frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 05 July 2023
DOI 10.3389/fmolb.2023.1204157

Check for updates

OPEN ACCESS

EDITED BY
Annalisa Pastore,
King's College London, United Kingdom

REVIEWED BY
Carlos Bueno,
Rice University, United States
Xinyu Gu,
Rice University, United States

*CORRESPONDENCE
Martijn A. Huynen,
martijn.huijnen@radboudumc.nl

RECEIVED 11 April 2023
ACCEPTED 22 June 2023
PUBLISHED 05 July 2023

CITATION
Ramakrishnan G, Baakman C, Heijl S, Vroling B, van Horck R, Hiraki J, Xue LC and Huynen MA (2023), Understanding structure-guided variant effect predictions using 3D convolutional neural networks.
Front. Mol. Biosci. 10:1204157.
doi: 10.3389/fmolb.2023.1204157

COPYRIGHT
© 2023 Ramakrishnan, Baakman, Heijl, Vroling, van Horck, Hiraki, Xue and Huynen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Understanding structure-guided variant effect predictions using 3D convolutional neural networks

Gayatri Ramakrishnan[1], Coos Baakman[1], Stephan Heijl[2], Bas Vroling[2], Ragna van Horck[3], Jeffrey Hiraki[3], Li C. Xue[1] and Martijn A. Huynen[1]*

[1]Department of Medical Biosciences, Radboud University Medical Center, Nijmegen, Netherlands, [2]BioProdict, Nijmegen, Netherlands, [3]Vartion, Malden, Netherlands

Predicting pathogenicity of missense variants in molecular diagnostics remains a challenge despite the available wealth of data, such as evolutionary information, and the wealth of tools to integrate that data. We describe DeepRank-Mut, a configurable framework designed to extract and learn from physicochemically relevant features of amino acids surrounding missense variants in 3D space. For each variant, various atomic and residue-level features are extracted from its structural environment, including sequence conservation scores of the surrounding amino acids, and stored in multi-channel 3D voxel grids which are then used to train a 3D convolutional neural network (3D-CNN). The resultant model gives a probabilistic estimate of whether a given input variant is disease-causing or benign. We find that the performance of our 3D-CNN model, on independent test datasets, is comparable to other widely used resources which also combine sequence and structural features. Based on the 10-fold cross-validation experiments, we achieve an average accuracy of 0.77 on the independent test datasets. We discuss the contribution of the variant neighborhood in the model's predictive power, in addition to the impact of individual features on the model's performance. Two key features: evolutionary information of residues in the variant neighborhood and their solvent accessibilities were observed to influence the predictions. We also highlight how predictions are impacted by the underlying disease mechanisms of missense mutations and offer insights into understanding these to improve pathogenicity predictions. Our study presents aspects to take into consideration when adopting deep learning approaches for protein structure-guided pathogenicity predictions.

KEYWORDS

protein structure, 3D CNN, missense variant, machine learning, gain-of-function, loss-of-function

1 Introduction

Numerous Mendelian diseases can be attributed to alterations in the coding regions of the DNA, i.e., missense variants (Kryukov et al., 2007). With rapid advances in sequencing technologies, the ease and ability to map a person's complete genome has dramatically aided in obtaining genetic diagnosis. Nevertheless, only a small fraction of the missense mutations is pathogenic (Lek et al., 2016) and for the majority of missense variants it is not clear whether the phenotypic outcome is pathogenic or neutral. Such variants are coined "variants

of uncertain significance" (VUS). Evidently, identifying and comprehending the functional effects of missense variants is of critical importance, not only to understand the etiology of the disease but also towards development of treatment regimens.

Significant advances have been made in the development of variant effect predictors that largely rely on evolutionary conservation, which is a strong signal for predicting pathogenicity. Such evolutionary cues in combination with physicochemical properties of amino acids form the base framework of several state-of-the-art techniques including SIFT (Ng and Henikoff, 2003), PolyPhen2 (Adzhubei et al., 2010), CADD (Kircher et al., 2014), and MutPred (Li et al., 2009). Although evolutionary information holds value in predicting pathogenicity, it does not provide mechanistic understanding. The mechanisms of the pathogenicity of missense variants are often attributable to perturbations in conformational and functional properties of three-dimensional structures (Wang and Moult, 2001; Iqbal et al., 2020), which can contribute to our understanding of the underlying molecular pathology. Several studies have thus incorporated features that leverage structural properties (Venselaar et al., 2010; Capriotti and Altman, 2011; Ittisoponpisan et al., 2019; Laskowski et al., 2020), protein dynamics (Ponzoni et al., 2020), protein-protein interaction networks (Yates et al., 2014), and protein structural stability (Ancien et al., 2018), to improve pathogenicity predictions on top of what can be achieved with sequence conservations. In the absence of experimental structural information, context-dependent sequence-based models have the potential to accurately capture intra-protein 3D contacts, i.e., via evolutionarily coupled residues (Morcos et al., 2011; Marks et al., 2012; Hopf et al., 2014). Utility of such models has shown reasonable improvement in distinguishing pathogenic missense variants from benign ones (Feinauer and Weigt, 2017; Hopf et al., 2017). A complete list of available resources and tools for variant effect prediction and their benchmark evaluation studies has been published elsewhere (Liu et al., 2011; Livesey and Marsh, 2022). Despite the significant advances, the challenge of distinguishing pathogenic variants from benign ones remains elusive with most methods exhibiting a wide spectrum of performances on different test datasets (Niroula and Vihinen, 2019; Livesey and Marsh, 2020).

Most knowledge-driven approaches that employ machine learning (ML) classifiers rely on various handcrafted features to predict variant effects, which could be time-consuming and laborious. This is compounded by heterogeneity in feature attributes that can pose challenges in data integration (Bagley and Altman, 1995). Deep learning accelerated approaches can help overcome such limitations. CNNs have gained prominence in the last decade due to their ability to automatically capture patterns from input data as well as the hierarchical representations therein (Krizhevsky et al., 2012), enabling them to capture relationships between different features. This aspect is particularly useful for analyzing high dimensional data such as protein structures.

Recent efforts have demonstrated the use of 3D-CNNs in exploiting protein structure data for several applications including the prediction of amino acids compatible with protein microenvironments (Torng and Altman, 2017; Pun et al., 2022), identification of novel gain-of-function mutations (Shroff et al., 2020), and the prediction of mutation-induced changes in protein stability (Li et al., 2020). We introduce DeepRank-Mut, a configurable 3D-CNN framework that predicts pathogenicity of missense variants using wildtype structural microenvironment surrounding the variants in 3D space. The base framework is derived from its parent DeepRank that distinguishes and ranks biologically relevant protein-protein interactions from those that arise due to crystallographic artifacts (Renaud et al., 2021). The underlying premise of our approach is that the functional outcome of any missense variant is often reflected in the properties of amino acids in the variant neighborhood, in addition to the properties of the variant amino acid itself. Our approach is similar to the method devised by Torng and Altman (2017), which, given a site, predicts the amino acids compatible with that specified site based on the surrounding protein microenvironment. In contrast, we train our model explicitly to learn label-specific (benign or pathogenic) features/patterns in the variant neighborhood. Given a missense variant, we first obtain the associated 3D protein structure, either from the protein itself or from a homolog, and calculate features including surface geometry, empirical energies, and atomic densities, in addition to the sequence conservation scores for the mutated site as well as the residues in its neighborhood. These features are mapped onto 3D grids parameterized using properties of the constituent atoms, followed by data augmentation to enrich the input dataset. We then use the power of 3D-CNNs to automatically discern spatially proximal features within these representations.

DeepRank-Mut achieves a performance comparable to techniques that efficiently combine sequence and structure-based features. We analyze the contribution of each of the features to the model's predictive ability, as well as how the neighborhood contributes to the performance. To better understand predictor accuracy, we explore underlying mechanisms of pathogenic mutations and show that the features identify autosomal recessive mutations better than autosomal dominant mutations. We discuss the overall generalizability of our method and provide avenues for better 3D-based missense variant prioritization.

