CDMPred: a tool for predicting cancer driver missense mutations with high-quality passenger mutations

Lihua Wang[1,2], Haiyang Sun[3], Zhenyu Yue[4], Junfeng Xia[1] and Xiaoyan Li[1]

[1] Information Materials and Intelligent Sensing Laboratory of Anhui Province, Institutes of Physical Science and Information Technology, Anhui University, Hefei, Anhui, China
[2] School of Information Engineering, Huangshan University, Huangshan, Anhui, China
[3] State Key Laboratory of Medicinal Chemical Biology, NanKai University, Tianjin, Tianjin, China
[4] School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei, Anhui, China

ABSTRACT

Most computational methods for predicting driver mutations have been trained using positive samples, while negative samples are typically derived from statistical methods or putative samples. The representativeness of these negative samples in capturing the diversity of passenger mutations remains to be determined. To tackle these issues, we curated a balanced dataset comprising driver mutations sourced from the COSMIC database and high-quality passenger mutations obtained from the Cancer Passenger Mutation database. Subsequently, we encoded the distinctive features of these mutations. Utilizing feature correlation analysis, we developed a cancer driver missense mutation predictor called CDMPred employing feature selection through the ensemble learning technique XGBoost. The proposed CDMPred method, utilizing the top 10 features and XGBoost, achieved an area under the receiver operating characteristic curve (AUC) value of 0.83 and 0.80 on the training and independent test sets, respectively. Furthermore, CDMPred demonstrated superior performance compared to existing state-of-the-art methods for cancer-specific and general diseases, as measured by AUC and area under the precision-recall curve. Including high-quality passenger mutations in the training data proves advantageous for CDMPred's prediction performance. We anticipate that CDMPred will be a valuable tool for predicting cancer driver mutations, furthering our understanding of personalized therapy.

Subjects Bioinformatics, Data Mining and Machine Learning
Keywords Cancer, Machine learning, Driver missense mutation prediction, Benchmark quality, XGBoost

Submitted 10 April 2024
Accepted 7 August 2024
Published 6 September 2024

Corresponding author
Xiaoyan Li, lixiaoyan@ahu.edu.cn

Academic editor
Rogerio Sotelo-Mundo

Additional Information and Declarations can be found on page 14

DOI **10.7717/peerj.17991**

INTRODUCTION

Cancer is a leading cause of death and suffering in humans worldwide, resulting in nearly 20 million new cases alongside 9.7 million deaths in 2022 (*Bray et al., 2024*). Researchers have confirmed that cancer is a multifaceted genetic disease caused by the accumulation of numerous mutations in the genome (*Wood et al., 2007*; *Tomasetti et al., 2015*; *Xi et al., 2020*). However, the tumorigenesis and development of most cancers are primarily driven by a small number of critical mutations (*Hanahan & Weinberg, 2011*; *Muiños et al., 2021*; *Ostroverkhova, Przytycka & Panchenko, 2023*), while the remaining mutations are

How to cite this article Wang L, Sun H, Yue Z, Xia J, Li X. 2024. CDMPred: a tool for predicting cancer driver missense mutations with high-quality passenger mutations. *PeerJ* **12**:e17991 http://doi.org/10.7717/peerj.17991

considered neutral (passengers). Identifying driver mutations from passenger mutations holds significant importance, as drivers are commonly utilized as diagnostic and prognostic biomarkers and potential drug targets for cancer treatment (*Xi et al., 2020*; *Cheng et al., 2024*).

Vogelstein et al. (2013) observed that most protein-coding mutations in cancer genomes were missense changes. Consequently, our focus in this study is on cancer driver missense mutations. To date, numerous computational methods have been developed to predict driver missense mutations, such as boostDM (*Muiños et al., 2021*), Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM) (*Carter et al., 2009*), Transformed Functional Impact score for Cancer (transFIC) (*Gonzalez-Perez, Deu-Pons & Lopez-Bigas, 2012*), Cancer Driver Annotation (CanDrA) (*Mao et al., 2013*), Functional Analysis through Hidden Markov Models (FATHMM) (*Shihab et al., 2013*), CScape-somatic (*Rogers, Gaunt & Campbell, 2020*), and CHASMplus (*Tokheim & Karchin, 2019*). Additionally, some methods are focused on identifying driver mutations at critical sites, such as protein allosteric sites (*Song et al., 2019*; *Song et al., 2023*; *Shen et al., 2017*). These methods typically utilize positive samples obtained from cancer-related databases, such as the Catalogue of Somatic Mutations in Cancer (COSMIC) database (*Carter et al., 2009*; *Gonzalez-Perez, Deu-Pons & Lopez-Bigas, 2012*; *Mao et al., 2013*; *Shihab et al., 2013*; *Rogers, Gaunt & Campbell, 2020*), while negative samples are commonly derived from statistical methods (*Muiños et al., 2021*; *Carter et al., 2009*) or putative samples (*Gonzalez-Perez, Deu-Pons & Lopez-Bigas, 2012*; *Mao et al., 2013*; *Shihab et al., 2013*; *Rogers, Gaunt & Campbell, 2020*; *Tokheim & Karchin, 2019*).

This study evaluated the potential for improved driver prediction by investigating high-quality passenger mutations. We then proposed a predictor, CDMPred, which incorporates high-quality passenger mutations and utilizes the eXtreme Gradient Boosting (XGBoost) algorithm. Initially, we conducted comparative analyses of the Cancer Passenger Mutations database (dbCPM), which comprises highly curated passenger mutations (*Yue, Zhao & Xia, 2020*). The results indicated that the dbCPM data aligns with other negative datasets regarding most classical features, while exhibiting specificity for cancer-related features (*Yue, Zhao & Xia, 2020*; *Wong et al., 2011*). Subsequently, we employed the high-quality passenger mutation data for model training and encoded 65 features. We used feature importance to identify the top 10 features from the 65 features mentioned above and evaluated the performance of various machine learning algorithms on the training set. Ultimately, we employed the optimal model (CDMPred) with an XGBoost classifier and the top 10 features. The results obtained from the training and independent test sets demonstrated that CDMPred exhibited superior performance compared to several state-of-the-art methods for both cancer-specific and general diseases, as assessed by two threshold-independent metrics: the area under the receiver operating characteristic curve (AUC) and the area under the precision–recall curve (AUPR).

MATERIALS & METHODS

Figure 1 presents the flowchart of the CDMPred method. Portions of this text were previously published as part of a preprint (https://www.researchsquare.com/article/rs-

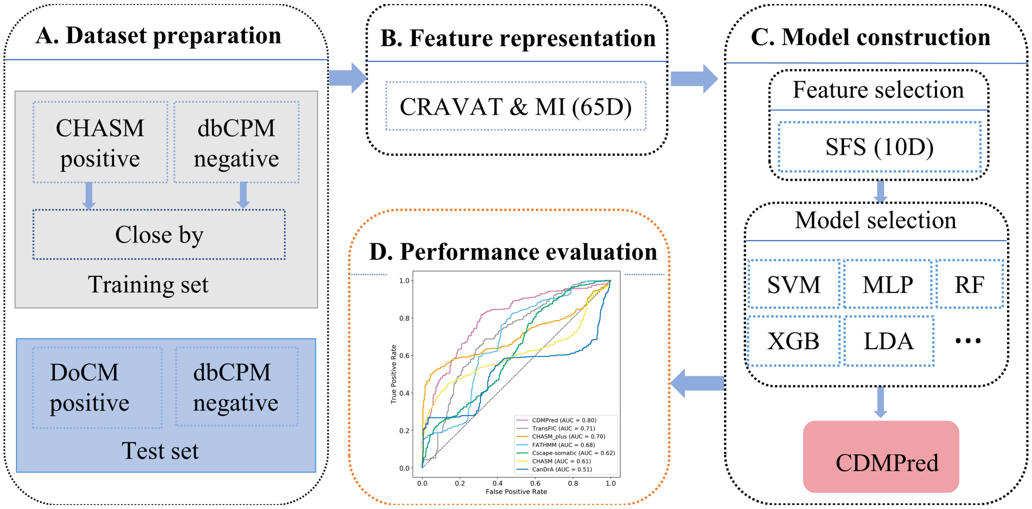

Figure 1 **Flowchart of the proposed method.**

Full-size DOI: 10.7717/peerj.17991/fig-1

1350438/v1). The procedure consists of four steps: dataset preparation, feature representation, model construction, and performance evaluation. Each step is explained in detail below.

