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], 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

Corresponding Author:

Xiaoyan Li [1]

No.111 Jiulong Road, Hefei, Anhui, 230601, China

Email address: lixiaoyan@ahu.edu.cn

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.

Deleted: uncertain.

Deleted:

Deleted: both

Deleted: The inclusion of

Deleted: as

Keywords

Cancer; Machine learning; Driver missense mutation prediction; Benchmark quality; XGBoost

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 [1]. Researchers have confirmed that cancer is a multifaceted genetic disease caused by the accumulation of numerous mutations in the genome [2-4]. However, the tumorigenesis and development of most cancers are primarily driven by a small number of critical mutations [5-7], while the remaining mutations are 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 [8, 9].

Deleted: key

Deleted: The identification of

Deleted: , as well as

Vogelstein et al. observed that most protein-coding mutations in cancer genomes were missense changes [10]. 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 [6], Cancer-specific High-

Deleted: the majority of

throughput Annotation of Somatic Mutations (CHASM) [11], Transformed Functional Impact score for Cancer (transFIC) [12], Cancer Driver Annotation (CanDrA) [13], Functional Analysis through Hidden Markov Models (FATHMM) [14], CScape-somatic [15], and CHASMplus [16]. Additionally, some methods are focused on identifying driver mutations at critical sites, such as protein allosteric sites [17-19]. These methods typically utilize positive samples obtained from cancer-related databases, such as the Catalogue of Somatic Mutations in Cancer (COSMIC) database [11-15], while negative samples are commonly derived from statistical methods [6, 11] or putative samples [12-16].

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 [20]. The results indicated that the dbCPM data aligns with other negative datasets regarding most classical features while exhibiting specificity for cancer-related features [20, 21]. 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).

Deleted: In this

Deleted: , we

Deleted: in terms of

Deleted: ,

Deleted: aforementioned

Deleted: in conjunction

Materials & Methods

Figure 1 presents the flowchart of the CDMPred method. The procedure consists of four steps: dataset preparation, feature representation, model construction, and performance evaluation. Each step is explained in detail below.

Deleted: The

Deleted: is presented in Figure 1

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) [22], which is a reliable source that aggregates functionally validated mutations in cancer. For negative samples, we gathered three passenger

Deleted: There were one

Formatted: Font color: Auto

datasets. The dataset dbCPM (v1.1) consists of 1,919 passenger mutations, including 1,634 distinct missense mutations [20]. 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 [23, 24], and the dataset CHASM (v3.1) consists of synthetic passenger mutations in the CHASM training set [11]. We removed the mutations simultaneously in DoCM in each passenger mutation dataset. The details are presented in Additional file 1: 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 harmful 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 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 [25], 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 [25, 26] from a dockerized tool, the Cancer-Related Analysis of Variants Toolkit [27] (CRAVAT, version 5.2.3). To prepare the input data, we curated the transcript information using Ensembl GRCh37 [28] 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 [19, 20], 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.

Deleted: 919

Deleted: Specifically

Deleted: originally

Deleted: were

Deleted: further

Deleted: which were

Deleted: detail is

Deleted: negative

Deleted: To

Deleted: , we obtained an independent test set

Deleted: Subsequent to

Deleted: applying

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 Additional File 1: Table S2.

Deleted: And the Amino

Deleted: file

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 [29]. 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 [30]. 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 (maxDepth), and the subsample ratio of columns when constructing each tree (colsample_bytree).

Deleted: conducted a comprehensive evaluation of

Deleted: aforementioned

Deleted: key

Performance evaluation

As quantitative measurements of prediction results, we employed two threshold-independent measures: AUC and AUPR [11, 31]. Additionally, we used two qualitative measures, namely sensitivity (or actual positive rate) and specificity (true negative rate), for model performance analysis, as previously described in research [32, 33]. These measures are defined as follows:

Deleted: employed

Deleted: true

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

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

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.

