JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

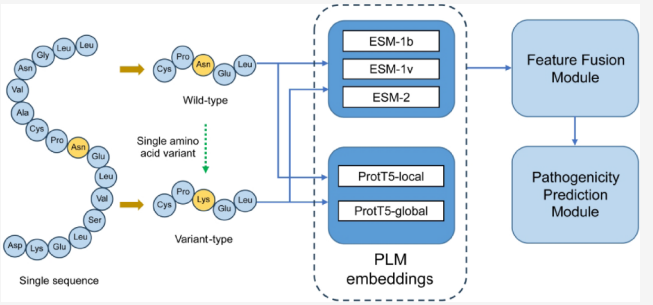
TransEFVP: A Two-Stage Approach for the Prediction of Human Pathogenic Variants Based on Protein Sequence Embedding Fusion

Zihao Yan,○ Fang Ge,○ Yan Liu, Yumeng Zhang, Fuyi Li, Jiangning Song,* and Dong-Jun Yu*

Cite This: J. Chem. Inf. Model. 2024, 64, 1407−1418 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Studying the effect of single amino acid variations (SAVs) on protein structure and function is integral to advancing our understanding of molecular processes, evolutionary biology, and disease mechanisms. Screening for deleterious variants is one of the crucial issues in precision medicine. Here, we propose a novel computational approach, TransEFVP, based on large-scale protein language model embeddings and a transformer-based neural network to predict disease-associated SAVs. The model adopts a two-stage architecture: the first stage is designed to fuse different feature embeddings through a transformer encoder. In the second stage, a support vector machine model is employed to quantify the pathogenicity of SAVs after dimensionality reduction. The prediction performance of TransEFVP on blind test data achieves a Matthews correlation coefficient of 0.751, an F_1-score of 0.846, and an area under the receiver operating characteristic curve of 0.871, higher than the existing state-of-the-art methods. The benchmark results demonstrate that TransEFVP can be explored as an accurate and effective SAV pathogenicity prediction method. The data and codes for TransEFVP are available at https://github.com/yzh9607/TransEFVP/tree/master for academic use.



1. INTRODUCTION

Each person has a collection of different genomic variants, including most single nucleotide polymorphisms (SNPs), insertion-deletion (Indels), and other types of variants.[1] The SNPs located in protein-coding regions sometimes result in substitutions of a single amino acid in the corresponding protein sequence that refer to single amino acid variations (SAVs), thus affecting protein structure and functions, and even leading to human diseases.[2,3] Although many SAVs are not harmful to human health, nearly one-third of SAVs were directly or indirectly associated with multiple diseases, such as thalassemia, hereditary breast cancer, schizophrenia, and SARS-CoV-2.[4−7] Due to the constraints in existing studies, the actual number of pathogenic SAVs present is considerably higher.[8] Therefore, it is critical to investigate the disease-related effects of SAVs on protein function and structure.

With the advancement of next-generation sequencing (NGS), the amount of genomic variation data has increased exponentially. For the uncharacterized data, experimental methods like gene probe, polymerase chain reaction (PCR), and restriction fragment length polymorphism (RFLP) are undoubtedly the most accurate to assess the variant effects,[9] but such experimental methods have the disadvantage of high consumption of time and money. Compared with traditional experimental methods, computational methods have obviously become a new auxiliary method to help researchers understand and study the pathogenesis of genetic diseases when dealing with large amounts of biological samples.[10,11] For example, Golgi_DF[12] and Phage_UniR_LGBM[13] are new classification methods for Golgi proteins and virion proteins. This is critical for the diagnosis and treatment of genetic diseases.

Various biocomputing methods have been developed and improved in the past years. SIFT[14] and PROVEAN[15] used sequence alignment-based protein conservation analysis to determine the risk of SAVs. SNPs&GO[16] introduced a score of association between gene ontology (GO)[17] annotations and the pathogenicity of missense variants and encoded into LGO features to improve prediction performance. SuSPect[18] applied the protein−protein interaction (PPI) network model to predict pathogenic SAVs. STRUM[19] and FoldX[20] used protein stability to approximate the effect of amino acid variation.[21] Many classical machine learning (ML) algorithms have also been applied to predicting the pathogenicity of SAVs, including support vector machines (SVMs),[16,22] random forest (RF),[23−26] naive Bayes classifier,[27,28] gradient tree boosting,[29] and ensemble ML models.[30] In recent years, deep learning (DL) models have arisen in this field.[31] MutPred2[32] integrated 6 different feature encoding methods of genetic and molecular

Received: December 19, 2023
Revised: January 30, 2024
Accepted: January 31, 2024
Published: February 9, 2024

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418

data, trained the model in a bagged ensemble architecture of 30-layer feed-forward neural network, and sorted the pathogenic probability of the variants. DeepSAV[33] incorporated population-level and gene-level information and used a convolutional neural network (CNN) to predict SAV pathogenicity based on input features of sequence, structure, and functional information. Pei et al. also developed the DBSAV database that reports gene tolerance of rare SAV (GTS) scores of human genes and DeepSAV scores of SAVs in the human proteome.[34] In addition, Pred-MutHTP,[35] mCSM-membrane,[36] and MutTMPredictor[37] were developed to predict the pathogenicity of variation in membrane proteins. The data sets used by these methods for training and testing are mostly extracted from ClinVar[38] and HUMSAVAR[39] databases, and the remaining are from precompiled variant databases like VariBench.[40]

Natural language processing (NLP) has been working on feeding large text corpora into DL-based language models (LMs). The rapid rise of large language models (LLMs) has led researchers to focus on whether this technology can be applied to biological sequences with language characteristics. The protein language model (PLM) results from applying natural language processing methods in bioinformatics and has succeeded in some directions.[41,42] Large-scale PLM-based embeddings learned underlying biological information from billions of protein sequences, focusing on expressing comprehensive features based only on protein sequences. The manipulation of PLMs typically involves transfer learning, where a pretrained LM is fine-tuned for a specific downstream task.[43] The PLM embeddings can be used even when resources are limited, greatly facilitating researchers' study and analysis. Several PLMs have been used for a wide range of biological tasks, including protein generation,[44] remote homology detection,[42] supervised low-N function prediction,[45] and binding residues for ligand prediction.[46] SSA[47] is based on bidirectional long short-term memory (LSTM) models trained with a two-part feedback mechanism to learn useful position-specific embeddings. CPCProt[48] divided protein sequences into fixed-size fragments and trained an autoregressive model to distinguish protein fragments. ESM-1b Transformer[49] collected 250 million sequences from the UniProt Archive (UniParc) database[50] and trained a model with ∼650m parameters using the Transformer model. ProtTrans[51] updated 6 models trained with different algorithms based on the big fantastic database (BFD).[52] MSA Transformer[53] combined multiple sequence alignment (MSA) and Transformer to realize information interaction between multiple coevolutionary sequences. ProGen[54] is conditioned on taxonomic and keyword tags such as molecular function and cellular component, trained a 1.2B-parameter language model on ∼280 M protein sequences. Their development team then scaled up the ProGen2[44] model to 6.4B parameters in databases of more than one billion proteins.

Although PLM embeddings have made outstanding contributions in many fields, research on predicting disease-causing variants is still scarce. SHINE[55] combined ESM-1b and MSA Transformers to predict the pathogenicity of short inframe insertion and deletion variants. ProtT5cons used a single PLM embedding as input to train a CNN model for conservative variation prediction and used its output to train a balanced logistic regression (LR) ensemble method to predict the effect score of SAVs.[56] E-SNP&GO[57] assembled ESM-1b, ProtT5, and GO functional annotations and used SVM to predict whether SAVs are associated with disease. These approaches have shown initial success but are stuck at using a single source of embeddings and directly feeding the embeddings into the classifier for training.

In this paper, we attempt to exploit the potential of multiple PLM embeddings. Some improvements are made in the following aspects: (1) The input encoding is divided into two parts, one of which includes ESM-1b,[49] ESM-1v,[43] and ESM-2[58] (the team's latest update), and the other is a pretrained model ProtT5[51] from ProtTrans project. It does not require any handcrafted features other than protein sequences, such as MSA and protein function annotation while increasing the information richness of the embeddings and not consuming too many computing resources. (2) As reported, the functional impact of one variant site is related to itself and several neighboring sites.[59] Therefore, we combine the local features of the variant site and global features of the surrounding contexual information in ProtT5 embedding. In this way, the "microenvironment" is constructed to reflect the natural characteristics of the variant site. (3) In protein property prediction tasks, different features are often combined in series as input to predictors. However, this simple and easy-to-implement method is often not the best for predictive performance. Here, we propose a transformer-based two-stage embedding fusion variant predictor named TransEFVP. Different PLM embeddings are fused through the encoder in the transformer model, and the output is used as the input feature of the second segment of the SVM classifier. Finally, the pathogenicity of SAVs is determined by the predicted probabilities. TransEFVP achieves a Matthews correlation coefficient (MCC) value of 0.751 on the benchmark data set, outperforming several existing state-of-the-art SAV pathogenicity prediction methods. Moreover, since we only use pretrained embeddings as input, which does not require much computation time and resources, our method is suitable for large-scale variation annotation.