2 Methods

2.1 Datasets

A total of 193,714 missense variants (164,574 benign, 29,140 disease-causing) were collected from ClinVar (Landrum et al., 2018), gnomAD (Karczewski et al., 2020) and Dutch genome diagnostic laboratories (VKGL, 2019), which could be linked to protein structures, either directly or through homology with a sequence identity cut-off of 40%. This cutoff was selected based on previous research that suggests that a 40% identity corresponds to a good likelihood of functional equivalence (Pearson, 2013). Missense variants were mapped onto protein structures using 3DM systems as a guide (Kuipers et al., 2010). Independent test datasets were obtained from studies based on BRCA1 (Findlay et al., 2018), Gunning et al. (2021) and the InSIGHT database (Thompson et al., 2014). This resulted in a total of 217,679 missense variants that could reliably be mapped onto 57,551 structures; 25,856 structures were mapped to 40,369 pathogenic variants, and 31,695 structures were mapped to 177,310 benign variants. It should be noted that, at this stage, the structures are mapped regardless of the experimental method used for their determination. Missense variants from ClinVar were incorporated if they had a review status of at least one star, excluding those with conflicting

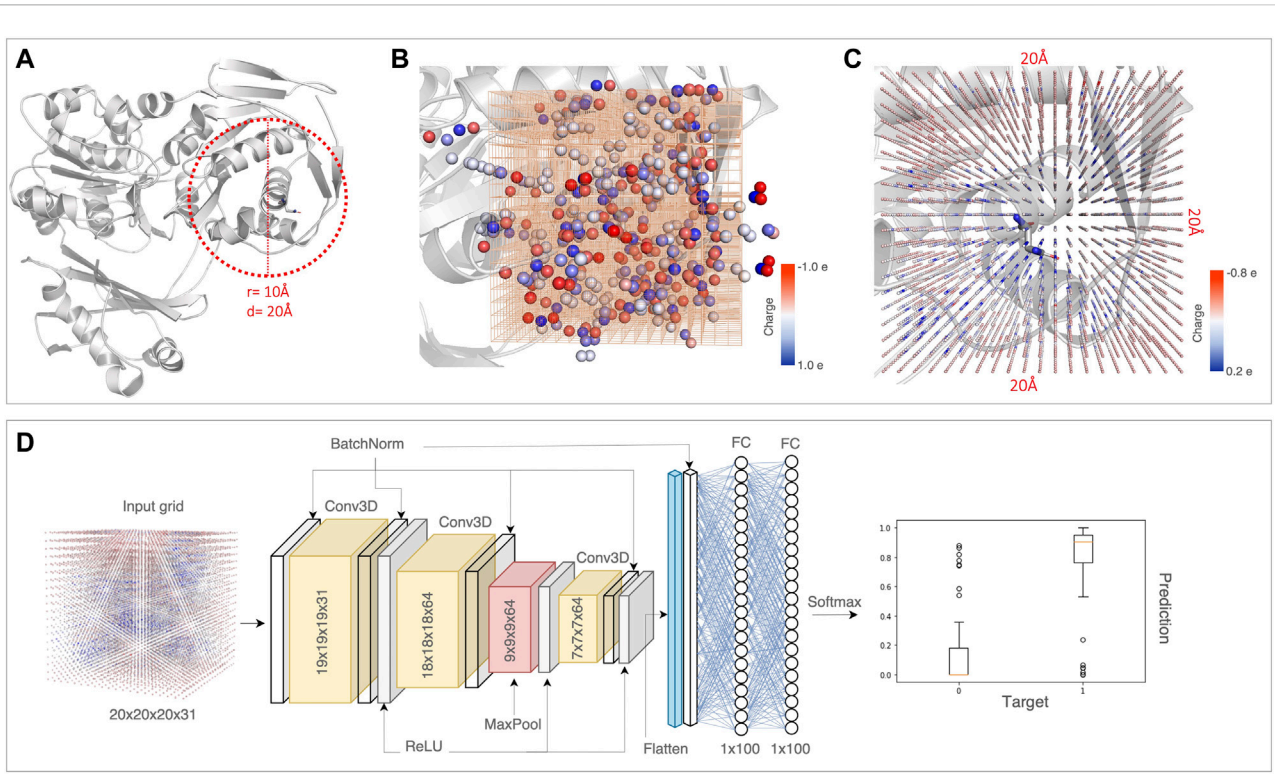

FIGURE 1
A schematic of the DeepRank-Mut framework. **(A)** The first step includes extraction of the variant environment, where residues within a radius of 10Å (diameter of 20Å) around the variant are drawn. As an example case, the crystal structure of phosphoglucomutase (PDB: 1C4G) with the missense variant Asn37 is depicted. This is followed by the feature calculation step where structural properties and PSSM scores are computed for the variant site and the residues in its environment. All features are localized on atoms as illustrated in **(B)**. For simplicity, one structural property (charge), localized on atoms, is shown. A 3D grid of size 20 × 20 × 20 is centered at the Cα atom of the residue at variant site, discretized into voxels of 1Å. **(C)** Each of the features calculated are normalized using standardization and then mapped onto the grid using a Gaussian function. For simplicity, the Gaussian mapping of one feature, i.e., charge for all atoms within a 20Å box is depicted. In principle, a total of 31 calculated features are mapped to the 3D grid for a given variant. **(D)** This 3D grid with mapped features of shape (20, 20, 20, 31) serves as an input for the 3D-CNN network. The final classification score takes a value between 0 and 1 for each class (benign and pathogenic).

interpretations. "Benign" and "Likely benign" ClinVar variants were included and categorized as benign, while "Pathogenic" and "Likely pathogenic" variants were incorporated and classified as pathogenic. The gnomAD variants with a minor allele frequency higher than 0.1% were selected and labeled as benign.

Our in-house database, HSSP (Touw et al., 2015) was consulted to obtain structure-based sequence alignments. Position-specific scoring matrices (PSSMs) were constructed for the alignments using PSI-BLAST (Altschul et al., 1997) with single iteration. Each of the PSSMs were then mapped back onto their respective structures using the PSSMGen package (https://github.com/DeepRank/PSSMGen).

2.2 Data pre-processing

2.2.1 Feature calculation and voxelization of the neighborhood

We use protein crystal structures of resolution better than 3Å in our study, as these provide details at the atomic level with high certainty (Zardecki et al., 2022). Consequently, variants that are mapped to structures solved using methods other than X-ray crystallography, such as NMR or cryo-EM, are excluded. For ease in data handling, we mapped each missense variant to a maximum of three crystal structures of the most similar sequences. For each variant mapped to a crystal structure, we first extract the local neighborhood with a radius of 10Å around the variant, which typically serves as a distance beyond which the strength of long-range non-bonded interaction energies gradually weakens (Pincus and Scheraga, 1977). We include residues whose atoms fall within this radius to obtain residue-based features. This is followed by calculation of atomic features such as densities and charges for the wildtype amino acid and the residues in its microenvironment. Pairwise Coulomb and van der Waals potentials are calculated between atoms of the wildtype residue and the residues in the neighborhood. For a given atom, these features are defined as the sum of all pairwise potentials between the atom and its contact atoms. Bonded pairs, i.e., pairs of atoms separated by up to 2 bonds are excluded from this measure. The atomic densities, charges and non-bonded energies are based on the OPLS force field (Jorgensen and Tirado-Rives, 1988), calculated in the same manner as in the parent DeepRank (Renaud et al., 2021) (see Supplementary

TABLE 1 List of features calculated for the residue at the mutation site and the residues in its neighborhood.

Features	Number of channels
Atomic densities (C, N, O, S)	4
Atomic charges	1
Solvent accessibility	1
Coulomb potential	1
van der Waals potential	1
Wildtype score: PSSM	1
Variant score: PSSM	1
Information content (PSSM)	1
PSSM profile	20
Total	31

All features, including residue-level features such as sequence conservation scores, are localized on atoms. The two sequence-based features (wildtype and variant probability) are mapped to the atoms of a given wildtype residue.

Methods). Solvent accessible surface area (SASA) is calculated using FreeSASA (v2.0.3) (Mitternacht, 2016). Water molecules in protein structures, when present, are not included in the analysis. In addition to the PSSM obtained for the wildtype and variant amino acids, we also include the PSSM profile for the residues in the variant microenvironment. Such residue-based feature values are assigned to the residue's constituent atoms. All feature values are localized on atoms, to be subsequently mapped on a 3D grid (see Figure 1); only those atoms that lie within 10Å radius of the variant are considered. At this stage, it should be noted that some structures in the PDB database may contain missing residues that fall within the variant environment radius, leading to errors in the feature mapping step. Such molecules are thus, excluded from the dataset.

We construct a 3D grid of size 20Å × 20Å × 20Å centered at the Cα atom of the amino acid at the variant site. This 20Å box is divided into voxels of 1Å, parameterized with 31 physicochemical property channels (Table 1). The properties are mapped on a 3D grid using Gaussian functions to approximate atom connectivity, as demonstrated previously in the parent DeepRank framework (Renaud et al., 2021). The contribution (w_k) of an atom k to a given grid point is determined based on Gaussian distance dependence, i.e., the contribution decreases with increasing distance between the atom and the grid point. This is given by the equation:

$$w_k(r) = v_k \, exp\left(\|r - r_k\|^2 \,/ 2\sigma^2\right) \quad (1)$$

where v_k is the feature value, r denotes position of the grid point and r_k denotes atomic coordinates (x, y, z). The standard deviation σ denotes the van der Waals radius of the associated atom.

The feature maps are stacked to create a tensor of shape (20, 20, 20, 31) that then serves as an input to the neural network. We also normalize features of the input data using standardization prior to training. To optimize for speed and efficient handling of large volumes of data, we developed a distributed data preprocessing framework with GPU support, which enabled faster preprocessing times and scalability during numerous iterations of experiments (see Supplementary Methods, Supplementary Figure S1).