Deleted: where

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

Deleted: just

Deleted: in terms of

repeated the second and the third steps 1,000 times. Finally, we calculated the empirical *p*-value by comparing the distribution of the 1000 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 Additional file 1: Table S1, which comprehensively represents the biological impacts of the mutation in the human genome [21]. 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, 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) [34, 35]. Subsequently, we identified three features based on their P-values, and the RMS score distribution of all samples is presented in Figure 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) [36, 37]. 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 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 Additional file 1: Figure S1. Subsequently, we utilized the feature selection method with XGBoost (using default parameters) to determine the importance of the features. We employed sequential

Deleted: 1000
Formatted: Font: Italic
Deleted: represent
Deleted: Using these features, we
Deleted: samples from
Deleted: the notion
Deleted: In order to
Deleted: and

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 Additional file 1: 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 (Figure 5). The results indicate that excluding features like ExonSnpDensity and ExonHapMapSnpDensity, which quantify 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 [7, 38, 39].

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 [40, 41]. 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 [42, 43]. 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 AUC values on the training set with 10-fold cross-validation are shown in Additional file 1: 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 3.00GHz with eight cores and running on a 64-bit Windows 10 system. The permutation test incurred a time cost of approximately 966 seconds.

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 [44], Mutation Assessor [45], PolyPhen-2 [46], CADD [47], MetaLR [31], MetaSVM [31], DANN [48], REVEL [49], M-CAP [50], and MVP [51]. For the

Deleted: utilized

Deleted: that

Deleted: influence

Deleted: the feature

Deleted: Significantly, the

Deleted: strong

Formatted: Font: Italic

Deleted: 8

Deleted: ,

Deleted: In order to

Deleted: several

Deleted: specifically

Deleted: 10

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 Javascript to retrieve the prediction results from the database [52]. All comparisons were conducted while disregarding any missing values from the tools. Figure 6 and Figure 7 depict the ROC and PR curves, respectively. The results demonstrated that CDMPred exhibited the highest performance in terms of AUROC and AUPR. The Delong tests [53] were conducted to assess whether the CDMPred's performance was significantly different from that of other cancer-specific methods (Additional file 1: Table S4) and general-purpose methods (Additional file 1: 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.

Deleted: website

Deleted: https://sites.google.com/site/jpopgen/dbNSFP) and utilized a Java script

Deleted: both

Deleted: compared

Formatted: Font: Italic

Formatted: Font: Italic

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.

Deleted: , as well as

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 [54]. Experimental evidence has shown that p.A1065T, located within the activation loop, induces constitutive autophosphorylation on tyrosine independent of vascular endothelial growth factor stimulation. Additionally, kinase inhibitors effectively suppressed its activity [54, 55]. 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

Deleted: plays a

Deleted: role

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 [56]. 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).

Deleted: as

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 sixth 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).

Deleted: ,

Deleted: ,

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 [11]. 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 [25]. 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 harmful data in the CHASM training set. Moreover, the opposite was observed for the actual 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.

Deleted: , whereas

Deleted: fully

Deleted: negative

Deleted: true

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 [13, 57]. 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 [58]. For instance, one of the true positives predicted by PolyPhen-2, “GATA2:p.R398W”, is associated with acute myeloid leukemia and alveolar proteinosis [59, 60].

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 [61, 62]. Therefore, we directed our attention to the genes corresponding to the actual 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 [63]. 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.

Deleted: not present

Deleted: wide

Deleted: Among the general-disease deleterious mutation predictors, …

Deleted: high

Deleted: not only

Deleted: but also with

Deleted: true

Deleted: , as well as

Deleted: ,

Deleted: aforementioned

Deleted: , while

Deleted: provide

Deleted: for

Deleted: -

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 [7, 64]. 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 [65, 66].

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 [67, 68].

Acknowledgments

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

References

1. Bray, Freddie, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal, *Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: A Cancer Journal for Clinicians, 2024. **74**(3): p. 229-263.

