Methods 235 (2025) 55–63

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

ZFP-CanPred: Predicting the effect of mutations in zinc-finger proteins in cancers using protein language models☆

Amit Phogat [a], Sowmya Ramaswamy Krishnan [a], Medha Pandey [a], M. Michael Gromiha [a,b,*]

[a] Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036 India
[b] International Research Frontiers Initiative, School of Computing, Tokyo Institute of Technology, Yokohama 226-8501 Japan

ARTICLE INFO

Keywords:
ZFP-CanPred
Zinc-fingers
Cancer
Driver
Neutral
Mutations
Neural network
Protein language model

ABSTRACT

Zinc-finger proteins (ZNFs) constitute the largest family of transcription factors and play crucial roles in various cellular processes. Missense mutations in ZNFs significantly alter protein-DNA interactions, potentially leading to the development of various types of cancers. This study presents ZFP-CanPred, a novel deep learning-based model for predicting cancer-associated driver mutations in ZNFs. The representations derived from protein language models (PLMs) from the structural neighbourhood of mutated sites were utilized to train ZFP-CanPred for differentiating between cancer-causing and neutral mutations. ZFP-CanPred, achieved a superior performance with an accuracy of 0.72, F1-score of 0.79, and area under the Receiver Operating Characteristics (ROC) Curve (AUC) of 0.74, on an independent test set. In a comparative analysis against 11 existing prediction tools using a curated dataset of 331 mutations, ZFP-CanPred demonstrated the highest AU-ROC of 0.74, outperforming both generic and cancer-specific methods. The model's balanced performance across specificity and sensitivity addresses a significant limitation of current methodologies. The source code and other related files are available on GitHub at https://github.com/amitphogat/ZFP-CanPred.git. We envisage that the present study contributes to understand the oncogenic processes and developing targeted therapeutic strategies.

1. Introduction

Protein-DNA interactions play an important role in various cellular processes, such as gene expression, regulation, methylation, DNA replication and repair. These interactions involve specific recognition of DNA sequences by DNA-binding proteins (DBPs) [1]. The DNA-binding proteins interact with DNA using specialized domains known as DNA-binding domains with primarily two types of interactions: direct or water-mediated hydrogen bonds and van der Waals interactions with the major groove of the DNA double helix. Additionally, they recognize specific DNA sequences known as motifs. There are various types of motifs involved in DNA binding, including helix-turn-helix (HTH), zinc-finger (ZF), leucine-zipper, helix-loop-helix (HLH), and high mobility group (HMG) [2]. Zinc-finger containing proteins (ZNFs) belong to the largest transcription factor (TFs) family. The transcription factors recognize and bind to DNA sequences for regulating transcription of many genes. The zinc-finger domain typically comprises conserved cysteine and histidine residues that coordinate a Zn^{2+} ion. This coordination forms a stable, finger-like structure through a combination of alpha-helical and beta-sheet folding patterns. Upon binding to its target site, the zinc-finger domain aligns three base pairs of DNA with specific amino acids in the α-helix structure. The amino acid composition at the contact site determines the DNA sequence recognition specificity of zinc-fingers [3]. Missense mutations in ZNFs alter the protein structure and conformation at the protein-DNA interface, affecting the expression of many genes and leading to diseases such as cancer.

ZNFs play significant roles in the development of various types of cancer. Specific ZNFs have been identified as oncogenes or tumor suppressors in different malignancies. For instance, ZNF322A and ZNF251 have been characterized as oncogenes that promote lung carcinogenesis. In breast cancer, ZNF711, ZNF143, and ZNF224 have been implicated in tumor progression. Hepatocellular carcinoma development has been associated with alterations in ZHX1 and ZHX2 expression. Furthermore, ZNF479 and ZNF281 have been linked to gastric cancer, while ZNF350 and ZNF703 have been shown to contribute to colorectal cancer development [4]. Munro et al., [5] analyzed the transcription factors

☆ This article is part of a special issue entitled: 'Natural language processing - YMETH' published in Methods.

* Corresponding author at: Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, India.
E-mail address: gromiha@iitm.ac.in (M.M. Gromiha).

https://doi.org/10.1016/j.ymeth.2025.01.020
Received 13 July 2024; Received in revised form 21 January 2025; Accepted 27 January 2025
Available online 3 February 2025

containing Cys2His2 type zinc-finger domain and showed that arginine at the 9th position responsible for dimer formation, is frequently mutated to isoleucine, leading to the development of uterine and colorectal carcinomas, as well as histidine at the 11th position coordinating with Zn^{2+} is frequently mutated to tyrosine in multiple melanomas. The CCCTC-binding factor (CTCF) is a protein with 11 zinc-finger domains responsible for gene regulation and development of various types of cancers. Several mutations in zinc-fingers 3 to 7 of CTCF including, R339Q, S354T, Q418R, R448Q, R377C, and R377H, abolish the hydrogen-bonding and electrostatic interactions at the interface, resulting in an unstable protein-DNA complex and affecting downstream gene regulation [6]. These findings underscore diverse effects of mutations in ZNFs, and highlight their potential as diagnostic biomarkers or therapeutic targets to treat cancers. Given the effects of various missense mutations and critical functions of ZNFs in cancer development, it is essential to identify cancer-causing mutations in these proteins to understand the molecular basis of diseases and the development of therapeutic strategies. Identifying cancer-causing mutations experimentally is a time-consuming and labor-intensive task [7,8]. Hence, computational methods need to be developed for predicting the effect of mutations in cancer development, to accelerate target identification. With emerging efforts to develop drugs against specific mutant proteins in cancers such as p53 [9], KRAS [10], and IDH1 [11], predictive models to identify cancer-causing mutations in cancer-associated proteins can significantly contribute toward drug development.

Numerous computational methods or tools have been developed to predict the impact of mutations in diseases. These tools utilize diverse features, including conservation metrics, spatial localization of mutations within protein structures, physicochemical properties, multiple sequence alignment, and various other structural parameters for model development. Some methods integrate multiple prediction scores from other tools, leveraging ensemble techniques to enhance predictive accuracy [12]. These methods can be categorized into three main groups: rule-based algorithms, machine learning-based methods, and deep learning-based methods. Rule-based algorithms include SIFT4G [13], FATHMM [14], MutationAssessor [15], and PROVEAN [16]. Machine learning-based methods comprise MutationTaster [17], PolyPhen-2 [18], MetaSVM [19], MetaLR [19], DEOGEN2 [20], and M−CAP [21]. Deep learning-based methods consist of MVP [22], PrimateAI [23], AlphaMissense [24], and ESM1b [25]. Buel and Walters [26] reported that AlphaFold2.0 is not capable of efficiently predicting the impact of missense mutations in protein three-dimensional structures. In addition, various cancer-specific tools have also been developed for predicting the effect of cancer specific mutations [27–30]. ZNFs consist of multiple zinc-finger domains that function synergistically to achieve sequence-specific DNA recognition. Limited availability of mutational data for ZNFs as compared to other DNA-binding protein families has led to poor performance of existing methods on ZNFs [3].

