Briefings in Bioinformatics, 2025, **26(2)**, bbaf119
https://doi.org/10.1093/bib/bbaf119
Problem Solving Protocol

DS-MVP: identifying disease-specific pathogenicity of missense variants by pre-training representation

Qiufeng Chen[1,‡], Lijun Quan[1,2,*,‡], Lexin Cao[1], Bei Zhang[1], Zhijun Zhang[1], Liangchen Peng[1], Junkai Wang[1], Yelu Jiang[1], Liangpeng Nie[1], Geng Li[1], Tingfang Wu[1,2,*], Qiang Lyu[1,2,*]

[1]School of Computer Science and Technology, Soochow University, Jiangsu 215006, China
[2]Collaborative Innovation Center of Novel Software Technology and Industrialization, Jiangsu 210000, China

*Corresponding authors. School of Computer Science and Technology, Soochow University, Jiangsu 215006, China. (Lijun Quan, E-mail: ljquan@suda.edu.cn; Tingfang Wu, E-mail: tfwu@suda.edu.cn; Qiang Lyu, E-mail: qiang@suda.edu.cn)
‡Qiufeng Chen and Lijun Quan contributed equally to this work.

Abstract

Accurately predicting the pathogenicity of missense variants is crucial for improving disease diagnosis and advancing clinical research. However, existing computational methods primarily focus on general pathogenicity predictions, overlooking assessments of disease-specific conditions. In this study, we propose DS-MVP, a method capable of predicting disease-specific pathogenicity of missense variants in human genomes. DS-MVP first leverages a deep learning model pre-trained on a large general pathogenicity dataset to learn rich representation of missense variants. It then fine-tunes these representations with an XGBoost model on smaller datasets for specific diseases. We evaluated the learned representation by testing it on multiple binary pathogenicity datasets and gene-level statistics, demonstrating that DS-MVP outperforms existing state-of-the-art methods, such as MetaRNN and AlphaMissense. Additionally, DS-MVP excels in multi-label and multi-class classification, effectively classifying disease-specific pathogenic missense variants based on disease conditions. It further enhances predictions by fine-tuning the pre-trained model on disease-specific datasets. Finally, we analyzed the contributions of the pre-trained model and various feature types, with gene description corpus features from large language model and genetic feature fusion contributing the most. These results underscore that DS-MVP represents a broader perspective on pathogenicity prediction and holds potential as an effective tool for disease diagnosis.

Keywords: disease-specific pathogenicity; missense variants; pre-training representation; Transformer

Introduction

Next-generation sequencing [1] provides a powerful tool for studying variants by enabling the high-throughput sequencing of DNA fragments. Among these variants, a missense variant represents a common single nucleotide variation, where the substitution of a single nucleotide in the DNA sequence alters the corresponding codon, leading to an amino acid change in the encoded protein. The missense variant can significantly impact the structure and functionality of proteins, potentially contributing to the development of various diseases [2–4]. By accurately predicting the pathogenicity of missense variants, researches can identify which missense variants that are likely to precipitate diseases. Thus, it is crucial to predict the pathogenicity of missense variants in advancing genomics and clinical research, offering valuable insights into genetic variations and their implications for human health [5].

Numerous prediction tools have been developed to address the challenge of assessing the pathogenicity associated with missense variants [6], which can be categorized into independent predictors and meta predictors. Independent predictors, such as MAGPIE [7], VEST4 [8], AlphaMissense [9], and CADD [10], use diverse algorithms. For example, MAGPIE integrates multi-source features with LightGBM [11], VEST4 applies random forests [12], AlphaMissense utilizes protein structure alignments, and CADD employs logistic regression on unbiased datasets. Meta predictors like MetaRNN [13], MVP [14], and ClinPred [15] combine scores from multiple tools to boost accuracy. MetaRNN integrates over 30 predictors using a BiGRU model, MVP processes six feature types via residual network (ResNet) [16], and ClinPred combines random forest and gradient boosting models. While current methods have demonstrated commendable performance in binary classification, there remains scope for refinement.

General pathogenic variant identification methods only provide a binary judgment: pathogenic or benign. They are often insufficient when addressing specific diseases. It would be better if the pathogenic judgment could be further classified into a specific disease. The pathological mechanisms of different diseases vary, and the same variant may exhibit significantly different effects and manifestations in different diseases due to tissue specificity. For example, in cancer, a missense variant may drive tumor progression in one type of cancer, while having a much smaller impact, or even no effect, in another [17, 18]. This variation is not only due to the variant itself but is also related to the specific biological environments and gene regulatory mechanisms of various diseases. Therefore, it is necessary to further classify specific diseases of variants into single or multiple disease categories according to disease conditions to better

Received: December 10, 2024. **Revised:** January 26, 2025. **Accepted:** March 4, 2025

understand their precise pathogenic mechanisms within different disease background. Such detailed analysis can provide valuable insights for personalized diagnosis and treatment. However, to the best of our knowledge, no prediction models for disease-specific pathogenicity are currently available.

There are two types of disease-specific pathogenicity caused by missense variants: multi-label (ML) disease and multi-class (MC) disease. A missense variant may trigger possible several ML diseases under a specific disease type. For example, a missense variant in the BRCA gene increases the risk of developing both breast and ovarian cancers [19]. On the other hand, a missense variant may cause a MC disease exclusively within a specific disease type. A typical example is a missense variant in the CFTR gene predominantly causes cystic fibrosis, with minimal impact on other related conditions [20]. To distinguish from general pathogenicity binary classification task, we name our study DS-MVP, for disease-specific missense variants pathogenicity.

DS-MVP leverages pre-trained deep learning (DL) model based on large amount of data to extract complex representation of missense variants, followed by effective target training on small datasets using shallow machine learning algorithm to predict the disease-specific pathogenicity of missense variants. The method begins by constructing multiple datasets, including both general and disease-specific pathogenicity datasets, using a large collection of missense variant data from the ClinVar database. DS-MVP first performs pre-training on a general pathogenicity missense variant dataset, using various feature extraction modules—genetic feature fusion module (GFFM), DNA sequence representation module (DSeqRM), mutated amino acid embedding module (MAAEM), OpenAI based gene representation module (OBGRM), and DNA structure representation module (DStructRM)—to capture and represent multidimensional representation of missense variants, including genetic features, DNA sequence features, mutated amino acid features, gene description corpus features based on large language model, and structure features. Such representations of missense variants are then recombined with the raw genetic features and used to train an XGBoost model [21] to predict the disease-specific pathogenicity of missense variants.

Experimental results show that despite not directly incorporating the scores of other prediction tools as features, DS-MVP outperforms 20 existing meta and independent predictors across multiple evaluation metrics on general pathogenicity and disease-specific pathogenicity test datasets. Additionally, DS-MVP demonstrates outstanding performance in identifying ML/MC disease, particularly in distinguishing between different types of specific diseases. We further optimized the DS-MVP model by fine-tuning its modules of pre-trained model on disease-specific pathogenicity datasets, improving its prediction accuracy for particular diseases. Studies learning properties and ablation confirm that the knowledge representation learned by the pre-trained model is successfully transferred to disease-specific pathogenicity predictions, significantly improving model utilizations. Significantly, we are the first to propose using gene-related rich corpus embeddings extracted by pre-trained large language model, text-embedding-3-small model from OpenAI (https://platform.openai.com/docs/guides/embeddings/), which had a substantial positive impact on pathogenicity prediction of variants. In conclusion, the strategy of pre-train for representation and fine-tune for target tasks not only provides flexible new avenues for research in related fields but also holds broad potential for pathogenicity prediction in other disease-specific domains.

Materials and methods

Overview of DS-MVP

The architecture of DS-MVP is shown in Fig. 1. DS-MVP employs a two-stage framework: a pre-trained DL model to extract missense variant representations, followed by fine-tuning with XGBoost for disease-specific pathogenicity prediction (see "Methods"). The pre-trained model consists of five key modules: GFFM integrates genetic features such as conservation scores and allele frequencies (AFs) for biological context. DSeqRM utilizes BiLSTM [22], Transformer encoder [23], and CNN [24] to differentiate reference and alternative DNA sequences. MAAEM encodes reference and alternative amino acids at variant sites to capture changes. OBGRM leverages OpenAI's language model to derive gene description representations via natural language processing. DStructRM extracts structural features of reference and alternative DNA sequences using Deep DNAShape [25, 26]. These modules collectively generate comprehensive missense variant representations. These representations, combined with genetic features, were used with XGBoost for binary pathogenicity prediction. Additionally, DS-MVP performed ML and MC classification to predict disease-specific pathogenic variants under detailed conditions.

Dataset

The data utilized in our study were derived from the ClinVar database [27]. ClinVar database was downloaded through January 2024 under the GRCh38/hg38 genome assembly, which is a public clinical database containing extensive data on human genetic variants and their associations with diseases [27]. We utilized training general pathogenicity dataset (GPD) consisted 101 560 missense variants regardless of AF, and assembled four test datasets, including rare missense variants test set (RareD), hard target prediction (HardD), disease-specific pathogenicity datasets binary cardiomyopathy-specific pathogenicity dataset (CSPD_Bi), and binary neurodegenerative disease-specific pathogenicity dataset (NSPD_Bi). In addition, we further classifying pathogenic missense variants into detailed disease conditions for specific disease datasets, containing ML disease type prediction for cardiomyopathy disease (CSPD_ML) and eye disease (ESPD_ML), and MC disease type prediction for neurodegenerative disease (NSPD_MC) and endocrine disease (EcSPD_MC). The proportion of disease datasets for each type are illustrated in Fig. 2. For detailed data processing procedures, see Supplementary Information (SI) Text S1.

The model architecture of DS-MVP

Pre-trained deep learning model for extracting representation of missense variants

The DS-MVP DL model extracts generalized features related to variant pathogenicity from five feature types: genetic features, DNA sequence, mutated amino acid, gene description corpus, and DNA structure (see SI Text S2). Below is a detailed description of each module.

Genetic feature fusion module. Genetic features were transformed using linear layers. Initially, the eight conservation scores and two AFs features, which were previously imputed, and processed separately through linear layers. These features were then fused using another linear layer to produce a 128-dimensional vector.

