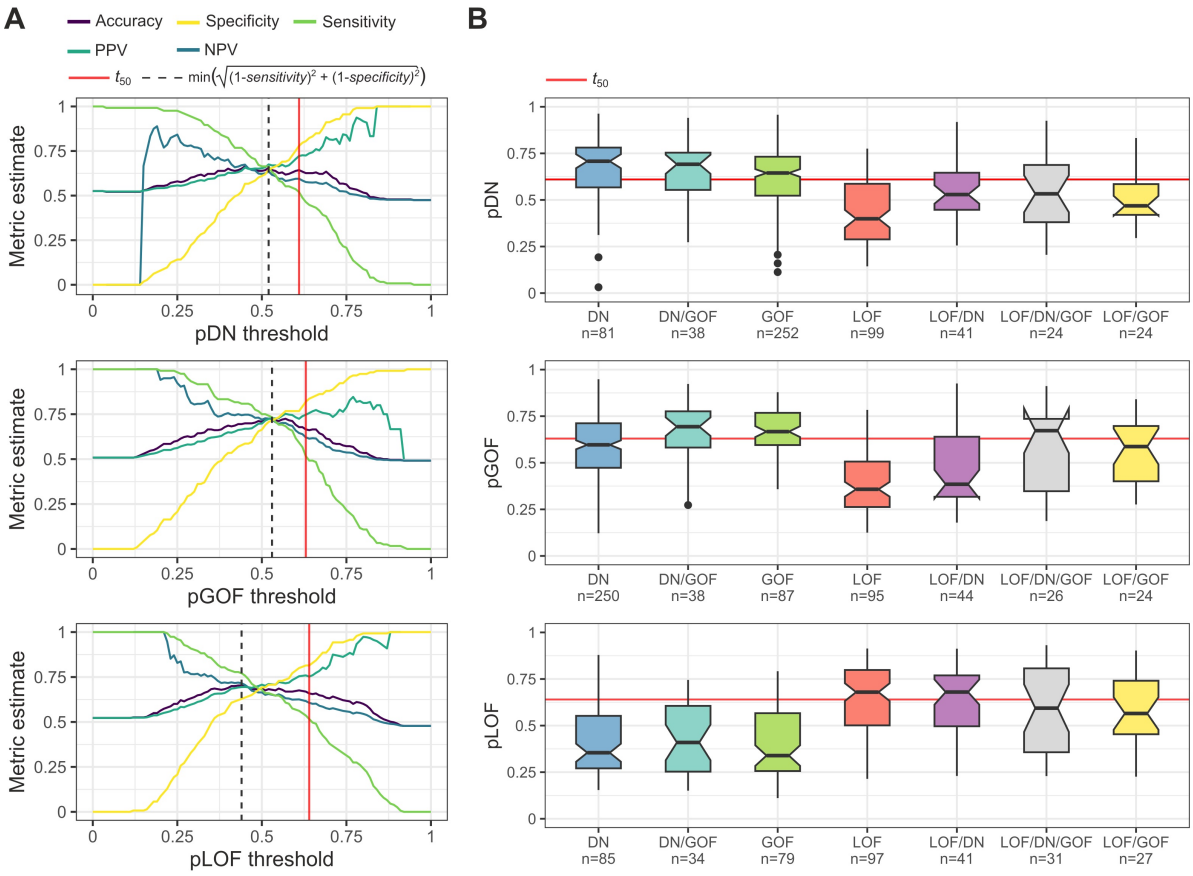
Figure 4



Local interpretations of model class probabilities.

Plots show the worst (left) and best (right) predicted genes of the training sets and Shapley values of the top 10 most influential features in 20 permutations. Bars show the mean, boxes denote data within 25th and 75th percentiles and the middle line is the median. Whiskers extend from the box to 1.5x the interquartile range.