Dataset preparation

The datasets were divided into two groups: one for feature analysis and the other for model construction and performance evaluation. One cancer driver mutation dataset and three passenger mutation datasets were used for feature analysis. For positive samples, we selected 1,248 driver missense mutations from the Database of Curated Mutations (DoCM) (v3.2) (*Ainscough et al., 2016*), which is a reliable source that aggregates functionally validated mutations in cancer. For negative samples, we gathered three passenger datasets. The dataset dbCPM (v1.1) consists of 1,919 passenger mutations, including 1,634 distinct missense mutations (*Yue, Zhao & Xia, 2020*). The other two datasets are oriented from classic prediction tools for cancer-specific driver mutations. Expressly, the dataset FATHMM was initially obtained from the UniProt database, which was taken as negative samples in the FATHMM training set (*Shihab et al., 2013*; *Apweiler et al., 2004*), and the dataset CHASM (v3.1) consists of synthetic passenger mutations in the CHASM training set (*Carter et al., 2009*). We removed the mutations simultaneously in DoCM in each passenger mutation dataset. The details are presented in Table S1.

The datasets utilized for model construction are described as follows. Out of the 1,634 missense passenger mutations in dbCPM v1.1, 1,104 items from dbCPM v1.0 were used as negative samples in our training set. We filtered the 13,235 positive samples in the CHASM (v3.1) training set to avoid overlap with samples from dbCPM v1.0. Next, we included only positive samples within 50 bp of a passenger mutation on the same transcript to address the imbalance and potential bias towards positive samples. As a result, our training set retained 2,151 driver missense mutations. We obtained an independent test

Table 1 **Summary of mutation datasets used for model construction and evaluation.**

	Training set		Independent test set	
	Positive	**Negative**	**Positive**	**Negative**
Source	CHASM v3.1	dbCPM v1.0	DoCM v3.2 nonoverlap	dbCPM v1.1 nonoverlap
Number	2151	1104	567	530

Notes.

DoCM v3.2 nonoverlap, data in DoCM v3.2 but not in CHASM v3.1; dbCPM v1.1 nonoverlap, data in dbCPM v1.1 except for dbCPM v1.0.

set to benchmark performance against state-of-the-art prediction tools. First, we collected missense mutations in dbCPM v1.1 reported after the initial database update (dbCPM v1.0) to serve as negative samples. Secondly, we considered all 1,248 driver mutations in DoCM as our positive samples. To prevent type 1 circularity (*Grimm et al., 2015*), which can cause overfitting from overlapping training and evaluation datasets, we excluded overlapping data with the training set, resulting in 567 driver mutations. The datasets utilized for model construction and performance evaluation are detailed in Table 1.

Feature representation

Considering both the significance of the protein's functions and conservation, seven feature groups were provided to capture the specific characteristics of cancer driver mutations, comprising protein physicochemical properties, evolutionary conservation scores, exon features, protein local structures, regional composition, amino acid residue triplet features, and UniProt annotations. For each missense mutation in the datasets mentioned above, the features were encoded with the 85 pre-computed features available in SNVBox (*Wong et al., 2011*; *Won et al., 2021*) from a dockerized tool, the Cancer-Related Analysis of Variants Toolkit (*Masica et al., 2017*) (CRAVAT, version 5.2.3). To prepare the input data, we curated the transcript information using Ensembl GRCh37 (*Flicek et al., 2014*) as a reference. Each feature underwent scaling by subtracting the mean value and dividing it by the root mean square (RMS) value, utilizing pre-computed values for the entire genome. After CHASM (*Shen et al., 2017*; *Yue, Zhao & Xia, 2020*), we applied the information gain method to remove irrelevant features among the 85 candidate features. By using a uniform threshold, we selected 65 predictive features that possessed a minimum of 0.001 bits of mutual information Specifically, 13 out of 16 protein physicochemical properties, all six evolutionary conservation score features, all three exon features, 11 out of 12 protein local structure features, six out of 11 regional composition features, and 26 out of 28 UniProt annotations were included. The amino acid residue trimer features were also excluded. A detailed list is indicated in Table S2.

Model construction

We utilized feature importance with XGBoost to select an optimal subset of features. Subsequently, we comprehensively evaluated multiple algorithms on the training set using a 10-fold cross-validation (*Buske et al., 2013*). We selected eight classifiers, namely random forest (RF), support vector machine (SVM), multilayer perceptron (MLP), gradient boosting decision tree (GBDT), linear discriminant analysis (LDA), logistic regression (LR), naïve Bayes (NB), and XGBoost (*Chen & Guestrin, 2016*). All the algorithms mentioned above were implemented using scikit-learn (v0.22.2) and Python 3.7. The classifiers were implemented with parameters optimized through grid search, utilizing the 10-fold cross-validation results of the training set. Specifically, we optimized three critical parameters in XGBoost: the boosting learning rate (learning_rate), the maximum depth of the tree (max_depth), and the subsample ratio of columns when constructing each tree (colsample_bytree).

Performance evaluation

As quantitative measurements of prediction results, we employed two threshold-independent measures: AUC and AUPR (*Carter et al., 2009*; *Dong et al., 2015*). Additionally, we used two qualitative measures, namely sensitivity (or true positive rate) and specificity (true negative rate), for model performance analysis, as previously described in research (*Cheng et al., 2020*; *Zeng et al., 2018*). These measures are defined as follows:

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

where TP (true positive) means the number of correctly predicted cancer driver mutations, FP (false positive) represents the number of passenger mutations predicted as drivers, TN (true negative) represents the number of correctly predicted passenger mutations, and FN (false negative) indicates the number of cancer driver mutations predicted as passengers.

The permutation test was conducted on CDMPred to demonstrate that the model learned more than noise. Specifically, we first trained the CDMPred model on the data and saved the AUC value of 10-fold cross-validation. Secondly, we randomly permuted the class labels in the dataset and trained a new model called "CDMPred_random". Thirdly, we assessed the performance of "CDMPred_random" regarding AUC. We repeated the second and the third steps 1,000 times. Finally, we calculated the empirical *p*-value by comparing the distribution of the 1,000 values to the corresponding value from the original CDMPred. The permutation test algorithm was implemented with the function named permutation_test_score in scikit-learn.

RESULTS

Analysis of features between different datasets

We quantified 85 features for all datasets presented in Table S1, which comprehensively represents the biological impacts of the mutation in the human genome (*Wong et al., 2011*). We statistically analyzed the dbCPM samples using these features in the nonparametric Wilcoxon signed rank hypothesis test. Figure 2A displays the significant features ($p < 0.05$) of positive samples obtained from DoCM, dbCPM, and other negative samples. Figure 2B illustrates the significant features among all negative samples. Our findings indicate that dbCPM data closely resemble other negative samples in terms of most classical features,

