PreMode predicts mode of action of missense variants by deep graph representation learning of protein sequence and structural context

Guojie Zhong[1,2], Yige Zhao[1,2], Demi Zhuang[1], Wendy K Chung[3], Yufeng Shen[1,4,5]

[1]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY 10032
[2]The Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University, New York, NY 10032
[3]Department of Pediatrics, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115
[4]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032
[5]JP Sulzberger Columbia Genome Center, Columbia University, New York, NY 10032

Abstract
Accurate prediction of the functional impact of missense variants is fundamentally important for disease gene discovery, clinical genetic diagnostics, therapeutic strategies, and protein engineering. Previous efforts have focused on predicting a binary pathogenicity classification, but the functional impact of missense variants is multi-dimensional. Pathogenic missense variants in the same gene may act through different modes of action (i.e., gain/loss-of-function) by affecting multiple protein biochemical properties. They may result in distinct clinical conditions that require different treatments. We developed a new method, PreMode, to perform gene-specific mode-of-action predictions. PreMode models effects of coding sequence variants using SE(3)-equivariant graph neural networks on protein sequences and structures. Using the largest-to-date set of mode-of-action-labeled missense variants, we show that PreMode reaches state-of-the-art performance in multiple types of mode-of-action predictions by efficient transfer-learning. Additionally, PreMode prediction of G/LoF variants in a kinase is consistent with inactive-active conformation transition. Finally, we show that PreMode enables improved mutagenesis analysis, clinical diagnosis and more broadly, artificial GoF engineering of proteins.

Main

Accurate and comprehensive prediction of variant effects has been a long-standing fundamental problem in genetics and protein biology. Single amino acid (missense) variants are the most common type of coding variants that contribute to many human diseases and conditions[1-5]. The functional impact of most missense variants remains uncertain. At the molecular level, missense variants in only 40 human genes have been screened in saturated mutagenesis experiments[6]. At the genetic level, only about 2% of clinically observed missense variants are classified as pathogenic or benign, while the majority remain of uncertain clinical significance[7]. Such limitations make it challenging for accurate clinical diagnosis and timely clinical interventions. Furthermore, understanding the functional impact of mutations is important to protein engineering, especially in directed evolution methods, where proteins are iteratively mutated to optimize function or fitness[8]. Such efforts were often limited by the high cost and explosion of sequence space. It remains a challenge to understand and predict the fitness landscape of mutants to reduce the search space and improve the efficiency of engineering[9,10].

In the past decade, many computational methods have been developed (such as Polyphen2[11], SIFT[12], CADD[13], REVEL[14], MPC[15], EVE[16], PrimateAI[17,18], gMVP[19], AlphaMissense[20]) to predict variant effects in a binary manner aiming at distinguishing pathogenic and benign variants. These methods showed that pathogenicity can be predicted by manually encoded or self-learned features including sequence conservation, protein secondary structures, and population allele frequency. Moreover, recently developed methods based on protein language models, leveraging Transformer architectures and self-supervised training on billions of protein sequences from UniProt[21], have demonstrated their capability to serve as versatile predictors of various protein features[22-24]. The embeddings from these models can offer zero-shot predictive potential for variant pathogenicity[25,26]. While those methods are helpful in genetic analyses, it remains a challenging problem in clinical application. Predicting mode of action is also critically important in certain applications.

We use "mode of action" as a generic term to encapsulate the multi-dimensional molecular and genetic mechanisms through which pathogenic variants impact protein functionality and increase the risk of diseases, respectively. More specifically, molecularly, pathogenic variants can change the biochemical properties of a protein in different ways. For example, decreasing/maintaining protein stability[27,28] or enzymatic activity[28,29]. Generically, variants can be categorized into two major types, gain or loss of function (G/LoF). For example, GoF variants encompass alterations that augment ion transfer activity, modify ion channel selectivity in ion channels[30,31], confer constitutive activation, increased sensitivity, and reduce specificity in signaling proteins[32]. Likewise, LoF variants can damage protein function in diverse ways, including diminishing net ion flow in ion channel genes[30], or eliminating function or destroying interaction domains[28,29,33-35]. Gain and loss of function variants usually result in markedly different clinical phenotypes[30,36-39], necessitating entirely distinct therapeutic approaches[30,33,40,41].

While numerous methods have demonstrated the potential to predict pathogenicity on a genome-wide scale, the effort in G/LoF prediction has been limited. Stein et al attempted to

predict genome-wide G/LoF variants via assembly of human curated features[42]. However, we note that mode of action centers around how a variant disrupts the normal function of a protein. Given the inherent diversity of protein functions, attempting to define a universally applicable predictive task for all G/LoF variants across all proteins could lead to conceptual ambiguity. Therefore, we propose that such predictive tasks should be defined within the context of individual proteins or protein families that share similar functions, like what Heyne et al did in predicting G/LoF variants in Na+/Ca2+ ion channel genes[30]. The main challenge is the limited availability of data for most genes and protein families.

We developed a new method, PreMode (Pretrained Model for Predicting Mode of action), to address these challenges with deep learning models. PreMode is built on SE(3)-equivariant graph attention transformers, utilizing protein language model embeddings[23] and protein structures[22] as inputs. We curated the largest-to-date labeled missense variants with mode-of-action annotations from clinical databases, genetic inference, and experimental assays. We applied PreMode to mode-of-action predictions of 17 genes. PreMode reached state-of-the-art performance at mode-of-action predictions compared to existing models. We further demonstrate PreMode's practical utility in both improving deep mutational scan experiments and assisting protein engineering by significantly reducing the size of mutants for screening via active learning.

Result

Overview

We propose a framework for predicting the mode of action at the molecular level and genetic level. Molecularly, the effect of a missense variant is about the change in biochemical properties of a protein, such as enzyme activity, stability, and the regulatory processes upon protein-protein interactions (**Figure 1a**). These changes can be measured by deep mutational scan experiments (**Figure 1a**). Genetically, the overall outcome of molecular effects results in different types of missense variants. One common categorization is "loss of function" (LoF) and "gain of function" (GoF) variants (**Figure 1a**). To conceptualize this framework to variant effect prediction models, we introduced two parameters: "distance from wild type" (denoted as 'r') and the "direction of change" (notated as 'θ'). The distance parameter distinguishes between pathogenic and benign variants (**Figure 1a**) and is shared across all genes. The direction parameter takes on different meanings both molecularly and genetically within various genes. Therefore, we propose that a mode-of-action predictor would make separate predictions of r and θ utilizing different datasets. It would first learn r prediction using labeled pathogenic and benign variants for all genes, just like conventional variant effect predictors, then learn θ prediction using protein or protein family specific datasets via transfer learning (**Figure 1b**).

Curation and characterization of mode-of-action labeled missense variants

We curated the largest-to-date mode-of-action labeled missense variant datasets annotated at both molecular and genetic levels, including 41,081 missense variants in eight genes with multi-dimensional measurements of different properties by deep mutational scan experiments, 2083 gain- and 7910 loss-of-function missense variants in ~200 genes. The gain and loss of function-

labeled variants were collected from literature searches[12,39,43,44], cancer hotspots[45-47] and published databases[48] (**Methods**).

We first investigated global properties of GoF and LoF variants using the curated data set. GoF variants are more likely to be located in regions with lower AlphaFold2 prediction confidence (pLDDT) than LoF variants and are more likely to be on protein surfaces (**Figure 2a**). In contrast, LoF variants have an overall bigger impact on protein folding energy than GoF variants (Figure 2a), although both types of variants confer a greater folding energy change to protein folding than benign variants (**Figure 2a**). As expected, both LoF and GoF variants are more likely to be in conserved regions than non-conserved regions, as shown in the density plot of conservation represented by entropy of amino acid frequencies across species in multiple sequence alignments (MSA) (**Figure 2a**), while benign variants are mostly located in non-conserved regions (**Figure 2a**). Additionally, GoF variants in general are more likely to be located in disordered regions without specific secondary structures than LoF variants (**Figure 2b**). However, we note this pattern is different across protein families. For example, in Na+/Ca2+ channel genes, GoF variants are more enriched in alpha helixes that are critical for ion transport and selectivity domains than LoF variants (**Figure 2b**). Finally, the number of GoF and LoF variants are not evenly distributed across genes, with only a few of the genes having more than 15 GoF and LoF variants (**Figure 2c**). Overall, those results showed that protein structure, energy and evolutionary features could help predict G/LoF variants while underscoring the necessity for the development of protein- and protein family-specific predictive models using limited data.

A deep learning model for mode-of-action predictions

We developed PreMode, a model pre-trained on pathogenicity task and optimized for transfer learning to mode of action prediction tasks. PreMode takes both explicit and implicit input features of amino acid biochemical properties, protein contexts, and cross-species conservation. PreMode explicitly models protein 3D context structure with SE(3)-equivariant graph neural networks (**Figure 3a**). PreMode was designed to not only capture the relative importance between residues by taking both backbone torsion angles and side chain directions into consideration, but also maintaining awareness of geometric equivariance so that rotation of the atom coordinates does not affect the predictions. PreMode's SE(3)-equivariant learning ability was achieved by using a graph representation of protein 3D structures, where each residue was represented as nodes with features that explicitly represent local biochemical properties and evolutionary conservation including secondary structures[49,50], pLDDT[22], amino acid frequencies in MSA[51], and relative coordinates of all atoms in sidechain with respect to alpha carbons (**Methods**). We also included protein sequence language model (ESM2) embeddings[23] into node embeddings, which implicitly capture similar structural and evolution information. Such implicit representation could serve as a compensation of possible missing information limited by secondary structure annotation or MSA generation algorithms. For each edge connecting two residues, the features include Euclidean vector of two corresponding beta-carbons (for glycine we use alpha-carbon instead) to encapsulate static contacts. Additionally, we incorporate co-evolutional strength calculated from MSA and contact strength predictions from MSA transformer[52] to represent potential dynamic contacts[53] or inter-homopolymer contacts that are missed in a static structure (**Methods**). PreMode applies SE(3)-equivariant mechanisms on edge

features and node features first through a star-graph that connects variant site with all other residues to capture the direct impact of amino acid alterations on other residues, then through the k-nearest neighbor (KNN)-graph that connects each residue with its closest neighbor residues to capture the second order impact, finally through a star-graph module to aggregate the impacts (**Supplementary Figure 1**).