DNA sequence representation module. We used a sliding window approach [28] to encode DNA sequences, creating a vocabulary of

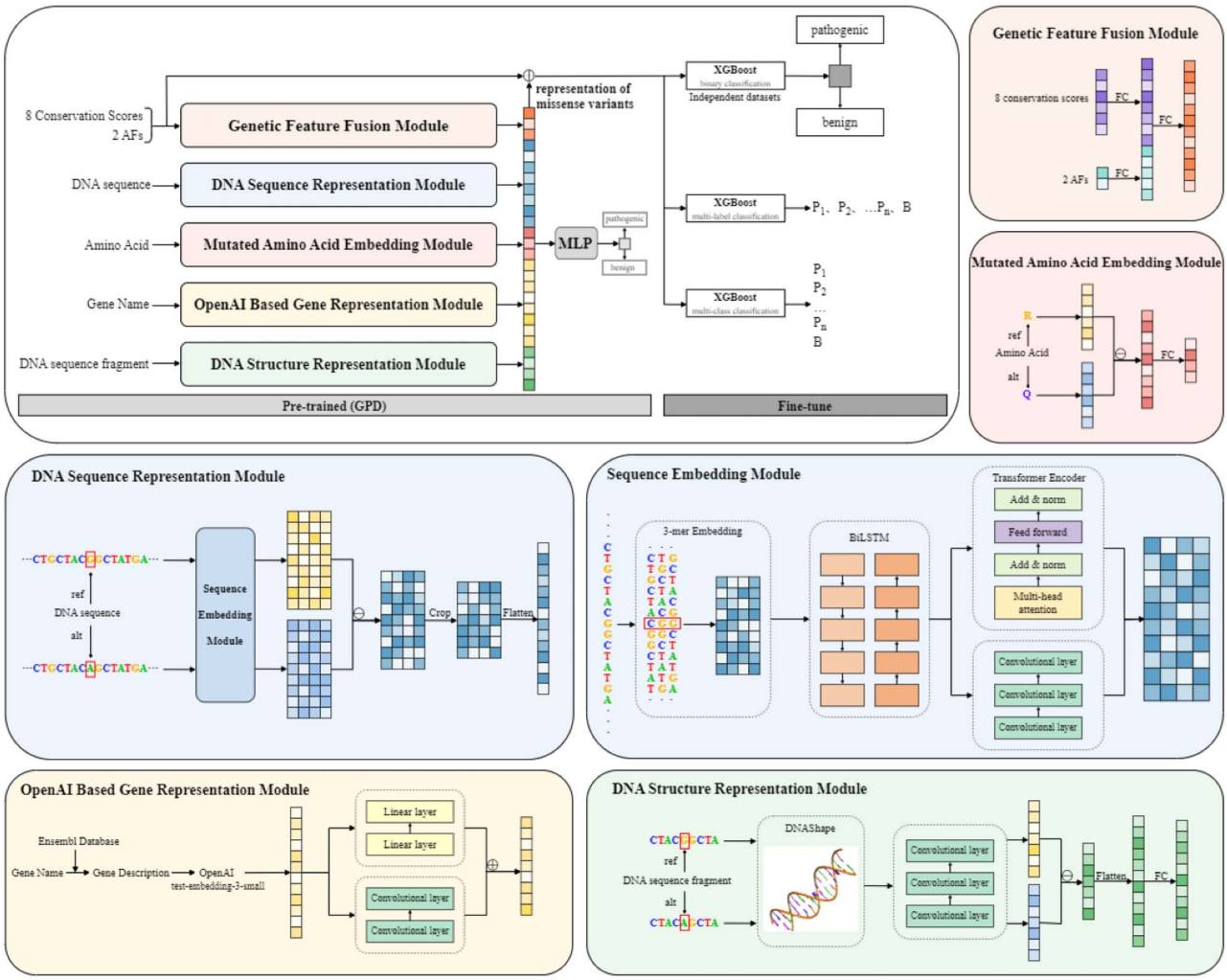

Figure 1. The architecture of DS-MVP. DS-MVP utilized a two-stage design includes pre-train for learning representation of missense variants based on GPD binary classification by a DL model, and used XGBoost model to fine-tune learned representation and genetic features for general and disease-specific prediction. The pre-trained DL model comprised five modules: GFFM, DSeqRM, MAAEM, OBGRM, and DStructRM. The XGBoost model aims to quickly predict the pathogenicity of missense variants for binary, ML, and MC classification.

nucleotides A, T, G, C, and N (for unknowns). We grouped three nucleotides into "words" to reflect the triplet coding of amino acids, assigning each a unique numerical ID. This resulted in a 125-word vocabulary, with a step size matching the window to preserve the genetic code's reading frame. A 99-base sequence was segmented into 97-word segments, each vectorized into a 128-dimensional space, forming a 97×128 matrix. This matrix was fed into a BiLSTM [22] for contextual analysis and a Transformer encoder [23] and CNNs [24] for feature extraction. The Transformer encoder learned the importance of each base within the whole sequence, while CNNs captured local variant site features. Outputs were merged to enhance feature abstraction and representation. A shared parameter mechanism processed reference and alternative sequences, computing differential features. Then a 7-length sequence was extracted around variants, flattened, and concatenated with other features for local effect analysis.

Mutated amino acid embedding module. Reference and alternative amino acids at the variant site were processed using a shared parameter mechanism. These amino acids were encoded into a 64-dimensional vector using an embedding technique. The features of the reference and alternative amino acids were subtracted to quantify the impact of variant and then combined through a linear layer to produce a 64-dimensional vector feature.

OpenAI based gene representation module. Gene descriptions extracted from gene names were input into the text-embedding-3-small model provided by OpenAI, resulting in a 1536-dimensional vector feature. The feature was then processed through two linear layers and two convolutional layers to capture more higher-level features and reduce it to 512 dimensions. The features from both reduction processes were combined to create a comprehensive 512-dimensional vector.

DNA structure representation module. Features from reference and alternative sequence fragments were processed using a shared parameter mechanism. Three layers of convolution neural networks with kernels of varying sizes were applied to enhance information capture across different scales and obtain abstract feature representations. The features were subtracted to emphasize differences, flattened for dimensionality reduction, and processed through a linear layer to achieve compact and informative feature representations.

Multi-layer perceptron. The outputs from the five modules were concatenated as representation of missense variants, then processed through fully connected layers with ReLU activation for feature dimensionality reduction and dropout for generalization. A final sigmoid activation provided the classification probability.

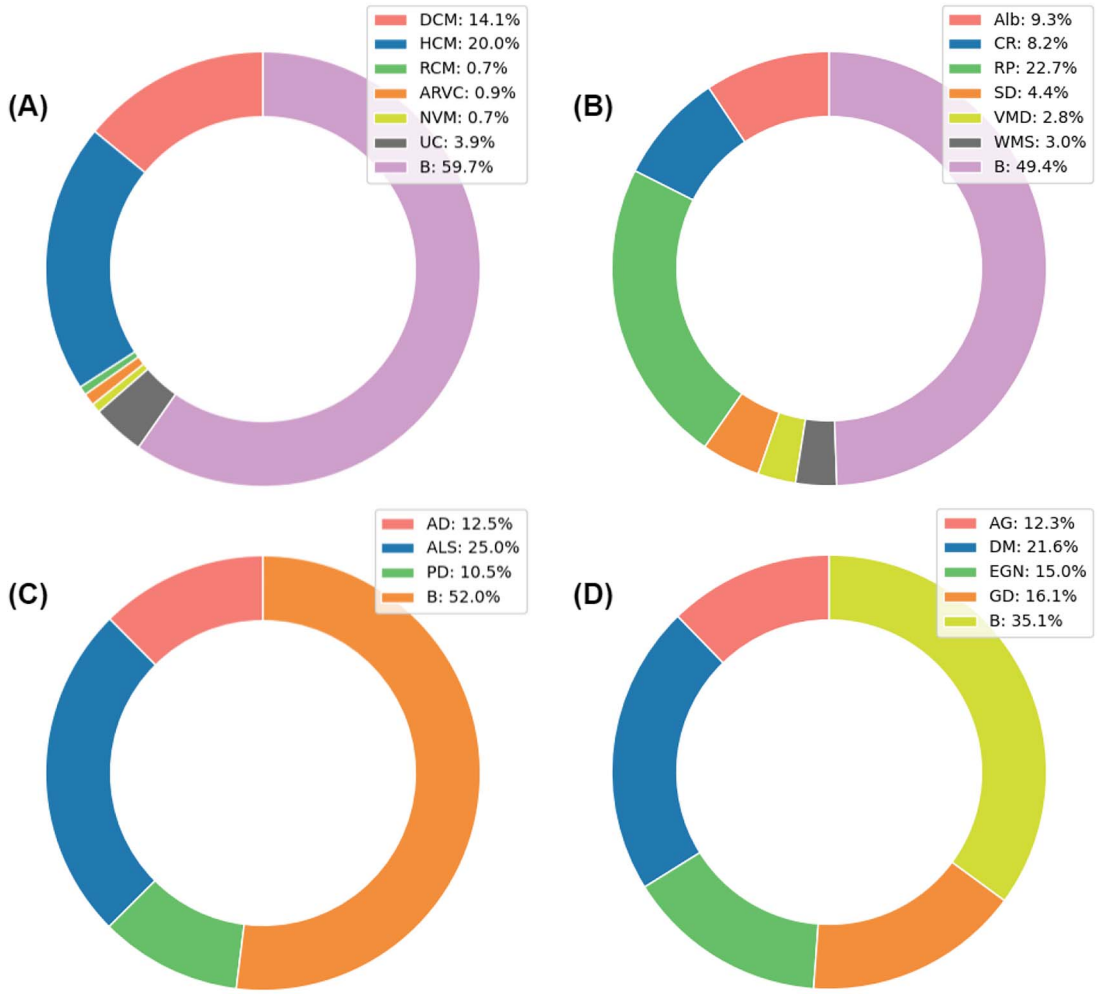

Figure 2. Percentage of disease data. (**A**) The percentage of cardiomyopathy data, including categories DCM, HCM, RCM, ARVC (arrhythmogenic right ventricular cardiomyopathy), NVM, UC (unclassified), and B (benign), with a total of 2161 data. (**B**) The distribution of eye disease data, including Alb, CR, RP, SD, VMD, WMS, and B, totaling 821 data. (**C**) The distribution of neurodegenerative disease data, including AD, ALS, PD, and B, totaling 891 data.(**D**) The distribution of endocrine disease data, including AG (adrenal gland diseases), DM (diabetes mellitus), EGN (endocrine gland neoplasms), GD (gonadal disorders), and B, totaling 921 data.