2. MATERIALS AND METHODS

2.1. Data Set. For a more intuitive comparison, TransEFVP adopts the same training and blind test set as E-SNP&GO. These samples were collected separately from two public sources: HUMSAVAR and ClinVar. And we only keep the SAVs that are obviously related to the diseases listed in OMIM[60] and MONDO.[61] Both databases classify SAVs into the following classes: pathogenic or likely pathogenic (P/LP), benign or likely benign (B/LB), and uncertain significance (US).[57] Overall, the data set contains 111,412 SAVs in 13 661 protein sequences, including 43 895 P/LP SAVs in 3603 proteins and 67 517 B/LB SAVs in 13 229 proteins.

As in E-SNP&GO, we clustered the protein sequences using the MMseqs2 program,[62] limiting a minimum sequence identity of 25% within at least 40% of the pairwise alignment coverage to avoid bias between the training and test sets. We randomly selected 10% of the data to build a blind test set as a benchmark test set to test the generalization performance of our model and compare it with other methods. The remaining 90% of the data set was further randomly split into 10 subsets (ensure an equal distribution of positive and negative samples in each subset) that were used in a 10-fold cross-validation to train and optimize the input embedding and hyperparameters of the model. Table 1 provides a statistical summary of the SAV data sets.

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418

Table 1. Statistical Summary of the Benchmark Data Sets

	the number of variants		the number of proteins	
data set	pathogenic SAVs	neutral SAVs	pathogenic SAVs	neutral SAVs
training set	39 812	61 334	3279	11 945
blind test set	4083	6183	324	1284
US data set	9165		2588	

As listed in Table 1, the final size of each data set: the training set contains a total of 12 347 protein sequences, including 39 812 pathogenic SAVs and 61 334 neutral SAVs; the blind test set contains a total of 1314 protein sequences, including 4083 pathogenic SAVs and 6183 neutral SAVs; and an additional US data set containing 9165 SAVs in 2588 protein sequences.

2.2. Feature Representation and TransEFVP Model Architecture. As Figure 1 shows, TransEFVP is a two-stage SAV pathogenicity predictor based on PLM embeddings. The model is mainly divided into three parts: an embedding-based input feature encoding module (Figure 1A), a transformer-based feature fusion module (Figure 1B), and a predictor and output module (Figure 1C). The feature encoding stage is shown in Figure 1A. The input data are SAVs from certain positions in the human protein sequences. Five embeddings were extracted for variant-type and wild-type of the variant position and the surrounding microenvironment using multiple pretrained PLM models from two projects as input for the first stage of training. For the wild-sequence and variant-sequence, five sets of features with lengths of 1280 and 1024 were extracted, respectively.

At the first stage of training, we take the embedding features from the same project as a group and use the transformer-based encoder to deep-fuse these features. The training at this stage will output a result at the end, and the parameters of the model will be determined according to the quality of the result. The obtained features are concatenated and used as the input for the next stage of training. The details of the encoder module for deep feature fusion are shown in Figure 1B. The feature fusion module borrows from the TransPPMP model.[63] Positional encoding provides information about the relative positions of residues along the sequence. Four encoder layers employing multihead attention complete the encoding and fusion of different feature embeddings. A three-layer fully connected neural network was used to output the pathogenicity score of the first training stage. The score is used only for comparison and determination of the model architecture. The output of the concatenate layer is the result of feature fusion and feed as the input of the second training stage.

The output and recognition processes are shown in Figure 1C. At the second stage of training, the 5120 features obtained by deep feature fusion are input into an SVM predictor after reducing the dimensionality by PCA. Whether the SAV is

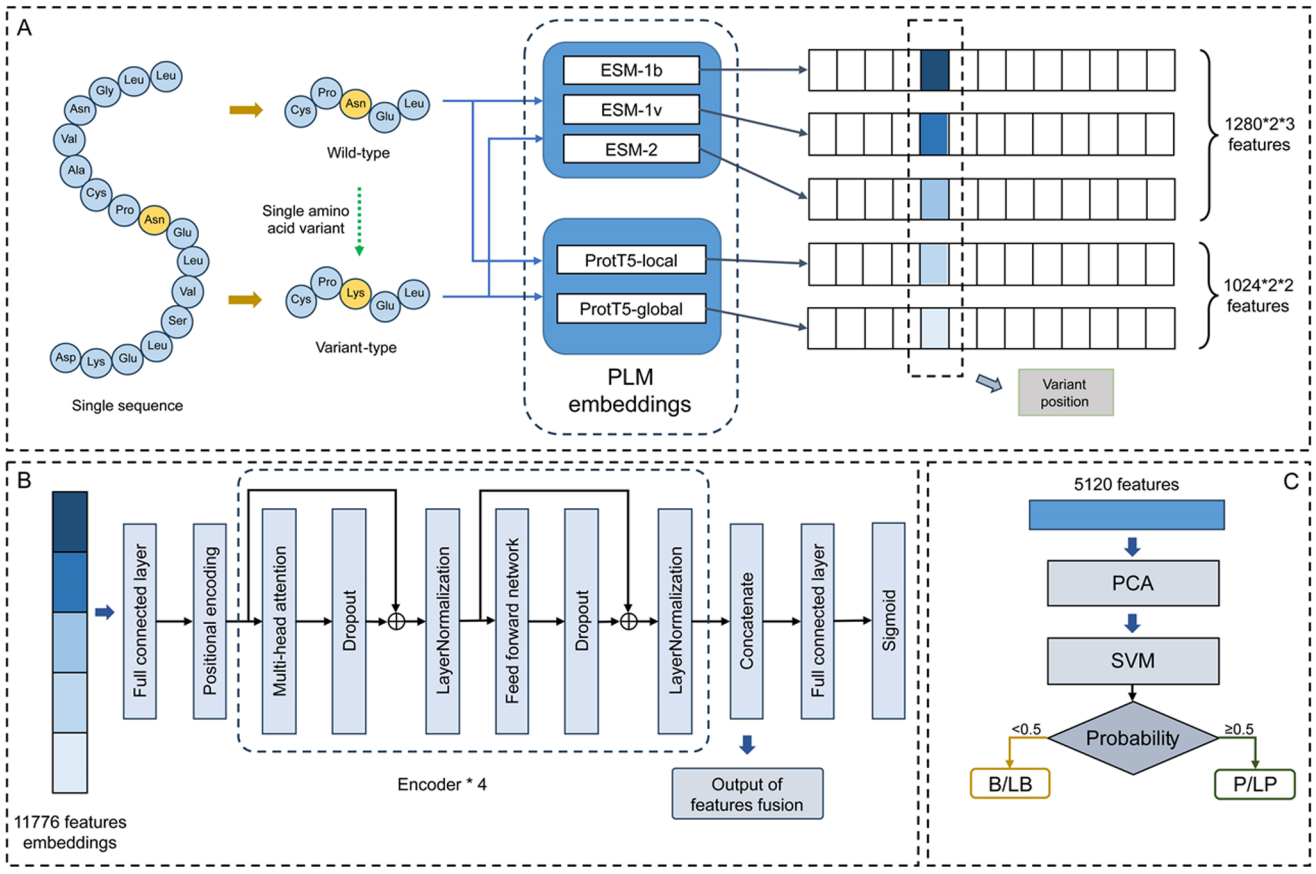


Figure 1. An overall workflow of TransEFVP: Panel a is the process of extracting PLM embeddings from wild-type and variant-type sequences. Panel B is the network architecture of the first stage of training. The extracted embeddings are input into the Transformer-based encoder, respectively, to complete the feature fusion. Panel C is the training process for the second stage. The fused features are reduced in dimension by PCA, and SVM completes the classification.

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418

pathogenic is judged according to the prediction score. Other details of the model will be described in the following sections.

2.2.1. Input Feature Representation. In the field of NLP, techniques based on self-supervision use the context in the text to predict missing words, which could generally represent the meaning of words.[64] PLM is the migration application of various language models in biochemistry. It inputs protein sequences and learns the underlying biochemical properties, secondary and tertiary structures, and internal laws of functions in the sequences. The learned representation space has a multiscale organization reflecting structure from the level of biochemical properties of amino acids to remote homology of proteins.[49] These representations are then applied as embeddings for downstream analysis tasks through transfer learning.