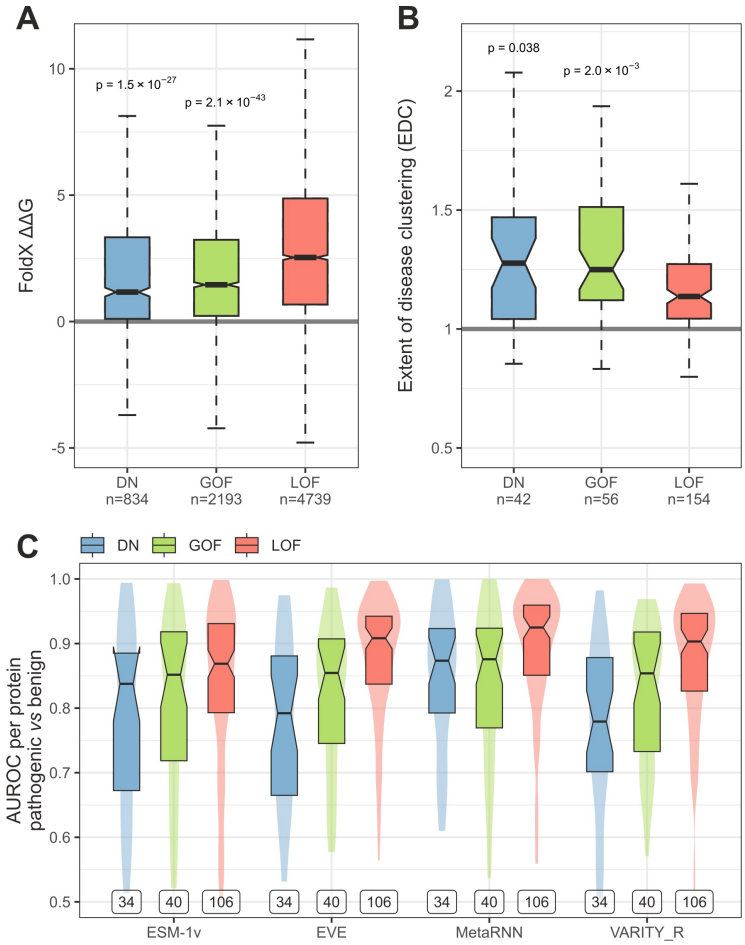
Figure 5



Threshold plots and test set class probabilities.