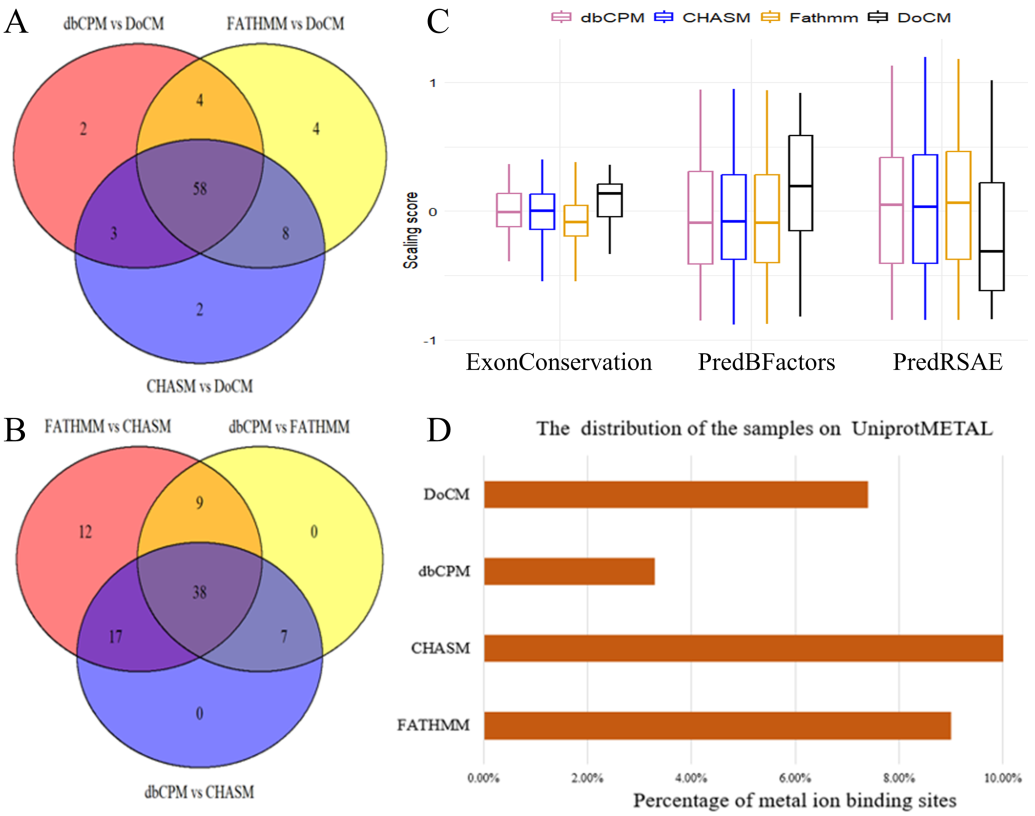

Figure 2 Statistical analysis of samples. (A) Overlap of features showing significant differences between negative and gold standard positive samples. (B) Overlap of features showing significant differences between negative samples. (C) RMS score distribution of all samples for three classical features. (D) RMS score distribution of all samples for the UniprotMETAL feature.

Full-size DOI: 10.7717/peerj.17991/fig-2

including 'ExonConservation' (conservation score for the entire exon calculated from the phylogenetic alignment of 46 species) and 'PredBFactorS' (probability that the residue backbone of wild type is stiff) (*Kent et al., 2002*; *Katzman et al., 2008*). Subsequently, we identified three features based on their *p*-values, and the RMS score distribution of all samples is presented in Fig. 2C. Therefore, the mutations in dbCPM were utilized as qualified negative samples for predicting disease-causing mutations. dbCPM exhibited distinguishable characteristics in cancer-specific features compared to other negative samples, including 'UniprotMETAL' (a binding site for a metal ion) and 'UniprotREP' (positions of repeated sequence motifs or domains) (*Ribeiro et al., 2004*; *Xu et al., 2016*). Figure 2D illustrates the distribution of RMS scores for the UniprotMETAL feature across all samples. These findings further support that dbCPM mutations are more representative than other negative samples in modeling a wide range of passenger mutations and are better suited for predicting cancer driver mutations.

Explorations for an optimal model

Figure 3 displays the AUC values of the training set for the eight classifiers. XGBoost outperformed all other classifiers, achieving an AUC value of 0.82. XGBoost was

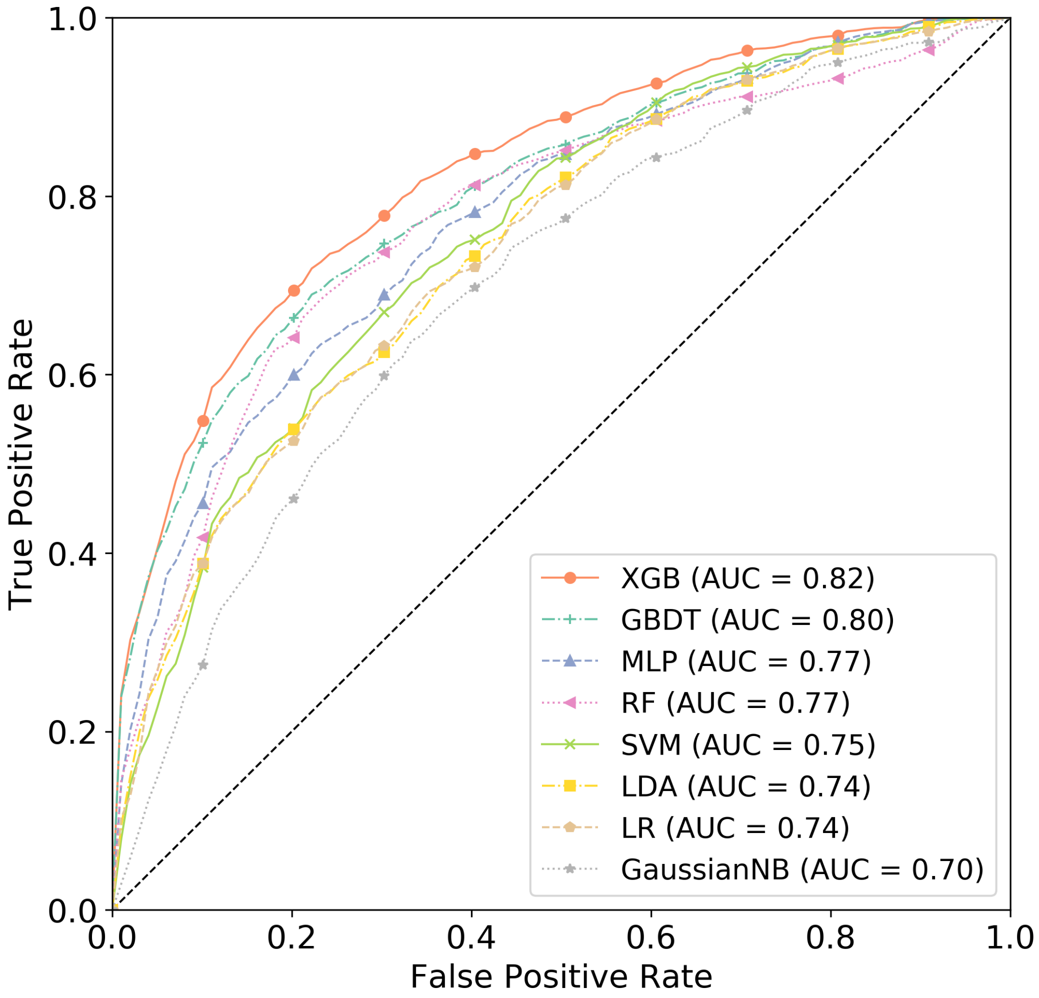

Figure 3 ROC curves of several machine learning methods with parameters tuned on the training set to obtain an optimal model.

Full-size DOI: 10.7717/peerj.17991/fig-3

applied with three optimized parameters: learning_rate = 0.04, max_depth = 4, and colsample_bytree = 0.2.