2.2.2 Data augmentation

Prior to the training step, we enrich each of the input 3D grids using data augmentation where a given grid is randomly rotated around its center, and features are mapped onto the grid subsequently. Such a strategy has been shown to improve the performance of CNNs (Shorten and Khoshgoftaar, 2019). For the current study, we used 5 augmentations based on hyper parameter tuning experiments (Supplementary Figure S3). We did not experiment with a higher number of augmentations due to the infeasible computational costs involved.

2.3 Network architecture

The network used in our study includes a sequential organization of three 3D convolutional layers, alternating with one 3D max pooling layer followed by two fully connected layers (Figure 1). We include batch normalization layers, in addition to dropout layers between the fully connected layers to regularize the model. Details of the architecture are provided in Table 2 and the complete schema is provided in Supplementary Figure S2. Each 3D convolution layer comprises a set of learnable filters that traverse the input space (depth, height and width) with a stride of 1, capturing local spatial patterns in the variant environment. The output from convolution operations, i.e., the computed feature maps are transformed by a rectified linear activation function (ReLU), which allows the network to identify and extract meaningful spatial features. This is followed by dimension reduction using max pooling operation and a final 3D convolutional layer with ReLU. The transformed output is then flattened to a one-dimensional vector that serves as an input to two fully connected layers. The two final layers integrate the features and apply a set of weights that are optimized during the training step to map extracted features to target classes. The output is then passed through the softmax function which provides the final classification score, a probability estimate between 0 and 1, each for benign and pathogenic classes.

2.4 Training

We performed 10-fold cross validation experiments while ensuring that the missense variants in the training and validation sets are from different proteins, to avoid type 1 circularity in predictions (Heijl et al., 2020). The test dataset included missense variants independent from the 10-fold training and validation sets. Most genetic variation is neutral, and it is therefore rather common to observe a higher number of benign variants than pathogenic variants in the training data, which has the potential to bias training and performance. We thus constructed balanced subsets of randomly sampled benign and pathogenic missense variants for each of the 10-fold runs. For efficient memory handling, we employed training in mini-batches of 256 variant instances which amounted to ~1,200 mini-batches per epoch. An epoch refers to a single pass through the complete training data during which the model weights are adjusted to minimize the error between predicted and true label for each input. With the input dataset, one epoch in our approach referred to one pass through more than 280,000 variant instances. We used the AdamW optimizer (Loshchilov and Hutter, 2019) with a

TABLE 2 Network architecture used in DeepRank-Mut.

Layer	Size	Output shape
Batch normalization layer 1	Input	20 × 20 × 20 × 31
3D convolutional layer 1	20 × 20 × 20, 31 filters, kernel size = 2, stride = 1	19 × 19 × 19 × 31
Batch normalization layer 2		19 × 19 × 19 × 31
3D convolutional layer 2	19 × 19 × 19, 64 filters, kernel size = 2, stride = 1	18 × 18 × 18 × 64
Batch normalization layer 3		18 × 18 × 18 × 64
3D max pooling layer	Stride = 2	9 × 9 × 9 × 64
3D convolutional layer 3	9 × 9 × 9, 64 filters, kernel size = 3, stride = 1	7 × 7 × 7 × 64
Batch normalization layer 4		7 × 7 × 7 × 64
Flatten	7 × 7 × 7 × 64	21,952
Batch normalization layer 5		21,952
Fully connected layer 1	21,952 × 100 neurons	100 neurons
Dropout (p = 0.5)		
Fully connected layer 2	100 × 100 neurons	100 neurons
Dropout (p = 0.5)		
Softmax	100 × 2	2 scores (benign, pathogenic)

learning rate of 0.001 and weight decay of 0.005 to train our model for 10 epochs. We used cross entropy loss during training, which attempts to minimize the differences in probability distributions between predicted and ground truth labels by adjusting weights. A dropout rate of 0.5 was used to regularize the model. The hyperparameters including number of convolutional layers, number of max pooling layers, grid size, were optimized based on performance on validation set across 10 folds, starting from default parameters of the parent DeepRank.

2.5 Evaluation metrics

Two metrics, Matthews Correlation Coefficient (MCC) and accuracy, were used to evaluate the performance of DeepRank-Mut. The primary metric used was MCC, as it offers a reliable statistical measure by taking all four categories-true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) into account, proportional to the size of the binary classes (Eq. 2). The usefulness of MCC over accuracy or F1 scores for binary classification has been demonstrated previously (Chicco and Jurman, 2020).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (2)$$

For comparative evaluation with popular state-of-the-art variant effect predictors, we used precomputed pathogenicity prediction scores of 8 algorithms from dbNSFP v4.3 database (Liu et al., 2020; 2011), including SIFT4G (Vaser et al., 2016), PolyPhen2 (Adzhubei et al., 2010), MutationTaster (Schwarz et al., 2014), MutationAssessor (Reva et al., 2011), FATHMM (Shihab et al., 2013), VEST4 (Carter et al., 2013), PROVEAN (Choi and Chan, 2015) and MutPred (Li et al., 2009), as well as prediction scores from Helix (Vroling and Heijl, 2021). Where available, we used "converted rankscores" from dbNSFP to ensure that a higher score always indicated higher likelihood of pathogenicity. We excluded meta predictors from this comparison, as well as those that combine annotations from other tools, to account for methods that rely on first principles to predict functional effects of missense variants.

3 Results

3.1 Overview of the datasets and DeepRank-Mut

Training, validation, and test sets are often generated using a simple random split. However, this can result in over fitting and misleading results due to data leakage between the training and evaluation sets (Heijl et al., 2020). Data splitting at the level of proteins or genes, where training sets never include any data samples from proteins that occur in the validation or test set is used to mitigate this. We split our dataset into 10 pairs of training and test sets, each containing 90% and 10% of the full dataset, respectively, allowing 10-fold cross validation on the full dataset. Independent test sets were gathered from three studies, as described in methods, to aid in the final assessment of the tool's performance. These test sets have been selected as they not only cover genes in-depth (Thompson et al., 2014; Findlay et al., 2018) but are also aimed at benchmarking pathogenicity predictors specifically (Gunning et al., 2021).

After splitting the data, the balanced subsets of randomly sampled benign and pathogenic variants, each mapped to at most three structures, comprised a total of ~50,000 instances in the training set, ~4,700 instances in validation and 6,571 in the test

set, per fold. The test set was kept identical across all cross-validation folds for an unbiased evaluation of the model.

DeepRank-Mut retains its modularity in implementing data pre-processing steps and training the deep neural network, similar to its parent DeepRank (Renaud et al., 2021). It allows for flexibility in tasks including feature calculations, setting the grid size and grid resolution, data augmentation, as well as optimizing hyperparameters of the neural network. The base requirements of DeepRank-Mut include a dataset of variants with labels (benign or pathogenic), a dataset of variant-structure maps where each variant is linked to a 3D structure (either experimentally determined or evolutionarily related), a dataset of 3D structures and an optional dataset of PSSM profiles derived for each structure. As detailed in the methods, the framework computes physicochemical properties of the amino acid at the variant site as well as its environment within a radius of 10Å, followed by voxelization to encode the atomic neighborhood of residues (Figures 1A, B). Our approach relies on leveraging local properties of sites characteristic of benign or pathogenic variants, as pathogenic variants generally tend to occur in regions important for structural/functional integrity of the protein (Iqbal et al., 2020), like its hydrophobic core. We thus compute a total of 31 features (Table 1), encompassing structural and sequence-based properties, for the residue at the variant site and residues spatially proximal to it. The computed features are mapped to a 3D grid where each voxel is parameterized with the feature channels (Figure 1C), which is then followed by data augmentation. As a given variant environment can differ in orientation within or across proteins, the data augmentation step accounts for rotational invariance, thereby improving the model's robustness to variations in input data (Supplementary Figure S3). From our dataset of structures and missense variants, we generated ~300,000 augmented grids per fold dataset, which were used as input to 3D-CNN (Figure 1D). Each augmented 3D grid is treated as a separate variant instance, thus our model outputs 6 predictions per missense variant (origin grid +5 augmented grids) which are averaged to give one final classification score.

3.2 Overall performance

Our approach achieved a mean accuracy of 0.77 and an average MCC score of 0.52 across the test datasets, with an average sensitivity (true positive rate) of 0.75 and an average specificity (true negative rate) of 0.78 (Figure 2; Table 3).