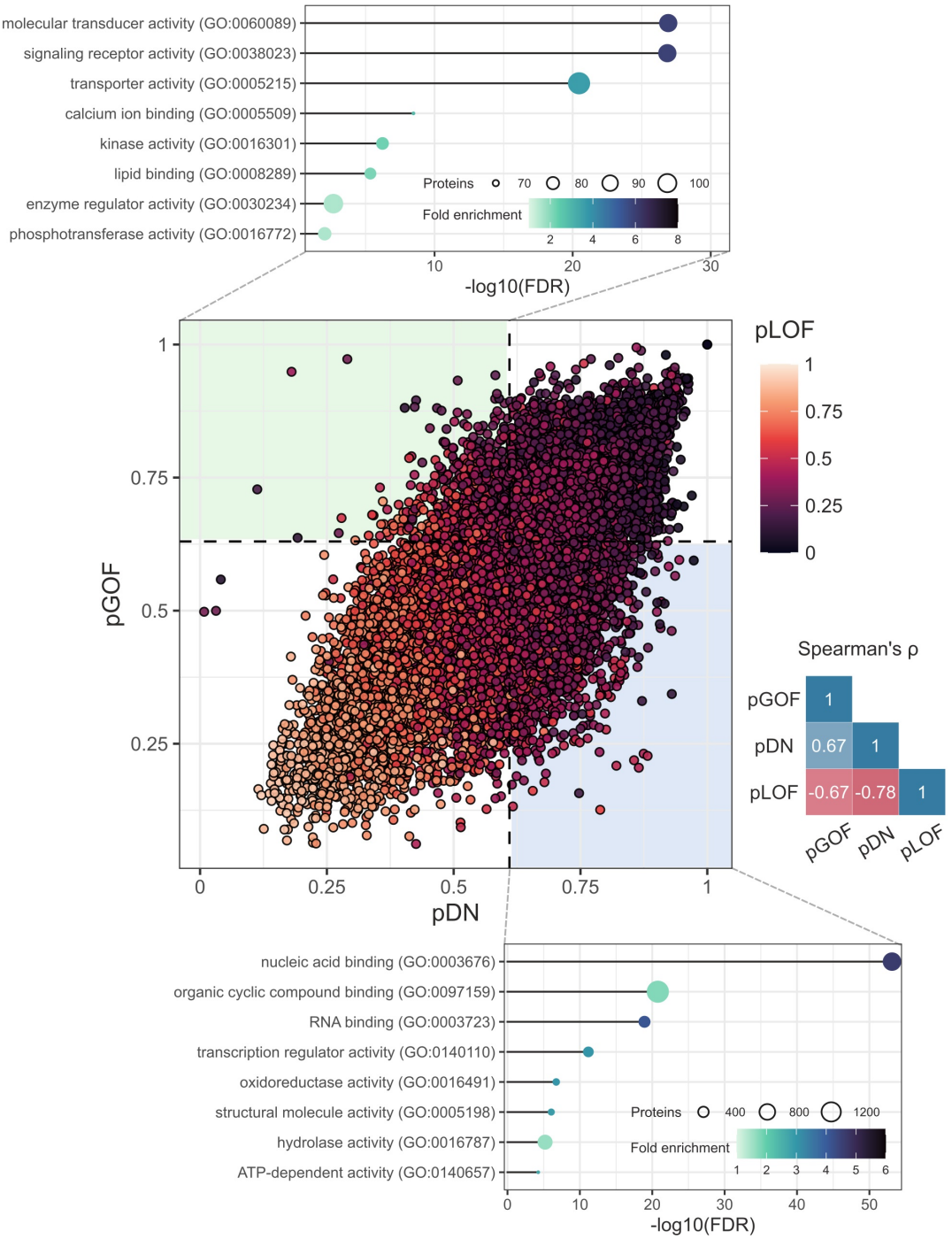
(**A**) Commonly used threshold metrics derived from the test sets for pDN, pGOF, and pLOF. NPV and PPV denote negative and positive predictive values, respectively. Estimates for these metrics are provided in **Table S2** for every 0.01 increase in class probability. (**B**) Model probabilities mapped to the test sets, classified by their ground truth classes, both unique and overlapping. Sample sizes indicate the number of proteins. Boxes denote data within 25th and 75th percentiles and the middle line represents the median. Whiskers extend from the box to 1.5x the interquartile range and the notch contains the 95% confidence interval of the median. Dashed lines run at the t_{50} probability thresholds.

Figure 6



Validation of the models through model-independent metrics on an unbiased analysis set. (**A**) FoldX-predicted ΔΔG of pathogenic missense mutations. Numbers below classes denote the number of mutations. Holm-Bonferroni corrected p-values above DN and GOF boxes are relative to the LOF group and were determined by one-sided Wilcoxon rank-sum test. Sample sizes indicate the number of variants. (**B**) Class probabilities of the analysis set *vs* EDC. Sample sizes indicate the number of proteins in each class. Holm-Bonferroni corrected p-values above DN and GOF boxes are relative to the LOF group and were determined by one-sided Wilcoxon rank-sum test. (**C**) Aggregated AUROC analysis of pathogenic *vs* benign variants in predicted molecular mechanism classes. Labels indicate the number of proteins in each class. Boxes denote data within 25th and 75th percentiles, the middle line represents the median and the notch contains the 95% confidence interval of the median. Violins show area-normalized distributions.

Figure 7



The functional landscape of proteins predicted to be exclusively DN or GOF.

Middle: pDN *vs* pGOF scatter plot coloured by pLOF. Correlation triangle on the right shows Spearman correlations between class probabilities. *Top*: The lollipop chart shows the most enriched molecular functions for genes predicted to be GOF below the t_{50} threshold for pDN (exclusively GOF). *Bottom*: Likewise, the chart shows the most enriched functions for DN genes below the t_{50} threshold for pGOF (exclusively DN). The *x*-axis shows the negative $\log_{10}$ false discovery rate.