We pretrained PreMode on the pathogenicity task to let the model learn general representation of the variant effects. We collected 83,844 benign variants from ClinVar and PrimateAI in 13,748 genes, and 64,480 likely pathogenic variants from ClinVar with at least one star confidence and HGMD in 3703 genes. We randomly selected 5% of the variants as a validation dataset and trained 10 epochs on the rest of the training data until validation loss stopped dropping (**Supplementary Figure 2**). While predicting pathogenicity is not the design goal of PreMode, we can still use pathogenicity prediction performance to investigate the contribution from various components of the model. As PreMode was pre-trained on ClinVar, using variants from the same resource as testing data can result in inflated performances during assessment. Instead, we used independent testing data to compare PreMode against other methods, including 533 pathogenic missense variants in cancer hotspots from cBioportal[45,47,54] and same number of benign variants in the same genes randomly selected from common variants in primates[17].

PreMode reached similar levels of performance as existing methods including AlphaMissense[20] and gMVP[19] on the test dataset with AUROC of 0.928 (**Figure 3b**). We further performed ablation analysis to assess the contribution of language model embeddings, structural information and MSA information to the prediction. Replacing ESM2 embeddings with one-hot encodings of amino acids results in a drop of AUROC to 0.907. Similarly, removing the MSA will drop the AUROC to 0.907 (**Figure 3c**). Removing the structure module will drop the performance slightly to 0.922 (**Figure 3c**). This showed that ESM2 embeddings, MSA module and SE(3)-equivariant module on AlphaFold2 predicted protein structures together provide non-redundant information for pathogenicity prediction.

PreMode reaches state-of-the-art in molecular mode-of-action predictions and facilitates interpretation of deep mutational scan experiments

We first investigated the utility of PreMode in predicting modes of action at the molecular level. While deep mutational scan is a promising high-throughput approach for measuring molecular effects of variants on protein fitness, only a few of them were designed to capture multi-dimensional functional effects. We obtained DMS data on eight genes (*PTEN, SNCA, CCR5, CXCR4, NUDT15, CYP2C9, GCK, ASPA*) from MAVEDB[6] with multiple assays of different biochemical properties. Typically, these assays span at least two aspects, the stability and function. These two assays are moderately correlated as protein function largely depends on protein stability. (**Supplementary Figure 3**).

We first compare the zero-shot prediction performance of PreMode and existing pathogenicity prediction tools (AlphaMissense[20], ESM1b.LLR[26], FoldXddG[55], gMVP[19], PrimateAI[17], REVEL[14]) on all the DMS assays and weighted sum the results based on the dataset sizes. We found that all of the pathogenicity prediction tools correlated better with functional assays than stability

assays, while FoldX did the opposite (**Figure 4a**). Although PreMode's zero-shot performance is as good as state-of-the-art methods when pretrained on a small dataset, it reached similar performance as AlphaMissense and ESM1b by simply increasing the pretrain data set sizes (**Methods**) and correlated better with stability assays (**Figure 4a**). While protein fitness is largely determined by its functional property, we note separate protein stability and functional measurements are important to help us understand the mechanisms of missense variants, especially in distinguishing dominant negative variants and loss of function variants, which often differ in protein stability. Therefore, we investigated the ability of PreMode to expand single-dimensional pathogenicity prediction to multi-dimensional protein stability and functional fitness predictions via transfer learning.

We split the data of each gene into 80% of training and 20% of testing data five times under different seeds and ran PreMode on each of them via transfer learning. While no other methods have been trained on such mode-of-action prediction tasks and their model weights are not available for us to do transfer learning, we compared PreMode against a baseline model utilizing ESM2 embeddings and a single layer perceptron (SLP, **Methods**) as approximated transfer learning with ESM2. PreMode outperforms the baseline method with higher spearman correlation (**Figure 4b**). A larger pretrain data set can slightly improve performance in both stability and functional fitness predictions (**Figure 4b, 4c**). Overall, after transfer learning, PreMode is able to predict both the multi-dimensional protein stability and functional fitness with a spearman correlation of 0.6 with experimental results (**Figure 4c**). Furthermore, we investigated the multi-dimensional transfer learning ability of PreMode under smaller sample sizes (148k), where we randomly subsample the training data and compared the performance in same testing dataset. We found that PreMode is able to accurately predict variant effects of all missense variants with around 40% (~2000) of variants measured inside one gene, after which increasing the number of data points will have minimal improvement on the performance in the testing dataset (**Figure 4d**).

We further investigated the utility of PreMode to improve the interpretation of experimental readouts in two applications. First, we hypothesized that by PreMode could examine the abnormal measurements in each experiment by transfer learning as it has implicitly modeled the fitness of variants in all proteins during pretraining. We used the stability DMS experiment[27] of *PTEN* as an example and trained PreMode on one of the biological replicates. We then compared the differences between PreMode's predictions and the experimental readouts. We showed that this difference value is highly correlated to the difference between the readout of the single biological replicate and average readouts in all experiments (**Supplementary Figure 5a**). The experimental readouts with large deviation from PreMode's predictions are more likely to be abnormal measurements (**Supplementary Figure 5b**). Next, we hypothesized that although the functional fitness landscape across genes is different, the stability fitness should generally be the same. Therefore, we applied PreMode to the largest stability measurement experiments in MAVEDB across >30 genes. We trained PreMode on 80% of the data and tested the other 20% of variants in completely different genes from training. PreMode trained on smaller pretrain data outperforms all other methods (**Supplementary Figure 6**). This result

indicates that PreMode could be further applied to novel genes by transfer learning on the stability fitness dataset.

PreMode is state-of-the-art in genetic mode-of-action predictions

We grouped the gain / loss of function variants dataset by genes and only kept those with ≥15 G/LoF variants (**Figure 2c**, **Methods**). We performed transfer learning on the selected genes using the pretrained model parameters as initial weights (**Methods**). We note there are two reasons to perform transfer-learning in individual genes rather than across all genes. First, G/LoF mechanisms are intrinsically different across genes, as they have different functions. Second, the number of G/LoF variants are extremely unbalanced across genes. A deep neural network model with transfer-learning across genes will potentially reach a local minimum where gene-properties dominate its predictions that are better at predicting likely G/LoF genes but do not distinguish G/LoF variants in the same gene (**Supplementary Figure 7**).

For each gene, we randomly split the gain and loss of function variants into training and testing. The total amount of data for training and testing in each gene is shown in **Supplementary Table 1**. There were tasks for nine genes (*ABCC8, BRAF, CACNA1A, FGFR2, KCNJ11, RET, SCN2A, SCN5A, TP53*; **Figure 2c**) in total. We compared PreMode against several baseline methods trained and tested on the same training/testing datasets. Overall, PreMode performed better in all the genes than baseline methods. It reached average AUC of 0.8~0.9 in genes *CACNA1A, KCNJ11*, and *RET,* and average AUCs of 0.7~0.8 in *ABCC8, BRAF*, and *TP53* (**Figure 5a**). Pretrained PreMode is better than non-pretrained model for all genes. Low-capacity models such as random forest and single layer perceptron fine-tuned ESM model is worse than pretrained PreMode except for *SCN2A*. We also compared it to LoGoFunc[42], a method trained on G/LoF variants across genes. PreMode is slightly worse than LoGoFunc in *SCN5A* but better in all other genes (**Supplementary Figure 8a**). Strikingly, while increasing pretrain data set size can increase the molecular mode-of-action prediction performances, it slightly reduces the performance of PreMode in distinguishing gain versus loss of function variants in all genes (**Supplementary Figure 8b**).

We next did ablation analysis on PreMode to identify the important features for Gain/Loss of function predictions. Removing structural input decreased the performance in five out of nine genes except for *SCN2A, SCN5A, RET* and *FGFR2* (**Figure 5b**). The LoF variants in those genes were located in slightly lower pLDDT region (**Figure 5c**). On the contrary, removing ESM embeddings always decreased the performance. We further found that adding post-translational-modification (PTM) information (PhosphoSitePlus[56]) into the model input can improve the predictions in most genes but decreases the performance in *CACNA1A* and *BRAF*. In these two genes, the GoF variants are closer to the PTM sites than LoF variants while the trend is opposite in the other genes (**Figure 5d**). Overall, PreMode with PTM information reached highest performance in the weighted sum of AUC across genes, followed by default setting PreMode and other methods (**Supplementary Figure 9**).

Next, we hypothesized that the gain and loss of function mechanisms are similar in the same domain across genes, and can use this to further improve the GoF/LoF predictions in each gene.

We split the data in each gene by domains and only selected the domains with ≥15 G/LoF variants for evaluation. We performed PreMode transfer learning within the domain using either gene-specific data or data across genes while testing on the same data within one gene. We observed increased performance in all of the domains when using data across genes rather than using the gene alone (**Supplementary Figure 10**).