In this paper, we integrated four different embeddings as input for feature fusion: ESM-1b, ESM-1v, ESM-2, and ProtT5. ESM-1b used the system optimization method to optimize the hyperparameters in the Transformer model and then pretrained on the UniRef50 database. Since structural information is difficult to extract from protein sequences, a masked language model objective is used for training the model. That is, some fragments in sequences are randomly masked, and the real residues in the masked part are predicted based on the remaining residues along the sequence. It can reflect structure−function by learning linkages between residues in sequences. After training, multidimensional protein-related information can be interpreted in the feature representation of the ESM-1b model, such as residue biochemical properties, sequence variation properties, distant homology, secondary structure, and tertiary structure. ESM-1v was trained on the larger UniRef90 database with the same structure as ESM-1b. Since the model learned sequential patterns across the entire evolutionary tree, it could perform zero-shot predictions, i.e., migrate directly to other tasks without additional training. ESM-1v releases five models generated by training with five different random seeds. Through experiments, we found that these five versions of the model had little effect on the training results. For the sake of saving computing resources and time, we use only the first model among them. ESM-2 improved the model architecture and training parameters and increased computational resources and data. The addition of relative positional embeddings enables generalization to arbitrary length sequences. Better performance than previous models can be obtained with fewer parameters. ProtT5 used the original transformer architecture proposed for machine translation, which consisted of an encoder that projected a source sequence to an embedding space and a decoder that generated a translation to a target sequence based on the embedding.[51] The model is first trained on the BFD database and fine-tuned on the smaller UniRef50 database.

In the variation encoding part, we used the above models to extract the local and global features of the protein sequence. Given a protein sequence with L residues, we intercept the variant position i as the center and 100 residues before and after to extract the microenvironment information. We encode protein variants in a window of 201 sequence length, and then the features of the variation position are intercepted as the input of the next step. When using ProtT5 for encoding, we first employ the same method to extract the features of the variant position and then encode the entire window sequence into the same dimension as the global feature of the network input. It is worth mentioning that when SAV is located at the edge of the protein or less than 100 residues from the edge, we take the variant position as the center, retain all of the residues on the edge side, and split 100 residues on the other side as the window sequence.

In general, we aggregate feature vectors of dimension 11 776, and the detailed composition is as follows:

- ESM-1b embedding: 1280 features embedded from position i of the variant-type sequence and 1280 features from the wild-type sequence.
- ESM-1v embedding: 1280 features embedded from position i of the variant-type sequence and 1280 features from the wild-type sequence.
- ESM-2 embedding: 1280 features embedded from position i of the variant-type sequence and 1280 features from the wild-type sequence.
- ProtT5-local embedding: 1024 features embedded from position i of the variant-type sequence and 1024 features from the wild-type sequence.
- ProtT5-global embedding: 1024 features embedded from a 201-long sequence window of the variant-type sequence and 1024 features from the wild-type sequence.

2.2.2. Feature Fusion with Encoder. Borrowing from the Transformer model, we use 4 encoder layers employing multihead attention. The dot-product attention function has three inputs: Query (Q), Key (K), and Value (V). First, the dot-product between Q and K is performed, and to prevent the result from being too large leading to a hard softmax, we divide by the scale $\sqrt{d_k}$ (d_k is the dimension of the K-vector). Then, the results are normalized to a probability distribution, that is, attention weights, by a softmax operation. Finally, the attention weights are multiplied by V to obtain the weighted sum representation.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

In the multihead attention module, the three matrices Q, K, and V come from the same input: the contextual information output from the previous network layer. Q, K, and V are integrated by the linear layer for feature integration to obtain a new representation of Q', K', and V', and then, Q', K', and V' are split into multiple heads, and the scaled dot-product attention function is applied to each head by a broadcast mechanism, after which the attention outputs of each head are concatenated and finally put into the linear layer to obtain the output.

In addition, the self-attention layer can be computed in parallel, which effectively improves the training efficiency. The encoder also uses a dropout layer to mitigate overfitting and improve model generalizability,[65] a layer normalization layer to expedite the convergence speed of the model,[66] and a residual skip connection to avoid vanishing gradient.

At this stage, we divide the embedding obtained above into two parts: ESM-1v, ESM-1b, and ESM-2 as one part and ProtT5-local and ProtT5-global as the other part. Then we "feed" them into the encoder to obtain vectors with dimensions of 3072 and 2048. Finally, these vectors are concatenated as inputs to the next stage predictor. For convenience of expression, we call the features fused through the encoder stage TransEPT, whose dimension is 5120.

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418

Table 2. Performance of Different Embeddings in the First Stage of Feature Fusion in Ten-Fold Cross-Validation[a]

embedding	ACC (%)	precision (%)	recall (%)	F_1 (%)	ROC-AUC	MCC
ESM-1b	68.1	71.2	32.0	44.2	0.792	0.302
ESM-1v	79.4	90.0	54.3	67.7	0.876	0.576
ESM-2	79.1	85.5	56.6	68.1	0.872	0.561
ESM-1b + ESM-1v	84.9	91.2	68.7	78.4	0.918	0.687
ESM-1b + ESM-2	84.8	90.5	68.9	78.3	0.917	0.683
ESM-1v + ESM-2	82.4	90.1	62.6	73.9	0.881	0.636
ESM-1b + ESM-1v-ESM-2	85.2	92.3	68.5	78.6	0.918	0.694
ProtT5-local	68.7	75.8	30.0	43.0	0.812	0.320
ProtT5-global	69.4	74.1	34.2	46.8	0.795	0.335
ProtT5-local + ProtT5-global	72.0	77.7	40.6	53.3	0.842	0.399
TransEPT	**86.7**	**91.8**	**73.2**	**81.4**	**0.948**	**0.724**

[a]Note: ESM-1b, ESM-1v, and ESM-2 contain 1280*2 = 2560-dimensional features. ProtT5-local and ProtT5-global contain 1024*2 = 2048-dimensional features.

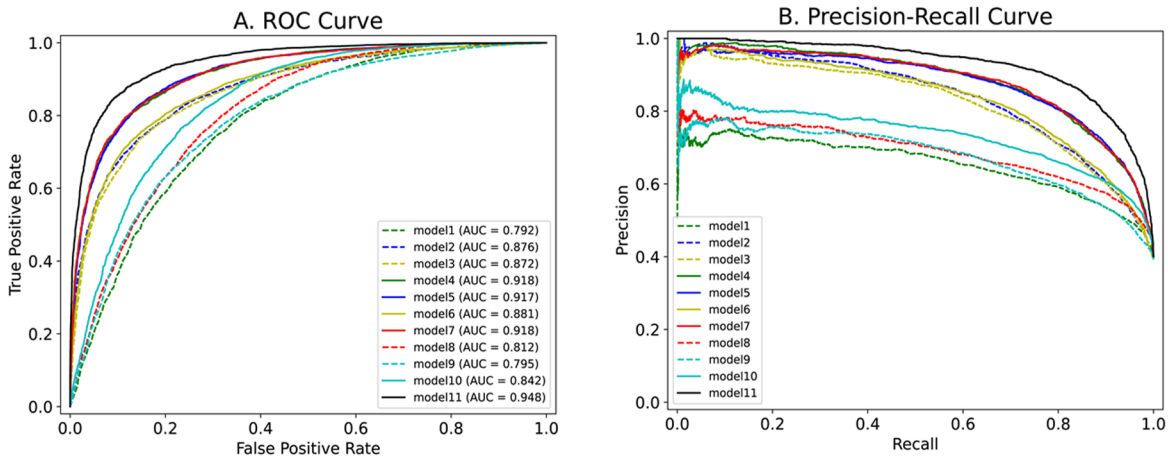


Figure 2. ROC and precision recall curves of different methods in the feature fusion part of the first training stage. Note: model1: ESM-1b; model2: ESM-1v; model3: ESM-2; model4: ESM-1b+ ESM-1v; model5: ESM-1b+ ESM-2; model6: ESM-1v+ ESM-2; model7: ESM-1b+ ESM-1v+ ESM-2; model8: ProtT5-local; model9: ProtT5-global; model10: ProtT5-local + ProtT5-global; model11: TransEPT.

2.2.3. Predictor and Output. In the second stage, for the concatenated fused features, principal component analysis (PCA) is used to reduce their dimensionality. The dimensionality-reduced features are then input into an SVM with a radial basis function (RBF) kernel for binary classification, and the output results classify variants into pathogenic and neutral. We optimized the hyperparameters of both methods by grid search, such as the parameter "n_components" of PCA, the penalty coefficient "C", and the kernel function parameter "gamma" of SVM. Through the experimental results, we finally determined that the value of parameter "n_components" of PCA most suitable for our task is 0.9, parameter "C" of SVM is 1.0, and the "gamma" is set to "scale."

These methods are implemented through the Scikit-learn library in Python. It is worth noting that the execution and parameter optimization of PCA are done in the process of cross-validation. The parameters are fixed and then projected to the vector of the test set in the reduced space.

In this work, our proposed two-stage predictor performs 10-fold cross-validation on the training set separately, and the specific process follows. The training set was divided into ten parts, and the training procedure was performed in ten randomized cycles. For each epoch, nine parts of the data set are used as the training set to train the model, while the remaining part is used to test the performance of the trained model. Then we calculated the average of ten cycles as the performance of the model.