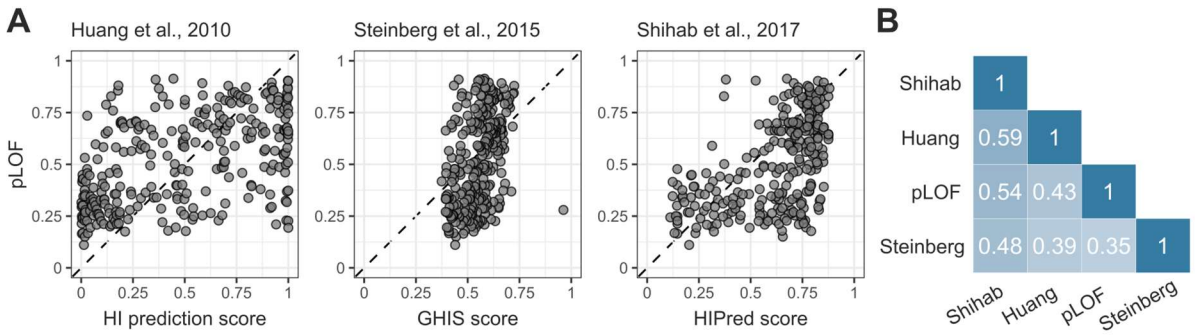
Table 1

Model	t_{50}	Sensitivity	Specificity	Accuracy	MCC	F1
DN *vs* LOF	0.61	**0.519** [0.448,0.584]	**0.778** [0.717,0.841]	**0.642** [0.592,0.693]	**0.306** [0.208,0.408]	**0.603** [0.539,0.664]
GOF *vs* LOF	0.63	**0.519** [0.45,0.592]	**0.818** [0.759,0.879]	**0.666** [0.619,0.712]	**0.353** [0.258,0.449]	**0.612** [0.547,0.676]
LOF *vs* non-LOF	0.64	**0.519** [0.451,0.588]	**0.815** [0.764,0.871]	**0.66** [0.618,0.703]	**0.348** [0.263,0.435]	**0.614** [0.556,0.672]

Threshold-dependent performance metrics.

Metrics derived for the probability thresholds at which the models have approximately 50% specificity on the test sets (t_{50}). Numbers in bold denote point estimates and those in brackets are 90% bootstrap confidence intervals.