Model training. Based on the DS-MVP method, we employed BCELoss as the loss function and selected the widely used Adam optimizer, which is a stochastic gradient descent algorithm. The learning rate was set to 0.0005, and the weight decay set to 0.001. Detailed hyperparameter configurations, including network architecture and activation functions, are provided in the SI Table S1.

XGBoost model training for predicting the general and disease-specific pathogenicity of missense variants

The representation of missense variants extracted from pre-trained DL model, in conjunction with the genetic features, was utilized as inputs for the machine learning. By integrating genetic features, which contain eight conservation scores and two AFs, and the representation of missense variants ("⊕" as shown in Fig. 1), we re-learned those genetic features at the fine-tuning stage. This strategy allowed the model to better capture and leverage these critical features, which significantly improved performance. The idea was inspired by the ResNet architecture [16], which ensures efficient transmission of these features while preventing the loss of effect of critical raw features during back propagation. The XGBoost model [21] was deployed to fast predict the pathogenicity of variants, which is a gradient boosting model that is highly efficient in prediction among machine learning methods [29, 30]. We used the XGBoost model for rapid pathogenicity prediction, optimizing it with grid search [31] for best parameter settings. Our prediction tasks included binary classification for general and disease-specific pathogenicity, and detailed classification of pathogenic missense variants into specific disease conditions for ML or MC classification. The detailed pseudocode for the model description can be found in SI Text S3.

Furthermore, we also fine-tuning pre-trained DL model for disease-specific pathogenicity of missense variants, the detailed description seen in SI Text S4. In our ML classification, the model used BCEWithLogitsLoss to weight labels according to their data distribution, ensuring attention to less frequent labels. For MC classification, CrossEntropyLoss was employed, optimizing performance by categorizing each sample into a single class from a predefined set.

Results

To validate the potential of representation of missense variants for downstream applications, we conducted evaluations on both general and disease-specific pathogenicity datasets.

These pre-trained representations were then applied to smaller, disease-specific datasets, fine-tuned with XGBoost for ML and MC classification. Based on this, we also conducted analysis of the contributions of various features and their corresponding modules in predicting specific disease types of missense variants, exploring the efficacy and applicability of the proposed approach in addressing the target problem.

DS-MVP performance on identifying general pathogenicity of missense variants based on pre-trained representation

To evaluate the effectiveness of the representation of missense variants from pre-trained pathogenicity model, we fed these representations along with genetic features into an XGBoost algorithm to train a shallow model for predicting general pathogenicity of missense variants. We then conducted a series of experiments across different datasets to evaluate and compare the performance of our method with existing predictors.

Here, we compared DS-MVP with 20 other predictors. We extracted the predicted scores of the other predictors from the dbNSFP database [32, 33] and divided the 20 predictors into two categories: independent predictors and meta predictors. The independent predictors consist of MAGPIE [7], AplhaMissense [9], LIST-S2 [34], CADD [10], FATHMM-XF [35], PrimateAI [36], DEOGEN2 [37], MutationTaster [38], VEST4 [8], and MutationAssessor [39], totaling 10 predictors. Meta predictors aggregate the predicted scores from various independent predictors. There are 10 components within the meta predictors, including MetaRNN [13], MVP [14], MutPred [40], ClinPred [15], BayesDel_addAF [41], REVEL [42], Eigen [43], M-CAP [44], MetaSVM [45], and MetaLR [45]. A detailed comparison of all predictors is provided in SI Table S2. Additionally, the calculation methods for all evaluation metrics can be found in SI Text S5.

Prediction accuracy of DS-MVP for rare missense variants

To ensure a fair comparison of DS-MVP with other predictors, we utilized a classic test dataset from MetaRNN: RareD for rare missense variants. DS-MVP outperformed all other predictors, achieving the highest area under the curve (AUC) of 0.9528 and area under the precision versus recall curve (AUPR) of 0.9643, as shown in Fig. 3A and B and SI Table S3. Among the existing meta predictors, MetaRNN performed the best with an AUC of 0.9328 and an AUPR of 0.9320, while MAGPIE showed the best performance with an AUC of 0.8920 and an AUPR of 0.9055 among independent predictors. Additionally, it was evident that apart from our model DS-MVP, the accuracy of predictions from other independent predictors are significantly lower than the top three meta predictors, with reductions of 2.94% and 1.07% in metrics AUC and AUPR, respectively. Furthermore, it was noteworthy that the results from other methods calculated using the dbNSFP database (SI Table S3) differ slightly from those provided in the MetaRNN paper (SI Fig. S1), but these differences were minimal. This minor divergence indicates that the results from the dbNSFP database are reliable.

We further analyzed the distribution of predicted scores for different pathogenicity of missense variants—pathogenic and benign as shown in Fig. 3D–F. Although there is still room for improvement in predicting certain benign variants, overall, our method provides significantly better scores in distinguishing whether missense variants are pathogenic or not.

Prediction accuracy of DS-MVP for hard target missense variants

To further assess the robustness of DS-MVP, we utilized a subset of the ClinVar database, HardD, as our test dataset, which has shown poor performance with other predictors regardless of AF. As shown in Fig. 3A and B, DS-MVP achieved the best performance on this dataset with an AUC of 0.8370 and an AUPR of 0.8091, which represented an improvement of 1.92% and 0.17% in metrics AUC and AUPR compared to the best-performing meta predictor ClinPred, and compared to the best-performing independent predictor AlphaMissense, DS-MVP showed improvements of 7.80% and 4.03% in metrics AUC and AUPR. The detailed are shown in SI Table S3. The results on this challenging test dataset demonstrated that DS-MVP enhances prediction accuracy, confirming it as a reliable tool for pathogenicity prediction.

Prediction accuracy of DS-MVP for disease-specific missense variants

We initially collected missense variant data related to cardiomyopathy and neurodegenerative diseases, CSPD_Bi and NSPD_Bi, to extend the evaluation of the pre-trained model for specific diseases. Among all predictors, DS-MVP demonstrated the best performance in CSPD_Bi dataset, achieving an AUC of 0.9692 and an AUPR of 0.9749 (see Fig. 3A and B and SI Table S4). In contrast, other major predictors such as BayesDel_addAF, MetaRNN, and ClinPred had AUC values ranging from 0.8596 to 0.9619, and AUPR values between 0.8194 and 0.9594. Additionally, on the NSPD_Bi dataset, DS-MVP exhibited a similar advantage, outperforming other methods by a value of 0.79% and 3.41% on metrics AUC and AUPR, respectively.

To evaluate prediction accuracy of DS-MVP based on genes, we categorized all test datasets (RareD, HardD, CSPD_Bi, and NSPD_Bi) based on gene names and conducted statistical and indicator computations for genes with at least 15 missense variants. The results are visually represented in Fig. 3C and the analysis shown in SI Text S6. Further analysis of these results revealed that DS-MVP's higher AUC scores across a larger set of genes underscore its robustness in identifying the pathogenicity of missense variants.

Identifying disease-specific pathogenicity of missense variants based on pre-trained representation

To explore the potential of our model in predicting pathogenicity for specific diseases, we collected missense variant data related to cardiomyopathy, eye, neurodegenerative, and endocrine diseases. On the basis of dividing the variants into pathogenic and benign, the pathogenic data are further classified according to their detailed disease conditions. We extracted high-dimensional representation of missense variants using a pre-trained model, which were then combined with genetic features used to retrain by XGBoost on the CSPD_ML, ESPD_ML, NSPD_MC, and EcSPD_MC datasets for the pathogenicity prediction of ML and MC disease type.

However, no existing predictors have been applied to ML and MC for pathogenicity prediction tasks at the nucleotide level. To address this gap, we selected three high-performing machine learning boosting algorithms as benchmarks: XGBoost [21], LightGBM [11], and CatBoost [46]. The input features used in these models were the same as those in DS-MVP, including genetic features, DNA sequence features, etc. All evaluation metrics for ML and MC disease type prediction can be found in SI Text S7 and