3.2.1 Impact of individual features and the variant environment on the performance

To investigate the contribution of neighborhood in the predictor accuracies, we compared the performance of our 3D-CNN model trained on all features to those trained separately on-a) PSSM features, b) structural features, c) variant site-specific PSSMs (Figure 3A). The model trained on PSSM features included PSSMs for the residues in the 3D neighborhood as well as the scores for wildtype and variant amino acids, while the model trained on variant site-specific PSSMs was devoid of the neighborhood profile. As illustrated in the figure, the features derived from the neighborhood, in the 3D context, seemingly hold more information than the site-specific features. This aspect was also observed during hyperparameter tuning experiments, where a range of different sizes of 3D grids were tested to find the optimal grid size. Models with smaller variant neighborhoods (grid sizes = 7Å, 8Å) performed poorly on validation sets as compared to the models with grid size of 15Å and 20Å (Supplementary Figure S4). It has been reported earlier that the atomic details do not provide significant information for local protein environments beyond a 20Å cutoff (Bagley and Altman, 1995). An optimal grid size of 20Å was thus chosen for all experiments. Additionally, we investigated the apparent contribution of individual structural features in prediction accuracies, as illustrated in Figure 3B. We note that solvent accessibility of residues has the most predictive capacity amongst all structural features. Residues buried in the hydrophobic core of the protein are often associated with pathogenicity, while solvent-exposed missense variants are often found to be enriched in populations, as also exemplified by Iqbal et al. (2020).

Additionally, we also performed leave-one-feature-out analysis to assess redundancy in our feature selection. Figure 3C illustrates similarity in ROC curves of models trained without pairwise potentials (Coulomb + van der Waals), atomic charges and atomic densities. The contributions of these features in prediction accuracies are similar as also noted in Figure 3B, suggesting redundancies in features employed. Subsequently, we tested our model's performance by excluding seemingly redundant features, such as atomic densities and charges from the feature set (Figure 3D). Although minimal, the contribution of each of the structural features holds value in the overall performance. Significantly, solvent accessibility and PSSMs show considerable impact on the model's performance.

3.2.2 Comparison with state-of-the-art resources

We used precomputed pathogenicity scores of 8 algorithms from dbNSFP database as well as scores from the Helix for the test dataset used in the study. In the case of PolyPhen2, we used scores from the HumVar-trained models as recommended by the authors for the purpose of distinguishing variants with drastic functional effects from benign ones (Adzhubei et al., 2013). Figure 4 illustrates the ROC curves drawn from these scores along with those from DeepRank-Mut for the variant predictions available for each algorithm. While the performance of our approach is seemingly comparable to other widely-used resources that incorporate sequence conservation and structural features, such as MutPred (Li et al., 2009) and PolyPhen2 (Adzhubei et al., 2010), it must be noted that the available variant predictions for these tools constitute 62% and 72% of the total test set, respectively (n in Figure 4, Supplementary Table S1). Both these ML-based tools incorporate several handcrafted features, aside from sequence conservation, including secondary structural assignments, normalized B-factors, and various annotations of functional sites; the only overlapping features with DeepRank-Mut being SASA and sequence conservation. Helix, built on proprietary structure-based sequence alignments (Kuipers et al., 2010; Vroling and Heijl, 2021), and VEST4, a variant prioritization tool that explores enrichment of functional variants across disease exomes (Carter et al., 2013), were notably the top performers.

3.3 3D-CNNs appear less powered to identify outcome of solvent-exposed variants

We examined our model's predictive ability by analyzing missense variants in the test-set that were consistently predicted

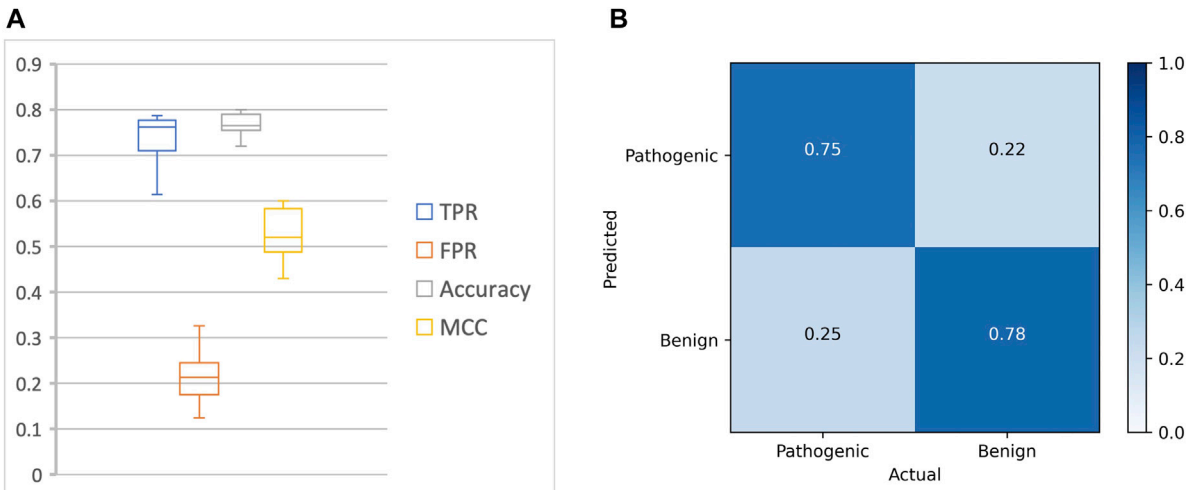

FIGURE 2
Overall performance of DeepRank-Mut. **(A)** The performance metrics of DeepRank-Mut on test sets across 10 folds are depicted as boxplots for true positive rate (TPR), false positive rate (FPR), accuracy and MCC. **(B)** Confusion matrix depicting the average of TP, FP, FN, and TN across 10 folds.

TABLE 3 Details of the performance metrics of DeepRank-Mut on test sets across 10 folds.

Fold	TPR	FPR	Accuracy	MCC
1	0.613	0.142	0.78	0.49
2	0.765	0.241	0.76	0.52
3	0.70	0.124	0.79	0.59
4	0.768	0.258	0.76	0.51
5	0.777	0.24	0.77	0.52
6	0.787	0.188	0.80	0.60
7	0.713	0.235	0.74	0.48
8	0.777	0.191	0.79	0.58
9	0.758	0.326	0.72	0.43
10	0.716	0.186	0.76	0.53
Average	0.737	0.213	0.77	0.52

incorrectly across all 10 folds. We explored the aspects that promoted incorrect classification. A total of 2,883 missense variants were found to be incorrectly classified across the cross-validation experiments, of which more than half (1,732) consisted of misclassified pathogenic variants. We computed relative solvent accessibilities (RSA) for each variant residue, by dividing their absolute solvent accessibilities in Å^2 by their maximum allowed solvent accessibilities obtained from Rost and Sander (Rost and Sander, 1994). Residues were categorized as solvent-exposed if the RSA values were >20% and buried if below 20%. Using these a substantial proportion of the misclassified pathogenic variants was found to be solvent-exposed (Supplementary Table S2).

We constructed 2 × 2 contingency tables based on the correct and incorrect classifications with respect to solvent accessibility of the associated variants. Figure 5 illustrates the role of solvent accessibility in the predicted outcomes. The misclassified variants pertained to solvent-exposed pathogenic variants and buried benign variants (Figure 5A, odds ratio = 0.27). That we are relatively successful in predicting pathogenicity in buried variants is consistent with the notion of buried enrichment of pathogenic variants (Iqbal et al., 2020; Savojardo et al., 2020). The distribution of raw atom-level solvent accessibility values across benign and pathogenic classes calculated in our approach is illustrated in Supplementary Figure S9. Two reasons for the quality of the predictions could be postulated: a) considering the contribution of SASA in the model's performance, it is likely that the model is unable to generalize on missense variants that fall outside the purview of typical SASA distribution observed in benign and pathogenic variants, or b) the 3D input grids for solvent-exposed missense variants are sparsely populated which leads to a lack of discernible patterns/features for the model to learn from.