To be effective in applications in genetic analysis, a machine learning method should be able to perform transfer learning with limited amount of data, as in most of the genes there are fewer than 10 known G/LoF variants (**Figure 2c**). We therefore examined PreMode's performance with subsampled training data in *ABCC8, BRAF, CACNA1A, KCNJ11, RET, TP53*, where PreMode reached ≥0.75 AUC with full training data. We found that PreMode could reach an AUC above 0.75 with less than 10 G/LoF variants data in *ABCC8, KCNJ11* and *RET* (**Figure 5e**). For the other three genes low-capacity methods like random forest will reach equal performance as PreMode under the low training sample size. Overall, this result suggested that PreMode is still useful in genes with small sample sizes.

PreMode predicted mode-of-action landscapes in individual proteins through in silico saturated mutagenesis.

We further applied PreMode to infer the modes of action of all possible variants in *BRAF, TP53, PTEN, RET*, and *KCNJ11*. We used a gradient boosting tree to combine the prediction scores of all five models that trained on different subsets of data and applied to the other variants in corresponding gene.

In *BRAF*, PreMode identified two regions that enriched GoF variants. One closer to the N-terminal at a phorbol-ester/DAG-type zinc-finger domain and the other on the kinase domain which includes the well-known V600 GoF position[57,58] (**Figure 6a**). The variants on the two domains act through different gain/loss of function mechanisms. The phorbol-ester/DAG-type zinc-finger domain auto-inhibits BRAF activity through binding with 14-3-3[59] while also cooperating with Ras-binding domain (RBD) to bind with Ras and activate BRAF activity[60]. To validate PreMode's predictions, we obtained B-Raf/14-3-3 and B-Raf/K-Ras binding structures from PDB (7MFD) and AlphaFold2 (Colabfold[61] implementation) predictions, respectively (**Figure 6b**). We then calculated the energy change on both structures upon mutations using FoldX[55]. We found that the LoF variants predicted by PreMode confer large energy changes that make BRAF binding to both 14-3-3 and K-Ras unstable, while GoF variants only destabilize B-Raf binding to 14-3-3 and maintain the binding stability to K-Ras at similar level as benign variants (**Figure 6c**). The ddG landscape suggests that the GoF variants in this region mostly act by abolishing its inhibitory regulation[62]. We further investigated the G/LoF variants predicted by PreMode in the kinase domain. The kinase domain has both active (PDB: 4MNE) and inactive (PDB: 4EHE) conformations with large and small enzyme pocket sizes, respectively[63] (**Figure 6d**). The FoldX ddG results are consistent with previous findings that GoF variants V600E/D can destabilize the inactive state while stabilizing the active state[64] (**Figure 6e**). Similarly, we found that the PreMode predicted LoF variants destabilize both conformations as well as the complex of BRAF-MEK1 while GoF variants only destabilize the inactive conformation (**Figure 6e, Supplementary Figure 11**).

TP53 is a tumor suppressor gene, and most of the pathogenic variants act through LoF. However, there are a few regions enriched with GoF variants identified by PreMode, all on the DNA binding domain of the p53 protein (**Supplementary Figure 12a**). Among those regions, sites 291 and 292 are essential post-translational modification sites for p53 ubiquitination and subsequent degradation[65]. A previous study showed the variants at sites 121 and 290-292 increase the ability to induce apoptosis in cultured cells[66] (**Supplementary Figure 12a**).

PreMode also identified several GoF enriched regions in *RET* and *KCNJ11* (**Supplementary Figure 12b, 12c**). In *RET*, PreMode identified GoF enriched regions both in the kinase domain and regulatory signaling domain near the extracellular binding sites. In *KCNJ11*, the regions are located at transmembrane domains and cytoplasmic domains (44-60, 160-190, 200-210,225-230,291-296, 320-340, PDB: 5YWB). Although they are spatially distinct to each other, the regions spanning position 160-190 and 291-296 form the core part of potassium channel in the tetramer (**Supplementary Figure 13**), especially residues L164 and F168 that form the inner helix gate[67].

PTEN is a tumor suppressor gene and an essential gene in fetal development. Loss of function in *PTEN* can lead to multiple syndromes. As shown by deep mutational scan experiments[29,68], some variants in *PTEN* can have different impact on stability and enzyme activities. We applied PreMode on 80% of the data from both functional assays to predict the effect of all possible variants in *PTEN*. Although most of variants decrease both the protein stability and enzyme function, PreMode identified variants that only disrupt stability but not enzyme function (**Supplementary Figure 14a**, blue), and such variants are located all over the protein. Similarly, PreMode identified variants that only disrupt enzyme function but not stability, all located in the phosphatase domain (**Supplementary Figure 14a**, red). These variants may have dominant negative effects. In fact, three known dominant negative variants in *PTEN* (C124S, G129E and R130G)[68,69] were successfully predicted by PreMode to maintain stability while causing loss of enzyme function, among which G129E was not in the training data. PreMode identified 4 regions enriched for such variants (**Supplementary Figure 14a**, pink). Further analysis showed those regions were spatially close to each other and form the enzyme pocket around the phosphatase site (**Supplementary Figure 14b**). Notably, PreMode can identify similar dominant-negative variants enriched in a region with only 20% (398 points) of the dataset (**Supplementary Figure 15**).

PreMode trained with active learning allows efficient few-shot transfer.

Deep mutational scan and directed evolution-based protein design experiments often incur expensive time costs upon scaling up. We asked if PreMode can be used to help with experimental design by active learning[70]. In an active learning framework, PreMode is iteratively trained on a set of experimental data and prioritizes the variants to be measured in the next rounds of experiments (Supplementary Figure X). We applied this framework to the protein design of green fluorescence protein (GFP) on fluorescence strength. PreMode was able to predict the fitness landscape of GFP with spearman correlation above state-of-the-art performance (0.69)[71,72] to the experimental data using only 40% of the training data by adaptive

learning, which is much more efficient than randomly subsampling data (**Supplementary Figure 16**)

Discussion

Previous methods for predicting the effect of missense variant have been focused on pathogenicity, which is a binary label. However, different pathogenic variants in the same gene can change the protein function in different or sometimes opposite ways. We introduce a new method, PreMode, to predict mode of actions of missense variants. We curated and characterized the largest-to-date labeled missense variants that act through different modes, especially for gain and loss of function variants. We discovered that gain of function variants tends to be located in low pLDDT regions and surfaces in Alphafold2 predicted structures. The low pLDDT regions are likely to be intrinsically disordered[73] which implies conformational heterogeneity and dynamics which could be associated with protein binding and regulatory activities, a context in which gain-of-function variants appear particularly influential. On the contrary, loss of function variants tend to increase protein folding energy, suggesting that some of these variants destabilize protein structures, while gain-of-function variants tend to preserve protein folding integrity[33]. Overall, these findings show how variant effects are associated with the protein structures and functions qualitatively. Gain-of-function variants often act through preserving overall protein structure while operating through diverse mechanisms by targeting specific structural domains, whereas LoF variants often destabilize protein structures and tend to be distributed across structured regions. Furthermore, we emphasize the need for protein-specific rather than genome-wide Mode-of-Action predictions for two reasons. First, the Mode-of-Action mechanisms are intrinsically different, making it conceptually ambiguous to predict gain/loss-of-function variants across genes. Second, the data are heavily uneven across genes, making deep learning algorithms easily stuck at local minimum and gene properties to distinguish genes that tend to act through G/LoF mechanisms while hard to distinguish G/LoF variants within same gene. We propose a framework to disentangle the mode-of-action predictions at molecular/genetic levels and selected 17 genes with sufficient deep mutation scan data or labeled gain/loss-of-function missense variants for model development and evaluation.

The input feature of PreMode includes both protein language model embeddings and representation of protein structure. PreMode performs gene-specific mode-of-action predictions through a genome-wide pretrain stage and a gene-specific transfer learning stage. At the molecular level, PreMode can simultaneously predict the multi-dimensional biochemical impact of single missense variants, which can reveal potential dominant negative variants that reduce protein function but maintain stability. We show that PreMode is efficient at transfer learning and that it can capture the fitness landscape of all possible variants within one protein using around 40% of mutagenesis data. At a genetic level, PreMode can efficiently utilize a small amount labeled data (a few dozens) to accurately distinguish G/LoF variants with an AUC of around 0.75 in most proteins. Furthermore, We demonstrate PreMode's application. First, PreMode can improve deep mutational scan experiments by detecting noisy data points in single measurement. Furthermore, PreMode could facilitate mutagenesis-based protein directed evolution through adaptive learning by efficiently lowering library sizes.

PreMode currently predicts a binary prediction of gain or loss of function. This is a limitation as it does not capture the complexity of protein functional changes. For instance, ion channel proteins undergo complex conformation changes and regulation to perform normal physiological functions[41]. Large scale functional studies of these genes may provide additional data that enable training of improved models. Furthermore, gain/loss of function could be further divided into amorph, hypomorph, hypermorph, antimorph and neomorph according to muller's morphs[74]. Accordingly, additional labeled data may facilitate the training of more accurate and comprehensive models.

PreMode can be potentially improved by considering protein dynamics. As ablation experiments showed it has lower performance for genes where both G/LoF variants are located in regions with relatively low pLDDT values. A static protein structure is not sufficient to model the variant effects in those regions. Furthermore, we notice that although simply increasing the data set size in the pretrain stage can improve the zero-shot performances of PreMode, it does not guarantee increased performance in transfer-learning for gene-specific mode-of-action predictions. Overall, increasing pretrain data set sizes can help PreMode better distinguish benign and different loss-of-function variants as we include more potential benign variants in the pretrain data. It does not help in distinguishing gain/loss-of-function variants.

Overall, our work suggests a potential direction for this new era of variant effect predictor development. PreMode has the potential to accelerate our understanding of mode-of-actions, contribute to better clinical diagnosis, therapeutic development and more broadly, artificial GoF engineering of proteins.