Deleted: both

Deleted: is not without

Deleted:

Deleted: offer enhanced

Deleted: it is crucial to include

Deleted: consisting

Deleted: designed to distinguish

Deleted: and

Deleted: to

Deleted: Acknowledgements¶

Deleted: valuable

2. Wood, L. D., D. W. Parsons, S. Jones, J. Lin, T. Sjoblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. V. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. K. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, and B. Vogelstein, *The Genomic Landscapes of Human Breast and Colorectal Cancers.* Science, 2007. **318**(5853): p. 1108-1113.

3. Tomasetti, Cristian, Luigi Marchionni, Martin A. Nowak, Giovanni Parmigiani, and Bert Vogelstein, *Only three driver gene mutations are required to develop lung and colorectal cancers.* Proceedings of the National Academy of Sciences, 2015. **112**(1): p. 118-123.

4. Xi, Jianing, Xiguo Yuan, Minghui Wang, Ao Li, Xuelong Li, and Qinghua Huang, *Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication.* Bioinformatics, 2019. **36**(6): p. 1855-1863.

5. Hanahan, Douglas and Robert A. Weinberg, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. **144**(5): p. 646-674.

6. Muiños, Ferran, Francisco Martínez-Jiménez, Oriol Pich, Abel Gonzalez-Perez, and Nuria Lopez-Bigas, *In silico saturation mutagenesis of cancer genes.* Nature, 2021. **596**(7872): p. 428-432.

7. Ostroverkhova, D., T. M. Przytycka, and A. R. Panchenko, *Cancer driver mutations: predictions and reality.* Trends in Molecular Medicine, 2023. **29**(7): p. 554-566.

8. Xi, Jianing, Xiguo Yuan, Minghui Wang, Ao Li, Xuelong Li, and Qinghua Huang, *Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication.* Bioinformatics, 2020. **36**(6): p. 1855-1863.

9. Cheng, Na, Chuanmei Bi, Yong Shi, Mengya Liu, Anqi Cao, Mengkun Ren, Junfeng Xia, and Zhen Liang, *Effect Predictor of Driver Synonymous Mutations Based on Multi-Feature Fusion*

Deleted: *for the development of*

and Iterative Feature Representation Learning. IEEE Journal of Biomedical and Health Informatics, 2024. **28**(2): p. 1144-1151.

10. Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, Jr., and Kenneth W. Kinzler, *Cancer Genome Landscapes.* Science, 2013. **339**(6127): p. 1546-1558.

11. Carter, H., S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and R. Karchin, *Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.* Cancer Research, 2009. **69**(16): p. 6660-6667.

12. Gonzalez-Perez, Abel, Jordi Deu-Pons, and Nuria Lopez-Bigas, *Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation.* Genome Medicine, 2012. **4**(11): p. 89.

13. Mao, Y., H. Chen, H. Liang, F. Meric-Bernstam, G. B. Mills, and K. Chen, *CanDrA: cancer-specific driver missense mutation annotation with optimized features.* PloS One, 2013. **8**(10): p. e77945.

14. Shihab, H. A., J. Gough, D. N. Cooper, I. N. Day, and T. R. Gaunt, *Predicting the functional consequences of cancer-associated amino acid substitutions.* Bioinformatics, 2013. **29**(12): p. 1504-1510.

15. Rogers, Mark F., Tom R. Gaunt, and Colin Campbell, *CScape-somatic: distinguishing driver and passenger point mutations in the cancer genome.* Bioinformatics, 2020. **36**(12): p. 3637-3644.

16. Tokheim, Collin and Rachel Karchin, *CHASMplus Reveals the Scope of Somatic Missense Mutations Driving Human Cancers.* Cell Systems, 2019. **9**(1): p. 9-23.e8.