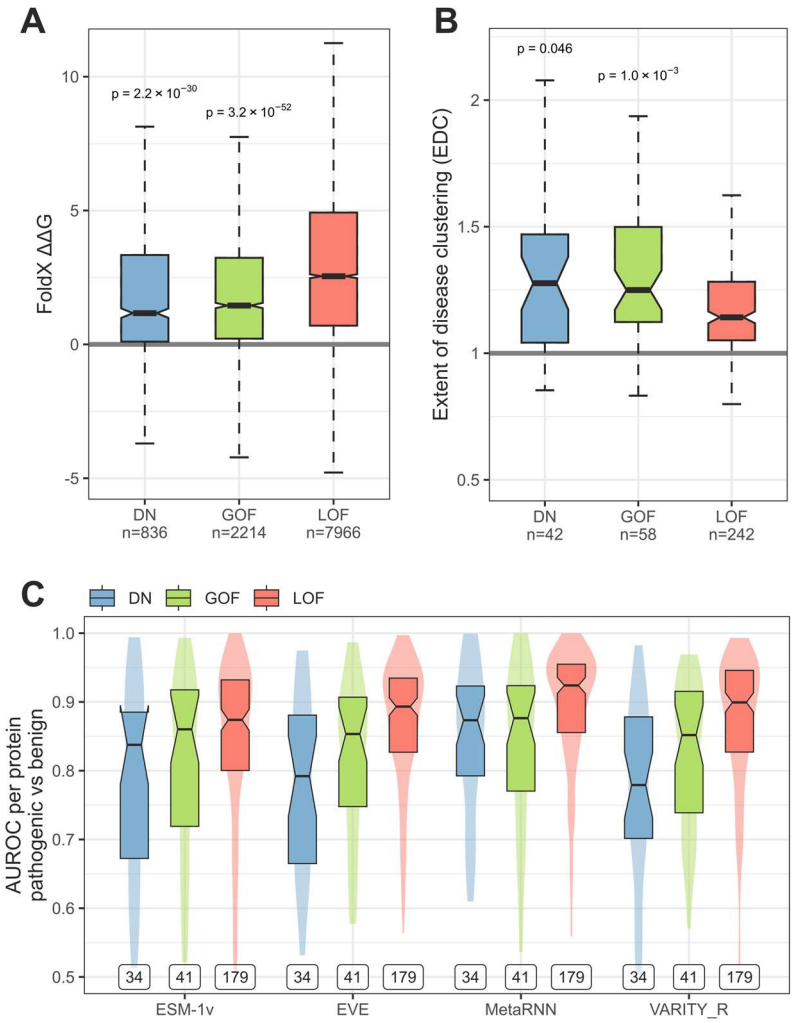
Figure S1



Comparison of pLOF with haploinsufficiency predictors.

(**A**) Scatter plots of haploinsufficiency scores from the indicated studies *vs* pLOF, assessed on the LOF *vs* non-LOF test set. pLOF has the highest agreement with the Shihab *et al*. predictions (Pearson's r = 0.54, $p = 4.5\times10^{-24}$) and the lowest with those of Steinberg *et al*. (r = 0.35, $p = 6.9\times10^{-19}$). (**B**) Pearson correlation triangle showing mutually shared values of all four metrics on human genes (n = 15,046). Although considerable variation exists even among haploinsufficiency predictors, low correlation with pLOF is expected, because haploinsufficiency predictors were not exclusively trained on dominant genes and their negative class comprised of haplo-sufficient genes of mostly no disease relevance. By contrast, pLOF values from the LOF *vs* non-LOF model were obtained by training the model to recognise haploinsufficient genes against a background of strictly dominant genes with molecular mechanisms other than simple LOF.

Figure S2



Supplemental analysis to Figure 6: Validation of the models through model-independent metrics on an unbiased analysis set, using rank-based classification without t_{50} cutoff.

(**A**) FoldX-predicted ΔΔG of pathogenic missense mutations. Numbers below classes denote the number of mutations. Holm-Bonferroni corrected p-values above DN and GOF boxes are relative to the LOF group and were determined by one-sided Wilcoxon rank-sum test. Sample sizes indicate the number of variants. (**B**) Class probabilities of the analysis set vs EDC. Sample sizes indicate the number of proteins in each class. Holm-Bonferroni corrected p-values above DN and GOF boxes are relative to the LOF group and were determined by one-sided Wilcoxon rank-sum test. (**C**) Aggregated AUROC analysis of pathogenic *vs* benign variants in predicted molecular mechanism classes. Labels indicate the number of proteins in each class. Boxes denote data within 25th and 75th percentiles, the middle line represents the median and the notch contains the 95% confidence interval of the median. Violins show area-normalized distributions.

References

1. Veitia RA. Exploring the etiology of haploinsufficiency. BioEssays. 2002;24: 175–184. doi:10.1002/bies.10023

2. Backwell L, Marsh JA. Diverse Molecular Mechanisms Underlying Pathogenic Protein Mutations: Beyond the Loss-of-Function Paradigm. Annual review of genomics and human genetics. 2022;23. doi:10.1146/ANNUREV-GENOM-111221-103208

3. Fowler DM, Adams DJ, Gloyn AL, Hahn WC, Marks DS, Muffley LA, et al. An Atlas of Variant Effects to understand the genome at nucleotide resolution. Genome Biology. 2023;24: 147. doi:10.1186/s13059-023-02986-x

4. Fomin A, Gärtner A, Cyganek L, Tiburcy M, Tuleta I, Wellers L, et al. Truncated titin proteins and titin haploinsufficiency are targets for functional recovery in human cardiomyopathy due to TTN mutations. Science Translational Medicine. 2021;13: eabd3079. doi:10.1126/scitranslmed.abd3079

5. Gerasimavicius L, Livesey BJ, Marsh JA. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. Nature Communications 2022 13:1. 2022;13: 1–15. doi:10.1038/s41467-022-31686-6

6. Badonyi M, Marsh JA. Buffering of genetic dominance by allele-specific protein complex assembly. Science Advances. 2023;9: eadf9845. doi:10.1126/sciadv.adf9845