Data and code availability

Data and code used in model training and analysis could be found at GitHub: https://github.com/ShenLab/PreMode

Acknowledgement

This work was supported by NIH grants (R35GM149527, P50HD109879) and Simons Foundation (SFARI #1019623). We thank Dr. Mohammed AlQuraishi, Dr. David Knowles, Dr. Haicang Zhang, Dr. Haiqing Zhao, Xi Fu, Dr. Xiao Fan, Dr. Na Zhu and Dr. Chao Hou for helpful discussions and suggestions.

Methods

Training and testing datasets

We curated labeled pathogenic/benign variants, gain/loss of function variants, deep mutational scan experiments from public databases and publications.

For gain/loss of function prediction tasks, we collected: 765 gain and 4,571 loss of function missense variants from Barak, et al[39]; 669 gain and 1,232 loss of function variants from Clinical Knowledge Base[48]; 199 gain and 1,506 loss of function variants from cBioportal cancer hotspots[45]; 56 gain and 57 loss of function variants in *ABCC8, GCK, KCNJ11* from Flanagan, et al.[12]; 45 gain and 7 loss of function variants in *STAT1* from Kagawa, et al.[43]. More specifically, for variants in cBioportal, we first calculate cancer hotspots based on existing algorithms[54], then annotate variants in hotspots of oncogenic genes as gain of function and tumor suppressor genes as loss of function based on COSMIC database[46]. We didn't annotate genes with multiple cancer roles.

As the gain and loss of function variants were extremely biased across genes, we didn't split the training and testing dataset in common machine learning manner but split by protein-wise manner with the following steps:

1. For each gene, we select 20% of GoF and 20% of LoF variants as testing. We use the rest variants in the same gene as training.
2. We only kept genes with more than 15 GoF variants and 15 LoF variants for model evaluation.

For pathogenic task, we collected 148,324 variants for training, including: 51,494 benign variants from PrimateAI[17] and 32,350 non-overlapping benign variants from ClinVar with at least one-star non-conflict submits that labeled as "benign" or "likely-benign"; 64,480 pathogenic variants from ClinVar database with at least one-star non-conflict submits[7], non-overlapping variants from HGMD[75]. We collected 1,066 variants for testing, including 533 pathogenic missense variants from somatic missense hotspots in 153 cancer driver genes that not annotated above; and 533 benign variants from the same genes randomly selected from ClinVar and PrimateAI that not overlapped with training dataset.

For deep mutational scan assays, we collected following datasets from MAVEDB[6] (**Supplementary Table 2**).

Input features

For a missense variant of interest, we construct two graphs based on its protein context. For each variant, we consider a 250 amino acid window flanking the variant position and the residues as nodes; the first graph $\mathcal{G}_1$ is a non-directed star graph that connect only the variant node and the other nodes. The second graph $\mathcal{G}_2$ is a non-directed K-nearest neighbor graph that connect each node with its neighbors based on 3D Euclidean distance of the alpha carbon atoms.

Each node has a set of invariant features and structure features. The invariant features include: Embeddings from the last layer of ESM2[23] (d=1280); Dssp[49,50] annotated secondary structure and torsion angles from the AlphaFold2 predicted protein structures (d=12); pLDDT[22] for the AlphaFold2 prediction; Amino acid frequencies across 199 species that reflects evolutionary conservation (d=20). The variant node has additional invariant feature of the embedding of alternate amino acid. The structure features include a set of Euclidean vectors from the alpha carbon to all other non-hydrogen atoms of side chain (d=3x35). If an atom does not exist in the side chain, then it is set as 0.

Each edge has a set of invariant features and structure features. The invariant features include, Weighted covariance matrix between the amino acid frequencies of two residue sites in multiple sequence alignments (MSA) of 199 species[19] (ALGORITHM 1); Euclidean distances of beta carbons between two residues (For glycine we use alpha carbon), transformed by exponential smearing functions (ALGORITHM 2); The contact strength predicted by MSA transformer. The structure features include the Euclidean vector of beta carbons between two residues.

ALGORITHM 1: Weighted MSA Encoder

Input: For node i, let M_i be a one hot matrix with shape (199,21) indicating the amino acids for each species on the index i of the aligned protein sequences; W_M is a learnable parameter of length 199 that weight the importance of species.

Output: A vector C_{ij} that reflects the co-evolutional strength between node i, j

$\bar{W}_M = softmax(W_M)$

For each node i **do**:

$\bar{M}_i = M_i * \bar{W}_M$

For each edge (i, j) **do**:

$$C_{ij_1} = \sum_k \bar{W}_{M_k} * M_{ik} * M_{jk}$$

$$C_{ij_2} = \sum_k \bar{M}_{ik} * \bar{M}_{jk}$$

$$C_{ij} = C_{ij_1} - C_{ij_2}$$

$$C_{ij} = concat(C_{ij}, \sqrt{\sum C_{ij}^2})$$

ALGORITHM 2: Edge Feature Encoder

Input: For node i, let M_i be a one hot MSA matrix defined in ALGORITHM 1; C_{B_i} be the coordinates of beta carbon; $Cont_{ij}$ be the contact strength define.

Output: Edge invariant features E_{ij}, structure features $\vec{D}_{ij}$.

For each edge (i, j) **do**:

$C_{ij} = WeightedMSAEncoder(M_i, M_j)$

$\vec{D}_{ij} = C_{B_j} - C_{B_i}$

$$D_{ij} = e^{-W_\beta\left(e^{-W_\alpha\left\|\vec{D}_{ij}\right\|_2 + B_\alpha}\right)^2}$$

5 $E_{ij} = concat(C_{ij}, D_{ij}, Cont_{ij})$

Model architectures

PreMode is an SE(3)-equivariant graph attention neural network with 4 layers.
First layer is a feature embedding layer that encodes the input features into latent dimensions. This layer was designed separately for invariant features and structure features. For invariant features, we use GeLU[76] as the activation function and 512 dimensions as output. The variant node has an additional embedding layer to incorporate with the alternate amino acid embeddings (ALGORITHM 3). For structure features, we use a linear layer without bias term to ensure the equivariance. The latent dimension is 32 (Sup Figure 1). For edge features, we encode the pairwise MSA, Euclidean distances, contact predictions, relative positions into a 444 dim invariant feature vector.

ALGORITHM 3: Feature Embedding Layer

Input: For node i, let I_i be the input node invariant features; let $\vec{I_i}$ be the input node structural features. Let I_c be the center node invariant features.

Output: Node invariant features X_i, structural features $\vec{S_i}$.

1 **For** each node i **do**:

2 **If** (node i is variant node) **do**:

3 $I_i = GeLU(I_c W_c + B_c)$

4 $X_i = GeLU(I_i W_i + B_i)$

5 $\vec{S_i} = \vec{I_i} W_s$

Second and third layers are equivariant graph attention layers. We calculate the attention values only between the connected nodes in star or KNN graphs defined in the features section. We used a modified SE(3)-equivariant attention mechanism inspired from torchmd-net[77] and gMVP[19] to calculate the attention values that takes co-evolutional evolution, structure features into consideration. The invariant and structure features were updated separately but share information across each other based on the attention values to maintain equivariance (ALGORITHM 4). Those two layers were designed to capture the biochemical context for the residue with interest, where the second layer focuses more on the first order impact of amino acid change to all other residues, and the third layer implicitly models the second order consequences.

The last layer is a graph attention layer that designed to summarize the overall impact of the variant to protein. It only takes the invariant embeddings output from the third layer to calculate the attention values between center node to other nodes while don't take structural features.

ALGORITHM 4: SE(3)-equivariant Graph Attention Transformer Layer

Input: A graph $\mathcal{G}$. For node i, let X_i be the invariant features, $\vec{S_i}$ be its structure features; For edge between node i, j. Let E_{ij} be the invariant features, $\vec{D_{ij}}$ be structure features.

Output: A graph $\mathcal{G}$ with updated X_i and $\vec{S}_i$

For each edge (i, j) **do**:

$$Q_i = X_i W_q + B_q$$
$$K_j = X_j W_k + B_j$$
$$V_j = X_j W_v + B_j$$
$$dK_{ij} = SiLU\left(E_{ij} W_{dk} + B_{dk}\right)$$
$$dV_{ij} = SiLU\left(E_{ij} W_{dv} + B_{dv}\right)$$
$$A_{ij} = SiLU\left(Q_i \cdot K_j \cdot dK_{ij}\right)$$

For each node i **do**:

$$O_i = \sum_j A_{ij} * V_j$$
$$\vec{T}_i = \sum_j \left(X_j W_{t_1} + B_{t_1}\right) * \vec{S}_j + \left(X_j W_{t_2} + B_{t_2}\right) * \vec{D}_{ij}$$
$$dX_i = \left(\vec{V}_i W_{v_1}\right) \cdot \left(\vec{V}_i W_{v_2}\right) * \left(O_i W_{O_1} + B_{O_1}\right) + \left(O_i W_{O_2} + B_{O_2}\right)$$
$$\overrightarrow{dS}_i = \vec{V}_i W_{v_3} * \left(O_i W_{O_3} + B_{O_3}\right) + \vec{V}_i$$
$$X_i = X_i + drop_out(dX_i)$$
$$\vec{S}_i = \vec{S}_i + drop_out(\overrightarrow{dS}_i)$$

Model training and testing

We trained PreMode with Adam[78] algorithm. For pathogenicity task, we set learning rate to 1e-4 and batch size to 256 and trained 20 epochs. We randomly selected 5% of training data as validation and calculated the loss on it every 250 steps. We decrease the learning rate by 0.8 if observed a plateau on the validation loss until the minimum learning rate 1e-6. For transfer learning on gene and protein family specific tasks, we set the batch size to 8, evaluate validation loss every 100 batches and set the minimum learning rate to 1e-7. For pretrain pathogenicity task, we used 4 NVIDIA A40 GPUs, and it took about 50h. For gene specific mode-of-action tasks, we used 1 NVIDIA A40 GPU, and it took 1~6h depending on the data sizes.