We created separate training subsets of buried and solvent-exposed variants to understand 3D-CNN's generalizability to either subset. We observed that the predictions on pathogenic variants improved with the model trained on buried missense variants alone, however, this model misclassified much of the benign variants, whereas the model trained on solvent-exposed variants alone showed a performance comparable to that of the full model trained on all variants (Figure 5B; Supplementary Figure S5). It is possible that the presence of a large proportion of solvent-exposed variants in our training data may have impacted the performance (Supplementary Figure S9). Furthermore, to assess whether sparsity of 3D grids of solvent-exposed variants affected the model's performance, we calculated the ratio of solvent (void) voxels to atom-contained (non-void) voxels in the 3D grids in test dataset and compared the distribution of these ratios against the corresponding

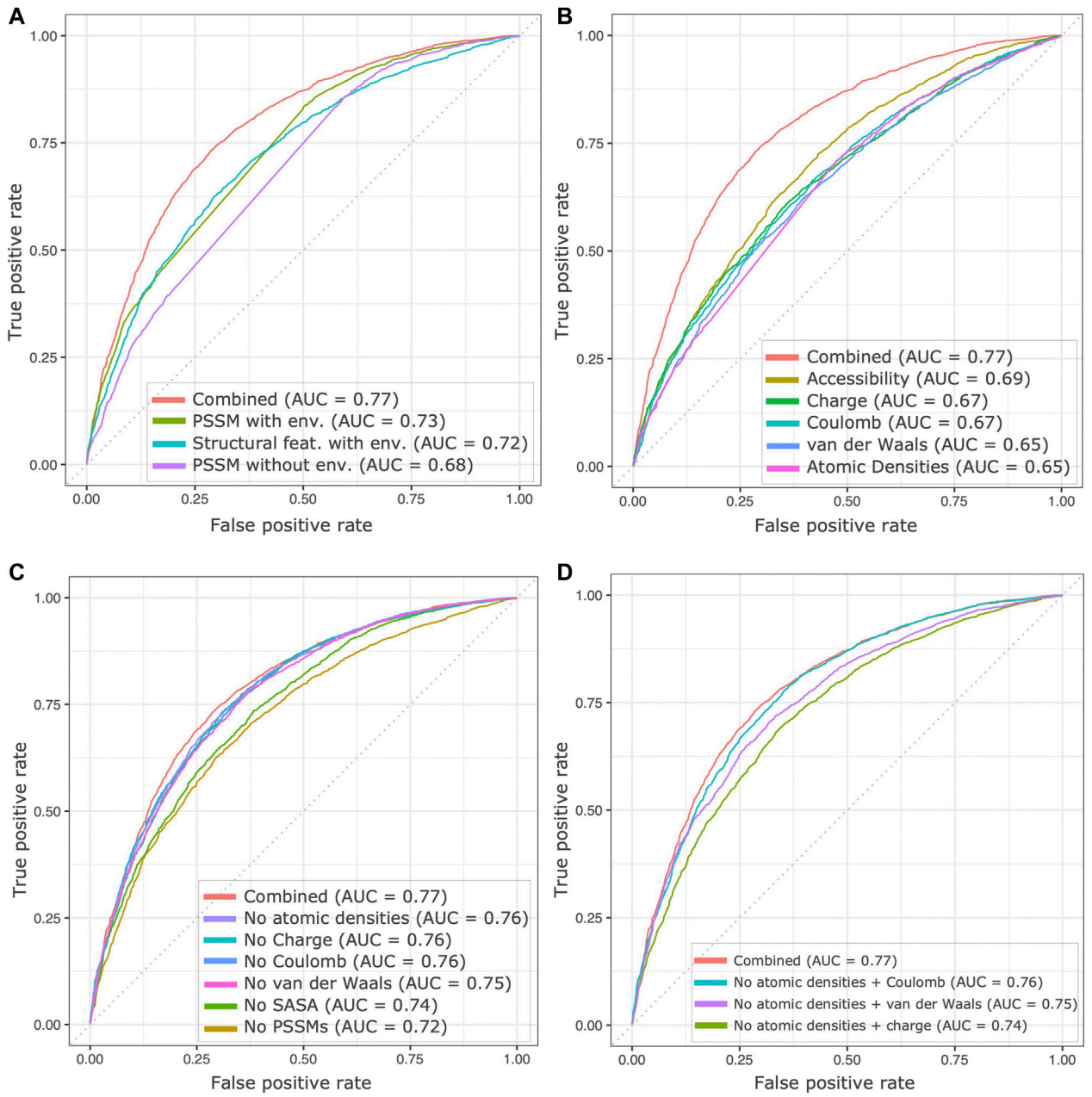

FIGURE 3
Contribution of the variant neighborhood and features to the model's predictive ability. **(A)** Receiver operating characteristic (ROC) curves are drawn from scores generated by the complete model (red) and by models separately trained only on the PSSM profile of residues in the neighborhood (green), only on the structural features from the neighborhood (cyan) and PSSM of mutation site alone (purple). The AUC values obtained illustrate the value of using 3D neighborhood information in the predictions. **(B)** ROC curves are drawn using scores generated by models separately trained on the individual structural features. **(C)** ROC curves for leave-one-feature-out analysis are drawn using scores from models trained without a specific feature. Model trained without PSSMs (beige) and the model trained on structural features from neighborhood (cyan) in the first panel **(A)** are identical. **(D)** ROC curves are drawn for scores from models where redundant features are removed. For consistency the model trained on all 31 features is included in all panels. Total n = 6,571 instances, 3,804 pathogenic and 2,767 benign.

pathogenic and benign prediction scores. We find no correlation for pathogenic variants (Pearson's r = −0.09), while we find that the presence of void voxels is weakly indicative of correct classifications for benign variants (Pearson's r = −0.24) (Supplementary Figure S10). This overall suggests that grid sparsity has weak effect on the correct classification of benign variants, whereas the incorrect classifications of solvent-exposed pathogenic variants is possibly due to other reasons, such as lack of function-specific features, and/or incomplete knowledge of their interaction partners.

Since data augmentation and feature normalization strategies, typically used to circumvent lack of generalizability and potential biases, are already incorporated in our approach we experimented

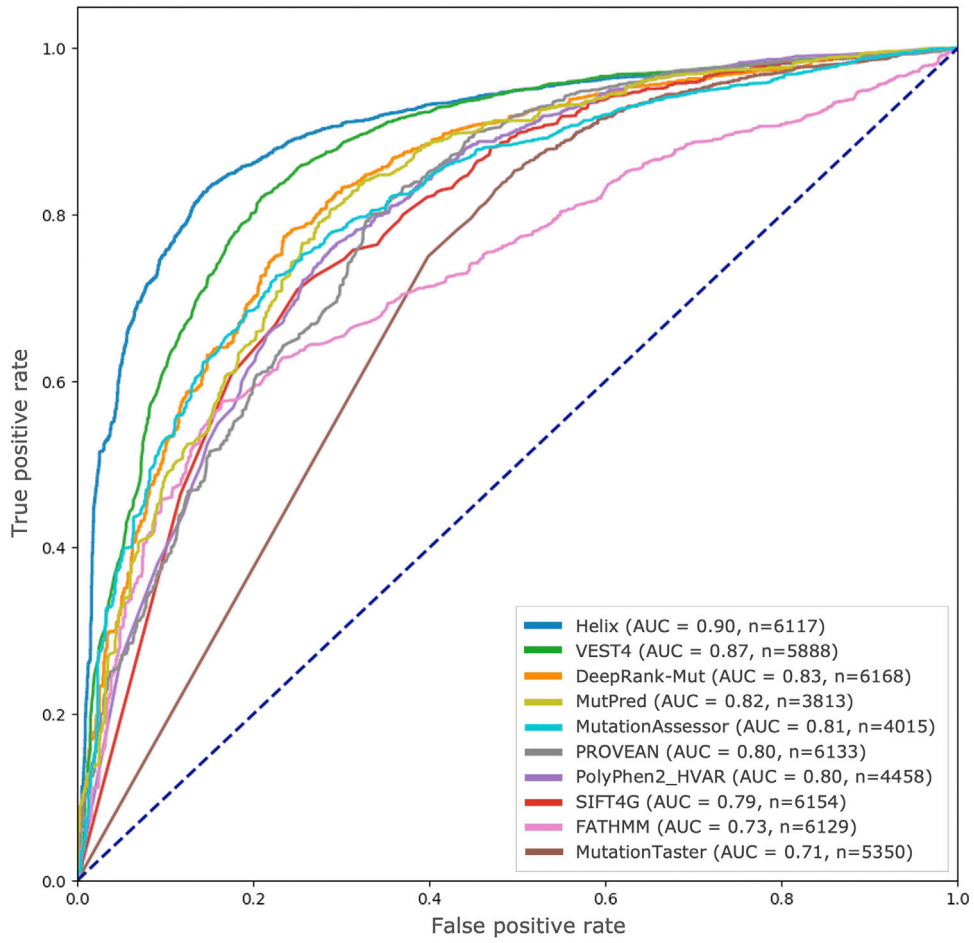

FIGURE 4
Comparison with other state-of-the-art resources. ROC curves drawn from scores generated by various pathogenicity predictors, including DeepRank-Mut, are shown based on the test variants available for each predictor in dbNSFP.