In this work, a model was trained on the dataset containing mutations specific to ZNFs. The resultant model, ZFP-CanPred, is a deep learning-based method for predicting the effect of mutations in cancer-associated zinc-finger proteins, using the representations from protein language models (PLMs) such as Evolutionary Scale Modeling (ESM-2) [31]. By leveraging the latent representations from the state-of-the-art LLMs, the model inherits a notion of similarity between different ZNFs. This is essential in aiding the model to distinguish between different mutants at the residue-level and predict their effect in cancer development. ZFP-CanPred achieved an accuracy of 0.72 and AU-ROC of 0.74 on a test dataset. In comparative analysis against existing computational tools, ZFP-CanPred showed an improved performance when evaluated on a standardized test dataset. We anticipate that this method will aid in developing targeted therapeutic strategies against mutant ZNFs in cancers.

2. Materials and methods

2.1. Dataset

The data available on cancer-causing mutations was collected from the COSMIC database v97 [32]. The frequency of each mutation was computed in the dataset and mutations observed at least three times in different samples were considered as driver mutations. Mutations specific to DNA-binding proteins were collected from UniProt [33] and mapped to disease-associated mutations from the COSMIC dataset. To remove redundancy from the dataset, the proteins were first clustered based on sequence identity (similarity cut-off of >20%) using CD-HIT [34]. The representative proteins obtained after clustering were taken and the corresponding mutations were used to create the final dataset. The DNA-binding domain information for all the proteins was taken from InterPro [35]. The neutral mutations for these proteins were extracted from Clinvar [36], dbSNP [37], HuVarBase [38] and dbCPM [39] databases. Mutations annotated as benign or likely benign, which do not alter protein function or contribute to disease pathogenesis, were classified as neutral mutations. We cross-referenced the neutral mutations against the driver mutations, excluding any overlapping variants between the two datasets. The final dataset was then stratified into training (80%), validation (10%), and test (10%) subsets using a protein-level splitting approach that prevents data leakage and maintains the independence of each subset, ensuring no proteins were shared across these datasets.

2.2. Feature extraction

2.2.1. Structural neighbor extraction

The protein structures were extracted from the AlphaFold [40] database, as the complete experimental structures were not available. The mutant structures were generated using the latest version of FoldX (v5.0) [41]. The structural neighbors of driver and neutral mutations were extracted using a distance cut-off of 8 Å, to include short-, medium-, and long-range interactions [42]. This dataset is termed as structural neighbors for driver and neutral mutations in this study.

2.2.2. Extracting representations

The protein language models (PLMs) were used to extract the features from wild-type and mutant structural neighbors. The model's output is a multi-dimensional vector reflecting biochemical properties and remote similarities between proteins. The following language models were tested for feature extraction:

a) ESM representation

ESM-2 [31] is a protein language model trained on 250 million protein sequences using unsupervised learning. The model takes a protein sequence as input, and provides a 1280-dimensional vector representation as an output. The learned representation space spans multiple scales, encompassing biochemical characteristics of amino acids and remote similarities between proteins.

b) ProteinBERT representations

ProteinBERT [43] was pre-trained using approximately 106 million protein sequences along with associated Gene Ontology (GO) annotations sourced from UniProtKB/UniRef90 [33,44]. The protein sequences are inputs for the model. The output of the model is a 1024-dimensional numerical vector.

c) ProtTrans representations

ProtTrans [45] is a collection of protein language models (Transformer-XL, XLNet, BERT, Albert, Electra, and T5) trained on up to 393

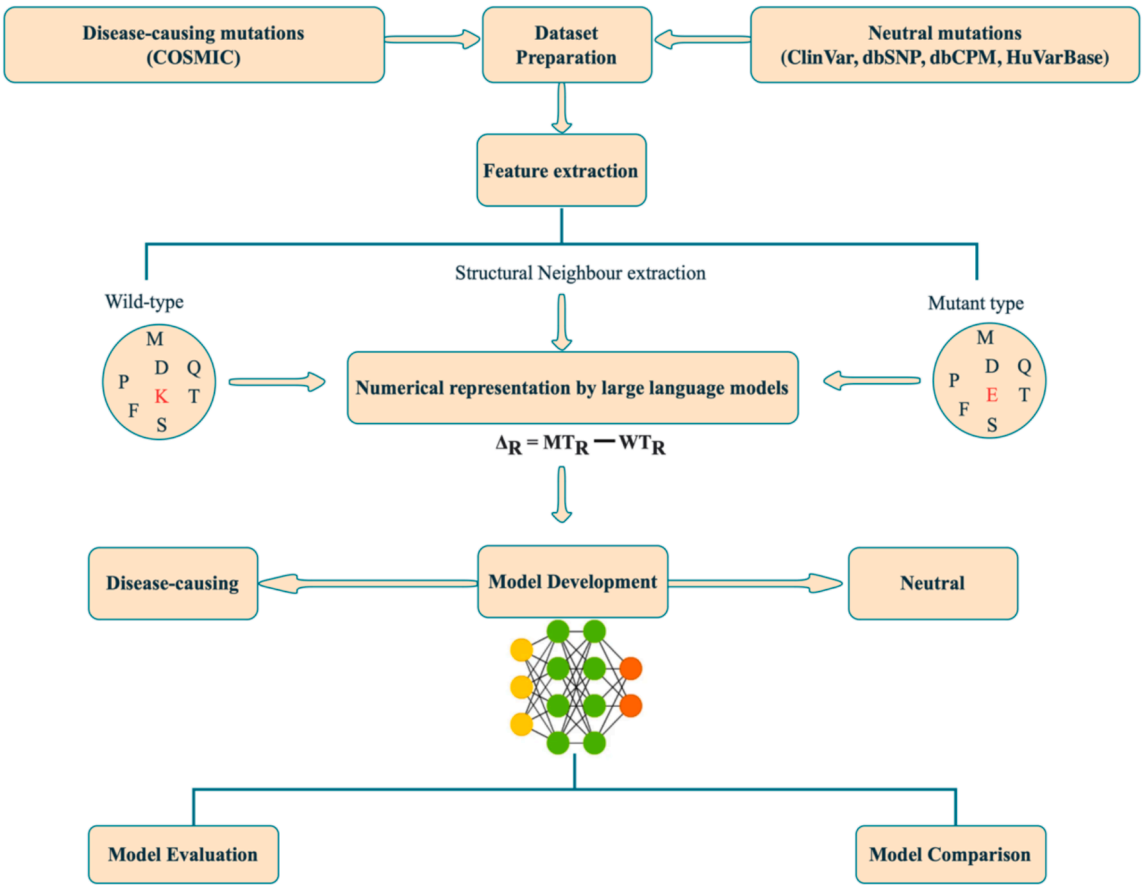

Fig. 1. Workflow of the development of ZFP-CanPred model.

billion amino acids from UniRef and BFD datasets. It generates vector representations of 1024 dimensions, which capture physicochemical properties.

d) ProtFlash representations

ProtFlash [46] is a lightweight protein language model with linear complexity. The model's architecture combines local and global context processing using chunk-based attention patterns and multiple positional encoding schemes. The model generates embedding vectors of 728 dimensions that encode semantic information from protein sequence.