We selected the model saved at the batch or epoch with minimum validation loss for testing. We calculated the AUROC (area under recall receiver operating characteristic curve) to assess the performances of models. For pretrain on the pathogenicity task, we only calculated one AUROC value. For transfer learning tasks, we calculated AUROC for all five random training/testing splits and calculated the average and standard deviation.

Baseline Methods and ablation analysis

For both molecular and genetics level mode-of-action predictions, we created several baseline methods to compare PreMode with. First, we built a single layer perceptron (SLP) model on top of the ESM embeddings (ESM + SLP). This model takes all invariant features that same as PreMode as inputs, followed by one linear layer and GeLU activation layer. Next, we implement several ablation analyses on PreMode. For structure feature ablation, we replaced all structural features with 0, and constructed the star and KNN graphs based on the 1D-distance on the sequences. For ESM embedding feature ablation, we replaced the ESM embeddings with one-

hot encodings of 20 amino acids. For MSA feature ablation, we replaced all MSA features with 0. All the models above were trained under same training configurations as PreMode at both pretrain and transfer learning stages. Lastly, we implemented capacity machine learning models using only biochemical properties as inputs, including conservation, secondary structure, reference, and alternate amino acid identities, pLDDT and ddG. For molecular-level mode-of-action predictions, we implemented it with elastic net linear regression model while for genetics-level mode-of-action predictions, we implemented it with random forest classification model.

Subsample of datasets

For deep mutational scan datasets, we subsampled the data to investigate how many points were required for sufficient adaptation to the tasks in transfer learning. For each of the multi-dimensional assays, we randomly subsampled training data to 10%, 20%, 40%, 60%, 80% of whole datasets 5 times with different random seeds and testing on the same 20% of whole datasets. We performed PreMode on each of the training data and evaluate the performance.

In silico saturated mutagenesis experiments

We selected genes with average AUROC above 0.75 for in silico saturated mutagenesis experiments, including BRAF, CACNA1A, KCNJ11, RET, TP53, PTEN. For each gene, we calculate two predictions r (pathogenicity score) and θ (Gain/Loss of function score) with pretrained model and corresponding transfer-learning models, respectively. For θ, there are 5 predicted scores from models trained on 5 training/testing splits under different random seeds. We selected the models with minimum validation loss under each training/testing split and ensemble the predictions by voting with a simple random forest model.

Active Learning experiments

We split the deep mutational scan experiment datasets to 80% of training data and 20% of testing data. Within the 80% of training data, we first perform PreMode transfer learning on 10% of the randomly selected data, then evaluate PreMode on the rest of 70% data. PreMode will output both a predicted value as well as model confidence values. Then we select the top 10% of data among the rest of training data where PreMode is most unconfident and add them to the next round of training. We repeat this procedure until all the training data were used.

Figure and Figure Legends

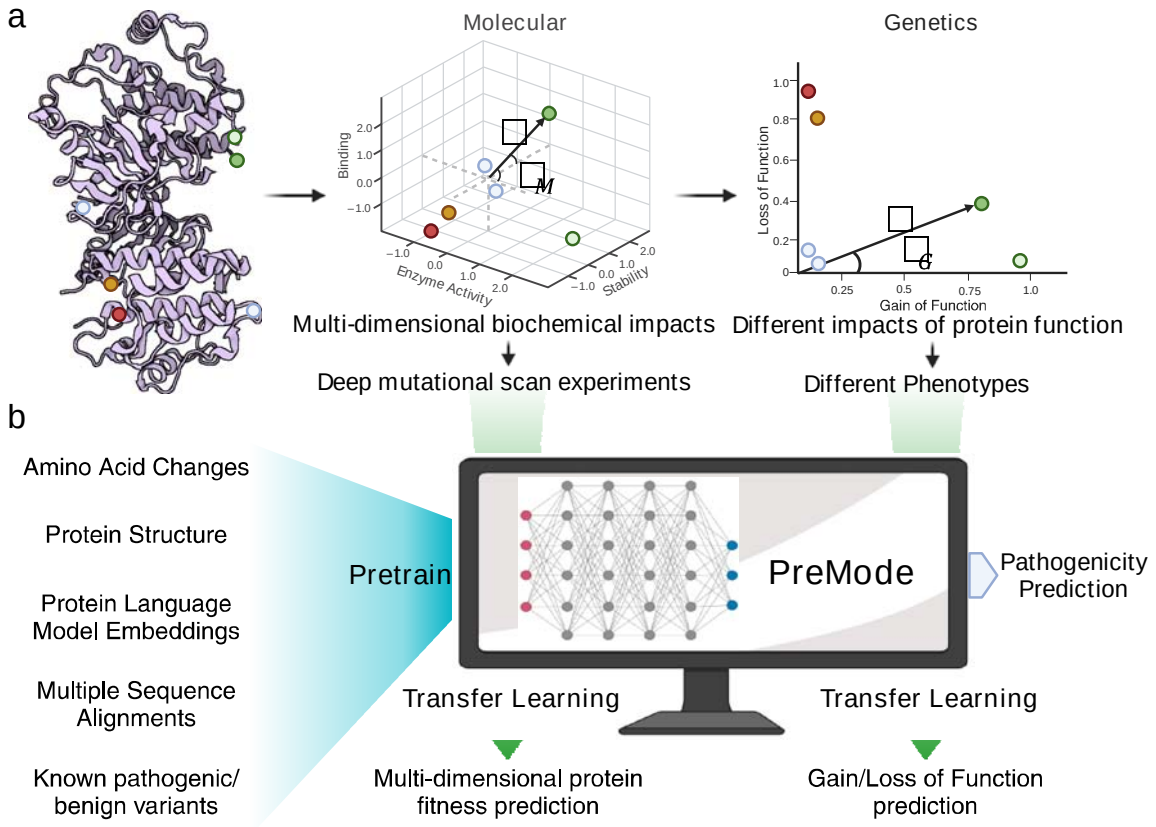

Figure 1. Mode of Action definition and PreMode framework. A) Mode of action definition at molecular level and genetic levels. B) PreMode framework. Blue arrows show the data flow and at pretrain stage for pathogenicity prediction, green arrows show transfer-learning stage for mode-of-action prediction.

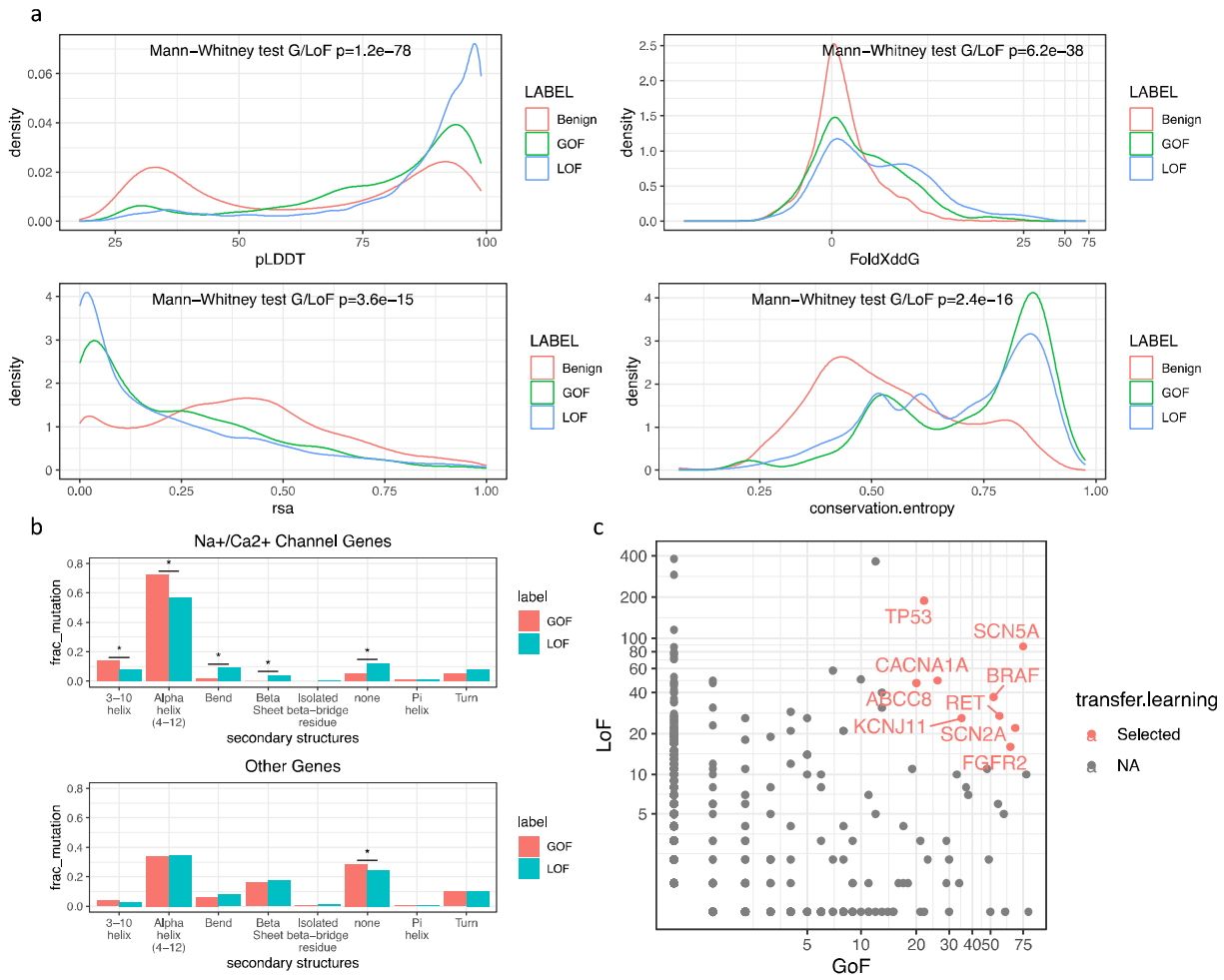

Figure 2. Characterization of gain-/loss-of-function variants. A) Genome wide comparison of gain- and loss-of-function variant sites pLDDT (AlphaFold2 prediction confidence), protein fold energy change, relevant solvent accessibility, and conservation. B) Protein family wise comparison of gain- and loss-of-Function variants enriched secondary structures. C). Gene wise comparison of gain- and loss-of-Function missense variants numbers.