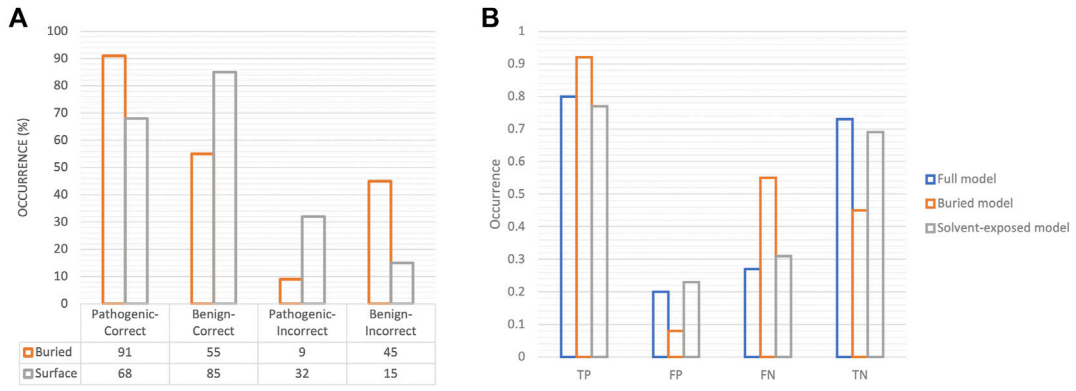

FIGURE 5
Association of solvent accessibility of variants in prediction outcomes. **(A)** Bar charts for correctly classified and misclassified variants with respect to their solvent accessibility are shown. **(B)** The performance metrics on test data in terms of TP, FP, FN, TN are depicted as bar charts for models trained on all variants (full model), on only buried variants (buried model) and on solvent-exposed variants alone (solvent-exposed model). The proportion of true positives, i.e., pathogenic variants in the model trained on buried variants is notably high.

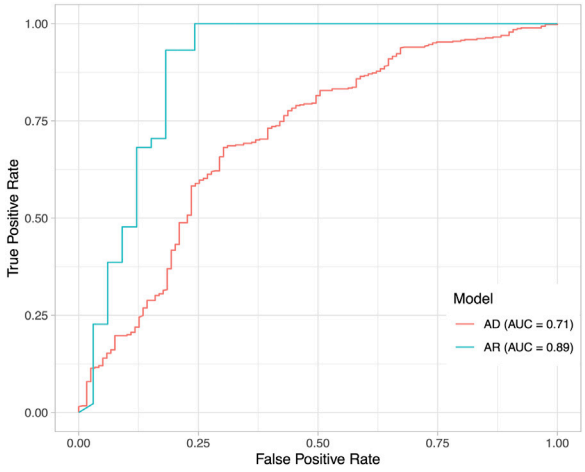

FIGURE 6
Impact of underlying disease mechanisms on pathogenicity predictions. Performance of DeepRank-Mut on two datasets that are divided based on mode of inheritance. ROC curves are drawn for scores generated from the model tested on variants with AD inheritance (n = 585), and from the model tested on variants with AR inheritance (n = 77). The AUC values are markedly different between the two datasets as depicted. It must be noted that the predictions are made for those variants that could be mapped to protein crystal structures.

with inclusion of other structural features: secondary structural content and normalized B-factors. The premise behind use of secondary structural content was based on the report by Abrusán and Marsh (Abrusán and Marsh, 2016), who showed differences in the ability of alpha helices and beta strands to tolerate mutations. Secondary structural assignments for protein structures were obtained from our in-house database (DSSP v.3.1.4) (Kabsch and Sander, 1983), and were stored as one-hot encoded features in 3D grids. B-factors or temperature factors are obtained from X-ray crystallography experiments that indicate atomic flexibility in the protein's crystalline state, and are known to correlate with flexible regions of the protein. Based on the earlier reports of active/functional sites associated with lower B-factors as compared to non-functional residues (Sun et al., 2019), we used normalized B-factors as a feature to potentially capture such differences. However, the two additional features did not serve as strong determinants of pathogenicity (Supplementary Figure S6). The relatively low quality of predictions for solvent-exposed pathogenic variants and buried benign variants could be due to lack of function-specific features.

3.4 Success of pathogenicity prediction depends on underlying disease mechanisms

We further investigated DeepRank-Mut's generalizability with respect to mutation mechanisms. Most available pathogenicity predictors do not make a distinction between different types of mutation mechanisms such as loss-of-function (LoF) or gain-of-function (GoF), that are often linked to mode of inheritance. LoFs are function-disrupting mutations that usually cause damage to protein structures and are straightforward to comprehend and identify, as they are generally not tolerated at sites of high structural and/or functional importance, and lead to degradation of the protein. In contrast, GoFs exhibit milder effects on protein stability while giving rise to altered protein functions that lead to diseases (Gerasimavicius et al., 2022). In terms of mode of inheritance, autosomal recessive (AR) diseases are predominantly linked to LoFs, while autosomal dominant (AD) diseases manifest through mechanisms such as GoFs, dominant-negative mutations (DN)

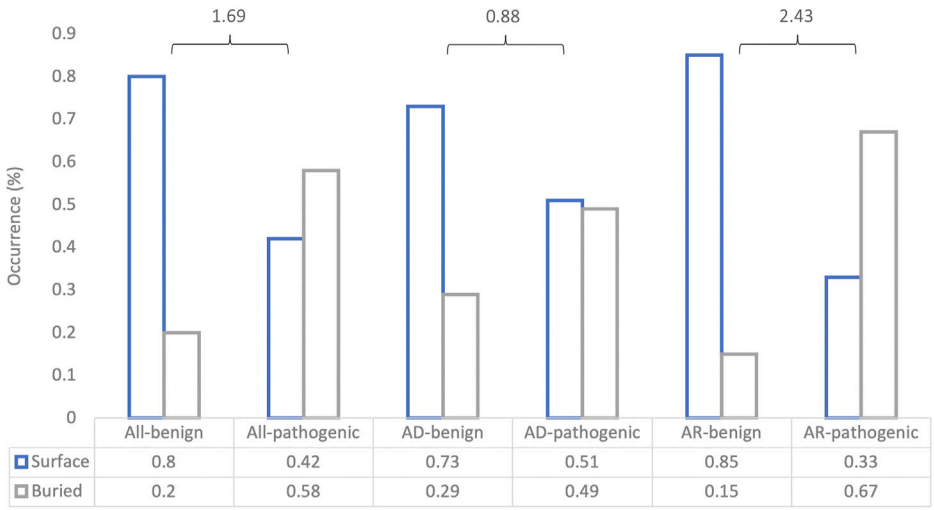

FIGURE 7
Association of missense variants across predictions on different test datasets with solvent accessibility. The bar plot shows the proportion of surface-exposed and buried missense variants in each of the binary outcomes for each of the datasets. "All" denotes all input variants, AD denotes mutations with autosomal dominant inheritance, and AR denotes mutations with AR inheritance. The log-odds ratio is calculated for each case to determine the strength of association between the binary feature (buried or surface-exposed) and the binary outcome (benign or pathogenic).

as well as through LoFs, i.e., haploin sufficiency (Veitia et al., 2018).

To understand how DeepRank-Mut generalizes on distinct modes of inheritance of pathogenic variants, we split our test datasets into variants with AD inheritance (n = 1,363; 550 benign, 813 pathogenic) and variants with AR inheritance (n = 563; 244 benign, 319 pathogenic), based on information obtained from ClinVar (Landrum et al., 2018). Only a smaller subset could be mapped to crystal structures: 585 structures mapped to 515 AD variants, and 77 structures mapped to 132 variants. We did not filter the AD dataset further to segregate mutations into haploinsufficient genes (LoFs) and non-LoFs (GoFs, DNs), due to lack of detailed annotations of non-LoFs in ClinVar. However, it is worth noting that mutations in the AD dataset could consist of higher proportion of LoFs than non-LoFs due to smaller mutational target for non-LoFs, i.e., fewer mutations alter protein function than disrupt it. Figure 6 illustrates a marked difference in the model's performance between the two datasets, suggesting dependence on underlying effects of the variant on the protein. It is apparent from the figure that our model is able to generalize AR mutations (LoFs) better than AD mutations (LoFs and non-LoFs). Details on the pathogenicity predictions obtained for AD and AR datasets, are provided in Supplementary Table S3. To further examine our relative success in correctly classifying buried pathogenic variants and AR variants we analyzed the distribution of solvent accessibility in the AD and AR datasets. Interestingly, the typical distribution of solvent-exposed benign variants and buried pathogenic variants was found to be more pronounced in AR datasets than in the AD datasets (Figure 7), explaining the relative success of our model in distinguishing LoF mutations from benign (Supplementary Figure S7, MCC = 0.67).

4 Discussion

Numerous efforts in the last decade have aided in the general understanding of effects of disease-causing mutations on the biophysical characteristics of proteins-including protein stability, dynamics, and protein-protein interactions (Kucukkal et al., 2015; Iqbal et al., 2020). It has been observed that pathogenic mutations are often associated with changes in local hydrogen-bonding network, electrostatic interactions, and overall side-chain geometry (Kucukkal et al., 2015). Although this knowledge has helped in the advancement of variant effect predictors that integrate various information on top of sequence-based features, the accurate prediction of a functional outcome of a missense variant is often fraught with challenges that we partly bring forth in this study.

We describe DeepRank-Mut, a structure-guided approach that leverages properties in the local variant neighborhood and uses 3D-CNNs to draw relationships between the spatially proximal features to distinguish pathogenic missense variants from benign. Our approach is robust to rotational variations, as we account for different orientations of a given variant environment through data augmentation steps. We did not experiment with larger augmentations due to large computational costs incurred. The performance of DeepRank-Mut was found to be comparable with other widely used predictors, such as PolyPhen2 which employs classical ML algorithm and relies on handcrafted features. Our investigations into the generalizability of our model revealed aspects that could be of interest to those who adopt deep learning techniques in structure-based variant effect predictions.