17. Song, Kun, Qian Li, Wei Gao, Shaoyong Lu, Qiancheng Shen, Xinyi Liu, Yongyan Wu, Binquan Wang, Houwen Lin, Guoqiang Chen, and Jian Zhang, *AlloDriver: a method for the identification and analysis of cancer driver targets.* Nucleic Acids Research, 2019. **47**(W1): p. W315-W321.

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

19. Shen, Qiancheng, Feixiong Cheng, Huili Song, Weiqiang Lu, Junfei Zhao, Xiaoli An, Mingyao Liu, Guoqiang Chen, Zhongming Zhao, and Jian Zhang, *Proteome-Scale Investigation of Protein Allosteric Regulation Perturbed by Somatic Mutations in 7,000 Cancer Genomes.* The American Journal of Human Genetics, 2017. **100**(1): p. 5-20.

20. Yue, Zhenyu, Le Zhao, and Junfeng Xia, *dbCPM: a manually curated database for exploring the cancer passenger mutations.* Briefings in Bioinformatics, 2020. **21**(1): p. 309–317.

21. Wong, Wing Chung, Dewey Kim, Hannah Carter, Mark Diekhans, Michael C. Ryan, and Rachel Karchin, *CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer.* Bioinformatics, 2011. **27**(15): p. 2147-2148.

22. Ainscough, Benjamin J., Malachi Griffith, Adam C. Coffman, Alex H. Wagner, Jason Kunisaki, Mayank N. K. Choudhary, Joshua F. McMichael, Robert S. Fulton, Richard K. Wilson, Obi L. Griffith, and Elaine R. Mardis, *DoCM: a database of curated mutations in cancer.* Nature Methods, 2016. **13**(10): p. 806-807.

23. Shihab, Hashem A., Julian Gough, David N. Cooper, Ian N. M. Day, and Tom R. Gaunt, *Predicting the functional consequences of cancer-associated amino acid substitutions.* Bioinformatics, 2013. **29**(12): p. 1504-1510.

24. Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Z. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. L. Yeh, *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Research, 2004. **32**: p. D115-D119.

25. Grimm, D. G., C. A. Azencott, F. Aicheler, U. Gieraths, D. G. MacArthur, K. E. Samocha, D. N. Cooper, P. D. Stenson, M. J. Daly, J. W. Smoller, L. E. Duncan, and K. M. Borgwardt, *The*

evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Human Mutation, 2015. **36**(5): p. 513-523.

26. Won, Dhong-Gun, Dong-Wook Kim, Junwoo Woo, Kyoungyeul Lee, and Tobias Marschall, *3Cnet: pathogenicity prediction of human variants using multitask learning with evolutionary constraints.* Bioinformatics, 2021. **37**(24): p. 4626-4634.

27. Masica, David L., Christopher Douville, Collin Tokheim, Rohit Bhattacharya, Ryangguk Kim, Kyle Moad, Michael C. Ryan, and Rachel Karchin, *CRAVAT 4: Cancer-Related Analysis of Variants Toolkit.* Cancer Research, 2017. **77**(21): p. e35-e38.

28. Flicek, Paul, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Eugene Kulesha, Fergal J. Martin, Thomas Maurel, William M. McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet S. Riat, Magali Ruffier, Daniel Sheppard, Kieron Taylor, Anja Thormann, Stephen J. Trevanion, Alessandro Vullo, Steven P. Wilder, Mark Wilson, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Jennifer Harrow, Javier Herrero, Tim J.P. Hubbard, Rhoda Kinsella, Matthieu Muffato, Anne Parker, Giulietta Spudich, Andy Yates, Daniel R. Zerbino, and Stephen M.J. Searle, *Ensembl 2014.* Nucleic Acids Research, 2014. **42**(D1): p. D749-D755.

29. Buske, Orion J., Ashokkumar Manickaraj, Seema Mital, Peter N. Ray, and Michael Brudno, *Identification of deleterious synonymous variants in human genomes.* Bioinformatics, 2013. **29**(15): p. 1843-1850.