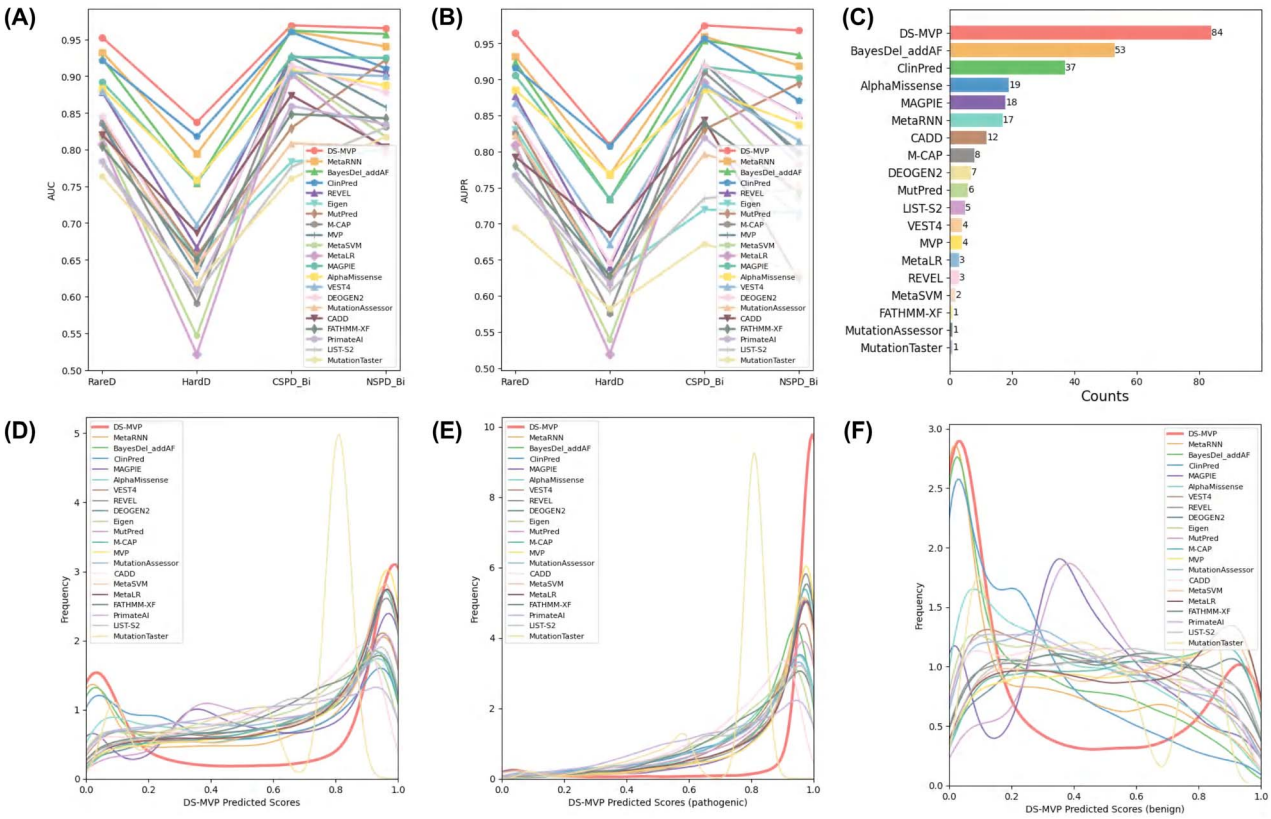

Figure 3. (**A**) The metrics of AUC on all test datasets. (**B**) The metrics of AUPR on all test datasets. (**C**) Number of genes with the best AUC. (**D**) Distribution of pathogenicity prediction scores of missense variants containing pathogenic and benign. (**E**) Distribution of pathogenicity prediction scores of pathogenic missense variants. (**F**) Distribution of pathogenicity prediction scores of benign missense variants.

Table 1. Performance of DS-MVP with XGBoost trained on the CSPD_ML and ESPD_ML dataset based on precision, recall, F1-score, and HammingLoss

Methods	CSPD_ML				ESPD_ML			
	Precision(↑)	Recall(↑)	F1-score(↑)	HammingLoss(↓)	Precision(↑)	Recall(↑)	F1-score(↑)	HammingLoss(↓)
XGBoost	0.8209	0.8008	0.8062	0.0591	0.6933	0.6738	0.6800	0.0929
LightGBM	0.8209	0.8008	0.8062	0.0585	0.6933	0.6738	0.6800	0.0955
CatBoost	0.8209	0.8008	0.8062	0.0591	0.6871	0.6677	0.6738	0.0938
DS-MVP	**0.8698**	**0.8519**	**0.8558**	**0.0372**	**0.8252**	**0.8241**	**0.8180**	**0.0465**

SI Text S8. Through this comparison, we demonstrated that the features and patterns learned from the pre-trained pathogenicity model can be transferred to a new, potentially smaller dataset for more complex tasks. This effectively enhanced the model's performance on new tasks, showcasing its strong generalization ability.

Performance of DS-MVP in multi-label classification

For the ML classification, we compared DS-MVP performance under both binary and ML evaluation schemes. In both evaluation schemes, the metrics of the three gradient boosting models were quite similar, while DS-MVP showed significant improvement in comparison as shown in Fig. 4 and Table 1.

For cardiomyopathy disease, DS-MVP achieved AUC values of 0.9147, 0.9643, 0.7814, 0.9704, 0.6368, 0.8379, and 0.9947, with corresponding AUPR scores of 0.7011, 0.8511, 0.1002, 0.4603, 0.0337, 0.6020, and 0.9968 across seven classes in binary classification evaluation scheme. Due to the limited availability of data for RCM (restrictive cardiomyopathy) and NVM (noncompaction of ventricular myocardium), the model's ability to effectively train on these types is restricted, resulting in reduced predictive performance for these specific disease types. For ML evaluation scheme, DS-MVP achieved scores of 0.8674, 0.8519, 0.8543, and 0.0385 for precision, recall, F1-score, and HammingLoss [47], respectively. DS-MVP outperformed the gradient boosting models across the first three metrics, with improvements of 4.65%, 5.11%, and 4.87%, respectively, and also surpassed them in the HammingLoss metric by more than 2.0%, where a lower HammingLoss indicates better performance.

In the prediction of the pathogenicity of eye diseases, the DS-MVP model demonstrated superior performance compared to three other gradient boosting algorithms within a binary classification scheme, achieving higher AUC values across five categories. Specifically, the DS-MVP model attained AUC scores of 0.9607 for Albinism (Alb), 0.9054 for cone-rod dystrophies

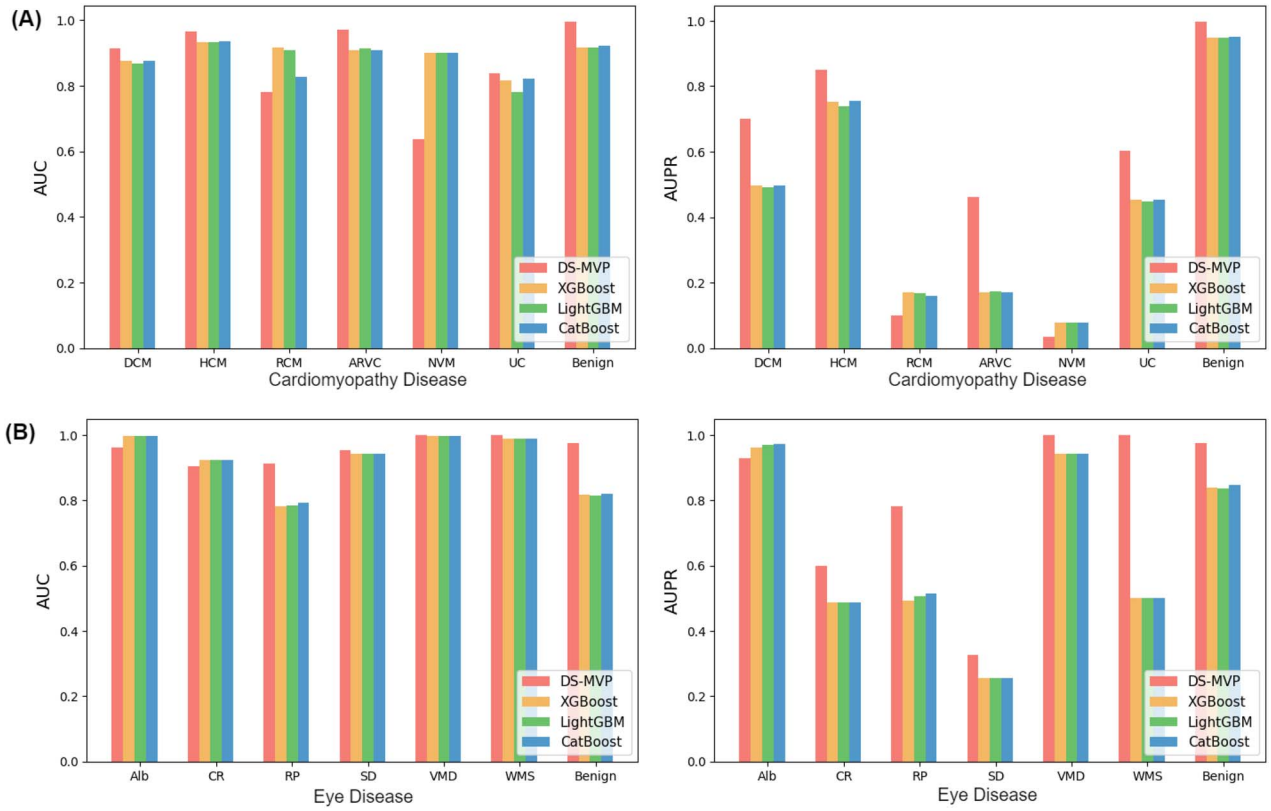

Figure 4. Performance of DS-MVP with XGBoost trained on ML classification. (**A**) Use AUC and AUPR as evaluation metrics on the CSPD_ML dataset. (**B**) Use AUC and AUPR as evaluation metrics on the ESPD_ML dataset.

(CR), 0.9138 for retinitis pigmentosa (RP), 0.9544 for Stargardt disease (SD), and 0.9753 for B. Notably, despite having fewer than five data points in the test set, the DS-MVP model accurately predicted pathogenicity in the vitelliform macular dystrophy (VMD) and Weill–Marchesani syndrome (WMS) categories. Given the data imbalance, we prioritized the AUPR metric. The DS-MVP model outperformed other models across six categories, highlighting its robustness in handling imbalanced data. In the ML evaluation scheme, DS-MVP achieved the highest precision (0.8252), recall (0.8241), F1-score (0.8180), and the lowest HammingLoss (0.0465), thereby demonstrating its comprehensive predictive performance. These outcomes underscored the advantages of our model, as well as the substantial contribution of pre-trained pathogenicity module to feature extraction.

Performance of DS-MVP in multi-class classification

For the MC classification, Fig. 5 presented a performance comparison between DS-MVP and gradient boosting models, the detailed results were displayed in SI Table S5. In neurodegenerative disease (Fig. 5A), DS-MVP showed superior performance across all categories, achieving an average accuracy of 0.9261, an average precision of 0.9385, an average recall of 0.8592, an average F1-score of 0.8913, an average AUC of 0.9790, and an average AUPR of 0.9372. Notably, in the critical AUPR metric, DS-MVP outperformed other methods with an average improvement of 8.66% in the Alzheimer disease (AD) category, 1.80% in the amyotrophic lateral sclerosis (ALS) category, 10.20% in the Parkinson disease (PD) category, and 5.28% in the B category, demonstrating significant advantages.