2.3. Model development

The complete workflow for ZFP-CanPred development is provided in Fig. 1. The representations are calculated for wild-type (WT_R) and mutant structural neighbours (MT_R). The difference between these representations (Δ_R) is used as the input feature for the deep neural network, which was created using the PyTorch library [47]. It consists of *n* number of layers including the input and output layers where, the input layer receives the feature vector (Δ_R). The outputs from each layer were processed through the *ReLU* activation function followed by a dropout layer to prevent overfitting. The output layer uses the *sigmoid* activation function to convert the final output into a probability value suitable for binary classification of the mutation of interest as driver or neutral.

2.3.1. Model optimization

The model training process employed the focal loss function [48] to address class imbalance in the training dataset. Hyperparameter optimization to finalize the architecture and training parameters was performed using the Optuna framework [49]. The hyperparameter space was searched, encompassing architectural parameters (number of hidden layers, nodes per layer), training dynamics (batch size, learning rate, number of epochs), and focal loss-specific parameters (alpha, gamma) (Table 1). Optuna utilizes search algorithms, such as tree-structured Parzen Estimators (TPE), which facilitate effective traversal of the high-dimensional hyperparameter space. This optimization approach was aimed to identify the optimal configuration that maximizes model performance while mitigating overfitting.

Table 1
The range of values of hyperparameters used for fine tuning the models.

Hyperparameter	Range
Batch size	64 to 300
Learning rate	1e-5 to 1e-1
Epochs	25 to 200
Alpha	0 to 1
Gamma	1 to 3

2.4. Model performance evaluation

The deep neural network model was evaluated for classifying cancer-causing and neutral mutations. The evaluation metrics used were:

$$Sensitivity = TP/(TP+FN) \quad (1)$$

$$Specificity = TN/(TN+FP) \quad (2)$$

$$Accuracy = (TP+TN)/(TP+FN+FP+TN) \quad (3)$$

Table 2
Occurrence of topmost 10 mutations in driver and neutral datasets.

Driver Mutations	Frequency	Neutral Mutations	Frequency
E → K	0.075	**P → S**	0.035
R → C	0.075	**P → L**	0.033
R → H	0.070	**A → T**	0.033
R → Q	0.055	**R → Q**	0.026
R → W	0.054	**A → V**	0.024
A → T	0.051	T → A	0.024
A → V	0.048	**R → H**	0.024
P → S	0.042	D → E	0.022
P → L	0.042	E → D	0.022
S → L	0.038	G → S	0.021

*__Bold:__ mutations present in both driver and neutral datasets.
*__Frequency__ is the ratio between number of individual mutations and total mutations in the dataset.

$$Balanced\,accuracy = (Sensitivity + Specificity)/2 \tag{4}$$

$$F1 - score = TP/\left(TP + \frac{(FP + FN)}{2}\right) \tag{5}$$

where, TP, TN, FN, and FP represent the number of true positives, true negatives, false negatives and false positives, respectively. The cancer-causing mutations were considered as positive class while neutral mutations were considered as negative class. The model performance was also assessed with F1-score and AU-ROC.

3. Results and discussion

3.1. Dataset statistics

The dataset of 147,518 mutations was extracted from the COSMIC database, corresponding to 1,920 DNA-binding proteins (DBPs). A non-redundant dataset comprising 1,180 representative protein sequences was generated through sequence clustering using CD-HIT [34] to eliminate redundancy. Further, domain annotation information from InterPro [35] was utilized to identify ZNFs. The resulting dataset contained 2,811 driver mutations and 723 neutral mutations, distributed across 208 distinct ZNFs. The topmost 10 preferred mutations in driver and neutral datasets are presented in Table 2. In the driver dataset, the predominant mutations were E→K, R→C, and R→H. Conversely, the neutral dataset exhibited a higher frequency of P→S, P→L, and A→T being the most prevalent. While some mutations were exclusive to either dataset, a considerable overlap was observed, with many mutations occurring in both datasets at varying frequencies. This poses a significant challenge in discriminating between driver and neutral mutations. We observed that the incorporation of structural neighbor information is capable of handling such cases for discriminating driver and neutral mutations (Section 3.2).

A comparative analysis of the spatial distribution of driver and neutral mutations showed that 10.3% driver and 6.2% neutral sites occur within the zinc-finger domain. To quantify the relative enrichment of driver mutations in domain regions, the odds ratio was calculated (Equation (6).

$$Oddsratio = \frac{N_{dz}/N_d}{N_{nz}/N_n} \tag{6}$$

where, Ndz, and Nnz denote the number of driver and neutral sites in zinc finger domain, respectively. Nd and Nn are total number of driver and neutral sites, respectively. The resulting odds ratio was 1.65 indicating a significant enrichment of driver mutations within the zinc-finger domain of ZNFs. The observation suggested that driver mutations are more likely to occur within zinc-finger domain as compared to neutral mutations.

A comparison between zinc-finger domain length and protein length was also performed. The majority of proteins were between 300 and 600 residues in length, whereas zinc-finger domains predominantly span between 50 and 100 residues.

The structural neighbors of wild-type and mutant sites are compared for driver and neutral datasets and the results are shown in Fig. 2A and B. The analysis showed that driver mutations exhibited a higher number of structural neighbors (15–25 residues) compared to neutral mutations (5–10 residues) with a p-value of close to zero. This observation suggested that driver mutations interact with a greater number of residues in zinc finger proteins, potentially disrupting more interactions within the protein structure. Conversely, neutral mutations interact with fewer surrounding residues, indicating less impact on the protein's structural integrity. These findings imply that the surrounding residues of mutated site play a significant role in determining the impact of mutations on protein function and stability.

3.2. Performance of protein language based models

The models were trained using the representations from ESM-2, ProteinBERT, ProtTrans and ProtFlash. Following hyperparameter optimization, the performance of the best-performing models for each PLM was evaluated on an independent test dataset. The ESM-based model demonstrated a superior performance, achieving a balanced accuracy of 0.90 on training dataset and 0.70 on test dataset (Supplementary Table S1). The ProtFlash model showed comparable training performance with a balanced accuracy of 0.89, with a lower test set performance of 0.67. Similarly, the ProtTrans-based model achieved a

