BMC Bioinformatics

RESEARCH

Open Access

Network-based prediction approach for cancer-specific driver missense mutations using a graph neural network

Narumi Hatano[1], Mayumi Kamada[1*], Ryosuke Kojima[1] and Yasushi Okuno[1,2*]

*Correspondence:
kamada.mayumi.2c@kyoto-u.ac.jp;
okuno.yasushi.4c@kyoto-u.ac.jp

[1] Graduate School of Medicine, Kyoto University, Kyoto, Japan
[2] HPC- and AI-driven Drug Development Platform Division, RIKEN Center for Computational Science(R-CCS), Kobe, Japan

Abstract

Background: In cancer genomic medicine, finding driver mutations involved in cancer development and tumor growth is crucial. Machine-learning methods to predict driver missense mutations have been developed because variants are frequently detected by genomic sequencing. However, even though the abnormalities in molecular networks are associated with cancer, many of these methods focus on individual variants and do not consider molecular networks. Here we propose a new network-based method, Net-DMPred, to predict driver missense mutations considering molecular networks. Net-DMPred consists of the graph part and the prediction part. In the graph part, molecular networks are learned by a graph neural network (GNN). The prediction part learns whether variants are driver variants using features of individual variants combined with the graph features learned in the graph part.

Results: Net-DMPred, which considers molecular networks, performed better than conventional methods. Furthermore, the prediction performance differed by the molecular network structure used in learning, suggesting that it is important to consider not only the local network related to cancer but also the large-scale network in living organisms.

Conclusions: We propose a network-based machine learning method, Net-DMPred, for predicting cancer driver missense mutations. Our method enables us to consider the entire graph architecture representing the molecular network because it uses GNN. Net-DMPred is expected to detect driver mutations from a lot of missense mutations that are not known to be associated with cancer.

Keywords: Driver mutation prediction, Cancer missense mutation, Graph neural network, Molecular interaction

Background

Genomic sequencing studies have been a massive advancement in cancer genomic medicine. In conventional cancer medicine, treatment is uniformly determined by characteristics such as anatomical site and progression stage. However, some patients do not respond well to treatment. Cancer genomic medicine has become popular because gene

mutations are associated with cancer development. In this medicine, the best treatment is selected based on the genomic background of each patient. For this reason, cancer genomic medicine is expected to enhance therapeutic effects and reduce side effects.

One of the challenges in cancer genomic medicine is the clinical interpretation of variants. Although a lot of variants are detected by genomic analysis, most of them are passenger mutations not directly involved in cancer development. A small fraction of variants are driver mutations that are involved in cancer development [1]. Therefore, it is important to distinguish between driver mutations and passenger mutations.

It is time-consuming and expensive to validate whether variants are driver mutations, so machine-learning methods have been developed to predict if missense mutations are driver mutations. For example, CHASM [2] and CHASMplus [3] predict driver mutations by utilizing features obtained from conserved sequences and protein structure to characterize each variant. CanDrA [4] is an ensemble tool that uses the results of other prediction tools as variant features.

In living organisms, molecular networks are formed by various molecular interactions and signal transduction pathways, and abnormalities in molecular networks are associated with cancer. However, despite their importance, most existing tools for driver mutation prediction do not consider molecular networks. Although CHASMplus uses the number of molecular interactions for the amino acid site of a mutation as a variant feature, it only considers local molecular relationships and does not consider molecular networks. Molecular networks can be represented as a graph with molecules as nodes and molecular relationships as edges. Some prediction methods utilizing a graph to consider molecular interactions have been developed. Network&AA [5] is a driver prediction tool that considers the centrality of graphs representing molecular networks. However, the previous methods only consider one aggregated aspect of the graph, such as centrality, and not the entire graph structure.

Here we proposed a new network-based prediction method, Net-DMPred. This method uses a graph neural network (GNN) and learns a graph represented as molecular networks. GNN is a deep learning method for graphs and can learn the entire graphical structure. Net-DMPred predicts driver mutations using the features of the molecular networks by GNN as background knowledge and combining them with the features of individual variants used in conventional methods.