We also created a confusion matrix to illustrate the predicted percentage and distribution for NSPD_MC dataset in each class, as shown in Fig. 6. The confusion matrix revealed that the model can successfully classifies the four categories, with the highest classification ratio for B (1.00), followed by ALS (0.93), PD (0.78), and AD (0.73). Observing the confusion matrix confirmed the good performance of DS-MVP on test dataset. The detailed analysis is provided in the SI Text S9.

In the MC classification task for endocrine diseases, the DS-MVP model consistently exhibits outstanding performance, as illustrated in Fig. 5B. Specifically, the DS-MVP model achieved remarkable overall average results across various metrics: precision of 0.9467, recall of 0.9448, F1-score of 0.9429, AUC of 0.9953, AUPR of 0.9861, and ACC of 0.9448. These results suggest that the DS-MVP model performs robustly on small-scale disease datasets in MC classification tasks, indicating its potential for practical application.

We further analyzed the fine-tuning process on disease-specific datasets, which led to a more accurate distinction of missense variants across genes due to precise gene characterization. Furthermore, the deep sequence analysis enabled precise prediction of pathogenicity at different variant locations within genes. This also confirmed that the representation of missense variants was effectively learned during pre-training. The results are shown in SI Text S10.

Analysis of specific disease based on TNNT2 and FBXO7 genes

We conducted an in-depth analysis of genes that are clinically associated with specific diseases and evaluated the accuracy of our model predictions.

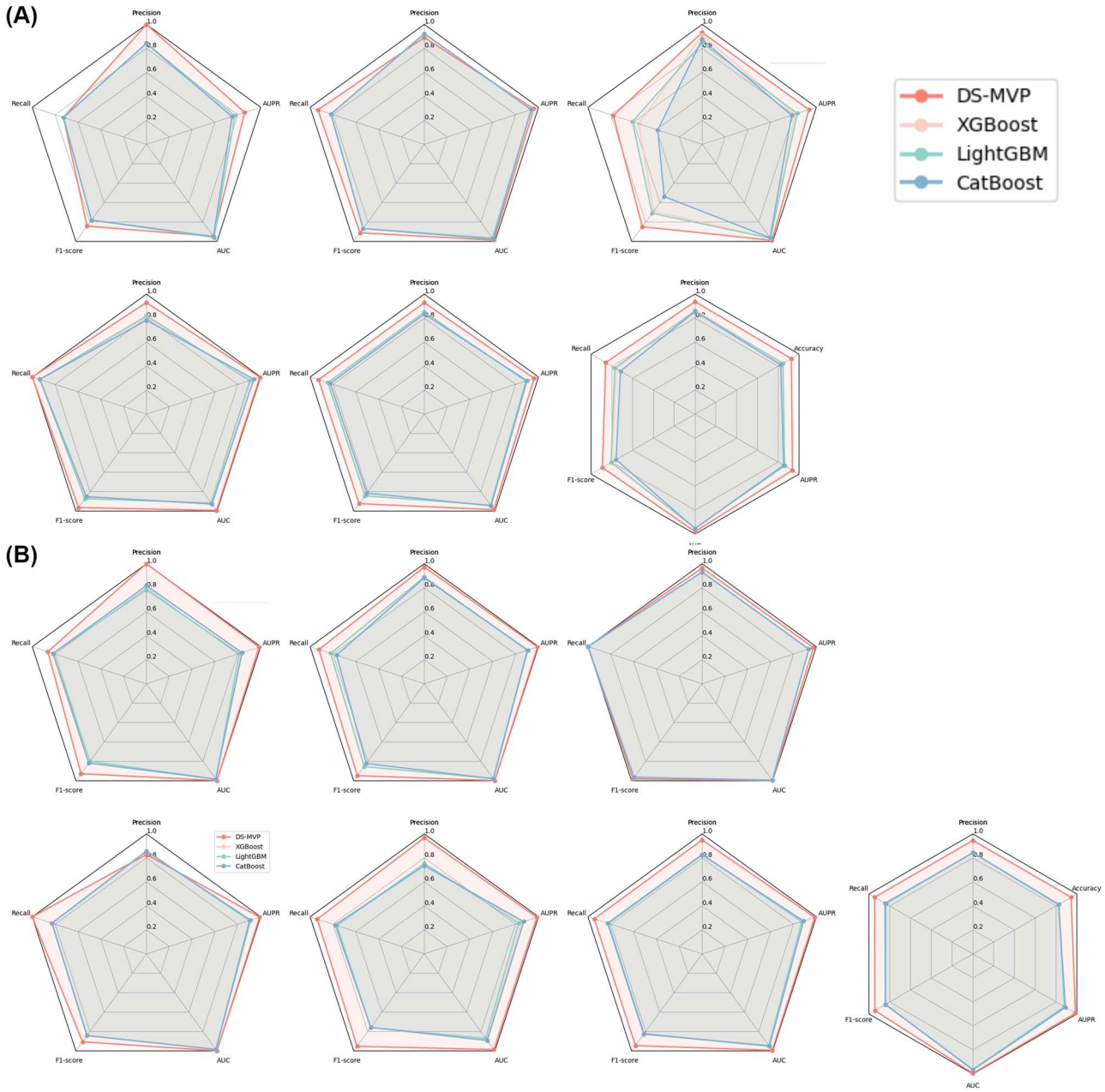

Figure 5. Performance of DS-MVP with XGBoost trained on the NSPD_MC and EcSPD_MC dataset based on precision, recall, F1-score, AUC, AUPR, and Accuracy. (**A**) The metrics of AD, ALS, PD, B, weighted average, macro average for NSPD_MC dataset. (**B**) The metrics of AG, DM, EGN, GD, B, weighted average, macro average for EcSPD_MC dataset.

The TNNT2 gene encodes cardiac troponin T, a vital component of the troponin complex responsible for regulating heart muscle contraction. Variants in this gene can impair the contractile function of cardiac muscles, leading to heart failure [48]. Clinically, TNNT2 gene variants are strongly linked to cardiomyopathies, particularly dilated cardiomyopathy (DCM) and hypertrophic cardiomyopathy (HCM) [49–51]. Our analysis focused on missense variants associated with both DCM and HCM, where DS-MVP successfully predicted missense variants based on TNNT2 gene labeled with two diseases. These predictions were validated against ClinVar database labels, which confirmed the association with DCM and HCM. This supports the key role of TNNT2 in cardiomyopathy disease while also demonstrating the accuracy of our model in predicting cardiomyopathy conditions.

Similarly, the FBXO7 gene plays a key role in the ubiquitin-proteasome pathway, which regulates protein degradation and is crucial for neuronal survival. Variants in FBXO7 are linked to early-onset PD [52, 53]. Using DS-MVP, we analyzed missense variants labeled as PD and corresponding to the gene FBXO7 in the test set. Our model accurately predicted these variants as being linked to PD, which is consistent with existing research highlighting the role of FBXO7 in the disease. This outcome affirming the effectiveness of DS-MVP in identifying disease-specific variants, further validating its precision in predictive.

These detailed analyses confirm DS-MVP's strong predictive power in diagnosing specific diseases, especially in tackling complex ML and MC classification. Its accuracy in prediction pathogenicity of disease-specific conditions underscores its

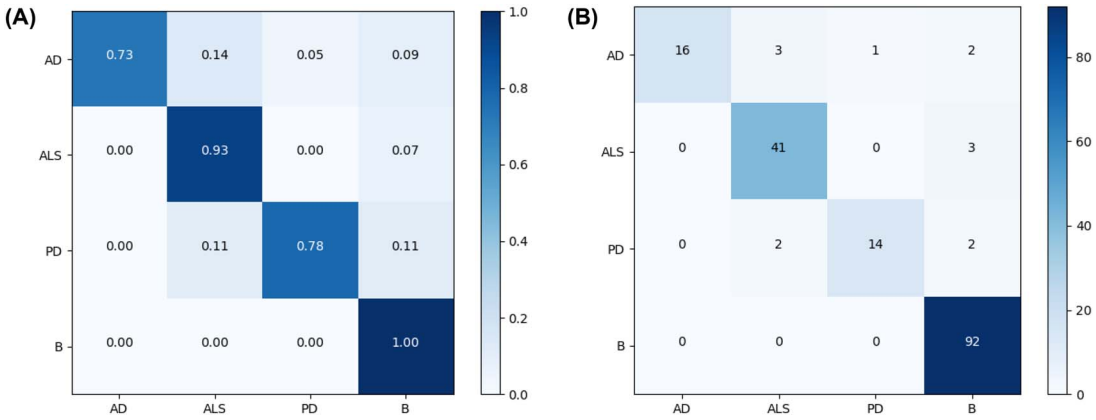

Figure 6. Confusion matrix of MC classification for NSPD_MC. (**A**) The normalized confusion matrix. (**B**) Confusion matrix shows detailed numbers.

reliability, while also offering new directions for future genetic research in various diseases.

Analysis of correlation between pathogenicity and conservation scores

The conservation score measures the degree of sequence preservation across evolution, with higher scores indicating evolutionary conservation, suggesting critical roles in organism function.

To investigate the link between variant pathogenicity and conservation scores, we plotted a scatter plot with DS-MVP predicted values on the x-axis and the average of eight conservation scores on the y-axis, as depicted in Fig. 7A and SI Fig. S2. This revealed a correlation between pathogenicity likelihood and conservation scores. However, benign variants showed a wide distribution across conservation score intervals, indicating a weak association. We also visualized conservation scores using nucleotide sequence logos for four missense variants, with site 8 as the variant site, as shown in Fig. 7B–E. The analysis of NM_017841.4:c.233G>A and NM_015164.4:c.1286G>A in Fig. 7B and C highlighted the significance of sequence conservation in identifying pathogenic variants. However, the relationship between pathogenicity and sequence conservation was inconsistent, as seen in Fig. 7D and E. To align pathogenicity with clinical practice, we found supporting evidence in the literature [54] for NM_000454.5:c.205T>C and protein structure, as shown in Fig. 7F, for NM_004415.4:c.6224G>A. These findings underscore the complexity of the pathogenicity-conservation relationship, necessitating further research. More detailed analysis can be found in SI Text S11.