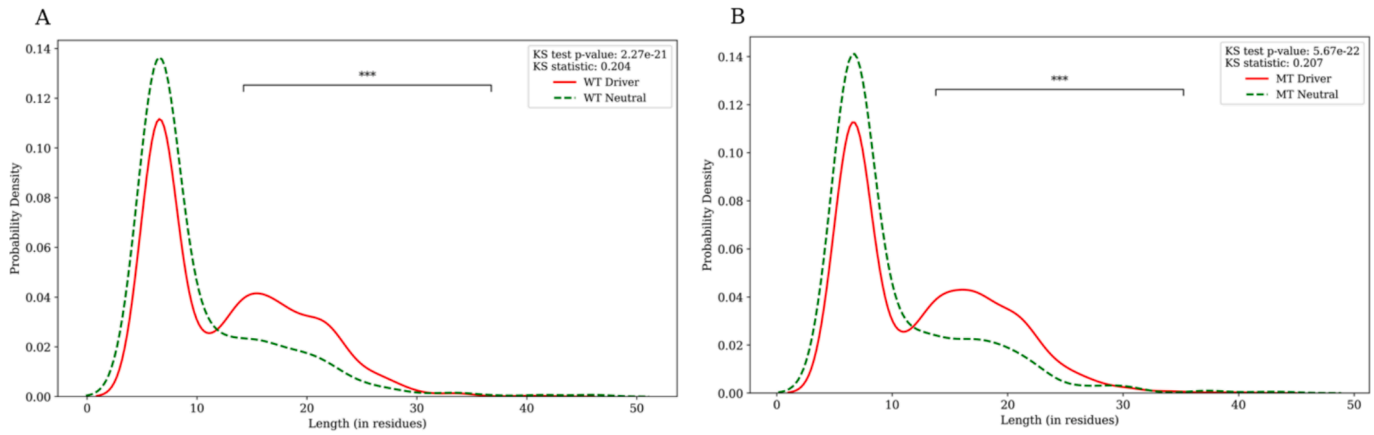

Fig. 2. Density plots showing the probability density for various lengths (in residues) of structural neighbors: (A) mutant driver and neutral sites and (B) wild-type driver and neutral sites. MT: Mutant, WT: Wild-Type.

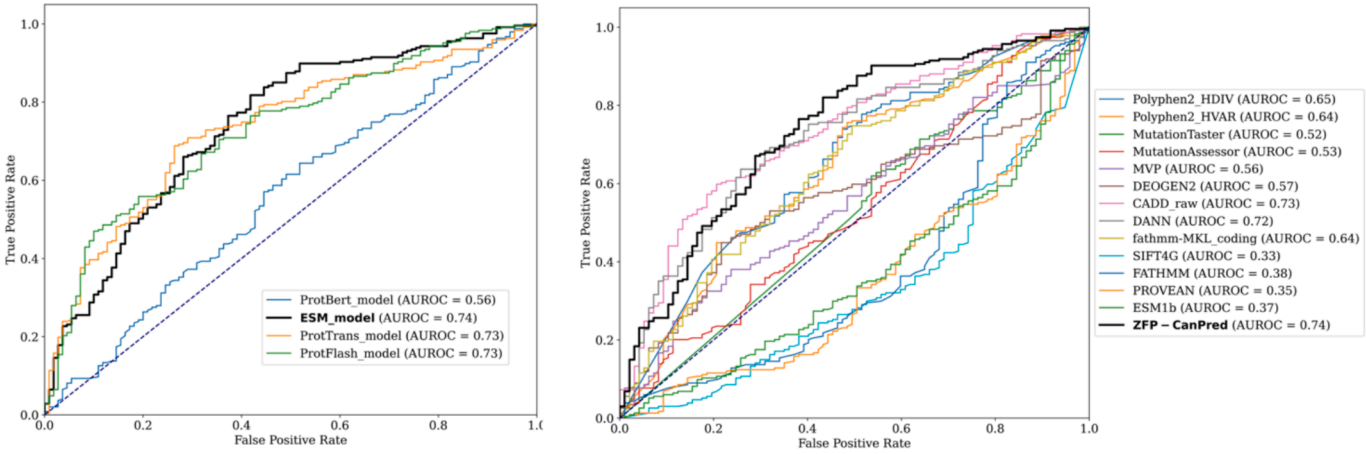

Fig. 3. Performance evaluation of ZFP-CanPred for predicting cancer driver mutations. (A) Comparison of ROC curves for ESM-based model (AUC = 0.74), ProtTrans (AUC = 0.73), ProtFlash (AUC = 0.73) and ProteinBERT-based model (AUC = 0.55). (B) Comparison of our ZFP-CanPred model (AUC = 0.74) against 11 existing prediction tools.

Table 3
Comparison prediction results of ZFP-CanPred with existing methods.

Tool	N_P	Specificity	Sensitivity	Accuracy (ACC)	Balanced ACC	AU-ROC
Polyphen2_HDIV	356	0.53 (0.61)	0.76 (0.76)	0.68 (0.72)	0.64 (0.66)	0.67 (0.74)
Polyphen2_HVAR	227	0.59 (0.58)	0.62 (0.75)	0.61 (0.7)	0.60 (0.64)	0.67 (0.71)
MutationTaster	351	0.03 (0.62)	1.00 (0.77)	0.70 (0.72)	0.52 (0.69)	0.54 (0.74)
MutationAssessor	356	0.27 (0.62)	0.79 (0.76)	0.63 (0.72)	0.53 (0.69)	0.55 (0.74)
MetaSVM	143	1.00 (0.59)	0.04 (0.79)	0.19 (0.76)	0.52 (0.69)	0.49 (0.69)
MetaLR	143	1.00 (0.59)	0.03 (0.79)	0.18 (0.76)	0.52 (0.69)	0.78 (0.69)
M−CAP	194	0.75 (0.57)	0.41 (0.77)	0.47 (0.73)	0.58 (0.67)	0.59 (0.69)
MVP	341	0.89 (0.61)	0.20 (0.77)	0.40 (0.72)	0.55 (0.69)	0.54 (0.74)
PrimateAI	221	0.80 (0.61)	0.43 (0.77)	0.52 (0.73)	0.62 (0.69)	0.68 (0.72)
DEOGEN2	357	0.84 (0.62)	0.31 (0.77)	0.47 (0.72)	0.57 (0.69)	0.58 (0.74)
LIST-S2	240	0.69 (0.63)	0.53 (0.75)	0.59 (0.71)	0.61 (0.69)	0.63 (0.76)
AlphaMissense	226	0.85 (0.65)	0.34 (0.75)	0.53 (0.71)	0.60 (0.7)	0.70 (0.76)
SIFT4G	357	0.75 (0.62)	0.53 (0.77)	0.60 (0.72)	0.64 (0.69)	0.32 (0.74)
FATHMM	141	0.99 (0.59)	0.01 (0.7)	0.40 (0.65)	0.50 (0.62)	0.31 (0.71)
PROVEAN	344	0.79 (0.62)	0.46 (0.77)	0.56 (0.72)	0.62 (0.69)	0.33 (0.75)
ESM1b	357	0.72 (0.62)	0.50 (0.77)	0.57 (0.72)	0.61 (0.69)	0.36 (0.74)
CADD	357	1.00 (0.62)	0.00 (0.77)	0.31 (0.72)	0.50 (0.69)	0.74 (0.74)
DANN	357	0.10 (0.62)	0.96 (0.77)	0.69 (0.72)	0.53 (0.69)	0.72 (0.74)
Fathmm-MKL	357	0.32 (0.62)	0.81 (0.77)	0.66 (0.72)	0.57 (0.69)	0.64 (0.74)
ZFP-CanPred	357	0.62	0.77	0.72	0.69	0.74

N_p: predicted number of mutations; the performance of ZFP-CanPred on the same dataset is shown in parentheses.

balanced accuracy of 0.83 on training dataset and 0.69 on test dataset. While the ProtTrans model's balanced accuracy were comparable to the ESM model, ESM outperformed ProtTrans with higher sensitivity (0.77) and overall accuracy (0.72). In contrast, the ProteinBERT-based model exhibited a balanced accuracy of 0.50 on both the training and test datasets, indicating poor generalization capabilities. The relative performance of these models is represented in the ROC curve (Fig. 3A). The comparative analysis of protein language models revealed the ESM-based model demonstrated superior performance across multiple metrics. The marked difference in performance between these models indicated that ESM-derived representations is more effective in capturing the relevant features for distinguishing driver and neutral mutations.