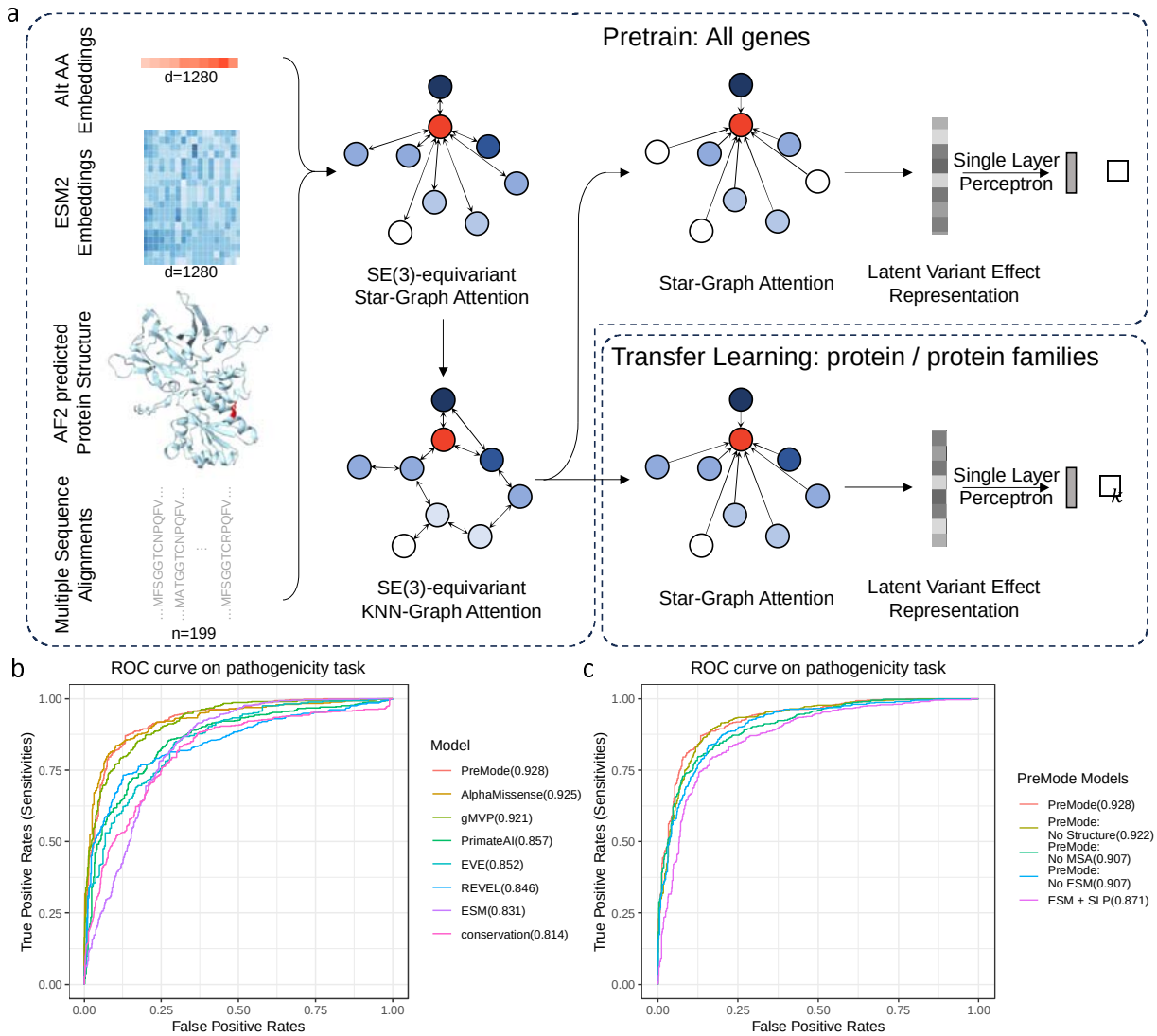

Figure 3. PreMode, a pretrained model for mode-of-action predictions. A) PreMode framework. PreMode takes amino acid changes, protein language model embeddings, alphafold2 predicted structures, multiple sequence alignments as inputs and outputs two parameters. PreMode first predicts pathogenicity (r) for all genes during pretrain, next predicts the mode of action parameters (θ) via transfer learning. B) AUROC of PreMode and other methods in pathogenicity prediction task. C) AUROC of PreMode in ablation analysis.

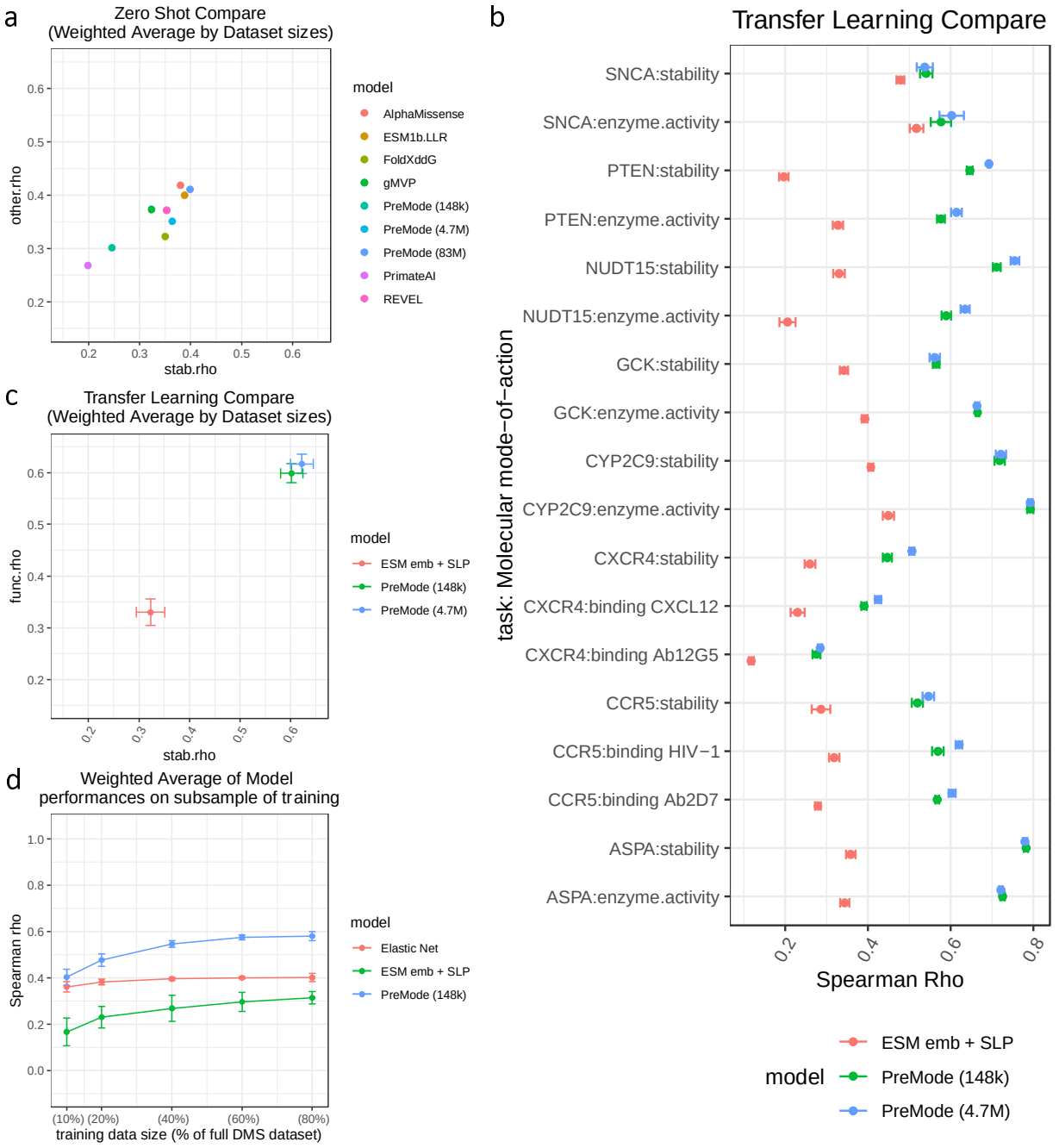

Figure 4. PreMode performances in molecular mode-of-action prediction tasks. A) Zero-shot performances of pathogenicity predictors on stability and functional fitness. B) Spearman correlations of PreMode in multiplexed deep mutational scan experiments of 8 genes compared to baseline method. C) Comparison of weighted average of spearman correlations across experiments based on dataset sizes. D) Comparison of weighted average of spearman correlations of PreMode and baseline methods under sub-sampled data.