3. RESULTS AND DISCUSSION

3.1. Performance of Different Embeddings in the First Stage. In the first stage of training, we evaluated the effects of different feature embeddings and their combinations on the training set. We took these embeddings and their combination as the input of the encoder and applied a sigmoid function to output the classification result. The feature fusion classification results for different embeddings and their combinations are listed in Table 2, and more detailed evaluation results are provided in Supplementary Table S1.

From the experimental results, individual embeddings do not exhibit promising performance, especially the global-based ProtT5 embedding, which lacks independent specificity. But when the combination of embeddings is adopted as the input of the encoder, the model performance begins to improve. Different combinations of embeddings can greatly improve the prediction results, and combinations of three or more embeddings can achieve better performance. Interestingly, when ESM-1b and global-based ProtT5 embedding, which have poor single embedding performance, are added to the fusion of multiple features, the recognition of features can still be improved. When the five embeddings are divided into two groups, the model obtains the best prediction effect where

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418

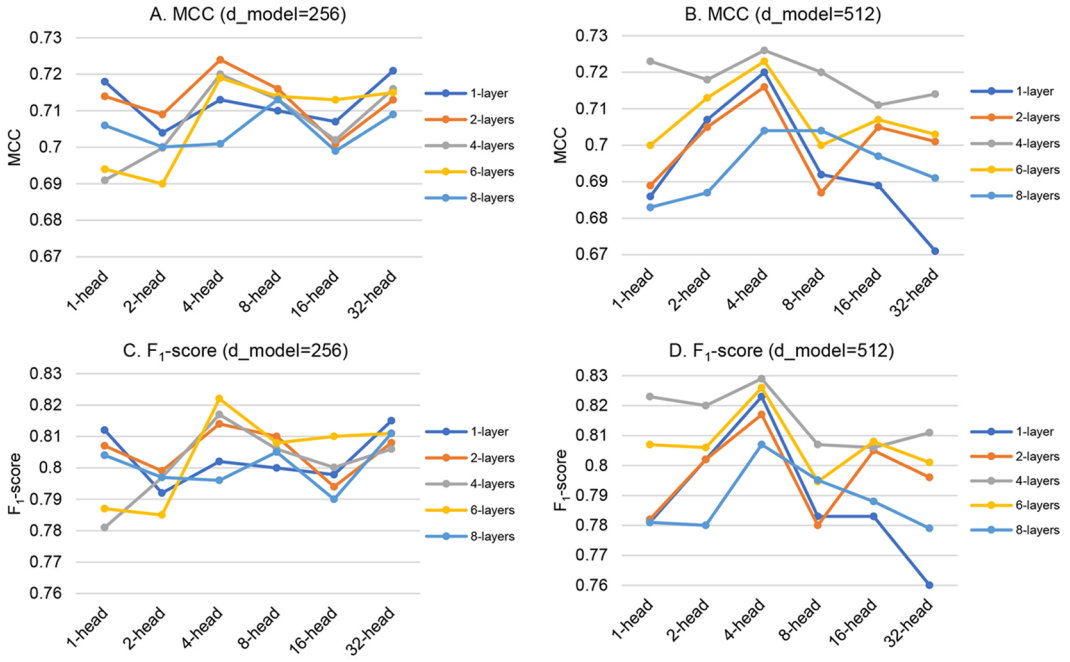


Figure 3. Model performance with different numbers of encoder layers and multiheads when num_layer is 256 and 512, respectively.

MCC reaches 0.724 and F_1-score reaches 0.814. We also plot the ROC and PR curves, as shown in Figure 2, which can reflect the performance of various embeddings more intuitively.

To clarify the mutual influence and redundancy among ESM-1b, ESM-1v, and ESM-2, we list the prediction performance on the blind test set of the models trained using these embeddings and their combinations in Supplementary Table S2. The results show that there is almost no redundancy among these three embeddings and no additional training cost is added. In addition, we also added other features and embeddings in the experiment, such as other versions of ESM-1v and ProtT5, position-specific scoring matrix (PSSM),[67,68] and features based on physicochemical properties. The results show that the addition of these features does not improve the performance of the model but increases the computational cost. We think this is because the PLM-based embedding has already learned enough structural and evolutionary information from the protein sequence, so the additional features are redundant. Therefore, we finally determined that the feature fusion of ESM-1b, ESM-1v, ESM-2 and ProtT5-local, ProtT5-global is the best input strategy for the encoder.

3.2. Performance of Different Hyperparameters in Encoder. The main architecture of the encoder consists of two parts: multihead self-attention and feed-forward neural network. Multihead self-attention includes multiple sets of *Q*, *K*, and *V* weight matrices; each weight matrix is used to project the input vector to a different representation subspace.[69] The output weight matrix becomes the input of the feed-forward neural network after compression. In order to solve the problem of vanishing gradient, the residual neural network backbone is used in the encoders. That is, the input of each feed-forward neural network not only includes the output of the above self-attention but also comprises the most original input. The parameter num_layer denotes the numbers of such identical encoder layers. The parameter d_model represents the hidden layer dimension of the model and determines the expressive power of the model. The parameter num_heads is the number of heads in the self-attention layers, i.e., the number of subspaces.

Figure 3 shows the classification performance of different numbers of encoder layers and multihead when d_model is 256 and 512, respectively, with fixed input features. The results show that when d_model is 512, and the number of encoder layers is 4, the model performance is substantially higher than in other cases. Among them, when the number of heads is 4, the effect is the most obvious. Overall, we finally determine that the values of the parameters num_layers, d_model, and num_heads are set to 4, 512, and 4, respectively.

3.3. Performance of Different Predictors in the Second Stage. In the second stage of training, we use the output features of the encoder in the previous step as the input and use the following methods to evaluate its impact on the final classification result: Encoder with the same structure as above, ResNet with three Convolution_block and Identity_block, SVM, *k*-nearest neighbors (KNN) (k = 10), decision tree (DT), RF, AdaBoost, and XGBoost.

During the experiment, we conducted multiple methods to find the most effective classification method and its parameters for our task. According to the training results, we fixed the key parameters of each predictor to the values shown in Supplementary Table S3 and performed the second cross-validation stage on the training set. The evaluation indicators obtained from the training of each method are shown in Table 3, and more calculation results and evaluation indicators are given in Supplementary Table S4. At the same time, in order to compare the prediction results of each predictor more intuitively, we plot the confusion matrix obtained from the experiment in Figure 4.

The experimental results listed in Table 3 show that the complex network structure cannot further improve the performance of the model. The addition of ResNet and Encoder did not significantly improve the prediction results, which shows that the PLM embeddings and transformer-based encoder we used in the first stage have already tapped out most

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418

Table 3. Performance of Different Predictors in the Second Stage via Ten-Fold Cross-Validation

predictor	ACC (%)	precision (%)	recall (%)	F_1 (%)	ROC-AUC	MCC
DT	86.0	83.0	81.0	82.0	0.892	0.706
ResNet	86.9	86.6	78.9	82.6	0.918	0.723
encoder	87.1	91.5	74.2	81.9	0.947	0.730
KNN	87.7	86.1	82.1	84.1	0.935	0.741
RF	87.9	85.7	83.1	84.4	0.944	0.745
AdaBoost	88.1	85.9	83.6	84.7	0.945	0.751
XGBoost	88.4	86.3	83.8	85.0	0.947	0.756
SVM	87.7	87.0	81.3	84.1	0.941	0.742

of the potential of the sequence information. And continuing to use complex deep neural networks does not further improve efficiency but only increases training time. DT, KNN, and RF these three classification algorithms that are relatively mature and have shown excellent performance in many fields have also achieved good results in our tasks but are not the best.

Both AdaBoost and XGBoost algorithms are based on the Boosting framework and have advantages in parallel computing efficiency, missing value processing, and prediction performance. They also showed excellent prediction performance in our task, especially XGBoost, which reached an MCC value of 0.756, an F_1-score value of 0.850, and an ROC-AUC value of 0.947, and was the best-performing predictor in the second stage of training. In this part of the training, SVM achieved a slightly lower score than XGBoost, with an MCC value of 0.742, an F_1-score value of 0.841, and an ROC-AUC value of 0.941. However, given the training after adding PCA and the experimental results on the blind test set, as well as the consideration of training time and cost, we finally selected SVM as the second-stage predictor, which will continue to be discussed in the next section.

3.4. Contribution of PCA to Prediction Performance.

Although our two-stage predictor has achieved good prediction performance, since the result of our first-stage feature fusion has a 5120-dimensional output, we consider whether we can use PCA as a dimensionality reduction method to remove redundant information. After experiments, we finally fixed the n_components parameter of PCA to 0.9, reduced the dimensionality of the features obtained in the first stage of training, and input them into the second stage.