3.3. Model performance and validation

The ESM-based model was selected as the final model based on its superior performance over the ProteinBERT-based model (Fig. 3**A**). The neural network architecture comprised an input layer, six hidden layers, and an output layer, with a total of 410,068 trainable parameters. To mitigate overfitting, a dropout layer with a rate of 0.42 was implemented after each hidden layer. Training was conducted using a batch size of 64, a learning rate of 2.28×10^{-5}, and 99 epochs. The focal loss function was employed with $\alpha = 0.462$ and $\gamma = 2.51$. The model exhibited robust performance, achieving an average training and validation loss of 3.33 and 0.42, respectively, indicating effective learning and generalization. Ten-fold cross-validation resulted in an average validation accuracy of 0.69 and an AU-ROC of 0.67, which aligned with the model's performance on the test dataset, indicating robust generalizability to unseen data (Supplementary Table S2).

3.4. Model comparison

A comparative analysis of ZFP-CanPred was performed with existing methods and the results are presented in Table 3. The comparison was done with 18 different tools including SIFT4G [13], FATHMM [14], MutationAssessor [15], PROVEAN [16], MutationTaster [17], PolyPhen-2 [18], MetaSVM [19], MetaLR [19], DEOGEN2 [20], M−CAP [21], MVP [22], PrimateAI [23], AlphaMissense [24], ESM1b [25] and cancer-specific methods including CADD [27], DANN [28], LIST-S2 [29], and fathmm-MKL_coding [30]. For evaluation, the prediction score of all methods were taken from the dbNSFP database (v4) [50] and

Table 4
Comparison of ZFP-CanPred with other existing tools based on balanced accuracy for different subtypes of missense mutations.

Tools	H-H	H-PC	H-PU	PC-H	PC-PC	PC-PU	PU-H	PU-PC	PU-PU
Polyphen2_HDIV	0.56	0.61	0.62	0.64	0.52	0.52	0.73	0.48	0.56
Polyphen2_HVAR	0.52	0.45	0.57	0.78	0.53	0.57	0.66	0.48	0.48
MutationTaster	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
MutationAssessor	0.49	0.53	0.50	0.42	0.46	0.60	0.64	0.34	0.54
MVP	0.52	0.47	0.54	0.70	0.54	0.52	0.47	0.52	0.44
DEOGEN2	0.55	0.36	0.64	0.38	0.52	0.57	0.59	0.62	0.60
CADD	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
DANN	0.49	0.44	0.54	0.50	0.54	0.50	0.56	0.50	0.60
Fathmm-MKL	0.50	0.34	0.54	0.61	0.60	0.44	0.66	0.55	**0.64**
SIFT4G	0.61	0.62	0.60	0.78	0.55	0.56	0.74	**0.62**	0.42
FATHMM	0.48	0.50	0.50	0.50	0.50	0.51	0.50	0.50	0.58
PROVEAN	0.57	0.50	**0.66**	0.52	0.61	**0.70**	0.60	0.45	0.50
ESM1b	0.54	0.47	0.54	0.75	0.70	0.50	0.60	0.38	0.40
ZFP-CanPred	**0.62**	**0.64**	0.52	**0.94**	**0.72**	0.52	**0.77**	0.36	0.62

* H = Hydrophobic, PC = Polar Charged, PU = Polar Uncharged.

assessed against ZFP-CanPred using a curated test dataset comprising 247 driver and 110 neutral mutations. To ensure unbiased evaluation, mutations present in the training datasets of FATHMM and PolyPhen-2 were excluded from our test dataset. For other tools such as SIFT4G, PROVEAN, MutationTaster, MutationAssessor, MetaSVM, and MetaLR, the training datasets were not accessible. Several tools including ESM1b, AlphaMissense, PrimateAI, LIST-S2, CADD, DANN, and DEOGEN2 utilize extensively large training datasets, making direct comparison impractical. The performance of tools was evaluated using a set of metrics, including accuracy, specificity, sensitivity, and AU-ROC.

A significant limitation observed in existing tools was their inability to predict the impact of all 357 mutations in the test dataset. The evaluation was conducted on a subset of mutations for which prediction data was available from each respective tool. The classification cut-off 15 was used for CADD [51]. Moreover, most tools exhibited a trade-off between sensitivity and specificity. For instance, FATHMM, MetaSVM, MetaLR, CADD, and MVP showed the highest specificity at the cost of extremely low sensitivity. Conversely, MutationTaster and DANN achieved almost 100% sensitivity with the specificity of 0 and 0.1, respectively. As compared to other tools, ZFP-CanPred stands out with a specificity of 0.62 and sensitivity of 0.77, resulting in an accuracy of 0.72 and a balanced accuracy of 0.69. ZFP-CanPred maintained a more balanced performance across these metrics compared to other tools and achieved maximum AU-ROC of 0.74 on test dataset. The Receiver Operating Characteristic (ROC) curve illustrated in Fig. 3B was constructed using the prediction scores by selecting 331 (234 driver and 97 neutral) mutations removing prediction tools with over 100 missing predictions ensuring consistent comparison across the predictive models. Further, we compared ZFP-CanPred's performance with existing tools for different subtypes of missense mutations categorized by changes in amino acid properties using the curated dataset of 331 mutations previously used for AU-ROC analysis (Table 4). The analysis revealed that ZFP-CanPred demonstrates superior to existing tools. ZFP-CanPred achieved a high balanced accuracy for polar charged to hydrophobic (PC-H: 0.94) and polar uncharged to hydrophobic (PU-H: 0.77) substitutions, outperforming all existing tools in these categories. For polar charged to polar charged (PC-PC) mutations, ZFP-CanPred showed strong performance with an balanced accuracy of 0.72, followed by ESM1b and PROVEAN with 0.7 and 0.61 respectively. While ZFP-CanPred's performance for hydrophobic to polar uncharged (H-PU: 0.52) and polar charged to polar uncharged (PC-PU: 0.52) substitutions was moderate, it achieved competitive balanced accuracy for polar uncharged to polar uncharged (PU-PU: 0.62) mutations.

These results demonstrate ZFP-CanPred's consistent and reliable performance across different types of amino acid property changes.