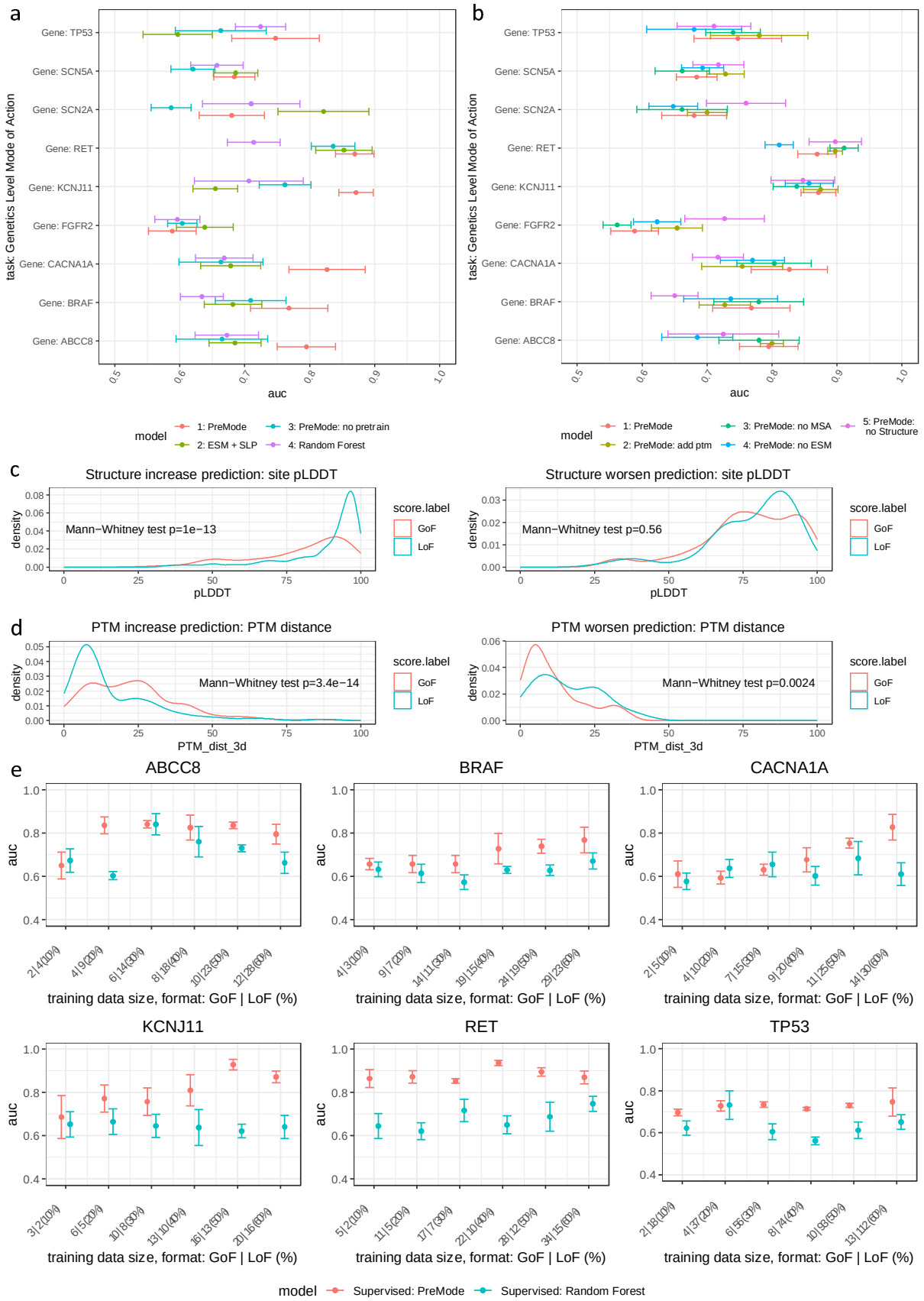
a
b
Gene: TP53
Gene: SCN5A
Gene: SCN2A
Gene: RET
Gene: KCNJ11
Gene: FGFR2
Gene: CACNA1A
Gene: BRAF
Gene: ABCC8
task: Genetics Level Mode of Action
auc
model
1: PreMode
2: ESM + SLP
3: PreMode: no pretrain
4: Random Forest
2: PreMode: add ptm
3: PreMode: no MSA
4: PreMode: no ESM
5: PreMode: no Structure
c
Structure increase prediction: site pLDDT
Mann−Whitney test p=1e−13
Structure worsen prediction: site pLDDT
Mann−Whitney test p=0.56
density
pLDDT
score.label
GoF
LoF
d
PTM increase prediction: PTM distance
Mann−Whitney test p=3.4e−14
PTM worsen prediction: PTM distance
Mann−Whitney test p=0.0024
PTM_dist_3d
e
ABCC8
BRAF
CACNA1A
KCNJ11
RET
TP53
training data size, format: GoF | LoF (%)
Supervised: PreMode
Supervised: Random Forest

Figure 5. PreMode performances in genetic mode-of-action prediction tasks. A) Performances of PreMode and existing methods on gain- and loss-of-function predictions in 9 genes. B) Ablation analysis of PreMode on gain- and loss-of-function predictions in 9 genes. C) Comparison of variant site pLDDT on *BRAF, TP53, ABCC8, KCNJ11, CACNA1A* (left) versus *SCN2A, FGFR2, RET, SCN5A* (right). D) Comparison of distances between variant site to the closest PTM site on *SCN2A, FGFR2, RET, SCN5A, TP53, ABCC8, KCNJ11* (left) versus *BRAF, CACNA1A* (right). E) Few-shot transfer learning of PreMode and base line method on 6 genes.

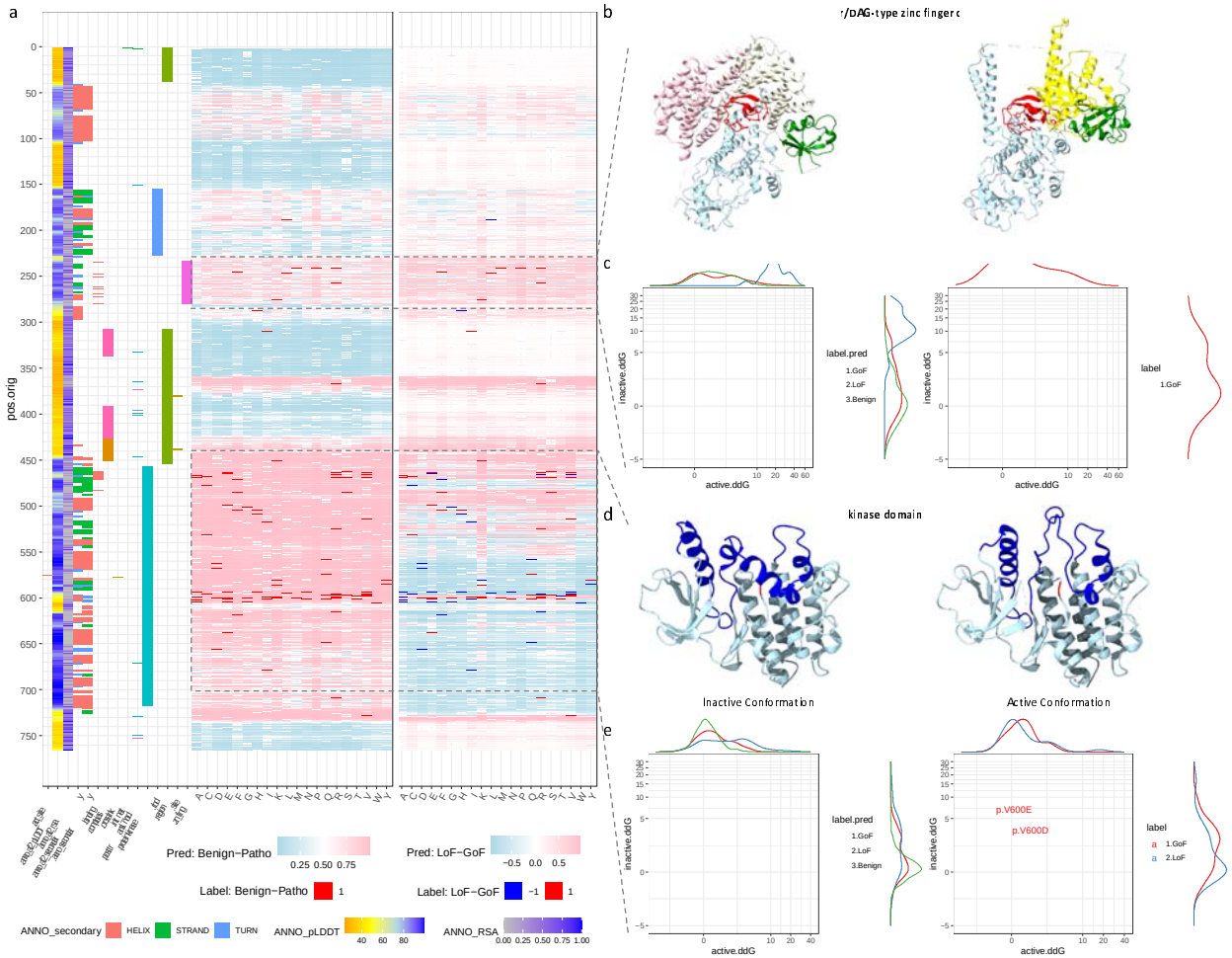

Figure 6. PreMode in-silico mutagenesis mode-of-action predictions on *BRAF*. A) PreMode predictions visualized with BRAF protein domains, x-axis are annotations and amino acid changes, y-axis are amino acid positions. The first few columns indicate Alphafold2 pLDDT, relevant solvent accessibility, secondary structures, protein domains, respectively. The left panel showed PreMode's pathogenicity predictions and labels used in training, pink indicates predicted pathogenic, light blue indicates predicted benign, red indicates pathogenic and blue indicates benign in training data, white indicates reference amino acid. The right panel showed PreMode's G/LoF predictions and labels used in training, pink indicates predicted GoF, light blue indicates predicted LoF, white indicates predicted benign. Red indicates GoF and blue indicates LoF in training data. B) Braf binding to 14-3-3 (left) and Kras (right) structures. Red is Braf phorbol-ester/DAG-type zinc finger domain, green is Braf Ras binding domain (RBD), light blue is the rest of Braf, pink and white are two 14-3-3 proteins, yellow is Kras. C) Energy change upon mutations of Braf phorbol-ester/DAG-type zinc finger domain on the inactive state (Braf-14-3-3 binding, y-axis) versus active state (Braf-Kras binding, x-axis), colored by PreMode predictions (left panel) and labels in training (right panel). D) Braf kinase domain inactive state conformation (left panel) and active state conformation (right panel), red is the active site, blue shows the flexible region, light blue shows the rest of kinase domain. E) Energy change upon

mutations of Braf kinase domain on the inactive state (y-axis) versus active state (x-axis), colored by PreMode predictions (left panel) and labels in training (right panel).