To explore the possibility of further refining the features selected from mutual information, we examined the correlations among the 65 features. We identified several highly related features in UniProt, as highlighted in yellow in Fig. S1. Subsequently, we utilized the feature selection method with XGBoost (using default parameters) to determine the importance of the features. We employed sequential feature selection (SFS) and used the optimized parameters of XGBoost to train the data. Figure 4 illustrates the comparison of the AUC results for these features. The top 10 features (highlighted in bold in Table S2) achieved the highest mean AUC of 0.83 with 10-fold cross-validation. We conducted a performance comparison between the top 10 features and the absence of the top k (1 to 10) features using 10-fold cross-validation (Fig. 5). The results indicate that excluding features like ExonSnpDensity and ExonHapMapSnpDensity, which quantify

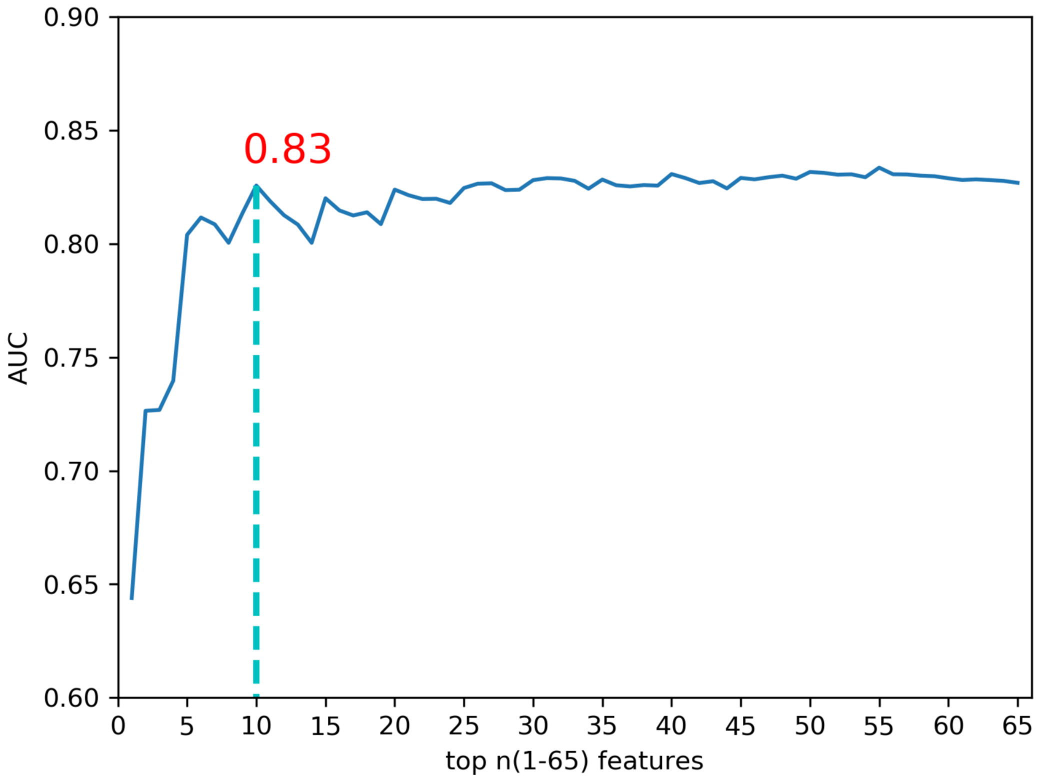

Figure 4 Comparison of AUC values with top n (1–65) features predicted by XGBoost feature importance on the training set. The optimal model was achieved with the top 10 features on the training set, including 'UniprotDOM_PostModEnz', 'MGAPHC', 'UniprotCARBOHYD', 'UniprotREP', 'ExonSnpDensity', 'ExonConservation', 'UniprotMETAL', 'MGAEntropy', 'ExonHapMapSnpDensity', 'UniprotDOM_MMBRBD'.

Full-size DOI: 10.7717/peerj.17991/fig-4

the density of SNPs and HapMap-verified SNPs in exons, resulted in a notable 4.8% and 4.3% decline in prediction performance, respectively. Although classified as exon features in CRAVAT, these features also relate to evolutionary conservation—a factor significantly influencing cancer driver prediction performance (*Ostroverkhova, Przytycka & Panchenko, 2023*; *Nourbakhsh et al., 2024*; *Rogers, Gaunt & Campbell, 2021*).

Additionally, the feature UniprotMETAL, which relates to the binding of metal ions at mutation sites, is crucial given the role of metal ions as protein cofactors in cellular processes linked to cancer development (*Xu et al., 2016*; *Ge et al., 2022*). Lastly, UniprotREP, which denotes genomic repetitive regions, is highlighted for its potential to induce genomic instability—a hallmark of cancer genomes, thereby strongly correlating with cancer occurrence (*Criscione et al., 2014*; *Liao et al., 2023*). Consequently, we chose XGBoost with the top 10 features and optimal parameters as the final CDMPred model.

Comparison with models trained on class labels using random permutation

To demonstrate that CDMPred acquired knowledge beyond random noise, we trained corresponding models of CDMPred_random. The mean values and standard deviations of

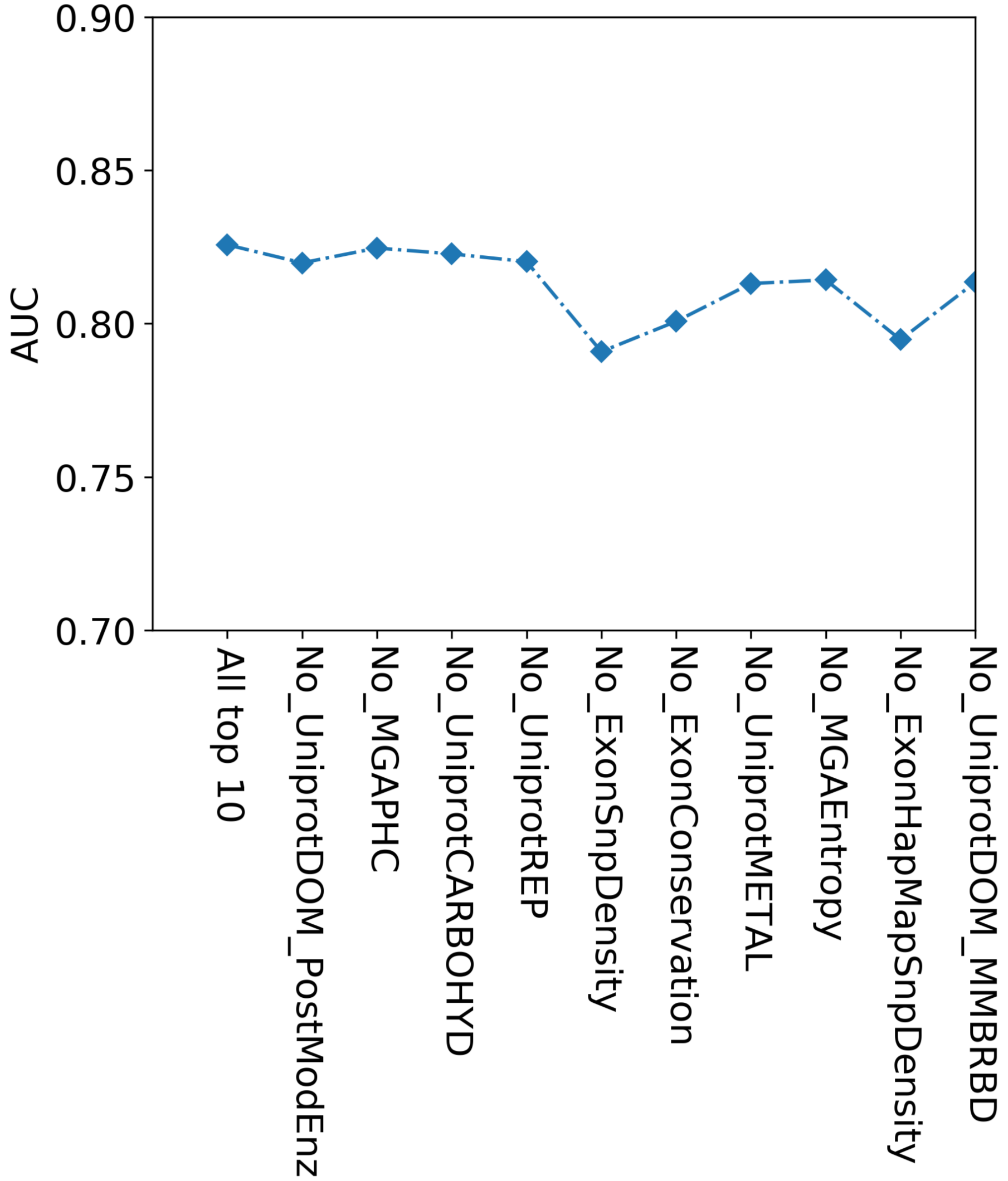

Figure 5 Comparison of performance with the top 10 features and without the top k (one to 10) features by 10-fold cross-validation on the training set.