Table 5
The number of driver-specific and neutral-specific features identified through the Integrated Gradient method.

Top n features	Driver-specific features	Neutral-specific features
100	7	0
300	38	11
500	72	26
700	124	50
900	194	81
1100	264	137
1280	357	193

3.5. Model interpretation

The ZFP-CanPred model, which utilizes protein structural neighbor representations derived from the ESM-2 (PLM), demonstrates superior predictive performance for classifying cancer-associated mutations in ZNFs. The model's high AUC of 0.74 indicates significant discriminative potential between driver and passenger mutations. We employed the Integrated Gradients method [52], a feature attribution technique implemented using the Captum library [53], which quantifies the contribution of individual input features to the model's output. We also considered Shapley analysis for model interpretation, prior to Integrated Gradients, but the computational complexity was very high due to the number of features present in our model as reported in previous studies [54].

Analysis of the test dataset revealed a subset of features exhibiting bipolar contributions: positive attribution values for driver mutations and negative for neutral mutations, or vice versa. With this approach we were able to identify the driver specific features and neutral specific features (Table 5), learned by the model. As the number of top features increases, there is a consistent trend of more features being associated with driver mutations compared to neutral mutations. For instance, among the top 1280 features, 357 features specifically contribute to predicting driver mutations, while only 193 are associated with neutral mutations. This difference in number of predictive features explains the observed higher sensitivity relative to specificity in the model's performance on test dataset.

We investigated whether the observed difference between sensitivity and specificity could be attributed to dataset imbalance. To address this, we attempted two approaches to balance the dataset: (i) augmenting the neutral mutations by including data from non-zinc finger proteins and (ii) reducing the number of driver mutations (Supplementary Table S3). However, neither approach yielded significant performance improvements. This may be due to two factors: (i) reducing the dataset size limited the neural network's learning capabilities and (ii) including neutral mutations from non-zinc finger proteins proved unsuitable due to their distinct structural and functional properties.

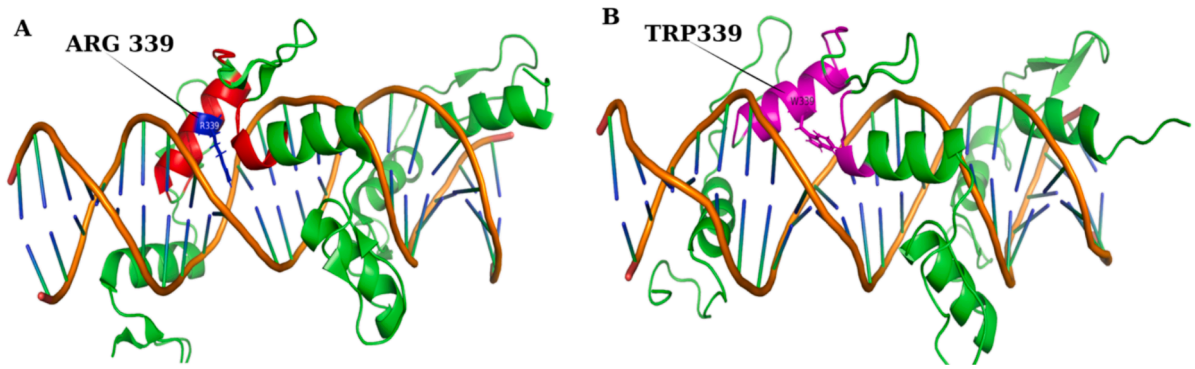

Fig. 4. The mutation R339W, A) Wild type residue Arginine at 339th position B) Mutant residue Tryptophan at 339th Position. *Red color is the neighbor residues our model is considering for predicting the mutation effect.

4. Application

We collected disease-causing mutations in the human CCCTC binding factor, (CTCF, UniProt ID: P49711) transcriptional repressor protein from published literature [55–57], ensuring that this protein was not represented in our training dataset. This independent test set comprised 35 cancer-associated mutations, of which ZFP-CanPred correctly classified 31 mutations as pathogenic, achieving an accuracy of 88.6% (Supplementary Table S5). This performance on a novel protein demonstrates the model's robust generalization capability beyond its training data.

Fig. 4 illustrates the impact of the R339W mutation within the zinc finger domain of CTCF, specifically in its DNA-binding region [56]. This substitution not only disrupts the direct protein-DNA interactions at the mutation site but also induces conformational changes that affect neighboring residues' interaction capabilities. The ZFP-CanPred captures these biochemical perturbations, demonstrating its ability to recognize both local and distributed effects of the mutation. This example particularly highlights the potential of our method for interpreting subtle changes in the protein's functional environment, as the replacement of a positively charged residue, Arg with the aromatic residue, Trp, likely alters the electrostatic landscape and structural dynamics of the DNA-binding interface.

We also tried to extend the applicability of ZFP-CanPred to non-zinc finger proteins such as High mobility group box domain and Basic-leucine zipper domain (Supplementary Table S4). The AU-ROC values (0.68–0.70) suggest a limited discriminative capability, indicating that the model's performance may be constrained by the inherent complexity of these protein families. This analysis showed that the performance is moderate, and method specific to each family could improve the performance.

5. Conclusion

The study presents a novel approach to the identification of driver mutations in zinc-finger proteins (ZNFs), a critical class of transcription factors implicated in numerous cellular processes and oncogenic pathways. By leveraging advanced protein language models, particularly the ESM-2 architecture, we have developed a robust neural network-based predictor, ZFP-CanPred, that demonstrates superior performance in distinguishing driver mutations from neutral variants in ZNFs specific to cancer. Our analysis of mutational patterns revealed critical insights into the nature of driver mutations in ZNFs, including their propensity to occur within functional domains and their association with extended neighboring sequence lengths. These findings contribute to our understanding of the structural and functional impacts of mutations on protein-DNA interactions. The superior performance of our model, as evidenced by its high accuracy (0.72), and AU-ROC (0.74), coupled with its ability to balance specificity and sensitivity, represents a significant improvement over existing methods of mutation effect prediction specific to ZNFs.

CRediT authorship contribution statement

Amit Phogat: Data curation, Methodology, Formal analysis, Software, Writing – original draft, Writing – review & editing. **Sowmya Ramaswamy Krishnan:** Methodology, Formal analysis, Writing – review & editing. **Medha Pandey:** Methodology, Formal analysis, Writing – review & editing. **M. Michael Gromiha:** Conceptualization, Methodology, Supervision, Formal analysis, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

SRK is an employee of Tata Consultancy Services at the time of manuscript submission.

Acknowledgements

We express our gratitude to IIT Madras for the computational facilities. Our sincere appreciation goes to the members of the Protein Bioinformatics Lab for their valuable suggestions. MMG acknowledges the Department of Biotechnology, Government of India for partial financial support to MMG (Grant No. BT/PR40156/BTIS/137/54/2023).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymeth.2025.01.020.

Data availability

Data will be made available on request.