Results and discussion

Overview of Net-DMPred

Net-DMPred consists of the graph and prediction part (Fig. 1). The graph part learns background knowledge and the prediction part learns combining individual information and background knowledge learned by the graph part. In this study, molecular networks were used as background knowledge and individual variant features as individual information.

In the graph part, graphs representing molecular networks are learned using GNN, and feature vectors for each molecular node are computed. GNN can consider the entire architecture of graph. In the prediction part, the variants are predicted to be driver or passenger by Random Forest with individual variant features and graph node features

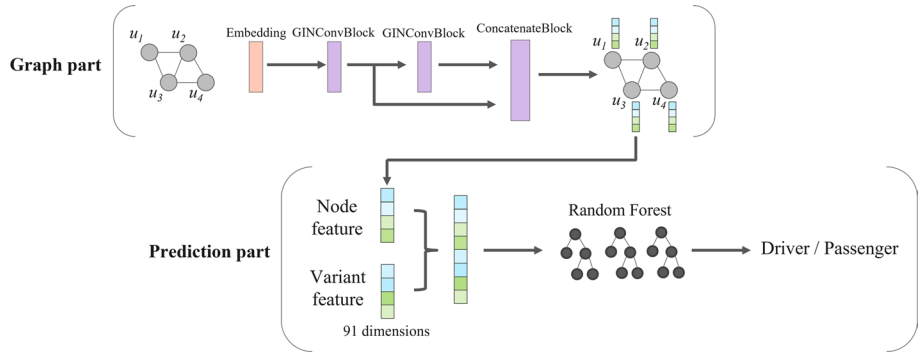


Fig. 1 Model architecture

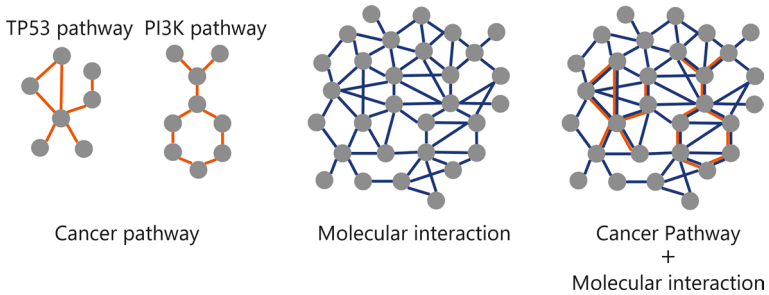


Fig. 2 Three knowledge graphs. "Cancer pathway" includes 10 pathways such as the TP53 pathway and the PI3K pathway

corresponding to the genes with variants. This framework enables us to utilize general molecular networks learned in the graph part as common background knowledge of individual variants in predicting driver mutation.

Performance evaluation of the training dataset

We used the training dataset of gene mutations provided by CHASMplus (http://karchinlab.org/data/CHASMplus/formatted_training_list.txt.gz, acquired on October 28, 2021). We performed the under sampling on this training dataset, and 928 driver and 3712 passenger mutations were used for training.

As features of gene mutations, we used 91 features obtained from SNVBox database [6] and the outputs of HotMAPS 1D method [7] following CHASMplus.

To investigate the relationship between the graph architecture and the prediction performance, we constructed three knowledge graphs representing molecular networks, "Cancer pathway," "Molecular interaction," and "Cancer pathway + Molecular interaction" (Fig. 2, Table 1). "Cancer pathway" was obtained from PathwayMapper [8] (acquired on November 1, 2021). PathwayMapper contains 10 public pathways [9] (such as TP53 pathway and the PI3K pathway) known to be associated with pancancer, and this study used these 10 pathways. "Molecular interaction" was obtained from Pathway Commons v12 [10]. Pathway Commons is public pathway and molecular interaction databases. "Cancer pathway + Molecular interaction" was combining "Cancer pathway" and "Molecular interaction". Moreover, to confirm the usefulness