The results of prediction using features reduced by PCA are listed in Table 4, the confusion matrix predicted by each predictor is plotted in Figure 5, and the ROC and PR curves are shown in Figure 6. More prediction results and experimental evaluation indicators are shown in Supplementary Table S5. Experimental results show that adding PCA has little impact on deep neural networks such as Resnet and Encoder. Likewise, several other machine learning methods were not significantly affected. However, the classification result of RF decreases after application of PCA, which may be related to the fact that random forest is more suitable for processing large-scale data sets and high-dimensional features.

Table 4. Contribution of PCA to Second-Stage Prediction Performance in Ten-Fold Cross-Validation

predictor	ACC (%)	precision (%)	recall (%)	F_1 (%)	ROC-AUC	MCC
DT	86.5	84.2	80.9	82.5	0.918	0.715
Resnet	87.0	91.6	73.7	81.7	0.920	0.728
encoder	87.4	91.0	75.3	82.4	0.947	0.735
KNN	87.5	86.2	81.4	83.7	0.932	0.737
RF	87.3	84.6	82.8	83.7	0.936	0.733
AdaBoost	88.2	85.3	84.7	85.0	0.946	0.753
XGBoost	88.3	85.9	84.0	84.9	0.947	0.753
SVM	**88.5**	**87.0**	**83.2**	**85.0**	**0.946**	**0.757**

It is worth noting that SVM is the prediction method most affected by the dimensionality reduction feature. The addition of PCA can increase the MCC of the classification result from 0.742 to 0.757 and the F_1-score from 0.841 to 0.850. Compared with XGBoost without PCA, combining SVM and

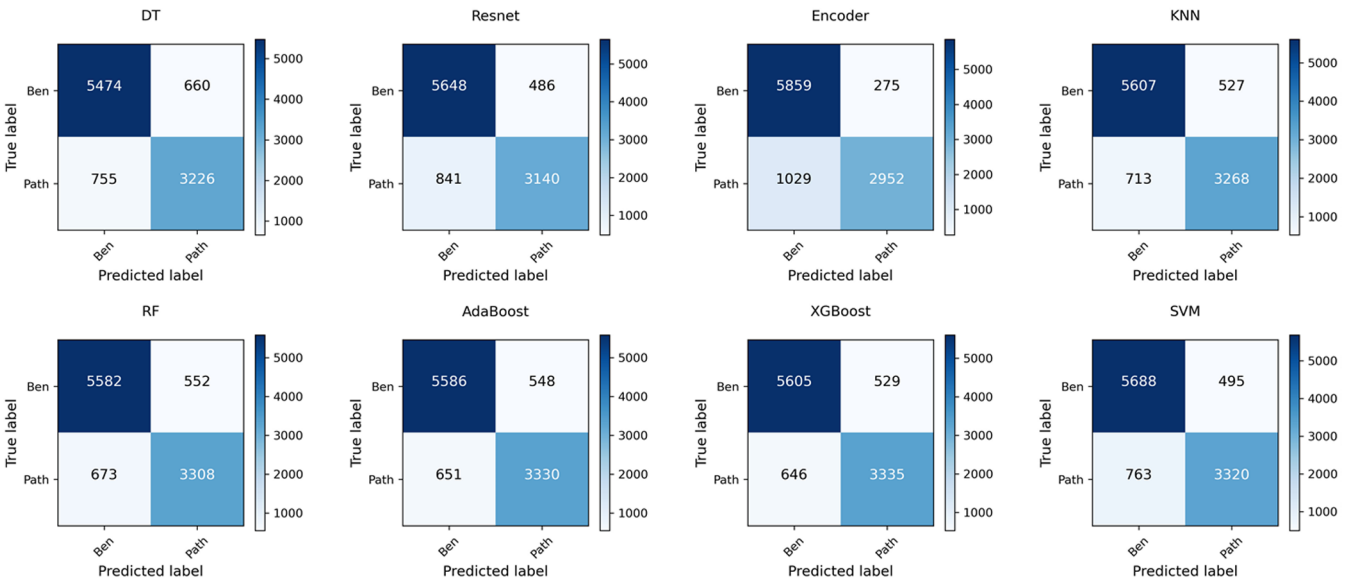


Figure 4. Confusion matrices of different predictors in the second stage via 10-fold cross-validation. In the confusion matrix, principal diagonal values represent TN and TP, while counterdiagonal values indicate FP and FN. Accordingly, the larger principal diagonal values indicate a more accurate prediction model.

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418

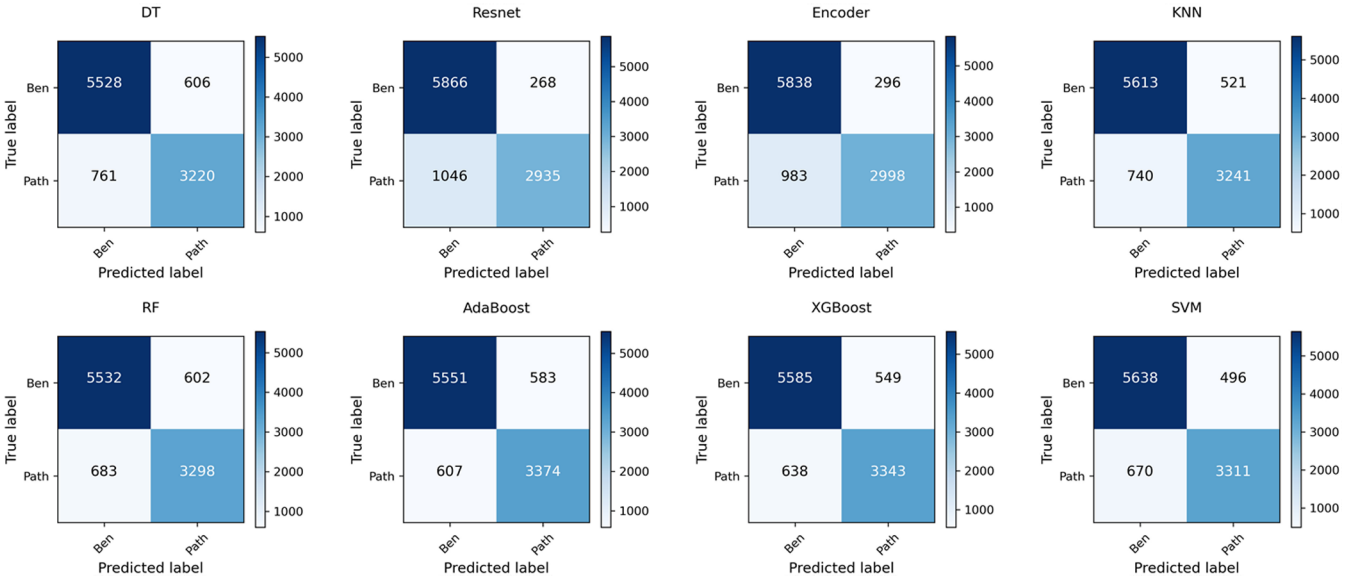


Figure 5. Confusion matrices of different predictiors after adding PCA in the second-stage via 10-fold cross-validation.

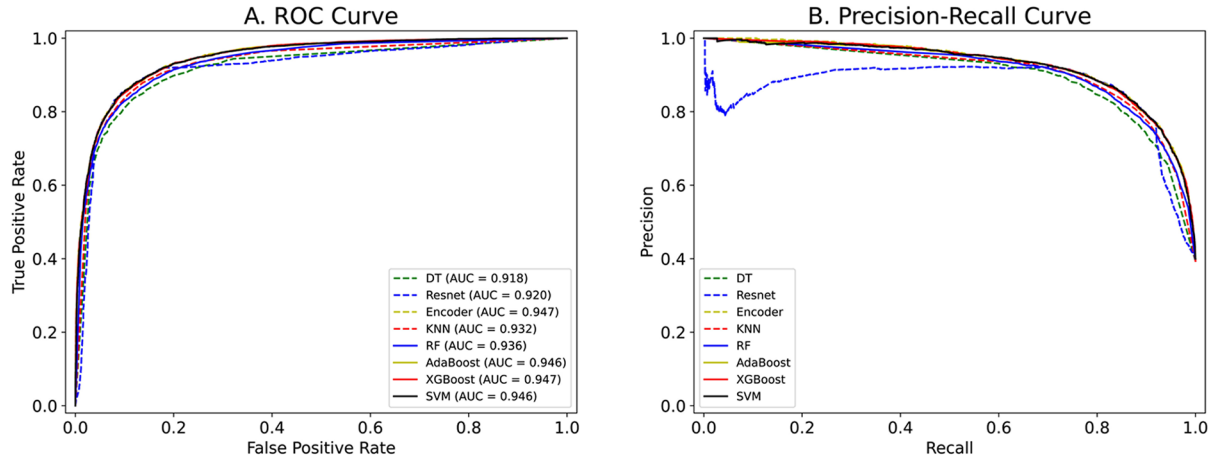


Figure 6. ROC curve of the contribution of PCA to the second-stage prediction performance.