Full-size DOI: 10.7717/peerj.17991/fig-5

AUC values on the training set with 10-fold cross-validation are shown in Table S3. The results illustrate an AUC value of 0.826 for the original CDMPred model. Nevertheless, the AUC value experienced a significant decrease upon random permutation of class labels for training the CDMPred_random model. Additionally, CDMPred exhibited statistical solid significance (with a p-value <0.001) compared to other models. The computational setup involved a system with 16 GB of memory, an Intel(R) Core (TM) i7-9700 CPU operating at

3.00 GHz with eight cores and running on a 64-bit Windows 10 system. The permutation test incurred a time cost of approximately 966 s.

Performance comparison with state-of-the-art predictors

To evaluate the performance of CDMPred on unseen samples, we assembled an independent test set. We utilized widely recognized tools designed explicitly for cancer-specific and general diseases, including CHASMplus, CHASM, CanDrA, FATHMM, TransFIC, and CScape-somatic. Additionally, we collected ten general disease predictors: SIFT (*Kumar, Henikoff & Ng, 2009*), Mutation Assessor (*Reva, Antipin & Sander, 2011*), PolyPhen-2 (*Adzhubei et al., 2010*), CADD (*Kircher et al., 2014*), MetaLR (*Dong et al., 2015*), MetaSVM (*Dong et al., 2015*), DANN (*Quang, Chen & Xie, 2015*), REVEL (*Ioannidis et al., 2016*), M-CAP (*Jagadeesh et al., 2016*), and MVP (*Qi et al., 2021*). For the cancer-specific methods, we submitted the test data to the respective websites of each tool to obtain the prediction results. As for the general disease predictors, we downloaded the dbNSFP4.1a software (https://sites.google.com/site/jpopgen/dbNSFP) and utilized a script written in Java to retrieve the prediction results from the database (*Liu et al., 2020*). All comparisons were conducted while disregarding any missing values from the tools. Figures 6 and 7 depict the ROC and PR curves, respectively. The results demonstrated that CDMPred exhibited the highest performance in terms of AUC and AUPR. The Delong tests (*De Long, De Long & Clarke-Pearson, 1988*) were conducted to assess whether the CDMPred's performance was significantly different from that of other cancer-specific methods (Table S4) and general-purpose methods (Table S5). The *p*-value of the AUC results indicated that CDMPred exhibited significantly superior performance to all cancer-specific methods and was superior to nine out of ten general-purpose methods, except CADD (*p*-value =0.09677, Delong's test). Furthermore, CDMPred demonstrated strong significance (with a *p*-value <0.001) compared to the other methods. It is worth noting that the AUPR value of CADD is 0.68 while that of CDMPred is 0.80. In total, the performance of CDMPred was robust.

Case study

The principal advantage of our computational approach lies in its ability to significantly broaden the scope of analysis while concurrently preserving efficiency in terms of time and cost. A particularly compelling feature is its potential to inform and direct future experimental research, adeptly pinpointing candidate cancer driver mutations that merit in-depth investigation. In this context, we presented two illustrative cases predicted by CDMPred, juxtaposed with the predictions from several leading-edge methods. These include the cancer driver predictors CHASMplus and CScape-somatic, and the pathogenic missense mutation predictors ESM1b and AlphaMissense.

The kinase insert domain receptor (KDR), a type III receptor tyrosine kinase, is pivotal in mediating proliferation, survival, and migration induced by vascular endothelial growth factor. Its involvement is implicated in several diseases, including lymphoma (*Rotunno et al., 2016*). Experimental evidence has shown that p.A1065T, located within the activation loop, induces constitutive autophosphorylation on tyrosine independent of vascular

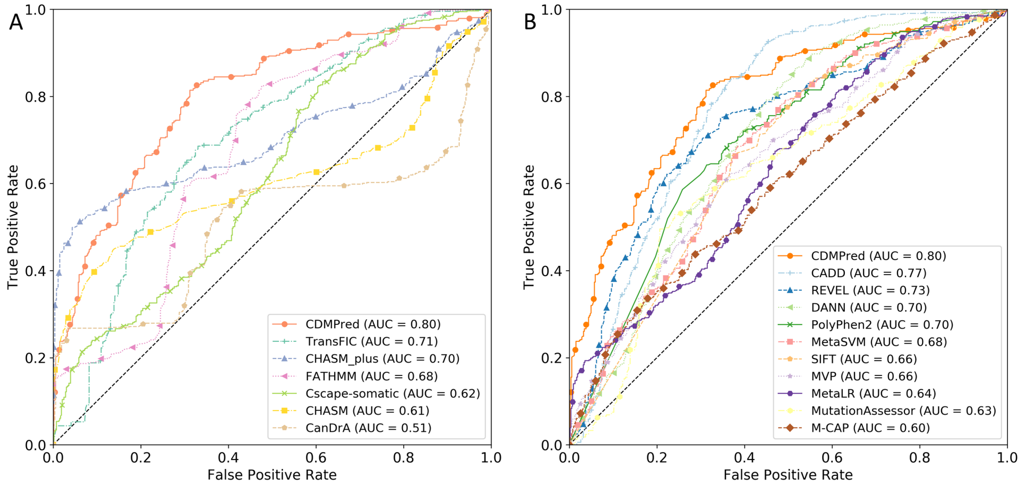

Figure 6 **ROC curves of CDMPred relative to state-of-the-art models on the independent test set.** (A) Comparison of performance between CDMPred and other methods designed to predict single nucleotide driver variants in cancer. (B) Comparison of performance between CDMPred and several general-purpose predictors.

Full-size DOI: 10.7717/peerj.17991/fig-6

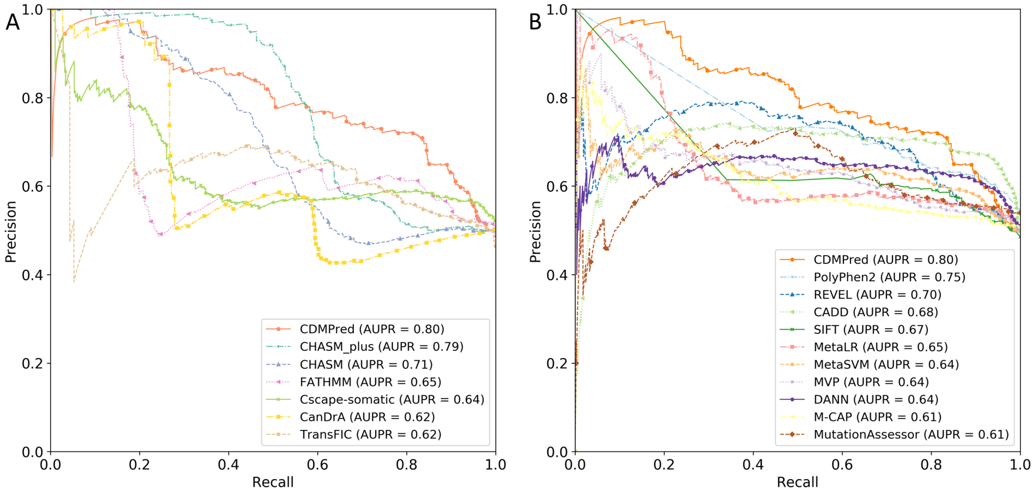

Figure 7 **Precision–recall (PR) curves of CDMPred relative to state-of-the-art models on the independent test set.** (A) Comparison of performance between CDMPred and methods designed to predict single nucleotide driver variants in cancer. (B) Comparison of performance between CDMPred and several general-purpose predictors.