We find that the evolutionary information (PSSM profile) of the variant neighborhood captures patterns in the 3D structural context of variant sites better than the individual structural properties themselves. In contrast, inclusion of variant site-specific conservation scores alone, devoid of the 3D context, render the 3D-CNN model myopic thereby affecting the overall predictive ability. This finding is of considerable significance as it shows that the model potentially draws context dependence in terms of evolutionarily coupled residues. Pairs of residues under structural and functional constraints can exhibit strong inter-residue correlations, and thus coevolve (de Juan et al., 2013). Such a property has been shown to be useful in capturing effects of genetic variations (Hopf et al., 2017). Without explicitly modeling such inter-residue correlations, the performance of our model trained only on the PSSM profile of the neighborhood illustrates the utility of 3D-CNNs in capturing complex relationships between residues. This is further strengthened by the leave-one-feature-out analysis, where exclusion of seemingly redundant features from the model affected its performance.

Solvent accessible surface area was identified as the second most important feature that contributed to the predictor accuracies. Considering earlier reports on the enrichment of solvent-exposed missense variants in populations and enrichment of pathogenic variants in the hydrophobic core of proteins (Iqbal et al., 2020; Savojardo et al., 2020), we sought to explore their distribution in missense variants which were consistently misclassified across our datasets. We note that a significant proportion of misclassified pathogenic variants were found to be solvent-exposed, which raises the question whether our model loses generalizability while prioritizing buried pathogenic variants. Our experiments with models separately trained on buried and solvent-exposed missense variants yielded interesting results. The buried model could correctly identify pathogenic variants, even those that are solvent-exposed, while misclassifying a significant proportion of benign variants. The solvent-exposed model, on the other hand, showed similar performance in comparison to the original full model trained on all variants. These findings necessitate incorporating function-specific features or use of other suitable representations of protein structures, such as graphs, to adequately capture the underlying differences within pathogenic missense variants. Achieving high classification scores on solvent-exposed variants do pose a challenge, yet may be overcome with the following strategies: a) ensemble learning, combining multiple models trained on different feature sets related to solvent-exposed variants, such as ligand binding sites or phosphorylation sites; b) active learning, iteratively selecting the most informative solvent-exposed variants for labeling and training the model; or c) self-supervised learning, training the model to predict masked residues. Moreover, it is also possible that the solvent-exposed pathogenic variant site is a part of a larger assembly or participates in protein-protein interactions, an aspect not considered in this study. Use of full protein complex structures for pathogenic variants, wherever applicable, or features that indicate their role in function could help improve classifications (Gerasimavicius et al., 2022). Overall, we find that the two main features: evolutionary information of residues in the

variant neighborhood and solvent accessibilities sufficiently capture most of the important traits around variant sites.

Consideration of disease mechanisms appears to be crucial in the quality of pathogenicity predictions, as exemplified in our study. Our approach could generalize on mutations linked to AR inheritance better than the mutations linked to AD inheritance, corroborating results from an earlier study by Gerasimavicius et al. (2022). This finding is primarily due to the underlying mechanisms of mutations where protein destabilizing LoFs, often associated with AR diseases, are more straightforward to identify than non-LoFs which tend to have milder impacts on protein stability. Moreover, distribution of solvent accessibility of variants was suggestive of notable differences in the proportion of buried and solvent-exposed pathogenic variants, across the datasets. The overall performance of AR datasets over AD dataset is potentially due to two plausible reasons: a) feature representations are sufficiently able to distinguish LoFs from benign, and not non-LoFs from benign and b) limited amount of data on variants with non-LoF mechanisms. Both these postulates hold true considering the damaging effects on protein structure caused by LoFs that are relatively straightforward to discern (Gerasimavicius et al., 2022), and considering the total size of missense variants with non-LoF mechanisms (GoF and DN) mapped onto protein structures (n = 972), which is insufficient for training using deep neural networks. Since we did not segregate the AD dataset further into non-LoFs (GoFs, DNs) and LoFs, i.e., mutations in haploinsufficient genes, it is not apparent how the PSSM profile of residues in a variant environment and their solvent accessibility impact the predictions made. Nevertheless, our analysis underscores the necessity of incorporating features related to non-LoFs in improving pathogenicity predictions. This can be achieved through scrutiny and inclusion of gene-level and protein-level features specific to each of the mutation mechanisms in question, as documented by Sevim Bayrak et al. (2021). In addition, proteins in both AD and AR datasets reportedly show significant differences in functional class prevalence (Gerasimavicius et al., 2022), necessitating function-specific analysis to delineate characteristics of the disease mechanisms of mutations (Iqbal et al., 2020).

Our current method does not include explicit modeling of mutations into the protein structure, nor inclusion of protein dynamics, an inherent property linked to protein function. Indeed, inclusion of such details can aid in the recognition of the extent of mutation-induced changes in intra-protein structural contacts, as well as changes in thermodynamic stability (Rodrigues et al., 2018). In combination with other relevant features, these may provide considerable insights into understanding different effects across different mutation types, even with limited protein structural data. While we acknowledge the limitations of training our model on static protein microenvironments, we understand that more features may not necessarily imply better performance with neural networks. With suitable representations of protein structures (graphs) and information on protein dynamics it is important to address fundamental problems, such as predicting functional sites (Chiang et al., 2022) or predicting structurally important sites to further our understanding of model-driven approaches. This can help gauge utility of protein dynamics-informed or physics-informed graph representations in predicting variant pathogenicity.

To summarize, we have described a structure-guided approach to predict functional outcomes of missense variants using 3D-CNNs. We analyze and demonstrate the contribution of different features on the predictive ability of the neural network. Of particular note is the influence of evolutionary information of the variant neighborhood and their solvent accessibilities in determining variant pathogenicity. We further provide detailed assessment of our model's generalizability on distinct mechanisms of mutations, which presents a complex but critical challenge in improving pathogenicity predictions. Our analysis presents lessons to consider when using model-driven approaches to address questions in structure-guided predictions of variant pathogenicity.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author. The source code and documentation of DeepRank-Mut are available at https://github.com/DeepRank/DeepRank-Mut/.

Author contributions

MH, LX, and BV conceived the project. CB and GR designed the algorithm from its parent. GR, CB, RvH, and JH implemented and evaluated the algorithm. BV and SH compiled and pruned datasets of missense variants and 3D structures. GR performed the analyses, interpretation of data and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the Europees Fonds voor Regionale Ontwikkeling (EFRO) (R0005582). LX acknowledges support from Hypatia Fellowship from RadboudUMC (Rv819.52706). The work was carried out on the National Computer Facilities (NWO-2021.047).

Acknowledgments

The authors acknowledge Dario Marzella for his inputs on grid feature visualizations. The authors also acknowledge Dr. Peter-Bram t'Hoen, Dr. Hanka Vensalaar, Daniel Rademaker and the reviewers for their useful suggestions.

Conflict of interest

Authors SH and BV were employed by Bio-Prodict. Authors RvH and JH were employed by Vartion.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1204157/full#supplementary-material

References

Abrusán, G., and Marsh, J. A. (2016). Alpha helices are more robust to mutations than beta strands. *PLOS Comput. Biol.* 12, e1005242. doi:10.1371/journal.pcbi.1005242

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi:10.1038/nmeth0410-248

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* 7, Unit7.20. doi:10.1002/0471142905.hg0720s76

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389

Ancien, F., Pucci, F., Godfroid, M., and Rooman, M. (2018). Prediction and interpretation of deleterious coding variants in terms of protein structural stability. *Sci. Rep.* 8, 4480. doi:10.1038/s41598-018-22531-2

Bagley, S. C., and Altman, R. B. (1995). Characterizing the microenvironment surrounding protein sites. *Protein Sci.* 4, 622–635. doi:10.1002/pro.5560040404

Capriotti, E., and Altman, R. B. (2011). Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinforma.* 12, S3. doi:10.1186/1471-2105-12-S4-S3

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., and Karchin, R. (2013). Identifying mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 14, S3. doi:10.1186/1471-2164-14-S3-S3

Chiang, Y., Hui, W.-H., and Chang, S.-W. (2022). Encoding protein dynamic information in graph representation for functional residue identification. *Cell Rep. Phys. Sci.* 3, 100975. doi:10.1016/j.xcrp.2022.100975

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, 6. doi:10.1186/s12864-019-6413-7

Choi, Y., and Chan, A. P. (2015). PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31, 2745–2747. doi:10.1093/bioinformatics/btv195

de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nat. Rev. Genet.* 14, 249–261. doi:10.1038/nrg3414