7. Bayrak CS, Stein D, Jain A, Chaudhary K, Nadkarni GN, Vleck TTV, et al. Identification of discriminative gene-level and protein-level features associated with pathogenic gain-of-function and loss-of-function variants. The American Journal of Human Genetics. 2021;108: 2301–2318. doi:10.1016/j.ajhg.2021.10.007

8. Flanagan SE, Patch A-M, Ellard S. Using SIFT and PolyPhen to Predict Loss-of-Function and Gain-of-Function Mutations. Genetic Testing and Molecular Biomarkers. 2010;14: 533–537. doi:10.1089/gtmb.2010.0036

9. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596: 590–596. doi:10.1038/s41586-021-03828-1

10. Agirre J, Atanasova M, Bagdonas H, Ballard CB, Baslé A, Beilsten-Edmands J, et al. The CCP4 suite: integrative software for macromolecular crystallography. Acta Cryst D. 2023;79: 449–461. doi:10.1107/S2059798323003595

11. Delgado J, Radusky LG, Cianferoni D, Serrano L. FoldX 5.0: Working with RNA, small molecules and a new graphical interface. Bioinformatics. 2019;35: 4168–4169. doi:10.1093/bioinformatics/btz184

12. Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. Bioinformatics. 2019;35: i343–i353. doi:10.1093/bioinformatics/btz324

13. Zhao B, Katuwawala A, Oldfield CJ, Dunker AK, Faraggi E, Gsponer J, et al. DescribePROT: database of amino acid-level protein structure and function predictions. Nucleic Acids Research. 2021;49: D298–D308. doi:10.1093/nar/gkaa931

14. Gane, A., Bileschi, M.L., Dohan, D., Speretta, E., Héliou, A., Meng-Papaxanthos, L., et al. ProtNLM: Model-based Natural Language Protein Annotation. 2022. Available: https://storage.googleapis.com/brain-genomics-public/research/proteins/protnlm/uniprot_2022_04/protnlm_preprint_draft.pdf

15. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011;7: 539. doi:10.1038/msb.2011.75

16. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. bioRxiv; 2021. p. 2021.07.09.450648. doi:10.1101/2021.07.09.450648

17. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599: 91–95. doi:10.1038/s41586-021-04043-8

18. Li C, Zhi D, Wang K, Liu X. MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. Genome Medicine. 2022;14: 115. doi:10.1186/s13073-022-01120-z

19. Wu Y, Liu H, Li R, Sun S, Weile J, Roth FP. Improved pathogenicity prediction for rare human missense variants. The American Journal of Human Genetics. 2021;108: 1891–1906. doi:10.1016/j.ajhg.2021.08.012

20. Livesey BJ, Marsh JA. Updated benchmarking of variant effect predictors using deep mutational scanning. Molecular Systems Biology. 2023;19: e11474. doi:10.15252/msb.202211474

21. R Core Team. R core team. R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria URL http://www R-project org. 2023.

22. Jung K, Lee J, Gupta V, Cho G. Comparison of Bootstrap Confidence Interval Methods for GSCA Using a Monte Carlo Simulation. Frontiers in Psychology. 2019;10. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02215

23. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Research. 2018;46: D1062–D1067. doi:10.1093/nar/gkx1153

24. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10: giab008. doi:10.1093/gigascience/giab008

25. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581: 434–443. doi:10.1038/s41586-020-2308-7

26. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17: 122. doi:10.1186/s13059-016-0974-4

27. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. Nucleic Acids Research. 2015;43: D789–D798. doi:10.1093/nar/gku1205

28. Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Research. 2021;49: D394–D403. doi:10.1093/nar/gkaa1106

29. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. Journal of Molecular Biology. 1987;196: 641–656. doi:10.1016/0022-2836(87)90038-6

30. Gerasimavicius L, Liu X, Marsh JA. Identification of pathogenic missense mutations using protein stability predictors. Sci Rep. 2020;10: 15387. doi:10.1038/s41598-020-72404-w