Table 5. Performance Comparison of Our Model and Other Advanced Predictors

model	ACC(%)	precision(%)	recall(%)	F_1 (%)	ROC-AUC	MCC
TransEFVP (SVM+PCA)	**88.2**	**87.5**	**81.9**	**84.6**	**0.871**	**0.751**
TransEFVP (XGBoost)	87.9	86.3	82.7	84.5	0.870	0.746
E-SNPsPC7&GO	86.8	85.7	80.1	82.8	0.856	0.72
MutPred2.0	85.6	78.6	87.7	82.9	0.859	0.71
PROVEAN	78.2	68.7	83.0	75.2	0.790	0.57

PCA can greatly shorten the training time and reduce the training costs. Through verification on the blind test set, we finally chose SVM and PCA, which performed best for classification. Compared to the independent training in the first stage, this combination can increase the MCC of the classification result by 3.3% points (from 0.724 to 0.757) and the F_1-score by 3.6% points (from 0.814 to 0.850). We analyze that the reason why PCA works is that there may be noise and redundant information among various PLM embeddings, which may mislead the weights of the model during training, thereby affecting the classification performance.

3.5. Prediction Results on Blind Test Set. We applied the trained two-stage SAV prediction model TransEFVP to the blind test set, and the results are listed in Table 5. At the same time, we list the most advanced SAV prediction tools, including E-SNPs&GO (one of the current state-of-the-art models dedicated to SAV prediction),[57] PROVEAN,[15] and MutPred2.[32] All of the methods used the same test set for a fair performance comparison. It should be noted that in the previous sections, we used predicted probabilities as elements for calculating ROC-AUC values. In the experiments of this section, to provide a more direct comparison with existing research, we used predicted values instead of the theoretical values as the basis for calculation.

As can be seen from the comparison results, our TransEFVP model achieves the state-of-the-art. Models based on SVM and

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418

PCA achieved slightly better performance than XGBoost on the blind test set. Considering the training cost and experiment performance, TransEFVP is the better model when SVM and PCA are used as predictors in the second stage. Compared with existing methods, each evaluation metric of TransEFVP is ahead of E-SNPs&GO; the more important indicators, F_1-score and MCC, increased by 3% points and 1.8% points, respectively. At the same time, both precision and recall (sensitivity) are higher than E-SNPs&GO, which demonstrates that our model has stronger generalizability and robustness.

The possible reason for the better results than E-SNPs&GO is that TransEFVP is more effective leveraging a deep attention network model to obtain information from high dimensional features of PLM. Although E-SNPs&GO uses similar PLM embeddings, some key sequence and structural information may be missed due to the relatively single embedding perspective used. Our TransEFVP model adds PLM embeddings from more perspectives and adds a deep attention neural network to mine deeper information. Thanks to the powerful performance of transformer, feature fusion in huge dimensions will not significantly increase the computing time and cost. This is one of the reasons why our two-stage classifier is superior to the existing models. Compared with MutPred2.0 and PROVEAN using conventional methods, our model reflects the superiority of PLM, which can obtain the best performance with less computing resources and costs. In summary, TransEFVP achieves the best performance in predicting the pathogenicity of SAVs while ensuring the use of fewer resources.

In order to more intuitively demonstrate the effectiveness and robustness of TransEFVP, we use the features shared with E-SNPs&GO for verification. Since both models use the feature embeddings of ESM-1v and ProtT5, we designed experiments to only use these two features to feed into their respective model architectures. Comparison of prediction results in 10-fold cross-validation and on the blind test set is shown in Supplementary Figure S2. Experimental results show that even if only the same fewer embeddings are used, the model performance of TransEFVP is better than that of E-SNPs&GO.

Next, we continue to explore the impact of feature fusion on prediction results. Compared with E-SNPs&GO, our TransEFVP model uses more comprehensive PLM embeddings and microenvironment information and adds a powerful feature fusion module. This difference can be further explored by visualizing the features. We reduce the 5120-dimensional features processed by the feature fusion module to 3 dimensions, plot these reduced-dimensional sequences into scatter plots, and color them according to pathogenicity classification. As shown in Figure 7, the vast majority of pathogenic data are distributed in completely different areas and can be effectively separated by the naked eye. For comparison, we visualized the features in E-SNPs&GO and the features in the TransEFVP model without feature fusion module using the same method and displayed them in Supplementary Figures S3 and S4. This means the PLM embeddings and feature fusion modules that we use can effectively improve classification efficiency. Due to the powerful performance of PLM embeddings and feature fusion module, we can use a simpler and more efficient classifier to complete the prediction model. Ultimately, high-precision prediction of SAVs pathogenicity is achieved using fewer resources and time.

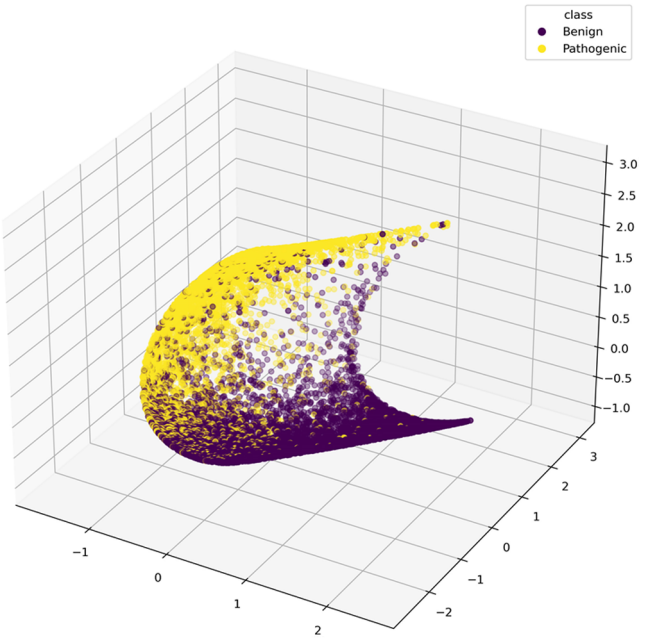


Figure 7. Dimensionality reduction and visualization. Reduce the 5120-dimensional features to 3 dimensions and visualize the 3D graphics as a scatter plot. The yellow points represent pathogenic data, and the purple points represent neutral data.

Additionally, we tested the TransEFVP model on the US data set containing 9165 SAVs, and the predicted results will be shown in Supplementary Figure S5. Since these SAVs are meaningless variants without labels, our prediction results can only be used as a reference. We hope that our model can be of greater help in more research.

4. CONCLUSIONS

Predicting the disease association of SAV is one of the important steps in understanding disease mechanisms. In the existing SAV pathogenicity prediction methods, either a single feature expression form is used or the effective information contained in the feature embedding is not fully mined. In this study, we developed a novel two-stage prediction model based on PLM embeddings to improve the performance of SAV pathogenicity assessment to make up for the above shortcomings. The first stage of the model extracts five embeddings based on the ESM and ProtT5 project from wild-type and variant protein sequences as input of the encoder. After the feature fusion through the transformer encoder with multihead attention mechanism, the fused embeddings are fed into the classifier of the next stage. In the second stage, these features are subjected to PCA dimensionality reduction processing and then input into the SVM classifier and the prediction results. Our model TransEFVP achieves excellent results on the training set and blind test set and outperforms existing mainstream prediction methods. We expect TransEFVP to be a powerful predictive tool, providing evidence and assistance for human SAVs' pathogenicity prediction.

In addition, even if our initial feature embedding dimension reaches more than ten thousand, thanks to the powerful fusion capability of the encoder, it will not cost too much computing resources and time. Without additional feature proof, the PLM embeddings have learned meaningful representations of protein sequences and can be applied to various tasks and

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418

scenarios. We believe that the great potential in PLM embeddings can still be used to contribute to human disease research. On the contrary, the exclusive use of PLM embeddings results in a lack of biological interpretability of predictive model. Adding more protein structure information to increase biological interpretability without increasing training costs is the direction of our next efforts.

■ ASSOCIATED CONTENT

Data Availability Statement

TransEFVP is released under an open source license at https://github.com/yzh9607/TransEFVP/tree/master and contains comprehensive data and codes in this work.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c02019.