Unveiling features with visual analysis in DS-MVP

DS-MVP extracted deeper representation of missense variants using DL during the pre-trained phase and subsequently trained XGBoost on the corresponding dataset using the representation along with the genetic features. To further explore the comprehensibility of the learned representation, we quantified the significance of various feature modalities and vividly illustrated the analysis through visualizations, offering a clear perspective on feature importances.

AFs and genetic features representation contribute the most

We used the function of XGBoost to get the feature importance weights for each dimension. Sorting dimensions and corresponding features individual to individual. The weights were averaged for each category of features to obtain the corresponding weight scores, the results are shown in Fig. 8A. From the figure, it can be seen that two AFs information weights accounting for the most important percentage of XGBoost. The genetic features representation consist of two AFs and eight conservation scores trained by DL were ranked as the second most important feature. Meanwhile, gene description corpus features, mutated amino acid features, and conservation scores closely followed in the rankings, holding positions of nearly equal significance. The features occupying the lowest rank on the importance roster were DNA sequence and DNA structure features.

Gene description corpus features hold considerable importance

To further analyze the importance of each feature in DL, we conducted ablation experiments by sequentially masking each class of features. We evaluated these models using the HardD test dataset regarding the AUC and AUPR as shown in Fig. 8B. We found that DS-MVP showed the best performance across both metrics compared with all other model setups. In particular, AFs, gene description corpus features and conservation scores play a great role in DL, which were consistent with the conclusions drawn from the XGBoost feature importance analysis. The utilization of a large language model for representing gene corpus features further enhances the model's capability to interpret biological information and distinguish between various missense variants. Furthermore, initially acquiring features through DL was a critical part of the model, as it significantly enhanced performance compared to models without DL.

Significance of two-stage training in disease-specific pathogenicity prediction

To validate our two-stage training approach for ML and MC tasks in disease-specific conditions, we compared three models: a single-stage DL model trained on disease-specific data, a model pre-trained on a general dataset then fine-tuned on disease-specific data, and DS-MVP. Results in SI Tables S6 and S7 show that DS-MVP significantly outperforms the others, effectively addressing the challenge of limited disease-specific data. By leveraging pre-training to capture general disease mechanisms, this approach improves prediction accuracy, adaptability, and generalization, enabling the rapid development of practical disease-specific predictive models.

The analysis of DNA sequence motif

Functional motifs of DNA sequence in the genome are short segments with specific biological functions and often appear in clusters, which are likely closely related to their regulatory roles

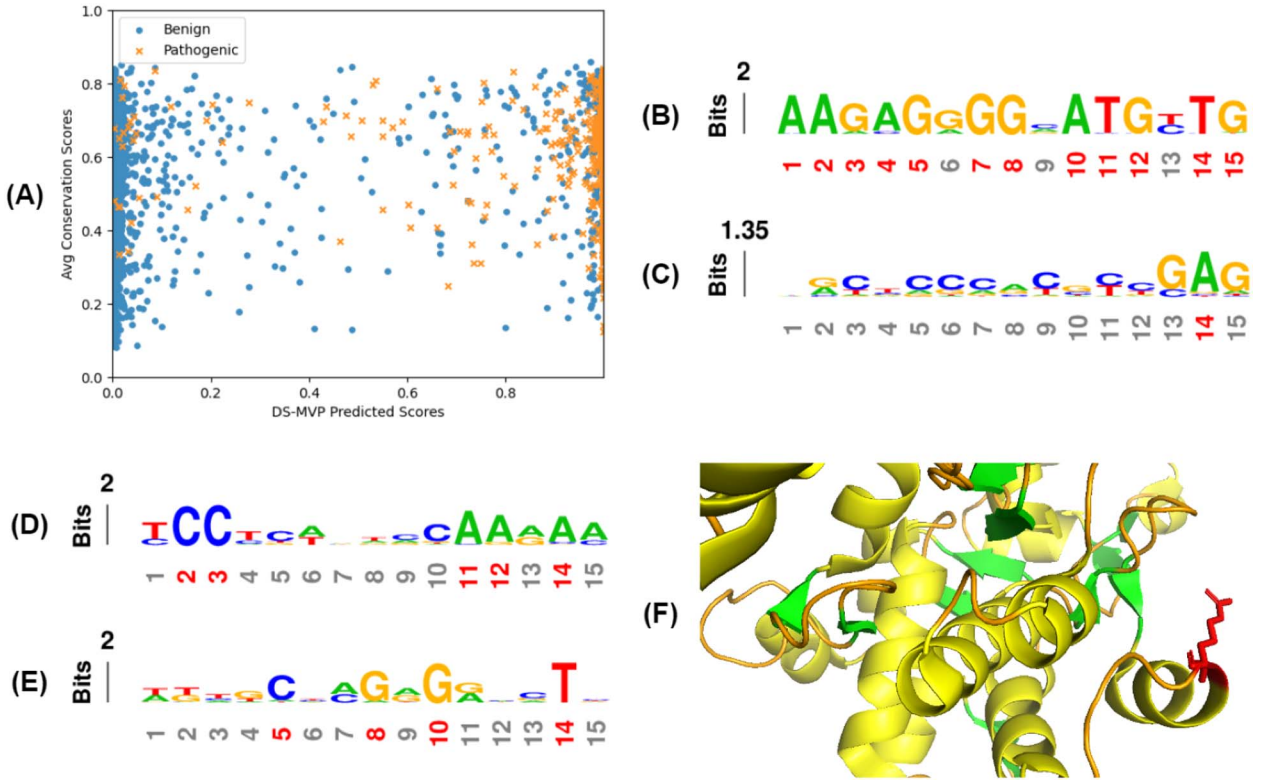

Figure 7. Correlation between pathogenicity and the average of conservation scores. (**A**) The overall distribution of DS-MVP predicted scores and the average of conservation scores on CSPD_Bi dataset. (**B**) A nucleotide sequence logo of pathogenic variant NM_017841.4:c.233G>A. (**C**) A nucleotide sequence logo of benign variant NM_015164.4:c.1286G>A. (**D**) A nucleotide sequence logo of pathogenic variant NM_000454.5:c.205T>C. (**E**) A nucleotide sequence logo of benign variant NM_004415.4:c.6224G>A.

[55]. Thus, studying functional motifs is crucial for understanding gene expression regulation mechanisms. To this end, we conducted a detailed analysis of DNA sequence features in our model. Specifically, we extracted sequence weight information from the LSTM layer, generated predicted motifs using the MEME tool [56], and compared their position weight matrices to known motifs in the JASPAR database [57] via the TOMTOM tool [58]. The results are shown in Fig. 9. Our analysis reveals that high-weight sequence segments in the model effectively match known motifs. This indicates that our model not only accurately identifies motifs in variant regions but also captures biologically significant motifs in surrounding contexts, thereby extracting functional biological information. These findings provide additional evidence supporting the scientific validity and methodological soundness of our model construction.

Pre-trained representation provides differential metric for general pathogenicity of missense variants

From feature ablation experiment, it was clear that representation of missense variants extracted through pre-trained pathogenicity model were superior to the direct use of source features for classification. To further verify support this observation, we applied t-SNE [59] to reduce the dimensionality of the representation of missense variants, both with and without DL, and plotted the results as scatterplots, respectively.

In Fig. 8C, yellow represents variants labeled as pathogenic, while green indicates variants labeled as benign. Comparing the two scatterplots, it was evident that the features extracted without DL were more mixed, with pathogenic and benign variants intermingled, making them harder to distinguish. In contrast, the features derived from DL exhibit clearer boundaries between pathogenic and benign variants. This visual comparison demonstrates that features pre-trained through DL were more effective for distinguishing between pathogenic and benign variants.

To evaluate the model's performance in ML and MC classification tasks, we applied t-SNE to analyze feature representations before and after training. For cardiomyopathy disease (Fig. 8D), the model effectively separated benign (red) and pathogenic variants, with DCM (brown) and HCM (sky blue) showing related clustering due to shared missense variants, consistent with theoretical expectations. Despite limited samples for other diseases, clear clustering was observed post-training. In the neurodegenerative disease (Fig. 8E), benign and pathogenic variants were distinctly separated, with three pathogenic subtypes forming well-defined clusters. These results demonstrate that our pretrained model captures disease-contextual semantic information, enhancing its accuracy in predicting the pathogenicity of missense variants.

Conclusion

In this study, we developed DS-MVP, a novel method for predicting the pathogenicity of DNA missense variants, with a focus on specific diseases. To build this model, we first collected a comprehensive dataset from the ClinVar database. DS-MVP leveraged pre-trained DL model to extract high-dimensional representation of missense variants in order to enrich the representation of general pathogenic mechanism, which were then combined with genetic features, and as inputs for XGBoost model to predict the pathogenicity of disease-specific missense variants. DS-MVP integrates different feature types, including genetic