References

[1] K. Harini, D. Kihara, M.M. Gromiha, PDA-Pred: Predicting the binding affinity of protein-DNA complexes using machine learning techniques and structural features, Methods 213 (2023 May) 10–17.
[2] O. Bonczek, L. Wang, S.V. Gnanasundram, S. Chen, L. Haronikova, F. Zavadil-Kokas, B. Vojtesek, DNA and RNA binding proteins: from motifs to roles in cancer, Int. J. Mol. Sci. 23 (16) (2022 Aug 18) 9329.
[3] M. Cassandri, A. Smirnov, F. Novelli, C. Pitolli, M. Agostini, M. Malewicz, G. Melino, G. Raschellà, Zinc-finger proteins in health and disease, Cell Death Discov. 13 (3) (2017 Nov) 17071.
[4] J. Zhao, D. Wen, S. Zhang, H. Jiang, X. Di, The role of zinc finger proteins in malignant tumors, FASEB J. 37 (9) (2023 Sep) e23157.

[5] D. Munro, D. Ghersi, M. Singh, Two critical positions in zinc finger domains are heavily mutated in three human cancer types, PLoS Comput. Biol. 14 (6) (2018 Jun 28) e1006290.

[6] A. Mao, C. Chen, S. Portillo-Ledesma, T. Schlick, Effect of Single-Residue Mutations on CTCF Binding to DNA: Insights from Molecular Dynamics Simulations, Int J Mol Sci. 24 (7) (2023 Mar 29) 6395.

[7] M. Luo, Y. Zhang, Z. Xu, S. Lv, Q. Wei, Q. Dang, Experimental analysis of bladder cancer-associated mutations in EP300 identifies EP300-R1627W as a driver mutation, Mol Med. 29 (1) (2023 Jan 16) 7.

[8] Y.K. Lam, J. Yu, H. Huang, X. Ding, A.M. Wong, H.H. Leung, A.W. Chan, K.K. Ng, M. Xu, X. Wang, N. Wong, TP53 R249S mutation in hepatic organoids captures the predisposing cancer risk, Hepatology 78 (3) (2023 Sep 1) 727–740.

[9] S. Nishikawa, T. Iwakuma, Drugs Targeting p53 Mutations with FDA Approval and in Clinical Trials, Cancers (Basel) 15 (2) (2023 Jan 9) 429.

[10] X. Xie, T. Yu, X. Li, N. Zhang, L.J. Foster, C. Peng, W. Huang, G. He, Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials, Signal Transduct Target Ther. 8 (1) (2023 Sep 6) 335.

[11] P. Montesinos, C. Recher, S. Vives, E. Zarzycka, J. Wang, G. Bertani, M. Heuser, R. T. Calado, A.C. Schuh, S.P. Yeh, S.R. Daigle, J. Hui, S.S. Pandya, D.A. Gianolio, S. de Botton, H. Döhner, Ivosidenib and Azacitidine in IDH1-Mutated Acute Myeloid Leukemia, N Engl J Med. 386 (16) (2022 Apr 21) 1519–1531.

[12] V. Suybeng, F. Koeppel, A. Harlé, E. Rouleau, Comparison of Pathogenicity Prediction Tools on Somatic Variants, J Mol Diagn. 22 (12) (2020 Dec) 1383–1392.

[13] R. Vaser, S. Adusumalli, S.N. Leng, M. Sikic, P.C. Ng, SIFT missense predictions for genomes, Nat Protoc. 11 (1) (2016 Jan) 1–9.

[14] M.F. Rogers, H.A. Shihab, M. Mort, D.N. Cooper, T.R. Gaunt, C. Campbell, FATHMM-XF: accurate prediction of pathogenic point mutations via extended features, Bioinformatics 34 (3) (2018 Feb 1) 511–513.

[15] B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics, Nucleic Acids Res. 39 (17) (2011 Sep 1) e118.

[16] Y. Choi, G.E. Sims, S. Murphy, J.R. Miller, A.P. Chan, Predicting the functional effect of amino acid substitutions and indels, PLoS One. 7 (10) (2012) e46688.

[17] J.M. Schwarz, C. Rödelsperger, M. Schuelke, D. Seelow, MutationTaster evaluates disease-causing potential of sequence alterations, Nat. Methods 7 (8) (2010 Aug) 575–576.

[18] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20.

[19] S. Kim, J.H. Jhong, J. Lee, J.Y. Koo, Meta-analytic support vector machine for integrating multiple omics data, BioData Min. 26 (10) (2017 Jan) 2.

[20] D. Raimondi, I. Tanyalcin, J. Ferté, A. Gazzo, G. Orlando, T. Lenaerts, M. Rooman, W. Vranken, DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins, Nucleic Acids Res. 45 (W1) (2017 Jul 3) W201–W206.

[21] K.A. Jagadeesh, A.M. Wenger, M.J. Berger, H. Guturu, P.D. Stenson, D.N. Cooper, J.A. Bernstein, G. Bejerano, M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity, Nat. Genet. 48 (12) (2016 Dec) 1581–1586.

[22] H. Qi, H. Zhang, Y. Zhao, C. Chen, J.J. Long, W.K. Chung, Y. Guan, Y. Shen, MVP predicts the pathogenicity of missense variants by deep learning, Nat Commun. 12 (1) (2021 Jan 21) 510.

[23] L. Sundaram, H. Gao, S.R. Padigepati, J.F. McRae, Y. Li, J.A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon, J. Xu, S. Batzoglou, X. Li, K.K. Farh, Predicting the clinical impact of human mutation with deep neural networks, Nat. Genet. 50 (8) (2018 Aug) 1161–1170.

[24] Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, Schneider RG, Senior AW, Jumper J, Hassabis D, Kohli P, Avsec Ž. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023 Sep 22;381(6664):eadg7492.

[25] N. Brandes, G. Goldman, C.H. Wang, C.J. Ye, V. Ntranos, Genome-wide prediction of disease variant effects with a deep protein language model, Nat. Genet. 55 (9) (2023 Sep) 1512–1522.

[26] G.R. Buel, K.J. Walters, Can AlphaFold2 predict the impact of missense mutations on structure? Nat Struct Mol Biol. 29 (1) (2022 Jan) 1–2.

[27] P. Rentzsch, D. Witten, G.M. Cooper, J. Shendure, M. Kircher, CADD: predicting the deleteriousness of variants throughout the human genome, Nucleic Acids Res. 47 (D1) (2019 Jan 8) D886–D894.

[28] D. Quang, Y. Chen, X. Xie, DANN: a deep learning approach for annotating the pathogenicity of genetic variants, Bioinformatics 31 (5) (2015 Mar 1) 761–763.

[29] N. Malhis, M. Jacobson, S.J.M. Jones, J. Gsponer, LIST-S2: taxonomy based sorting of deleterious missense mutations across species, Nucleic Acids Res. 48 (W1) (2020 Jul 2) W154–W161.

[30] H.A. Shihab, M.F. Rogers, J. Gough, M. Mort, D.N. Cooper, I.N. Day, T.R. Gaunt, C. Campbell, An integrative approach to predicting the functional effects of non-coding and coding sequence variation, Bioinformatics 31 (10) (2015 May 15) 1536–1543.