It contains additional experimental parameters, results, and comparisons of TransEFVP. (a) A brief description of the model performance evaluation metrics and all predictors used in the experiments, (b) additional evaluation metrics obtained from cross-validation and independent validation and their comparison, (c) comparison of prediction results of ESM-1b, ESM-1v, and ESM-2 embeddings on the blind test set, (d) key parameters of all predictors involved in the experiment, and (e) prediction results of TransEFVP model for VUS data set (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Jiangning Song − *Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Victoria 3800, Australia;* orcid.org/0000-0001-8031-9086; Email: jiangning.song@monash.edu

Dong-Jun Yu − *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, PR China;* orcid.org/0000-0002-6786-8053; Email: njyudj@njust.edu.cn

Authors

Zihao Yan − *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, PR China*

Fang Ge − *State Key Laboratory of Organic Electronics and Information Displays & Institute of Advanced Materials (IAM), Nanjing University of Posts & Telecommunications, Nanjing 210023, PR China*

Yan Liu − *Department of Computer Science, Yangzhou University, Yangzhou 225100, PR China*

Yumeng Zhang − *School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, PR China; Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Victoria 3800, Australia*

Fuyi Li − *South Australian immunoGENomics Cancer Institute (SAiGENCI), Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia; The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, Victoria 3000, Australia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c02019

Author Contributions

○Z.H.Y. and F.G. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (62372234 and 62072243), the Natural Science Foundation of Jiangsu (BK20201304), and the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY223062).

■ REFERENCES

(1) Sanchez-Mazas, A. A Review of HLA Allele and SNP Associations with Highly Prevalent Infectious Diseases in Human Populations. *Swiss Med. Wkly.* **2020**, *150*, w20214.

(2) Frazer, J.; Notin, P.; Dias, M.; Gomez, A.; Min, J. K.; Brock, K.; Gal, Y.; Marks, D. S. Disease Variant Prediction with Deep Generative Models of Evolutionary Data. *Nature* **2021**, *599* (7883), 91−95.

(3) Adhikari, P.; Jawad, B.; Rao, P.; Podgornik, R.; Ching, W.-Y. Delta Variant with P681r Critical Mutation Revealed by Ultra-Large Atomic-Scale Ab Initio Simulation: Implications for the Fundamentals of Biomolecular Interactions. *Viruses* **2022**, *14*, 465.

(4) Jaing, T.-H.; Chang, T.-Y.; Chen, S.-H.; Lin, C.-W.; Wen, Y.-C.; Chiu, C.-C. Molecular Genetics of B-Thalassemia: A Narrative Review. *Medicine* **2021**, *100* (45), No. e27522.

(5) Manahan, E. R.; Kuerer, H. M.; Sebastian, M.; Hughes, K. S.; Boughey, J. C.; Euhus, D. M.; Boolbol, S. K.; Taylor, W. A. Consensus Guidelines on Genetic Testing for Hereditary Breast Cancer from the American Society of Breast Surgeons. *Ann. Surg. Oncol.* **2019**, *26*, 3025−3031.

(6) Singh, T.; Poterba, T.; Curtis, D.; Akil, H.; Al Eissa, M.; Barchas, J. D.; Bass, N.; Bigdeli, T. B.; Breen, G.; Bromet, E. J.; et al.et al. Rare Coding Variants in Ten Genes Confer Substantial Risk for Schizophrenia. *Nature* **2022**, *604*, 509−516.

(7) Nie, C.; Sahoo, A. K.; Netz, R. R.; Herrmann, A.; Ballauff, M.; Haag, R. Charge Matters: Mutations in Omicron Variant Favor Binding to Cells. *ChemBioChem* **2022**, *23*, No. e202100681.

(8) Narasimhan, V. M.; Rahbari, R.; Scally, A.; Wuster, A.; Mason, D.; Xue, Y.; Wright, J.; Trembath, R. C.; Maher, E. R.; Van Heel, D. A.; et al.et al. Estimating the Human Mutation Rate from Autozygous Segments Reveals Population Differences in Human Mutational Processes. *Nat. Commun.* **2017**, *8*, 303.

(9) Levitsky, L. I.; Kuznetsova, K. G.; Kliuchnikova, A. A.; Ilina, I. Y.; Goncharov, A. O.; Lobas, A. A.; Ivanov, M. V.; Lazarev, V. N.; Ziganshin, R. H.; Gorshkov, M. V.; Moshkovskii, S. A. Validating Amino Acid Variants in Proteogenomics Using Sequence Coverage by Multiple Reads. *J. Proteome Res.* **2022**, *21* (6), 1438−1448.

(10) Brinkerhoff, H.; Kang, A. S. W.; Liu, J.; Aksimentiev, A.; Dekker, C. Multiple rereads of single proteins at single−amino acid resolution using nanopores. *Science* **2021**, *374* (6574), 1509−1513.

(11) Mahmud, M.; Kaiser, M. S.; McGinnity, T. M.; Hussain, A. Deep Learning in Mining Biological Data. *Cognit. Comput.* **2021**, *13*, 1−33.

(12) Bao, W.; Gu, Y.; Chen, B.; Yu, H. Golgi_DF: Golgi Proteins Classification with Deep Forest. *Front. Neurosci.* **2023**, *17*, 1197824.

(13) Bao, W.; Cui, Q.; Chen, B.; Yang, B.; Wei, L. Phage_UniR_LGBM: Phage Virion Proteins Classification with Unirep Features and LightGBM Model. *Comput. Math. Methods Med.* **2022**, *2022*, 1−8.

(14) Ng, P. C.; Henikoff, S. Predicting Deleterious Amino Acid Substitutions. *Genome Res.* **2001**, *11*, 863−874.