30. Chen, Tianqi and Carlos Guestrin, *Xgboost: A scalable tree boosting system[C].* ACM, 2016: p. 785-794.

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

32. Cheng, Na, Menglu Li, Le Zhao, Bo Zhang, Yuhua Yang, Chun-Hou Zheng, and Junfeng Xia, *Comparison and integration of computational methods for deleterious synonymous mutation prediction.* Briefings in Bioinformatics, 2020. **21**(3): p. 970-981.

33. Zeng, Xiangxiang, Li Liu, Linyuan Lü, and Quan Zou, *Prediction of potential disease-associated microRNAs using structural perturbation method.* Bioinformatics, 2018. **34**(14): p. 2425-2432.

34. Kent, W. James, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler, *The human genome browser at UCSC.* Genome research, 2002. **12**(6): p. 996-1006.

35. Katzman, Sol, Christian Barrett, Grant Thiltgen, Rachel Karchin, and Kevin Karplus, *PREDICT-2ND: a tool for generalized protein local structure prediction.* Bioinformatics, 2008. **24**(21): p. 2453-2459.

36. Ribeiro, G. R. H., G. Francisco, L. V. S. Teixeira, R. F. Romao-Correia, J. A. Sanches, C. F. Neto, and I. R. G. Ruiz, *Repetitive DNA alterations in human skin cancers.* Journal of Dermatological Science, 2004. **36**(2): p. 79-86.

37. Xu, David, Shadia I. Jalal, George W. Sledge, Jr., and Samy O. Meroueh, *Small-molecule binding sites to explore protein-protein interactions in the cancer proteome.* Molecular Biosystems, 2016. **12**(10): p. 3067-3087.

38. Nourbakhsh, Mona, Kristine Degn, Astrid Saksager, Matteo Tiberti, and Elena Papaleo, *Prediction of cancer driver genes and mutations: the potential of integrative computational frameworks.* Briefings in Bioinformatics, 2024. **25**(2): p. 1-16.

39. Rogers, Mark F., Tom R. Gaunt, and Colin Campbell, *Prediction of driver variants in the cancer genome via machine learning methodologies.* Briefings in Bioinformatics, 2021. **22**(4): p. bbaa250.

40. Xu, David, Shadia I. Jalal, George W. Sledge, and Samy O. Meroueh, *Small-molecule binding sites to explore protein–protein interactions in the cancer proteome.* Molecular Biosystems, 2016. **12**(10): p. 3067-3087.

41. Ge, Eva J., Ashley I. Bush, Angela Casini, Paul A. Cobine, Justin R. Cross, Gina M. DeNicola, Q. Ping Dou, Katherine J. Franz, Vishal M. Gohil, Sanjeev Gupta, Stephen G. Kaler, Svetlana Lutsenko, Vivek Mittal, Michael J. Petris, Roman Polishchuk, Martina Ralle, Michael L. Schilsky, Nicholas K. Tonks, Linda T. Vahdat, Linda Van Aelst, Dan Xi, Peng Yuan, Donita C. Brady, and Christopher J. Chang, *Connecting copper and cancer: from transition metal signalling to metalloplasia.* Nature Reviews Cancer, 2022. **22**(2): p. 102-113.

42. Criscione, Steven W., Yue Zhang, William Thompson, John M. Sedivy, and Nicola Neretti, *Transcriptional landscape of repetitive elements in normal and cancer human cells.* BMC Genomics, 2014. **15**(1): p. 583.

43. Liao, Xingyu, Wufei Zhu, Juexiao Zhou, Haoyang Li, Xiaopeng Xu, Bin Zhang, and Xin Gao, *Repetitive DNA sequence detection and its role in the human genome.* Communications Biology, 2023. **6**(1): p. 954.

44. Kumar, P., S. Henikoff, and P. C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nature Protocols, 2009. **4**(7): p. 1073-1081.