Full-size DOI: 10.7717/peerj.17991/fig-7

endothelial growth factor stimulation. Additionally, kinase inhibitors effectively suppressed its activity (*Rotunno et al., 2016*; *Flerlage et al., 2023*). Our computational approach, CDMPred, precisely identified the KDR-p.A1065T mutation as a significant driver with a high prediction score of 0.824. In stark contrast, the cancer driver predictors CHASMplus and CScape-somatic misclassified it as a passenger mutation, with substantially lower prediction scores of 0.119 and 0.139, respectively. The pathogenic missense mutation predictors ESM1b and AlphaMissense also provided divergent assessments, with ESM1b categorizing it as a tolerated mutation (score = 0.423) and AlphaMissense as a likely benign mutation (score = 0.335).

The Mitogen-Activated Protein Kinase Kinase 1 (MAP2K1) gene encodes MEK1, a pivotal protein kinase in the RAS/MAPK pathway that transduces extracellular chemical signals to the cell nucleus. This signaling pathway regulates fundamental cellular processes such as proliferation, differentiation, migration, and apoptosis. A recent clinical observation identified the p.E120D mutation in a non-small-cell lung cancer patient (*Wang et al., 2021*). CDMPred and CScape-somatic correctly predicted MAP2K1-p.E120D as a significant driver mutation, with prediction scores of 0.810 and 0.742, respectively. Conversely, CHASMplus misclassified this mutation with a borderline score of 0.499, suggesting it was a passenger mutation. Additionally, ESM1b and AlphaMissense provided divergent classifications, with ESM1b scoring it as a tolerated mutation (score = 0.334) and AlphaMissense deeming it an ambiguous mutation (score = 0.366).

DISCUSSION

For cancer-specific methods, TransFIC applied to PolyPhen-2 predictions due to the fewest missing values and achieved the second-highest AUC performance but ranked last in terms of AUPR. The CHASM prediction yielded an AUC of 0.61, sensitivity of 0.74, and specificity of 0.15. Similarly, CanDrA achieved an AUC of 0.51, sensitivity of 0.76, and specificity of 0.07. Therefore, both CHASM and CanDrA exhibited poor performance on the negative samples, indicating a severe imbalance that resulted in significantly low AUC values (as discussed below).

The CHASM training set comprised a balanced collection of positive and negative samples; however, there was only a 0.6% overlap at the transcript level (*Carter et al., 2009*). Therefore, we hypothesized that CHASM might be influenced by type 2 circularity, where the variant status was predominantly predicted based on other variants within the same protein (*Grimm et al., 2015*). As anticipated, 53% of false negatives in the CHASM predictions occurred in transcripts that completely overlapped with positive data in the CHASM training set. In contrast, only 0.9% were found in transcripts that entirely overlapped with negative data in the CHASM training set. Moreover, the opposite was observed for the true negatives of the CHASM predictions, with a higher number of samples found in transcripts that exclusively overlapped with negative data in the CHASM training set. Consequently, CHASM was influenced by type 2 circularity.

CanDrA proposed that driver mutations recurrently occurred in proximity (hotspots) in various types of cancer, whereas passenger mutations were not detected in any Cancer Gene Census (CGC) genes (*Mao et al., 2013*; *Futreal et al., 2004*). Based on our findings, we suspected the presence of type 2 circularity in CanDrA since it adhered to the screening criteria of the training set, resulting in minimal overlap between positive and negative samples at the transcript level. When the genes of the negative sample in the independent test set overlapped with the CGC genes, we identified shared genes in both sets. These genes were absent in the CanDrA training set, and the independent test set consisted of 95%

negative samples, of which only 3% were true negatives. Moreover, the genes exclusively present in the negative samples of the independent test set, which could potentially be the genes corresponding to negative samples in the CanDrA training set, comprised 5% of the negative samples in the independent test set, of which >80% were predicted to be true negatives. Therefore, CanDrA predicted the variant status by relying on other variants within the same protein, indicating the presence of type 2 circularity. We have shown that the low AUC values obtained by both CHASM and CanDrA can be primarily attributed to type 2 circularity. Furthermore, considering the quality of training data, we propose that negative samples used in CHASM and CanDrA fail to represent the broad spectrum of passenger mutations.

CDMPred demonstrated the highest comprehensive predictive capacity among the general-disease deleterious mutation predictors, followed by CADD, Polyphen-2, and REVEL. Interestingly, these methods also surpassed the second-best predictor specific to cancer. PolyPhen-2 achieved an AUPR of 0.75 and a sensitivity of 0.83, indicating a relatively higher predictive ability than CDMPred for positive samples. However, in both the positive and negative samples of the independent test set, numerous predictions made by PolyPhen-2 were classified as "positive", potentially corresponding to a range of diseases rather than solely cancer drivers (*Bertrand et al., 2018*). For instance, one of the true positives predicted by PolyPhen-2, "GATA2:p.R398W", is associated with acute myeloid leukemia and alveolar proteinosis (*Kazenwadel et al., 2012*; *Griese et al., 2015*).

Furthermore, one of the false negatives predicted by PolyPhen-2, "HMBS:p.D359N", is associated not only with cancer but also with acute intermittent porphyria (*Dorschner et al., 2013*; *Lewis, 2006*). Therefore, we directed our attention to the genes corresponding to the true negative and positive categories and the false negative and positive categories in the PolyPhen-2 predictions. We conducted enrichment analysis using the online tool DAVID to validate the suppositions mentioned above (*Huang, Sherman & Lempicki, 2009*). We gathered the pathways exclusively associated with general diseases, excluding cancer, and subsequently calculated the adjusted p-value (<0.05) using the hypergeometric test followed by the Benjamini–Hochberg test. Upon mapping the enrichment results at the mutation level, 65% of the results were associated with diseases present in both the true negatives and true positives of the PolyPhen-2 predictions. In comparison, 54% were associated with diseases present in both the false negatives and false positives of the PolyPhen-2 predictions. In conclusion, these findings support the presence of a systematic bias in driver mutation prediction by PolyPhen-2, even among general disease predictors.

CDMPred utilizes high-quality passenger mutations from dbCPM to distinguish between cancer missense driver mutations and passenger mutations. The results demonstrate that CDMPred achieved superior performance compared to various state-of-the-art methods for cancer-specific and general diseases. While our method offers significant insights, it has limitations. First, the curated datasets exhibit inherent biases, acknowledging that a mutation's role as a driver or passenger mutation can vary with tumor microenvironments, as noted in recent literature (*Ostroverkhova, Przytycka & Panchenko, 2023*; *Wodarz, Newell & Komarova, 2018*). Therefore, this introduces selection and information bias in our supervised learning model. Second, our current method lacks the exploration of advanced machine-learning techniques. Recent studies have demonstrated that deep learning and protein language models could enhance performance in identifying pathogenic missense mutations (*Cheng et al., 2023*; *Schubach et al., 2024*).

CONCLUSIONS

The predictive performance of machine learning methods relies heavily on the quality of the training data. Consequently, including well-defined positive and negative samples of known instances is crucial. This study introduces CDMPred, a novel predictor that distinguishes cancer missense driver mutations from passenger mutations. Specifically, high-quality passenger mutations from dbCPM, chosen for their superior representativeness in modeling the diverse range of passenger mutations, were utilized as negative samples in the training set. The results demonstrated that incorporating high-quality passenger mutations through an ensemble learning method enhanced the accuracy of algorithms in predicting driver mutations in human cancer. In the future, our research will expand to include a broader collection of experimentally verified negative samples and explore the utilization of ensemble deep learning methods further to refine the predictive model (*Xi et al., 2023*; *Deng et al., 2020*).

ACKNOWLEDGEMENTS

The authors thank the members of our laboratory for their valuable contributions to CDMPred and the reviewers for their useful comments.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Natural Science Foundation of China (U22A2038, 82101611, 32271283, and 62102004) and the Natural Science Foundation of the Anhui Higher Education Institutions of China (2023AH051392). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
the National Natural Science Foundation of China: U22A2038, 82101611, 32271283, 62102004.
The Natural Science Foundation of the Anhui Higher Education Institutions of China: 2023AH051392.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Lihua Wang performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

- Haiyang Sun performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Zhenyu Yue analyzed the data, prepared figures and/or tables, and approved the final draft.
- Junfeng Xia conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Xiaoyan Li conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data for training and testing and the code for CDMPred are available in the Supplementary Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.17991#supplemental-information.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* **7(4)**:248–249 DOI 10.1038/nmeth0410-248.

Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Mayank Choudhary NK, McMichael JF, Fulton RS, Wilson RK, Griffith OL, Mardis ER. 2016. DoCM: a database of curated mutations in cancer. *Nature Methods* **13(10)**:806–807 DOI 10.1038/nmeth.4000.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang HZ, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LSL. 2004. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **32**:D115–D119 DOI 10.1093/nar/gkh131.

Bertrand D, Drissler S, Chia BK, Koh JY, Li C, Suphavilai C, Tan IB, Nagarajan N. 2018. ConsensusDriver improves upon individual algorithms for predicting driver alterations in different cancer types and individual patients. *Cancer Research* **78(1)**:290–301 DOI 10.1158/0008-5472.CAN-17-1345.

Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **74(3)**:229–263 DOI 10.3322/caac.21834.

Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M. 2013. Identification of deleterious synonymous variants in human genomes. *Bioinformatics* **29(15)**:1843–1850 DOI 10.1093/bioinformatics/btt308.

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. 2009. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Research* **69**(**16**):6660–6667 DOI 10.1158/0008-5472.CAN-09-1133.

Chen T, Guestrin C. 2016. Xgboost: a scalable tree boosting system. In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 785–794 DOI 10.1145/2939672.2939785.

Cheng J, Novati G, Pan J, Bycroft C, Žemgulyte A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, Schneider RG, Senior AW, Jumper J, Hassabis D, Kohli P, Ž Avsec. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**(**6664**):eadg7492 DOI 10.1126/science.adg7492.

Cheng N, Bi C, Shi Y, Liu M, Cao A, Ren M, Xia J, Liang Z. 2024. Effect predictor of driver synonymous mutations based on multi-feature fusion and iterative feature representation learning. *IEEE Journal of Biomedical and Health Informatics* **28**(**2**):1144–1151 DOI 10.1109/JBHI.2023.3343075.

Cheng N, Li M, Zhao L, Zhang B, Yang Y, Zheng C-H, Xia J. 2020. Comparison and integration of computational methods for deleterious synonymous mutation prediction. *Briefings in Bioinformatics* **21**(**3**):970–981 DOI 10.1093/bib/bbz047.

Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. 2014. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15**(**1**):583 DOI 10.1186/1471-2164-15-583.

De Long ER, De Long DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating ch aracteristic curves: a nonparametric approach. *Biometrics* **44**(**3**):837–845 DOI 10.2307/2531595.

Deng Y, Xu X, Qiu Y, Xia J, Zhang W, Liu S. 2020. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics* **36**(**15**):4316–4322 DOI 10.1093/bioinformatics/btaa501.

Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. 2015. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics* **24**(**8**):2125–2137 DOI 10.1093/hmg/ddu733.

Dorschner MO, Amendola LM, Turner EH, Robertson PD, Shirts BH, Gallego CJ, Bennett RL, Jones KL, Tokita MJ, Bennett JT, Kim JH, Rosenthal EA, Kim DS, National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project, Tabor HK, Bamshad MJ, Motulsky AG, Ronald Scott C, Pritchard CC, Walsh T, Burke W, Raskind WH, Byers P, Hisama FM, Nickerson DA, Jarvik GP. 2013. Actionable, pathogenic incidental findings in 1,000 participants' exomes. *American Journal of Human Genetics* **93**(**4**):631–640 DOI 10.1016/j.ajhg.2013.08.006.

Flerlage JE, Myers JR, Maciaszek JL, Oak N, Rashkin SR, Hui Y, Wang Y-D, Chen W, Wu G, Chang T-C, Hamilton K, Tithi SS, Goldin LR, Rotunno M, Caporaso N, Vogt A, Flamish D, Wyatt K, Liu J, Tucker M, Hahn CN, Brown AL, Scott HS, Mullighan C, Nichols KE, Metzger ML, McMaster ML, Yang JJ, Rampersaud E. 2023. Discovery of novel predisposing coding and noncoding variants in familial Hodgkin lymphoma. *Blood* **141**(**11**):1293–1307 DOI 10.1182/blood.2022016056.

Flicek P, Ridwan Amode M, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, García Girón C, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kähäri AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ruffier M, Sheppard D, Taylor K, Thormann A, Trevanion SJ, Vullo A, Wilder SP, Wilson M, Zadissa A, Aken BL, Birney E, Cunningham F, Harrow J, Herrero J, Hubbard TJP, Kinsella R, Muffato M, Parker A, Spudich G, Yates A, Zerbino DR, Searle SMJ. 2014. Ensembl 2014. *Nucleic Acids Research* **42(D1)**:D749–D755 DOI 10.1093/nar/gkt1196.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. *Nature Reviews Cancer* **4(3)**:177–183 DOI 10.1038/nrc1299.

Ge EJ, Bush AI, Casini A, Cobine PA, Cross JR, De Nicola GM, Dou QPing, Franz KJ, Gohil VM, Gupta S, Kaler SG, Lutsenko S, Mittal V, Petris MJ, Polishchuk R, Ralle M, Schilsky ML, Tonks NK, Vahdat LT, Van Aelst L, Xi D, Yuan P, Brady DC, Chang CJ. 2022. Connecting copper and cancer: from transition metal signalling to metalloplasia. *Nature Reviews Cancer* **22(2)**:102–113 DOI 10.1038/s41568-021-00417-2.

Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. 2012. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Medicine* **4(11)**:89 DOI 10.1186/gm390.

Griese M, Zarbock R, Costabel U, Hildebrandt J, Theegarten D, Albert M, Thiel A, Schams A, Lange J, Krenke K, Wesselak T, Schön C, Kappler M, Blum H, Krebs S, Jung A, Kröner C, Klein C, Campo I, Luisetti M, Bonella F. 2015. GATA2 deficiency in children and adults with severe pulmonary alveolar proteinosis and hematologic disorders. *BMC Pulmonary Medicine* **15(1)**:87 DOI 10.1186/s12890-015-0083-2.

Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Borgwardt KM. 2015. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation* **36(5)**:513–523 DOI 10.1002/humu.22768.

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* **144(5)**:646–674 DOI 10.1016/j.cell.2011.02.013.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4(1)**:44–57 DOI 10.1038/nprot.2008.211.

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Powell IJ, Cussenot O, Cancel-Tassin G, Giles GG, MacInnis RJ, Maier C, Hsieh CL, Wiklund F, Catalona WJ, Foulkes WD, Mandal D, Eeles RA, Kote-Jarai Z, Bustamante CD, Schaid DJ, Hastie T, Ostrander EA, Bailey-Wilson JE, Radivojac P, Thibodeau SN, Whittemore AS, Sieh W. 2016. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *American Journal of Human Genetics* **99(4)**:877–885 DOI 10.1016/j.ajhg.2016.08.016.

Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G. 2016. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics* **48(12)**:1581–1586 DOI 10.1038/ng.3703.

Katzman S, Barrett C, Thiltgen G, Karchin R, Karplus K. 2008. PREDICT-2ND: a tool for generalized protein local structure prediction. *Bioinformatics* **24(21)**:2453–2459 DOI 10.1093/bioinformatics/btn438.

Kazenwadel J, Secker GA, Liu YJ, Rosenfeld JA, Wildin RS, Cuellar-Rodriguez J, Hsu AP, Dyack S, Fernandez CV, Chong C-E, Babic M, Bardy PG, Shimamura A, Zhang MY, Walsh T, Holland SM, Hickstein DD, Horwitz MS, Hahn CN, Scott HS, Harvey NL. 2012. Loss-of-function germline GATA2 mutations in patients with MDS/AML or MonoMAC syndrome and primary lymphedema reveal a key role for GATA2 in the lymphatic vasculature. *Blood* **119(5)**:1283–1291 DOI 10.1182/blood-2011-08-374363.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Research* **12(6)**:996–1006 DOI 10.1101/gr.229102.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46(3)**:310–315 DOI 10.1038/ng.2892.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* **4(7)**:1073–1081 DOI 10.1038/nprot.2009.86.