31. Zeng T, Spence JP, Mostafavi H, Pritchard JK. Bayesian estimation of gene constraint from an evolutionary model with gene features. bioRxiv; 2023. p. 2023.05.19.541520. doi:10.1101/2023.05.19.541520

32. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368: m441. doi:10.1136/bmj.m441

33. Lorena AC, de Carvalho ACPLF, Gama JMP. A review on the combination of binary classifiers in multiclass problems. Artif Intell Rev. 2009;30: 19. doi:10.1007/s10462-009-9114-9

34. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics. 1971;27: 857–871. doi:10.2307/2528823

35. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017. Available: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

36. Kuhn M. Futility Analysis in the Cross-Validation of Machine Learning Models. arXiv; 2014. doi:10.48550/arXiv.1405.6974

37. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. pp. 4768–4777.

38. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589. doi:10.1038/s41586-021-03819-2

39. Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A. The structural coverage of the human proteome before and after AlphaFold. PLOS Computational Biology. 2022;18: e1009818. doi:10.1371/journal.pcbi.1009818

40. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol Update for Large-scale genome and gene function analysis with PANTHER Classification System (v.14.0). Nat Protoc. 2019;14: 703–721. doi:10.1038/s41596-019-0128-8

41. Ziegler C, Martin J, Sinner C, Morcos F. Latent generative landscapes as maps of functional diversity in protein sequence space. Nat Commun. 2023;14: 2222. doi:10.1038/s41467-023-37958-z

42. Bertolini M, Fenzl K, Kats I, Wruck F, Tippmann F, Schmitt J, et al. Interactions between nascent proteins translated by adjacent ribosomes drive homomer assembly. Science. 2021;371: 57–64. doi:10.1126/science.abc7151

43. Boone PM, Chan YM, Hunter JV, Pottkotter LE, Davino NA, Yang Y, et al. Increased bone turnover, osteoporosis, progressive tibial bowing, fractures, and scoliosis in a patient with a final-exon SATB2 frameshift mutation. American Journal of Medical Genetics Part A. 2016;170: 3028–3032. doi:10.1002/ajmg.a.37847

44. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen — The Clinical Genome Resource. New England Journal of Medicine. 2015;372: 2235–2242. doi:10.1056/nejmsr1406261

45. McLean WHI, Moore CBT. Keratin disorders: from gene to therapy. Human Molecular Genetics. 2011;20: R189–R197. doi:10.1093/hmg/ddr379

46. Izumi K, Nakato R, Zhang Z, Edmondson AC, Noon S, Dulik MC, et al. Germline gain-of-function mutations in AFF4 cause a developmental syndrome functionally linking the super elongation complex and cohesin. Nat Genet. 2015;47: 338–344. doi:10.1038/ng.3229

47. Seidman JG, Seidman C. Transcription factor haploinsufficiency: when half a loaf is not enough. J Clin Invest. 2002;109: 451–455. doi:10.1172/JCI15043

48. Kim J, Jiang H, Park S, Eltorai AEM, Stewart FR, Yoon H, et al. Haploinsufficiency of Human APOE Reduces Amyloid Deposition in a Mouse Model of Amyloid-β Amyloidosis. J Neurosci. 2011;31: 18007–18012. doi:10.1523/JNEUROSCI.3773-11.2011

49. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and Predicting Haploinsufficiency in the Human Genome. Schierup MH, editor. PLoS Genet. 2010;6: e1001154. doi:10.1371/journal.pgen.1001154

50. Shihab HA, Rogers MF, Campbell C, Gaunt TR. HIPred: an integrative approach to predicting haploinsufficient genes. Bioinformatics. 2017;33: 1751–1757. doi:10.1093/bioinformatics/btx028

51. Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. Nucleic Acids Research. 2015;43: e101. doi:10.1093/nar/gkv474

52. Veitia RA. Exploring the molecular etiology of dominant-negative mutations. Plant Cell. 2007;19: 3843–3851. doi:10.1105/tpc.107.055053

53. Bergendahl LT, Marsh JA. Functional determinants of protein assembly into homomeric complexes. Sci Rep. 2017;7: 4932. doi:10.1038/s41598-017-05084-8