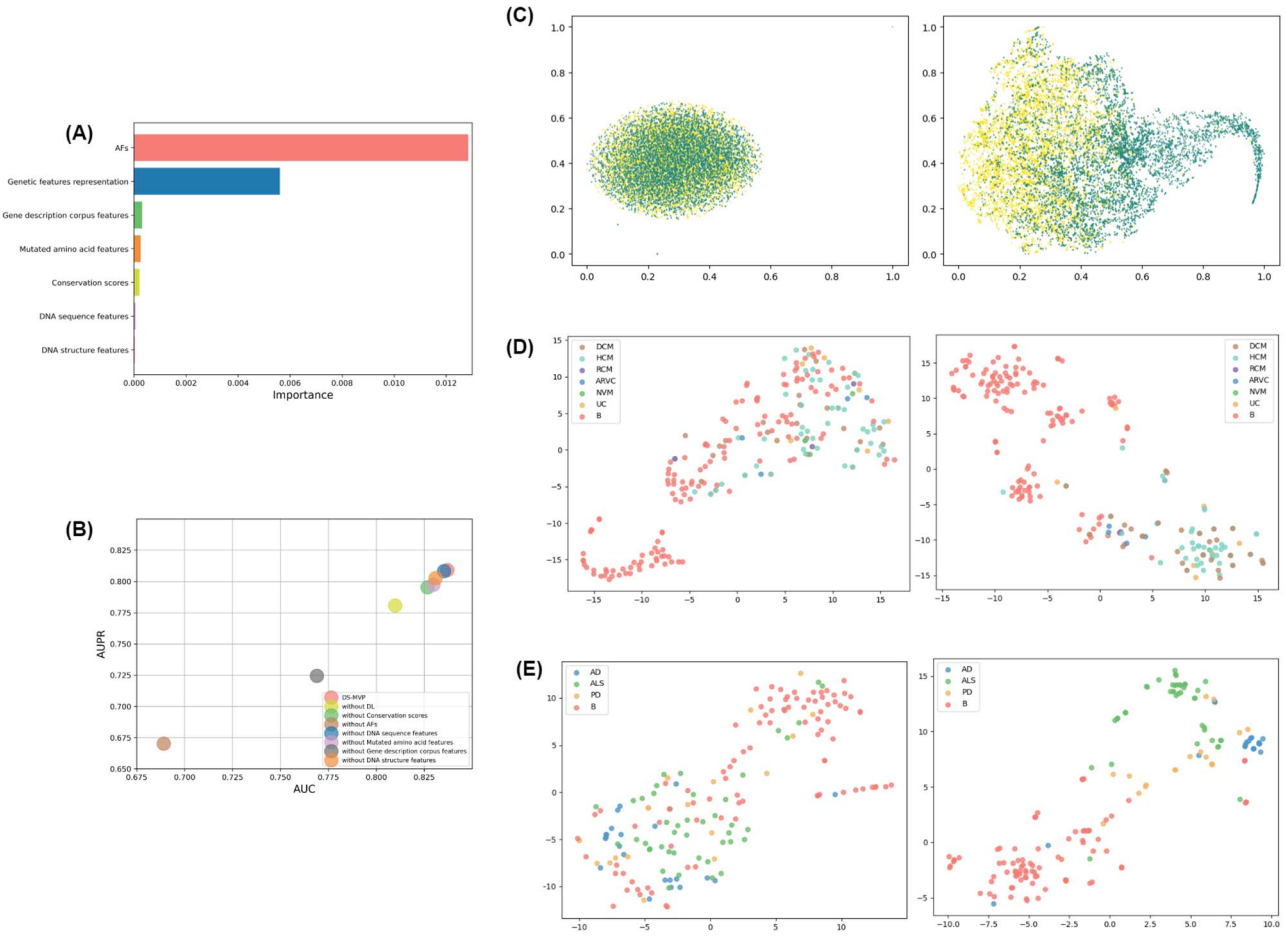

Figure 8. Analysis the importance of DS-MVP features and t-SNE visualization. (**A**) XGBoost feature importance ranking. (**B**) Impact of feature ablation on model performance. (**C**) Feature visualization of binary classification before and after training by DL. (**D**) Feature visualization of ML classification before and after training by DL. (**E**) Feature visualization of MC classification before and after training by DL.

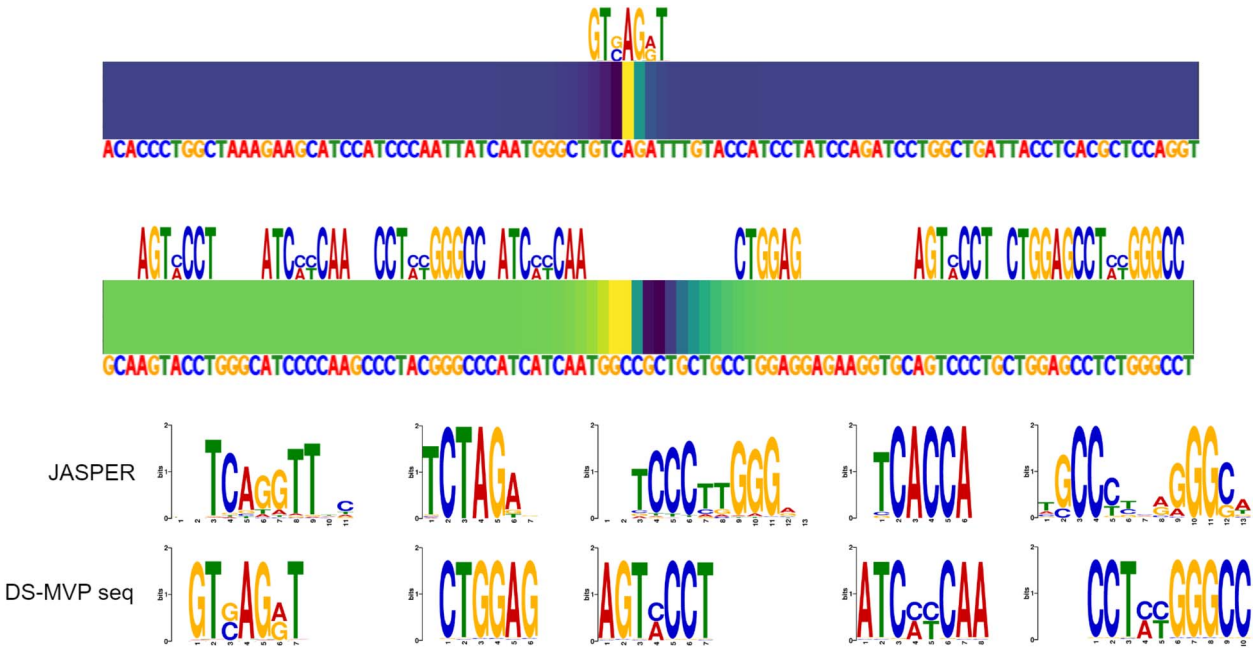

Figure 9. The analysis of functional motifs extracted from the sequences.

features (AFs and conservation scores), DNA sequence features, mutated amino acid features, gene description corpus features embedded by text-embedding-3-small model from OpenAI, and DNA structure features from Deep DNAShape. Each feature type is processed by specialized DL modules (GFFM, DSeqRM, MAAEM, OBGRM, DStructRM) to learn complex representations, which are concatenated to form a comprehensive variant representation. To demonstrate the validity of the learned representation and for application to downstream to the prediction of specific diseases, we first to conducted a general binary pathogenicity prediction to distinguish between pathogenic and benign. Then we further classified a pathogenic missense variants into detailed disease types for ML and MC pathogenicity prediction. Furthermore, we conducted a comprehensive visualization study and analysis of the model and features.

The primary contribution of DS-MVP is its novel approach to further classification of specific diseases according detail disease conditions based on pathogenicity in missense variants at the nucleotide level. This capability enriches our understanding of the complexity of genetic diseases and their varying effects on disease risk. Importantly, DS-MVP does not rely on the outputs of other pathogenicity prediction tools, ensuring its independence and robustness. Additionally, leveraging large language model from OpenAI for gene corpus representation proved invaluable, as it allowed for more nuanced characterization of gene information, improving overall prediction accuracy. Although our method demonstrates strong performance in predicting specific diseases, each disease currently requires separate modeling. In the future, we aim to develop a unified model to directly learning the representation of the specific diseases for disease-specific pathogenicity prediction. Moreover, predicting missense variants can facilitate a deeper understanding of research in other fields. By predicting missense variants, we can further infer their effects on protein function and structure, which can alter molecular interactions, affect drug-binding sites, and influence drug efficacy [60–62]. In summary, DS-MVP is essential for predicting missense variant pathogenicity and offers key insights into protein structure-function relationships, with significant implications for drug design, personalized medicine, and the study of genetic disease mechanisms.

Key Points

- DS-MVP, as an independent predictors, utilized computational methods to quickly predict pathogenicity of missense variants at the nucleotide level. It outperforms existing state-of-the-art methods, including meta predictors, making it crucial for improving disease diagnosis and advancing clinical research.
- DS-MVP is an innovative approach for assessing the pathogenicity and conditions of missense variants, specifically focusing on ML diseases and MC diseases.
- DS-MVP conducted pre-training on a large-scale dataset of general pathogenicity for missense variants to model fundamental representations of pathogenic mechanisms, enriching our understanding and enabling subsequent disease-specific pathogenicity predictions.
- DS-MVP leveraged sequence, structure, and biophysical properties, with the most influential ones being genetic features and the novel gene description feature. Among these, the gene description corpus was encoded using a large language model from OpenAI, enabling a more nuanced characterization of biological contexts.

Author contributions

QF.C. developed the methodology, coded, conducted experiments, and drafted the manuscript. LJ.Q. conceptualized, investigated, established methodology, and edited the manuscript. LX.C. and YL.J. aided in methodology development. B.Z. and ZJ.Z. curated the data. LC.P., JK.W., and LP.N. handled project visualization. G.L. managed software. TF.W. reviewed the manuscript. Q.Lyu. conceptualized, provided computing support and advice, and edited the manuscript.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported in part by the National Natural Science Foundation of China Fund under Grants (62272335, 62002251), in part by Jiangsu Colleges and Universities QingLan project, in part by University-Industry Collaborative Education Program, in part by the Natural Science Foundation of Jiangsu Province Youth Fund (BK20200856), in part by Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

Data availability

Source data/code is available at https://github.com/ljquanlab/DS-MVP.