45. Reva, B., Y. Antipin, and C. Sander, *Predicting the functional impact of protein mutations: application to cancer genomics.* Nucleic Acids Research, 2011. **39**(17): p. e118.

46. Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, *A method and server for predicting damaging missense mutations.* Nature Methods, 2010. **7**(4): p. 248-249.

47. Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, *A general framework for estimating the relative pathogenicity of human genetic variants.* Nature Genetics, 2014. **46**(3): p. 310-315.

48. Quang, Daniel, Yifei Chen, and Xiaohui Xie, *DANN: a deep learning approach for annotating the pathogenicity of genetic variants.* Bioinformatics, 2015. **31**(5): p. 761-763.

49. Ioannidis, N. M., J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, L. A. Cannon-Albright, C. C. Teerlink, J. L. Stanford, W. B. Isaacs, J. Xu, K. A. Cooney, E. M. Lange, J. Schleutker, J. D. Carpten, I. J. Powell, O. Cussenot,

G. Cancel-Tassin, G. G. Giles, R. J. MacInnis, C. Maier, C. L. Hsieh, F. Wiklund, W. J. Catalona, W. D. Foulkes, D. Mandal, R. A. Eeles, Z. Kote-Jarai, C. D. Bustamante, D. J. Schaid, T. Hastie, E. A. Ostrander, J. E. Bailey-Wilson, P. Radivojac, S. N. Thibodeau, A. S. Whittemore, and W. Sieh, *REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants.* American Journal of Human Genetics, 2016. **99**(4): p. 877-885.

50. Jagadeesh, K. A., A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, J. A. Bernstein, and G. Bejerano, *M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity.* Nature Genetics, 2016. **48**(12): p. 1581-1586.

51. Qi, Hongjian, Haicang Zhang, Yige Zhao, Chen Chen, John J. Long, Wendy K. Chung, Yongtao Guan, and Yufeng Shen, *MVP predicts the pathogenicity of missense variants by deep learning.* Nature Communications, 2021. **12**(1).

52. Liu, Xiaoming, Chang Li, Chengcheng Mou, Yibo Dong, and Yicheng Tu, *dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs.* Genome Medicine, 2020. **12**(1).

53. DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating ch aracteristic curves: a nonparametric approach.* Biometrics, 1988. **44**(3): p. 837-845.

54. Rotunno, M., M. L. McMaster, J. Boland, S. Bass, X. Zhang, L. Burdett, B. Hicks, S. Ravichandran, B. T. Luke, M. Yeager, L. Fontaine, P. L. Hyland, A. M. Goldstein, S. J. Chanock, N. E. Caporaso, M. A. Tucker, and L. R. Goldin, *Whole exome sequencing in families at high risk for Hodgkin lymphoma: identification of a predisposing mutation in the KDR gene.* Haematologica, 2016. **101**(7): p. 853-860.

55. Flerlage, Jamie E., Jason R. Myers, Jamie L. Maciaszek, Ninad Oak, Sara R. Rashkin, Yawei Hui, Yong-Dong Wang, Wenan Chen, Gang Wu, Ti-Cheng Chang, Kayla Hamilton, Saima S. Tithi, Lynn R. Goldin, Melissa Rotunno, Neil Caporaso, Aurélie Vogt, Deborah Flamish, Kathleen Wyatt, Jia Liu, Margaret Tucker, Christopher N. Hahn, Anna L. Brown, Hamish S. Scott, Charles Mullighan, Kim E. Nichols, Monika L. Metzger, Mary L. McMaster, Jun J. Yang,

and Evadnie Rampersaud, *Discovery of novel predisposing coding and noncoding variants in familial Hodgkin lymphoma.* Blood, 2023. **141**(11): p. 1293-1307.

56. Wang, W., C. Xu, D. Wang, Y. Zhu, W. Zhuang, M. Fang, T. Lv, and Y. Song, *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, 2021. **16**(3): p. S564.