Lewis PD. 2006. Novel human pathological mutations. *Human Genetics* **118(6)**:359–364.

Liao X, Zhu W, Zhou J, Li H, Xu X, Zhang B, Gao X. 2023. Repetitive DNA sequence detection and its role in the human genome. *Communications Biology* **6(1)**:954 DOI 10.1038/s42003-023-05322-y.

Liu X, Li C, Mou C, Dong Y, Tu Y. 2020. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine* **12(1)**:103 DOI 10.1186/s13073-020-00803-9.

Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K. 2013. CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PLOS ONE* **8(10)**:e77945 DOI 10.1371/journal.pone.0077945.

Masica DL, Douville C, Tokheim C, Bhattacharya R, Kim R, Moad K, Ryan MC, Karchin R. 2017. CRAVAT 4: cancer-related analysis of variants toolkit. *Cancer Research* **77(21)**:e35-e38 DOI 10.1158/1538-7445.AM2017-3538.

Muiños F, Martínez-Jiménez F, Pich O, Gonzalez-Perez A, Lopez-Bigas N. 2021. In silico saturation mutagenesis of cancer genes. *Nature* **596(7872)**:428–432 DOI 10.1038/s41586-021-03771-1.

Nourbakhsh M, Degn K, Saksager A, Tiberti M, Papaleo E. 2024. Prediction of cancer driver genes and mutations: the potential of integrative computational frameworks. *Briefings in Bioinformatics* **25**(**2**):1–16.

Ostroverkhova D, Przytycka TM, Panchenko AR. 2023. Cancer driver mutations: predictions and reality. *Trends in Molecular Medicine* **29**(**7**):554–566 DOI 10.1016/j.molmed.2023.03.007.

Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, Guan Y, Shen Y. 2021. MVP predicts the pathogenicity of missense variants by deep learning. *Nature Communications* **12**(**1**):510 DOI 10.1038/s41467-020-20847-0.

Quang D, Chen Y, Xie X. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**(**5**):761–763 DOI 10.1093/bioinformatics/btu703.

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research* **39**(**17**):e118 DOI 10.1093/nar/gkr407.

Ribeiro GRH, Francisco G, Teixeira LVS, Romao-Correia RF, Sanches JA, Neto CF, Ruiz IRG. 2004. Repetitive DNA alterations in human skin cancers. *Journal of Dermatological Science* **36**(**2**):79–86 DOI 10.1016/j.jdermsci.2004.08.003.

Rogers MF, Gaunt TR, Campbell C. 2020. CScape-somatic: distinguishing driver and passenger point mutations in the cancer genome. *Bioinformatics* **36**(**12**):3637–3644 DOI 10.1093/bioinformatics/btaa242.

Rogers MF, Gaunt TR, Campbell C. 2021. Prediction of driver variants in the cancer genome via machine learning methodologies. *Briefings in Bioinformatics* **22**(**4**):bbaa250 DOI 10.1093/bib/bbaa250.

Rotunno M, McMaster ML, Boland J, Bass S, Zhang X, Burdett L, Hicks B, Ravichandran S, Luke BT, Yeager M, Fontaine L, Hyland PL, Goldstein AM, Chanock SJ, Caporaso NE, Tucker MA, Goldin LR. 2016. Whole exome sequencing in families at high risk for Hodgkin lymphoma: identification of a predisposing mutation in the KDR gene. *Haematologica* **101**(**7**):853–860 DOI 10.3324/haematol.2015.135475.

Schubach M, Maass T, Nazaretyan L, Röner S, Kircher M. 2024. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Research* **52**(**D1**):D1143–D1154 DOI 10.1093/nar/gkad989.

Shen Q, Cheng F, Song H, Lu W, Zhao J, An X, Liu M, Chen G, Zhao Z, Zhang J. 2017. Proteome-scale investigation of protein allosteric regulation perturbed by somatic mutations in 7000 cancer genomes. *The American Journal of Human Genetics* **100**(**1**):5–20 DOI 10.1016/j.ajhg.2016.09.020.

Shihab HA, Gough J, Cooper DN, Day IN, Gaunt TR. 2013. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* **29**(**12**):1504–1510 DOI 10.1093/bioinformatics/btt182.

Song K, Li Q, Gao W, Lu S, Shen Q, Liu X, Wu Y, Wang B, Lin H, Chen G, Zhang J. 2019. AlloDriver: a method for the identification and analysis of cancer driver targets. *Nucleic Acids Research* **47**(**W1**):W315–W321 DOI 10.1093/nar/gkz350.

Song Q, Li M, Li Q, Lu X, Song K, Zhang Z, Wei J, Zhang L, Wei J, Ye Y, Zha J, Zhang Q, Gao Q, Long J, Liu X, Lu X, Zhang J. 2023. DeepAlloDriver: a deep learning-based strategy to predict cancer driver mutations. *Nucleic Acids Research* **51(W1)**:W129–W133 DOI 10.1093/nar/gkad295.

Tokheim C, Karchin R. 2019. CHASMplus reveals the scope of somatic missense mutations driving human cancers. *Cell Systems* **9(1)**:9–23.e8 DOI 10.1016/j.cels.2019.05.005.

Tomasetti C, Marchionni L, Nowak MA, Parmigiani G, Vogelstein B. 2015. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America* **112(1)**:118–123.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. 2013. Cancer genome landscapes. *Science* **339(6127)**:1546–1558 DOI 10.1126/science.1235122.

Wang W, Xu C, Wang D, Zhu Y, Zhuang W, Fang M, Lv T, Song Y. 2021. P70.05 The association between MAP2K1 mutation class and clinical features in MAP2K1-mutant east asian non-small cell lung cancer patients. *Journal of Thoracic Oncology* **16(3)**:S564 DOI 10.1016/j.jtho.2021.01.1016.

Wodarz D, Newell AC, Komarova NL. 2018. Passenger mutations can accelerate tumour suppressor gene inactivation in cancer evolution. *Journal of the Royal Society, Interface* **15(143)**:20170967 DOI 10.1098/rsif.2017.0967.

Won D-G, Kim D-W, Woo J, Lee K, Marschall T. 2021. 3Cnet: pathogenicity prediction of human variants using multitask learning with evolutionary constraints. *Bioinformatics* **37(24)**:4626–4634 DOI 10.1093/bioinformatics/btab529.

Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. 2011. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* **27(15)**:2147–2148 DOI 10.1093/bioinformatics/btr357.

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson KVJ, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PVK, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* **318(5853)**:1108–1113 DOI 10.1126/science.1145720.

Xi J, Sun D, Chang C, Zhou S, Huang Q. 2023. An omics-to-omics joint knowledge association subtensor model for radiogenomics cross-modal modules from genomics and ultrasonic images of breast cancers. *Computers in Biology and Medicine* **155**:106672 DOI 10.1016/j.compbiomed.2023.106672.

Xi J, Yuan X, Wang M, Li A, Li X, Huang Q. 2020. Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* **36(6)**:1855–1863 DOI 10.1093/bioinformatics/btz793.

Xu D, Jalal SI, Sledge GW, Meroueh SO. 2016. Small-molecule binding sites to explore protein–protein interactions in the cancer proteome. *Molecular Biosystems* **12**(**10**):3067–3087 DOI 10.1039/C6MB00231E.

Yue Z, Zhao L, Xia J. 2020. dbCPM: a manually curated database for exploring the cancer passenger mutations. *Briefings in Bioinformatics* **21**(**1**):309–317 DOI 10.1093/bib/bby105.

Zeng X, Liu L, Lü L, Zou Q. 2018. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* **34**(**14**):2425–2432 DOI 10.1093/bioinformatics/bty112.