References

1. Satam H, Joshi K, Mangrolia U. *et al.* Next-generation sequencing technology: current trends and advancements. *Biology* 2023;**12**:997.
2. Tokheim C, Bhattacharya R, Niknafs N. *et al.* Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res* 2016;**76**:3719–31. https://doi.org/10.1158/0008-5472.CAN-15-3190
3. Medina-Carmona E, Betancor-Fernández I, Santos J. *et al.* Insight into the specificity and severity of pathogenic mechanisms associated with missense mutations through experimental and structural perturbation analyses. *Hum Mol Genet* 2019;**28**:1–15. https://doi.org/10.1093/hmg/ddy323
4. Manzoor H, Aslam N, Pervez MT. *et al.* Evaluating accuracy of pathogenicity prediction methods for single nucleotide polymorphisms. *VFAST Trans Softw Eng* 2023;**11**:215–26. https://doi.org/10.21015/vtse.v11i2.1568
5. Stefl S, Nishi H, Petukh M. *et al.* Molecular mechanisms of disease-causing missense mutations. *J Mol Biol* 2013;**425**:3919–36. https://doi.org/10.1016/j.jmb.2013.07.014

6. de Oliveira Garcia FA, de Andrade ES, Palmero EI. Insights on variant analysis in silico tools for pathogenicity prediction. *Front Genet* 2022;**13**:1010327. https://doi.org/10.3389/fgene.2022.1010327
7. Liu Y, Zhang T, You N. *et al.* MAGPIE: accurate pathogenic prediction for multiple variant types using machine learning approach. *Genome Med* 2024;**16**:3. https://doi.org/10.1186/s13073-023-01274-4
8. Carter H, Douville C, Stenson PD. *et al.* Identifying mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 2013;**14**:1–16. https://doi.org/10.1186/1471-2164-14-S3-S3
9. Cheng J, Novati G, Pan J. *et al.* Accurate proteome-wide missense variant effect prediction with alphamissense. *Science* 2023;**381**:eadg7492. https://doi.org/10.1126/science.adg7492
10. Rentzsch P, Witten D, Cooper GM. *et al.* CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**:D886–94. https://doi.org/10.1093/nar/gky1016
11. Ke G, Meng Q, Finley T. *et al.* LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017;**30**:3146–54.
12. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324
13. Li C, Zhi D, Wang K. *et al.* MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. *Genome Med* 2022;**14**:115. https://doi.org/10.1186/s13073-022-01120-z
14. Qi H, Zhang H, Zhao Y. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nat Commun* 2021;**12**:1–9. https://doi.org/10.1038/s41467-020-20847-0
15. Alirezaie N, Kernohan KD, Hartley T. *et al.* ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *Am J Hum Genet* 2018;**103**:474–83. https://doi.org/10.1016/j.ajhg.2018.08.005
16. He K, Zhang X, Ren S. *et al.* Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA. pp. 770–8, 2016.
17. Fagerberg L, Hallström BM, Oksvold P. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 2014;**13**:397–406. https://doi.org/10.1074/mcp.M113.035600
18. Patel SA, Hirosue S, Rodrigues P. *et al.* The renal lineage factor PAX8 controls oncogenic signalling in kidney cancer. *Nature* 2022;**606**:999–1006. https://doi.org/10.1038/s41586-022-04809-8
19. Loboda AP, Adonin LS, Zvereva SD. *et al.* BRCA mutations-the achilles heel of breast, ovarian and other epithelial cancers. *Int J Mol Sci* 2023;**24**:4982. https://doi.org/10.3390/ijms24054982
20. Krasnova M, Efremova A, Bukhonin A. *et al.* The effect of complex alleles of the *CFTR* gene on the clinical manifestations of cystic fibrosis and the effectiveness of targeted therapy. *Int J Mol Sci* 2023;**25**:114. https://doi.org/10.3390/ijms25010114
21. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA. pp. 785–94, 2016.
22. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 2005;**18**:602–10. https://doi.org/10.1016/j.neunet.2005.06.042
23. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. *Advances in Neural Information Processing Systems* 2017;**30**:5998–8.
24. LeCun Y, Bottou L, Bengio Y. *et al.* Gradient-based learning applied to document recognition. *Proc IEEE* 1998;**86**:2278–324. https://doi.org/10.1109/5.726791
25. Li J, Chiu T-P, Rohs R. Predicting DNA structure using a deep learning method. *Nat Commun* 2024;**15**:1243. https://doi.org/10.1038/s41467-024-45191-5
26. Li J, Rohs R. Deep DNAshape webserver: prediction and real-time visualization of DNA shape considering extended k-mers. *Nucleic Acids Res* 2024;**52**:W7–W12.
27. Landrum MJ, Lee JM, Benson M. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;**46**:D1062–7. https://doi.org/10.1093/nar/gkx1153
28. Koç CK. Analysis of sliding window techniques for exponentiation. *Comput Math Appl* 1995;**30**:17–24. https://doi.org/10.1016/0898-1221(95)00153-P
29. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev* 2021;**54**:1937–67. https://doi.org/10.1007/s10462-020-09896-5
30. Nielsen D. *Tree Boosting with Xgboost-why Does Xgboost Win" every" Machine Learning Competition? Master's Thesis*. NTNU, 2016.
31. Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS arXiv 2019. https://doi.org/10.48550/arXiv.1912.06059
32. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* 2011;**32**:894–9. https://doi.org/10.1002/humu.21517
33. Liu X, Li C, Mou C. *et al.* dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med* 2020;**12**:1–8.
34. Malhis N, Jacobson M, Jones SJM. *et al.* LIST-S2: taxonomy based sorting of deleterious missense mutations across species. *Nucleic Acids Res* 2020;**48**:W154–61. https://doi.org/10.1093/nar/gkaa288
35. Rogers MF, Shihab HA, Mort M. *et al.* FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* 2018;**34**:511–3. https://doi.org/10.1093/bioinformatics/btx536
36. Sundaram L, Gao H, Padigepati SR. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* 2018;**50**:1161–70. https://doi.org/10.1038/s41588-018-0167-z
37. Raimondi D, Tanyalcin I, Ferté J. *et al.* DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res* 2017;**45**:W201–6. https://doi.org/10.1093/nar/gkx390
38. Schwarz JM, Cooper DN, Schuelke M. *et al.* MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* 2014;**11**:361–2. https://doi.org/10.1038/nmeth.2890
39. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011;**39**:e118–8. https://doi.org/10.1093/nar/gkr407
40. Pejaver V, Urresti J, Lugo-Martinez J. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat Commun* 2020;**11**:5918. https://doi.org/10.1038/s41467-020-19669-x
41. Feng B-J. PERCH: a unified framework for disease gene prioritization. *Hum Mutat* 2017;**38**:243–51. https://doi.org/10.1002/humu.23158
42. Ioannidis NM, Rothstein JH, Pejaver V. *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 2016;**99**:877–85. https://doi.org/10.1016/j.ajhg.2016.08.016

43. Ionita-Laza I, McCallum K, Bin X. *et al.* A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016;**48**:214–20. https://doi.org/10.1038/ng.3477
44. Jagadeesh KA, Wenger AM, Berger MJ. *et al.* M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 2016;**48**:1581–6. https://doi.org/10.1038/ng.3703
45. Dong C, Wei P, Jian X. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 2015;**24**:2125–37. https://doi.org/10.1093/hmg/ddu733
46. Prokhorenkova L, Gusev G, Vorobev A. *et al.* CatBoost: unbiased boosting with categorical features. *Adv Neural Inf Process Syst* 2018;**31**:6639–49.
47. Sorower MS. *A literature survey on algorithms for multi-label learning*, Oregon State University, Corvallis 2010;**18**:25.
48. Shafaattalab S, Li AY, Gunawan MG. *et al.* Mechanisms of arrhythmogenicity of hypertrophic cardiomyopathy-associated troponin t (TNNT2) variant I79N. *Front Cell Dev Biol* 2021;**9**:787581. https://doi.org/10.3389/fcell.2021.787581
49. Pham JH, Giudicessi JR, Tweet MS. *et al.* Tale of two hearts: a TNNT2 hypertrophic cardiomyopathy case report. *Front Cardiovasc Med* 2023;**10**:1167256. https://doi.org/10.3389/fcvm.2023.1167256
50. Li X, Luo R, Haiyong G. *et al.* Cardiac troponin T (TNNT2) mutations in Chinese dilated cardiomyopathy patients. *Biomed Res Int* 2014;**2014**:907360. https://doi.org/10.1155/2014/907360
51. Pettinato AM, Ladha FA, Mellert DJ. *et al.* Development of a cardiac sarcomere functional genomics platform to enable scalable interrogation of human TNNT2 variants. *Circulation* 2020;**142**:2262–75. https://doi.org/10.1161/CIRCULATIONAHA.120.047999
52. Lee SH, Lee YJ, Jung S. *et al.* E3 ligase adaptor FBXO7 contributes to ubiquitination and proteasomal degradation of SIRT7 and promotes cell death in response to hydrogen peroxide. *J Biol Chem* 2023;**299**:102909. https://doi.org/10.1016/j.jbc.2023.102909
53. Burchell VS, Nelson DE, Sanchez-Martinez A. *et al.* The Parkinson's disease–linked proteins Fbxo7 and Parkin interact to mediate mitophagy. *Nat Neurosci* 2013;**16**:1257–65. https://doi.org/10.1038/nn.3489
54. Fahmy N, Müller K, Andersen PM. *et al.* A novel homozygous p. Ser69Pro SOD1 mutation causes severe young-onset ALS with decreased enzyme activity. *J Neurol* 2023;**270**:1770–3.
55. D'haeseleer P. What are DNA sequence motifs? *Nat Biotechnol* 2006;**24**:423–5. https://doi.org/10.1038/nbt0406-423
56. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, Calif, USA. 1994;**2**:28–36.
57. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2021;**50**:D165–73. https://doi.org/10.1093/nar/gkab1113
58. Gupta S, Stamatoyannopoulos JA, Bailey TL. *et al.* Quantifying similarity between motifs. *Genome Biol* 2007;**8**:R24. https://doi.org/10.1186/gb-2007-8-2-r24
59. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–5.
60. Pandurangan AP, Blundell TL. Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning. *Protein Sci* 2020;**29**:247–57. https://doi.org/10.1002/pro.3774
61. Pennica C, Hanna G, Islam SA. *et al.* Missense3D-PPI: a web resource to predict the impact of missense variants at protein interfaces using 3D structural data. *J Mol Biol* 2023;**435**:168060. https://doi.org/10.1016/j.jmb.2023.168060
62. Huang L, Guo Z, Wang F. *et al.* KRAS mutation: from undruggable to druggable in cancer. *Signal Transduct Target Ther* 2021;**6**:386. https://doi.org/10.1038/s41392-021-00780-4

Briefings in Bioinformatics, 2025, **26(2)**, bbaf119
https://doi.org/10.1093/bib/bbaf119
Problem Solving Protocol