57. Futreal, P. A., L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, *A census of human cancer genes.* Nature Reviews Cancer, 2004. **4**(3): p. 177-183.

58. Bertrand, D., S. Drissler, B. K. Chia, J. Y. Koh, C. Li, C. Suphavilai, I. B. Tan, and N. Nagarajan, *ConsensusDriver Improves upon Individual Algorithms for Predicting Driver Alterations in Different Cancer Types and Individual Patients.* Cancer Research, 2018. **78**(1): p. 290-301.

59. Kazenwadel, Jan, Genevieve A. Secker, Yajuan J. Liu, Jill A. Rosenfeld, Robert S. Wildin, Jennifer Cuellar-Rodriguez, Amy P. Hsu, Sarah Dyack, Conrad V. Fernandez, Chan-Eng Chong, Milena Babic, Peter G. Bardy, Akiko Shimamura, Michael Y. Zhang, Tom Walsh, Steven M. Holland, Dennis D. Hickstein, Marshall S. Horwitz, Christopher N. Hahn, Hamish S. Scott, and Natasha L. Harvey, *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, 2012. **119**(5): p. 1283-1291.

60. Griese, Matthias, Ralf Zarbock, Ulrich Costabel, Jenna Hildebrandt, Dirk Theegarten, Michael Albert, Antonia Thiel, Andrea Schams, Joanna Lange, Katazyrna Krenke, Traudl Wesselak, Carola Schön, Matthias Kappler, Helmut Blum, Stefan Krebs, Andreas Jung, Carolin Kröner, Christoph Klein, Ilaria Campo, Maurizio Luisetti, and Francesco Bonella, *GATA2 deficiency in children and adults with severe pulmonary alveolar proteinosis and hematologic disorders.* BMC Pulmonary Medicine, 2015. **15**(1): p. 87.

61. Dorschner, Michael O., Laura M. Amendola, Emily H. Turner, Peggy D. Robertson, Brian H. Shirts, Carlos J. Gallego, Robin L. Bennett, Kelly L. Jones, Mari J. Tokita, James T. Bennett, Jerry H. Kim, Elisabeth A. Rosenthal, Daniel S. Kim, Holly K. Tabor, Michael J. Bamshad,

Arno G. Motulsky, C. Ronald Scott, Colin C. Pritchard, Tom Walsh, Wylie Burke, Wendy H. Raskind, Peter Byers, Fuld M. Hisama, Deborah A. Nickerson, Gail P. Jarvik, Lung Natl Heart, Inst Blood, and Sequencing Grand Opportunity Exome, *Actionable, Pathogenic Incidental Findings in 1,000 Participants' Exomes.* American Journal of Human Genetics, 2013. **93**(4): p. 631-640.

62. Lewis, P. D., *Novel human pathological mutations.* Human Genetics, 2006. **118**(6): p. 359-364.

63. Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 2009. **4**(1): p. 44-57.

64. Wodarz, D., A. C. Newell, and N. L. Komarova, *Passenger mutations can accelerate tumour suppressor gene inactivation in cancer evolution.* Journal of the Royal Society, Interface, 2018. **15**(143).

65. Cheng, Jun, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec, *Accurate proteome-wide missense variant effect prediction with AlphaMissense.* Science, 2023. **381**(6664): p. eadg7492.

66. Schubach, Max, Thorben Maass, Lusiné Nazaretyan, Sebastian Röner, and Martin Kircher, *CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions.* Nucleic Acids Research, 2024. **52**(D1): p. D1143-D1154.

67. Xi, Jianing, Donghui Sun, Cai Chang, Shichong Zhou, and Qinghua Huang, *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, 2023. **155**: p. 106672.

68. Deng, Yifan, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, and Shichao Liu, *A multimodal deep learning framework for predicting drug–drug interaction events.* Bioinformatics, 2020. **36**(15): p. 4316-4322.