Table 1 Details of three knowledge graphs

Graph	Node	Edge	Edge type	Data source
Cancer pathway	200	870	23	PathwayMapper [8]
Molecular interaction	30,899	3,672,040	13	Pathway commons [10]
Cancer pathway + molecular interaction	30,923	3,672,910	36	PathwayMapper [8], Pathway commons [10]

Number of nodes, number of edges, and types of edges in the three graphs. In the "Cancer pathway" graph, each node represents a protein, complex, and process, such as "apoptosis", and the edge type represents a type of relationship between them, such as "Activates" and "Inhibits." In the "Molecular interaction" graph, each node represents a protein and a small chemical compound, and the edge type represents a type of relationship between them defined by Pathway Commons, such as "controls-phosphorylation-of" and "controls-state-change-of"

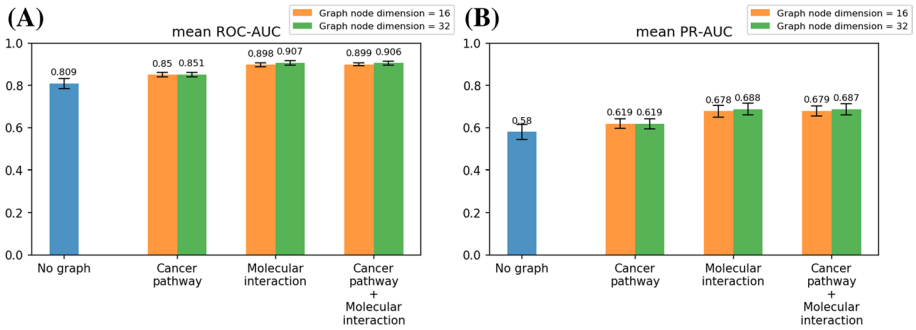


Fig. 3 Results of performance evaluation of the training dataset. **A** Mean value of ROC-AUC, **B** mean value of PR-AUC

of our framework, we performed the prediction using only individual variant features without the graph features (hereinafter referred to as "No graph"). It corresponds to the conventional prediction approach.

The models were evaluated the mean of ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and PR-AUC (Area Under the Precision-Recall Curve) with five-fold cross-validation. In this study, the dimension of the graph node vector was 16 and 32. Additional file 1: Figs. S1 and S2 present the ROC-AUC results obtained when using the Support Vector Machine and Multi-Layer Perceptron for prediction, as opposed to the Random Forest. Random Forest classifier performed better than other classification methods, achieving this across both graph node vector dimensions of 16 and 32. Therefore, we employed Random Forest for the prediction part in subsequent analyses of this study.

Figure 3 shows the results of the performance evaluation of the training dataset when using the Random Forest for prediction. Comparing the mean value of ROC-AUC and PR-AUC, the models with graphs, "Cancer pathway," "Molecular interaction," and "Cancer pathway + Molecular interaction" performed better than the model without a graph, "No graph." These results show that considering molecular networks is useful for driver mutation prediction.

Comparing the differences in graph architectures, "Molecular interaction" and "Cancer pathway + Molecular interaction" performed better than "Cancer pathway." This result demonstrates that the prediction performances were increased when not only the local network, "Cancer pathway," but also the large-scale network, "Molecular

interaction," were used as molecular networks. It also indicates that it is important to construct appropriate graphs because the prediction performances differed by graphs.

For the dimension of the node vectors, the prediction performance was slightly higher when 32, rather than 16 dimensions, were used. Therefore, we used the node features with 32 dimensions in the following analysis.