References

1. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-6 (2015).
2. Jin, S.C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet* **49**, 1593-1601 (2017).
3. Huang, K.L. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355-370 e14 (2018).
4. Satterstrom, F.K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **180**, 568-584 e23 (2020).
5. Zhou, X. *et al.* Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nat Genet* **54**, 1305-1319 (2022).
6. Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol* **20**, 223 (2019).
7. Landrum, M.J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018).
8. Romero, P.A. & Arnold, F.H. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* **10**, 866-76 (2009).
9. Yang, K.K., Wu, Z. & Arnold, F.H. Machine-learning-guided directed evolution for protein engineering. *Nat Methods* **16**, 687-694 (2019).
10. Wang, Y. *et al.* Directed Evolution: Methodologies and Applications. *Chem Rev* **121**, 12384-12444 (2021).
11. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
12. Flanagan, S.E., Patch, A.M. & Ellard, S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet Test Mol Biomarkers* **14**, 533-7 (2010).
13. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
14. 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).
15. Samocha, K.E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*, 148353 (2017).
16. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91-95 (2021).
17. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* **50**, 1161-1170 (2018).
18. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
19. Zhang, H., Xu, M.S., Fan, X., Chung, W.K. & Shen, Y. Predicting functional effect of missense variants using graph attention neural networks. *Nature Machine Intelligence* **4**, 1017-1028 (2022).
20. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

21. UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489 (2021).
22. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
23. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
24. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102-2110 (2022).
25. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021.07.09.450648 (2021).
26. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet* **55**, 1512-1522 (2023).
27. Matreyek, K.A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**, 874-882 (2018).
28. Suiter, C.C. *et al.* Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *Proc Natl Acad Sci U S A* **117**, 5394-5401 (2020).
29. Mighell, T.L., Evans-Dutson, S. & O'Roak, B.J. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *Am J Hum Genet* **102**, 943-955 (2018).
30. Heyne, H.O. *et al.* Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. *Sci Transl Med* **12**(2020).
31. Lester, H.A. & Karschin, A. Gain of function mutants: ion channels and G protein-coupled receptors. *Annu Rev Neurosci* **23**, 89-125 (2000).
32. Tao, Y.X. Constitutive activation of G protein-coupled receptors and diseases: insights into mechanisms of activation and therapeutics. *Pharmacol Ther* **120**, 129-48 (2008).
33. Gerasimavicius, L., Livesey, B.J. & Marsh, J.A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat Commun* **13**, 3895 (2022).
34. Iqbal, S. *et al.* Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc Natl Acad Sci U S A* **117**, 28201-28211 (2020).
35. Lage, K. Protein-protein interactions and genetic diseases: The interactome. *Biochim Biophys Acta* **1842**, 1971-1980 (2014).
36. Swan, H. *et al.* Gain-of-function mutation of the SCN5A gene causes exercise-induced polymorphic ventricular arrhythmias. *Circ Cardiovasc Genet* **7**, 771-81 (2014).
37. Blanchard, M.G. *et al.* De novo gain-of-function and loss-of-function mutations of SCN8A in patients with intellectual disabilities and epilepsy. *J Med Genet* **52**, 330-7 (2015).
38. McDonell, L.M., Kernohan, K.D., Boycott, K.M. & Sawyer, S.L. Receptor tyrosine kinase mutations in developmental syndromes and cancer: two sides of the same coin. *Hum Mol Genet* **24**, R60-6 (2015).
39. 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).

40. Zhang, C. *et al.* Gain-of-function mutant p53 in cancer progression and therapy. *J Mol Cell Biol* **12**, 674-687 (2020).
41. Sanders, S.J. *et al.* Progress in Understanding and Treating SCN2A-Mediated Disorders. *Trends Neurosci* **41**, 442-456 (2018).
42. Stein, D. *et al.* Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of a diverse feature set. *Genome Med* **15**, 103 (2023).
43. 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).
44. Li, Y., Zhang, Y., Li, X., Yi, S. & Xu, J. Gain-of-Function Mutations: An Emerging Advantage for Cancer Biology. *Trends Biochem Sci* **44**, 659-674 (2019).
45. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
46. Tate, J.G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947 (2019).
47. Chang, M.T. *et al.* Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer Discov* **8**, 174-183 (2018).
48. Patterson, S.E., Statz, C.M., Yin, T. & Mockus, S.M. Utility of the JAX Clinical Knowledgebase in capture and assessment of complex genomic cancer data. *NPJ Precis Oncol* **3**, 2 (2019).
49. Joosten, R.P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res* **39**, D411-9 (2011).
50. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637 (1983).
51. Valdar, W.S. Scoring residue conservation. *Proteins* **48**, 227-41 (2002).
52. Rao, R.M. *et al.* MSA Transformer. in *Proceedings of the 38th International Conference on Machine Learning* Vol. 139 (eds Marina, M. & Tong, Z.) 8844--8856 (PMLR, Proceedings of Machine Learning Research, 2021).
53. Fung, A., Koehl, A., Jagota, M. & Song, Y.S. The Impact of Protein Dynamics on Residue-Residue Coevolution and Contact Prediction. *bioRxiv*, 2022.10.16.512436 (2022).
54. Chang, M.T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol* **34**, 155-63 (2016).
55. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res* **33**, W382-8 (2005).
56. Hornbeck, P.V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**, D512-20 (2015).
57. Wan, P.T. *et al.* Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* **116**, 855-67 (2004).
58. Ng, P.K. *et al.* Systematic Functional Annotation of Somatic Mutations in Cancer. *Cancer Cell* **33**, 450-462 e10 (2018).
59. Martinez Fiesco, J.A., Durrant, D.E., Morrison, D.K. & Zhang, P. Structural insights into the BRAF monomer-to-dimer transition mediated by RAS binding. *Nat Commun* **13**, 486 (2022).

60. Tran, T.H. *et al.* KRAS interaction with RAF1 RAS-binding domain and cysteine-rich domain provides insights into RAS-mediated RAF activation. *Nat Commun* **12**, 1176 (2021).
61. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682 (2022).
62. Lu, H. *et al.* Engineering and Functional Characterization of Fusion Genes Identifies Novel Oncogenic Drivers of Cancer. *Cancer Res* **77**, 3502-3512 (2017).
63. Kiel, C., Benisty, H., Llorens-Rico, V. & Serrano, L. The yin-yang of kinase activation and unfolding explains the peculiarity of Val600 in the activation segment of BRAF. *Elife* **5**, e12814 (2016).
64. Maloney, R.C., Zhang, M., Jang, H. & Nussinov, R. The mechanism of activation of monomeric B-Raf V600E. *Comput Struct Biotechnol J* **19**, 3349-3363 (2021).
65. Lee, E.W. *et al.* Differential regulation of p53 and p21 by MKRN1 E3 ligase controls cell cycle arrest and apoptosis. *EMBO J* **28**, 2100-13 (2009).
66. Kakudo, Y., Shibata, H., Otsuka, K., Kato, S. & Ishioka, C. Lack of correlation between p53-dependent transcriptional activity and the ability to induce apoptosis among 179 mutant p53s. *Cancer Res* **65**, 2108-14 (2005).
67. Zhao, C. & MacKinnon, R. Molecular structure of an open human K(ATP) channel. *Proc Natl Acad Sci U S A* **118**(2021).
68. Matreyek, K.A., Stephany, J.J., Ahler, E. & Fowler, D.M. Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome Med* **13**, 165 (2021).
69. Papa, A. *et al.* Cancer-associated PTEN mutants act in a dominant-negative manner to suppress PTEN protein function. *Cell* **157**, 595-610 (2014).
70. Ren, P. *et al.* A Survey of Deep Active Learning. arXiv:2009.00236 (2020).
71. Rao, R. *et al.* Evaluating Protein Transfer Learning with TAPE. arXiv:1906.08230 (2019).
72. Shanehsazzadeh, A., Belanger, D. & Dohan, D. Is Transfer Learning Necessary for Protein Landscape Prediction? arXiv:2011.03443 (2020).
73. Ruff, K.M. & Pappu, R.V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J Mol Biol* **433**, 167208 (2021).
74. Muller, H.J. Further studies on the nature and causes of gene mutations. in *In Proceedings of the Sixth International Congress on Genetics* 215–255. (1932).
75. Stenson, P.D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* **136**, 665-677 (2017).
76. Page, D.J. *et al.* Whole Exome Sequencing Reveals the Major Genetic Contributors to Nonsyndromic Tetralogy of Fallot. *Circ Res* **124**, 553-563 (2019).
77. Thölke, P. & De Fabritiis, G. TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials. arXiv:2202.02541 (2022).
78. Kingma, D.P. & Ba, J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 (2014).