[31] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc Natl Acad Sci U S A. 118 (15) (2021 Apr 13) e2016239118.

[32] S.A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C.G. Cole, S. Ward, E. Dawson, L. Ponting, R. Stefancsik, B. Harsha, C.Y. Kok, M. Jia, H. Jubb, Z. Sondka, S. Thompson, T. De, P.J. Campbell, COSMIC: somatic cancer genetics at high-resolution, Nucleic Acids Res. 45 (D1) (2017 Jan 4) D777–D783.

[33] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A.J. Bridge, S. Poux, L. Bougueleret, I. Xenarios, UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View, Methods Mol Biol. 1374 (2016) 23–54.

[34] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (23) (2012 Dec 1) 3150–3152.

[35] T. Paysan-Lafosse, M. Blum, S. Chuguransky, T. Grego, B.L. Pinto, G.A. Salazar, M. L. Bileschi, P. Bork, A. Bridge, L. Colwell, J. Gough, D.H. Haft, I. Letunić, A. Marchler-Bauer, H. Mi, D.A. Natale, C.A. Orengo, A.P. Pandurangan, C. Rivoire, C.J.A. Sigrist, I. Sillitoe, N. Thanki, P.D. Thomas, S.C.E. Tosatto, C.H. Wu, A. Bateman, InterPro in 2022, Nucleic Acids Res. 51 (D1) (2023 Jan 6) D418–D427.

[36] M.J. Landrum, J.M. Lee, M. Benson, G.R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J.B. Holmes, B. L. Kattman, D.R. Maglott, ClinVar: improving access to variant interpretations and supporting evidence, Nucleic Acids Res. 46 (D1) (2018 Jan 4) D1062–D1067.

[37] E.M. Smigielski, K. Sirotkin, M. Ward, S.T. Sherry, dbSNP: a database of single nucleotide polymorphisms, Nucleic Acids Res. 28 (1) (2000 Jan 1) 352–355.

[38] K. Ganesan, A. Kulandaisamy, S. Binny Priya, M.M. Gromiha, HuVarBase: A human variant database with comprehensive information at gene and protein levels, PLoS One. 14 (1) (2019 Jan 31) e0210475.

[39] Z. Yue, L. Zhao, J. Xia, dbCPM: a manually curated database for exploring the cancer passenger mutations, Brief Bioinform. 21 (1) (2020 Jan 17) 309–317.

[40] M. Varadi, D. Bertoni, P. Magana, U. Paramval, I. Pidruchna, M. Radhakrishnan, M. Tsenkov, S. Nair, M. Mirdita, J. Yeo, O. Kovalevskiy, K. Tunyasuvunakool, A. Laydon, A. Žídek, H. Tomlinson, D. Hariharan, J. Abrahamson, T. Green, J. Jumper, E. Birney, M. Steinegger, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences, Nucleic Acids Res. 52 (D1) (2024 Jan 5) D368–D375.

[41] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, L. Serrano, The FoldX web server: an online force field, Nucleic Acids Res. (2005). Jul 1;33(Web Server issue):W382–8.

[42] R. Nagarajan, A. Archana, A.M. Thangakani, S. Jemimah, D. Velmurugan, M. M. Gromiha, PDBparam: Online Resource for Computing Structural Parameters of Proteins, Bioinform Biol Insights. 14 (10) (2016 Jun) 73–80.

[43] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics 38 (8) (2022 Apr 12) 2102–2110.

[44] B.E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C.H. Wu, UniRef: comprehensive and non-redundant UniProt reference clusters, Bioinformatics 23 (10) (2007 May 15) 1282–1288.

[45] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning, IEEE Trans Pattern Anal Mach Intell. 44 (10) (2022 Oct) 7112–7127.

[46] L. Wang, H. Zhang, W. Xu, Z. Xue, Y. Wang, Deciphering the protein landscape with ProtFlash, a lightweight language model, Cell Rep. Phys. Sci. 4 (10) (2023) 101600.

[47] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv [Preprint]. 2019 Dec 3: 1912.01703. Available from: https://arxiv.org/abs/1912.01703.

[48] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. arXiv [Preprint]. 2018 Aug 7:1708.02002. Available from: https://arxiv.org/abs/1708.02002.

[49] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv [Preprint]. 2019 Jul 25: 1907.10902. Available from: https://arxiv.org/abs/1907.10902.

[50] X. Liu, C. Li, C. Mou, Y. Dong, Y. Tu, dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs, Genome Med. 12 (1) (2020 Dec 2) 103.

[51] K.J. van der Velde, E.N. de Boer, C.C. van Diemen, B. Sikkema-Raddatz, K. M. Abbott, A. Knopperts, L. Franke, R.H. Sijmons, T.J. de Koning, C. Wijmenga, R. J. Sinke, M.A. Swertz, GAVIN: Gene-Aware Variant INterpretation for medical sequencing, Genome Biol. 18 (1) (2017 Jan 16) 6.

[52] Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. arXiv [Preprint]. 2017 Mar 4:1703.01365. Available from: https://arxiv.org/abs/1703.01365.

[53] Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, Melnikov A, Kliushkina N, Araya C, Yan S, Reblitz-Richardson O. Captum: A unified and generic model interpretability library for PyTorch. arXiv [Preprint]. 2020 Sep 16: 2009.07896. Available from: https://arxiv.org/abs/2009.07896.

[54] K. Aas, M. Jullum, A. Løland, Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, Artif Intell. 298 (2021) 103502.

[55] G.N. Filippova, C.F. Qi, J.E. Ulmer, J.M. Moore, M.D. Ward, Y.J. Hu, D.I. Loukinov, E.M. Pugacheva, E.M. Klenova, P.E. Grundy, A.P. Feinberg, A.M. Cleton-Jansen, E. W. Moerland, C.J. Cornelisse, H. Suzuki, A. Komiya, A. Lindblom, F. Dorion-Bonnet, P.E. Neiman, H.C. Morse 3rd, S.J. Collins, V.V. Lobanenkov, Tumor-

associated zinc finger mutations in the CTCF transcription factor selectively alter tts DNA-binding specificity, Cancer Res. 62 (1) (2002 Jan 1) 48–52.
[56] C.G. Bailey, S. Gupta, C. Metierre, P.M.S. Amarasekera, P. O'Young, W. Kyaw, T. Laletin, H. Francis, C. Semaan, M. Sharifi Tabar, K.P. Singh, C.G. Mullighan, O. Wolkenhauer, U. Schmitz, J.E.J. Rasko, Structure-function relationships explain CTCF zinc finger mutation phenotypes in cancer, Cell Mol Life Sci. 78 (23) (2021 Dec) 7519–7536.
[57] I.A. Voutsadakis, Molecular Lesions of Insulator CTCF and Its Paralogue CTCFL (BORIS) in Cancer: An Analysis from Published Genomic Studies, High Throughput. 7 (4) (2018 Oct 1) 30.