Feinauer, C., and Weigt, M. (2017). Context-aware prediction of pathogenicity of missense mutations involved in human disease. *Arxiv*. doi:10.48550/arXiv.1701.07246

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., et al. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 562, 217–222. doi:10.1038/s41586-018-0461-z

Gerasimavicius, L., Livesey, B. J., and Marsh, J. A. (2022). Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat. Commun.* 13, 3895. doi:10.1038/s41467-022-31686-6

Gunning, A. C., Fryer, V., Fasham, J., Crosby, A. H., Ellard, S., Baple, E. L., et al. (2021). Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J. Med. Genet.* 58, 547–555. doi:10.1136/jmedgenet-2020-107003

Heijl, S., Vroling, B., van den Bergh, T., and Joosten, H.-J. (2020). Mind the gap: Preventing circularity in missense variant prediction. *Biorxiv*. doi:10.1101/2020.05.06.080424

Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G., Kohlbacher, O., Sander, C., et al. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* 3, e03430. doi:10.7554/eLife.03430

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., et al. (2017). Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35, 128–135. doi:10.1038/nbt.3769

Iqbal, S., Pérez-Palma, E., Jespersen, J. B., May, P., Hoksza, D., Heyne, H. O., et al. (2020). 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. doi:10.1073/pnas.2002660117

Ittisoponpisan, S., Islam, S. A., Khanna, T., Alhuzimi, E., David, A., and Sternberg, M. J. E. (2019). Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated? *J. Mol. Biol.* 431, 2197–2212. doi:10.1016/j.jmb.2019.04.009

Jorgensen, W. L., and Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* 110, 1657–1666. doi:10.1021/ja00214a001

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi:10.1002/bip.360221211

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi:10.1038/s41586-020-2308-7

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi:10.1038/ng.2892

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems* (Curran Associates, Inc).

Kryukov, G. V., Pennacchio, L. A., and Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *Am. J. Hum. Genet.* 80, 727–739. doi:10.1086/513473

Kucukkal, T. G., Petukh, M., Li, L., and Alexov, E. (2015). Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Curr. Opin. Struct. Biol.* 32, 18–24. doi:10.1016/j.sbi.2015.01.003

Kuipers, R. K., Joosten, H.-J., van Berkel, W. J. H., Leferink, N. G. H., Rooijen, E., Ittmann, E., et al. (2010). 3DM: Systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Proteins Struct. Funct. Bioinforma.* 78, 2101–2113. doi:10.1002/prot.22725

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46, D1062–D1067. doi:10.1093/nar/gkx1153

Laskowski, R. A., Stephenson, J. D., Sillitoe, I., Orengo, C. A., and Thornton, J. M. (2020). VarSite: Disease variants and protein structure. *Protein Sci.* 29, 111–119. doi:10.1002/pro.3746

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi:10.1038/nature19057

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., et al. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25, 2744–2750. doi:10.1093/bioinformatics/btp528

Li, B., Yang, Y. T., Capra, J. A., and Gerstein, M. B. (2020). Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLoS Comput. Biol.* 16, e1008291. doi:10.1371/journal.pcbi.1008291

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 32, 894–899. doi:10.1002/humu.21517

Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* 12, 103. doi:10.1186/s13073-020-00803-9

Livesey, B. J., and Marsh, J. A. (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* 16, e9380. doi:10.15252/msb.20199380

Livesey, B. J., and Marsh, J. A. (2022). Interpreting protein variant effects with computational predictors and deep mutational scanning. *Dis. Models Mech.* 15, dmm049510. doi:10.1242/dmm.049510

Loshchilov, I., and Hutter, F. (2019). Decoupled weight decay regularization. *Arxiv*. doi:10.48550/arXiv.1711.05101

Marks, D. S., Hopf, T. A., and Sander, C. (2012). Protein structure prediction from sequence variation. *Nat. Biotechnol.* 30, 1072–1080. doi:10.1038/nbt.2419

Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res* 5, 189. doi:10.12688/f1000research.7931.1

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., et al. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* 108, E1293–E1301. doi:10.1073/pnas.1111471108

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi:10.1093/nar/gkg509

Niroula, A., and Vihinen, M. (2019). How good are pathogenicity predictors in detecting benign variants? *PLOS Comput. Biol.* 15, e1006481. doi:10.1371/journal.pcbi.1006481

Pearson, W. R. (2013). An introduction to sequence similarity ("Homology") searching. *Curr. Protoc. Bioinforma. 0* 3, 3.1.1, 3.1.8. doi:10.1002/0471250953.bi0301s42

Pincus, M. R., and Scheraga, H. A. (1977). An approximate treatment of long-range interactions in proteins. *J. Phys. Chem.* 81, 1579–1583. doi:10.1021/j100531a013

Ponzoni, L., Peñaherrera, D. A., Oltvai, Z. N., and Bahar, I. (2020). Rhapsody: Predicting the pathogenicity of human missense variants. *Bioinformatics* 36, 3084–3092. doi:10.1093/bioinformatics/btaa127

Pun, M. N., Ivanov, A., Bellamy, Q., Montague, Z., LaMont, C., Bradley, P., et al. (2022). Learning the shape of protein micro-environments with a holographic convolutional neural network. *Arxiv*. doi:10.1101/2022.10.31.514614

Renaud, N., Geng, C., Georgievska, S., Ambrosetti, F., Ridder, L., Marzella, D. F., et al. (2021). DeepRank: A deep learning framework for data mining 3D protein-protein interfaces. *Nat. Commun.* 12, 7068. doi:10.1038/s41467-021-27396-0

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.* 39, e118. doi:10.1093/nar/gkr407

Rodrigues, C. H., Pires, D. E., and Ascher, D. B. (2018). DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.* 46, W350–W355. doi:10.1093/nar/gky300

Rost, B., and Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins* 20, 216–226. doi:10.1002/prot.340200303

Savojardo, C., Manfredi, M., Martelli, P. L., and Casadio, R. (2020). Solvent accessibility of residues undergoing pathogenic variations in humans: From protein structures to protein sequences. *Front. Mol. Biosci.* 7, 626363. doi:10.3389/fmolb.2020.626363

Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: Mutation prediction for the deep-sequencing age. *Nat. Methods* 11, 361–362. doi:10.1038/nmeth.2890

Sevim Bayrak, C., Stein, D., Jain, A., Chaudhary, K., Nadkarni, G. N., Van Vleck, T. T., et al. (2021). 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. doi:10.1016/j.ajhg.2021.10.007

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* 34, 57–65. doi:10.1002/humu.22225

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 60. doi:10.1186/s40537-019-0197-0

Shroff, R., Cole, A. W., Diaz, D. J., Morrow, B. R., Donnell, I., Annapareddy, A., et al. (2020). Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS Synth. Biol.* 9, 2927–2935. doi:10.1021/acssynbio.0c00345

Sun, Z., Liu, Q., Qu, G., Feng, Y., and Reetz, M. T. (2019). Utility of B-factors in protein science: Interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chem. Rev.* 119, 1626–1665. doi:10.1021/acs.chemrev.8b00290

Thompson, B. A., Spurdle, A. B., Plazzer, J.-P., Greenblatt, M. S., Akagi, K., Al-Mulla, F., et al. (2014). Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat. Genet.* 46, 107–115. doi:10.1038/ng.2854

Torng, W., and Altman, R. B. (2017). 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinforma.* 18, 302. doi:10.1186/s12859-017-1702-0

Touw, W. G., Baakman, C., Black, J., te Beek, T. A. H., Krieger, E., Joosten, R. P., et al. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* 43, D364–D368. doi:10.1093/nar/gku1028

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9. doi:10.1038/nprot.2015.123

Veitia, R. A., Caburet, S., and Birchler, J. A. (2018). Mechanisms of mendelian dominance. *Clin. Genet.* 93, 419–428. doi:10.1111/cge.13107

Venselaar, H., Te Beek, T. A. H., Kuipers, R. K. P., Hekkelman, M. L., and Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinforma.* 11, 548. doi:10.1186/1471-2105-11-548

VKGL (2019). *Vereniging klinisch genetische laboratoriumdiagnostiek - home*. URL: https://www.vkgl.nl/nl/ (accessed October 3, 2019).

Vroling, B., and Heijl, S. (2021). White paper: The Helix pathogenicity prediction platform. *Arxiv*. doi:10.48550/arXiv.2104.01033

Wang, Z., and Moult, J. (2001). SNPs, protein structure, and disease. *Hum. Mutat.* 17, 263–270. doi:10.1002/humu.22

Yates, C. M., Filippis, I., Kelley, L. A., and Sternberg, M. J. E. (2014). SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol.* 426, 2692–2701. doi:10.1016/j.jmb.2014.04.026

Zardecki, C., Dutta, S., Goodsell, D. S., Lowe, R., Voigt, M., and Burley, S. K. (2022). PDB-101: Educational resources supporting molecular explorations through biology and medicine. *Protein Sci.* 31, 129–140. doi:10.1002/pro.4200