(15) Choi, Y.; Sims, G. E.; Murphy, S.; Miller, J. R.; Chan, A. P.; de Brevern, A. G. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PloS One* **2012**, *7*, No. e46688.
(16) Calabrese, R.; Capriotti, E.; Fariselli, P.; Martelli, P. L.; Casadio, R. Functional Annotations Improve the Predictive Score of Human Disease-Related Mutations in Proteins. *Hum. Mutat.* **2009**, *30*, 1237−1244.
(17) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al.et al. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000**, *25* (1), 25−29.
(18) Yates, C. M.; Filippis, I.; Kelley, L. A.; Sternberg, M. J. E. SuSPect: Enhanced Prediction of Single Amino Acid Variant (Sav) Phenotype Using Network Features. *J. Mol. Biol.* **2014**, *426*, 2692−2701.
(19) Quan, L.; Lv, Q.; Zhang, Y. STRUM: Structure-Based Prediction of Protein Stability Changes Upon Single-Point Mutation. *Bioinformatics* **2016**, *32* (19), 2936−2946.
(20) Iqbal, S.; Li, F.; Akutsu, T.; Ascher, D. B.; Webb, G. I.; Song, J. Assessing the Performance of Computational Predictors for Estimating Protein Stability Changes Upon Missense Mutations. *Briefings Bioinf.* **2021**, *22* (6), bbab184.
(21) Khan, S.; Vihinen, M. Performance of Protein Stability Predictors. *Hum. Mutat.* **2010**, *31* (6), 675−684.
(22) Kircher, M.; Witten, D. M.; Jain, P.; O'roak, B. J.; Cooper, G. M.; Shendure, J. A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants. *Nat. Genet.* **2014**, *46* (3), 310−315.
(23) Li, B.; Krishnan, V. G.; Mort, M. E.; Xin, F.; Kamati, K. K.; Cooper, D. N.; Mooney, S. D.; Radivojac, P. Automated Inference of Molecular Mechanisms of Disease from Amino Acid Substitutions. *Bioinformatics* **2009**, *25* (21), 2744−2750.
(24) Carter, H.; Douville, C.; Stenson, P. D.; Cooper, D. N.; Karchin, R. Identifying Mendelian Disease Genes with the Variant Effect Scoring Tool. *BMC Genomics* **2013**, *14* (S3), 1−16.
(25) Niroula, A.; Urolagin, S.; Vihinen, M.; Tosatto, S. C. E. PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. *PloS One* **2015**, *10* (2), No. e0117380.
(26) Raimondi, D.; Tanyalcin, I.; Ferté, J.; Gazzo, A.; Orlando, G.; Lenaerts, T.; Rooman, M.; Vranken, W. DEOGEN2: Prediction and Interactive Visualization of Single Amino Acid Variant Deleteriousness in Human Proteins. *Nucleic Acids Res.* **2017**, *45* (W1), W201−W206.
(27) Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; Sunyaev, S. R. A Method and Server for Predicting Damaging Missense Mutations. *Nat. Methods* **2010**, *7* (4), 248−249.
(28) Schwarz, J. M.; Rödelsperger, C.; Schuelke, M.; Seelow, D. Mutationtaster Evaluates Disease-Causing Potential of Sequence Alterations. *Nat. Methods* **2010**, *7* (8), 575−576.
(29) Jagadeesh, K. A.; Wenger, A. M.; Berger, M. J.; Guturu, H.; Stenson, P. D.; Cooper, D. N.; Bernstein, J. A.; Bejerano, G. M-CAP Eliminates a Majority of Variants of Uncertain Significance in Clinical Exomes at High Sensitivity. *Nat. Genet.* **2016**, *48* (12), 1581−1586.
(30) Liu, J.-J.; Yu, C.-S.; Wu, H.-W.; Chang, Y.-J.; Lin, C.-P.; Lu, C.-H. The Structure-Based Cancer-Related Single Amino Acid Variation Prediction. *Sci. Rep.* **2021**, *11* (1), 13599.
(31) Talukder, A.; Barham, C.; Li, X.; Hu, H. Interpretation of Deep Learning in Genomics and Epigenomics. *Briefings Bioinf.* **2021**, *22* (3), bbaa177.
(32) Pejaver, V.; Urresti, J.; Lugo-Martinez, J.; Pagel, K. A.; Lin, G. N.; Nam, H.-J.; Mort, M.; Cooper, D. N.; Sebat, J.; Iakoucheva, L. M.; Mooney, S. D.; Radivojac, P. Inferring the Molecular and Phenotypic Impact of Amino Acid Variants with Mutpred2. *Nat. Commun.* **2020**, *11* (1), 5918.
(33) Pei, J.; Kinch, L. N.; Otwinowski, Z.; Grishin, N. V.; Dokholyan, N. V. Mutation Severity Spectrum of Rare Alleles in the Human Genome Is Predictive of Disease Type. *PloS Comput. Biol.* **2020**, *16* (5), No. e1007775.
(34) Pei, J.; Grishin, N. V. The DBSAV Database: Predicting Deleteriousness of Single Amino Acid Variations in the Human Proteome. *J. Mol. Biol.* **2021**, *433*, 166915.
(35) Kulandaisamy, A.; Zaucha, J.; Sakthivel, R.; Frishman, D.; Michael Gromiha, M. Pred-MutHTP: Prediction of Disease-Causing and Neutral Mutations in Human Transmembrane Proteins. *Hum. Mutat.* **2020**, *41* (3), 581−590.
(36) Pires, D. E. V.; Rodrigues, C. H. M.; Ascher, D. B. mCSM-Membrane: Predicting the Effects of Mutations on Transmembrane Proteins. *Nucleic Acids Res.* **2020**, *48* (W1), W147−W153.
(37) Ge, F.; Zhu, Y.-H.; Xu, J.; Muhammad, A.; Song, J.; Yu, D.-J. MutTMPredictor: Robust and Accurate Cascade Xgboost Classifier for Prediction of Mutations in Transmembrane Proteins. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 6400−6416.
(38) Landrum, M. J.; Lee, J. M.; Benson, M.; Brown, G. R.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Jang, W.; et al.et al. ClinVar: Improving Access to Variant Interpretations and Supporting Evidence. *Nucleic Acids Res.* **2018**, *46* (D1), D1062−D1067.
(39) Consortium, T. U. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523−D531.
(40) Nair, P. S.; Vihinen, M. VariBench: A Benchmark Database for Variations. *Hum. Mutat.* **2013**, *34* (1), 42−49.
(41) Rao, R.; Meier, J.; Sercu, T.; Ovchinnikov, S.; Rives, A. Transformer Protein Language Models Are Unsupervised Structure Learners; Biorxiv, **2020**. .
(42) Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, P.; Canny, J.; Abbeel, P.; Song, Y. Evaluating Protein Transfer Learning with Tape. *Adv. Neural Inf. Process. Syst.* **2019**, *3296899701*.
(43) Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29287−29303.
(44) Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L., Jr.; Xiong, C.; Sun, Z. Z.; Socher, R.; Fraser, J. S.; Naik, N. Large Language Models Generate Functional Protein Sequences across Diverse Families. *Nat. Biotechnol.* **2023**, *41* (8), 1099−1106.
(45) Biswas, S.; Khimulya, G.; Alley, E. C.; Esvelt, K. M.; Church, G. M. Low-N Protein Engineering with Data-Efficient Deep Learning. *Nat. Methods* **2021**, *18* (4), 389−396.
(46) Littmann, M.; Heinzinger, M.; Dallago, C.; Weissenow, K.; Rost, B. Protein Embeddings and Deep Learning Predict Binding Residues for Various Ligand Classes. *Sci. Rep.* **2021**, *11* (1), 23916.
(47) Bepler, T.; Berger, B. Learning Protein Sequence Embeddings Using Information from Structure; arXiv preprint arXiv:1902.08661, 2019. DOI: .
(48) Lu, A. X.; Zhang, H.; Ghassemi, M.; Moses, A. Self-Supervised Contrastive Learning of Protein Representations by Mutual Information Maximization; BioRxiv, 2020. DOI: .
(49) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118* (15), No. e2016239118.
(50) UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480−D489.
(51) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M. ProtTrans: Toward Understanding the Language of Life through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7112−7127.
(52) Bryant, P.; Pozzati, G.; Elofsson, A. Improved Prediction of Protein-Protein Interactions Using AlphaFold2. *Nat. Commun.* **2022**, *13* (1), 1265.
(53) Rao, R. M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; Rives, A. MSA Transformer. In *International Conference on Machine Learning*; PMLR, 2021.
(54) Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L., Jr.; Xiong, C.; Sun, Z. Z.; Socher, R.;

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418

Fraser, J. S.; Nikhil, N. Deep Neural Language Modeling Enables Functional Protein Generation across Families; *BioRxiv*, **2021**. DOI: .
(55) Fan, X.; Pan, H.; Tian, A.; Chung, W. K.; Shen, Y. SHINE: Protein Language Model-Based Pathogenicity Prediction for Short Inframe Insertion and Deletion Variants. *Briefings Bioinf.* **2023**, *24* (1), bbac584.
(56) Marquet, C.; Heinzinger, M.; Olenyi, T.; Dallago, C.; Erckert, K.; Bernhofer, M.; Nechaev, D.; Rost, B. Embeddings from Protein Language Models Predict Conservation and Variant Effects. *Hum. Genet.* **2022**, *141* (10), 1629−1647.
(57) Manfredi, M.; Savojardo, C.; Martelli, P. L.; Casadio, R.; Boeva, V. E-SNPs&GO: Embedding of protein sequence and function improves the annotation of human pathogenic variants. *Bioinformatics* **2022**, *38* (23), 5168−5174.
(58) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al.et al. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123−1130.
(59) Ge, F.; Zhang, Y.; Xu, J.; Muhammad, A.; Song, J.; Yu, D.-J. Prediction of Disease-Associated Nssnps by Integrating Multi-Scale Resnet Models with Deep Feature Fusion. *Briefings Bioinf.* **2022**, *23* (1), bbab530.
(60) Amberger, J. S.; Bocchini, C. A.; Scott, A. F.; Hamosh, A. OMIM.org: Leveraging knowledge across phenotype−gene relationships. *Nucleic Acids Res.* **2019**, *47* (D1), D1038−D1043.
(61) Shefchek, K. A.; Harris, N. L.; Gargano, M.; Matentzoglu, N.; Unni, D.; Brush, M.; Keith, D.; Conlin, T.; Vasilevsky, N.; Zhang, X. A.; Balhoff, J. P.; Babb, L.; Bello, S. M.; Blau, H.; Bradford, Y.; Carbon, S.; Carmody, L.; Chan, L. E.; Cipriani, V.; Cuzick, A.; Della Rocca, M.; Dunn, N.; Essaid, S.; Fey, P.; Grove, C.; Gourdine, J.-P.; Hamosh, A.; Harris, M.; Helbig, I.; Hoatlin, M. The Monarch Initiative in 2019: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species. *Nucleic Acids Res.* **2020**, *48* (D1), D704−D715.
(62) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35* (11), 1026−1028.
(63) Nie, L.; Quan, L.; Wu, T.; He, R.; Lyu, Q.; Martelli, P. L. TransPPMP: Predicting Pathogenicity of Frameshift and Non-Sense Mutations by a Transformer Based on Protein Features. *Bioinformatics* **2022**, *38* (10), 2705−2711.
(64) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding; arXiv preprint arXiv:1810.04805, 2018. DOI: .
(65) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.
(66) Ba, J. L.; Kiros, J. R.; Hinton, G. E. Layer Normalization; arXiv preprint arXiv:1607.06450, 2016. DOI: .
(67) Yu, D.-J.; Hu, J.; Yan, H.; Yang, X.-B.; Yang, J.-Y.; Shen, H.-B. Enhancing Protein-Vitamin Binding Residues Prediction by Multiple Heterogeneous Subspace Svms Ensemble. *BMC Bioinf.* **2014**, *15* (1), 1−14.
(68) Ge, F.; Muhammad, A.; Yu, D.-J. DeepnsSNPs: Accurate Prediction of Non-Synonymous Single-Nucleotide Polymorphisms by Combining Multi-Scale Convolutional Neural Network and Residue Environment Information. *Chemom. Intell. Lab. Syst.* **2021**, *215*, 104326.
(69) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need; Proceedings of the 31st International Conference on Neural Information Processing System, 2017.

https://doi.org/10.1021/acs.jcim.3c02019
J. Chem. Inf. Model. 2024, 64, 1407−1418