Performance evaluation with the benchmark datasets

We compared our approach with existing methods using five benchmark datasets obtained from Tokheim and Karchin [3] (http://karchinlab.org/data/CHASMplus/Tokheim_Cell_Systems_2019.tar.gz, acquired on October 28, 2021); Kim et al. [11], IARC TP53 [12], Ng et al. [13], Gene panel (OncoKB) [14, 15], CGC-recurrent [16]. These datasets were derived from in vivo and in vitro experiments and literature. Each dataset had different criteria for positive (driver) and negative (passenger) mutations. Therefore, these five datasets allowed for the multifaceted evaluation of the prediction model. In these datasets, some gene mutations overlapped with the training dataset. To strictly evaluate the prediction model, gene mutations that overlapped with the training dataset were dropped from benchmark datasets. Table 2 shows counts of gene mutation and unique genes with mutations used in benchmark datasets.

Figure 4 shows the prediction performances of the proposed models and the existing prediction tools for cancer driver mutations; CHASM, TransFIC [17], CanDrA, ParsSNP [18], CHASMplus, Network&AA. (The performances of other tools are shown in Additional file 2: Table S1 and Additional file 3: Table S2. The results of accuracy, precision, recall and F1 are shown in Additional file 4: Tables S3–S6.) The prediction performance of our models with the graph, "Cancer pathway," "Molecular interaction," and "Cancer pathway + Molecular interaction," was better or comparable to the model without a graph, "No graph."

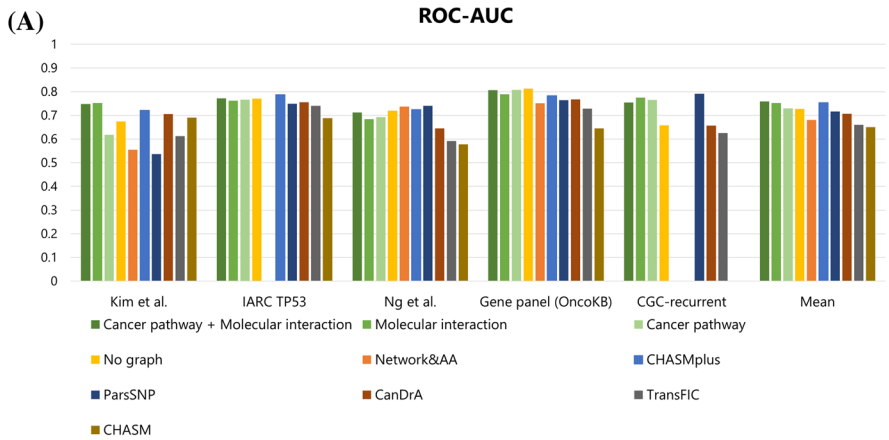
The proposed models showed higher performance than Network&AA, which considers the centralities of molecular networks. These results show that it is important to consider the entire molecular networks.

Moreover, in the proposed models, "Molecular interaction" and "Cancer pathway + Molecular interaction" showed overall higher performances than "Cancer pathway." These results show that the prediction performances were increased when not only the local network, "Cancer pathway," but also the large-scale network, "Molecular interaction," were used. The performance on some benchmark datasets was increased when the models with graphs, but others were not increased. It may be

Table 2 Benchmark datasets

Benchmark	Data source	Positive	Negative	Protein unique
Kim et al.	In vivo	8	17	11
IARC TP53	In vitro	375	1533	1
Ng et al.	In vitro	116	357	43
Gene panel (OncoKB)	Literature-based	329	38,596	410
CGC-recurrent	Literature-based	33	6805	3261

Number of positive and negative data and unique of proteins with mutations



(B)

		in vivo	in vitro		Literature-based	
		Kim et al.	IARC TP53	Ng et al.	Gene panel (OncoKB)	CGC-recurrent
Proposed methods	Cancer pathway + Molecular interaction	0.748	0.772	0.712	0.807	0.755
	Molecular interaction	0.752	0.762	0.684	0.789	0.775
	Cancer pathway	0.618	0.766	0.693	0.808	0.765
	No graph	0.674	0.771	0.72	0.813	0.658
	Network&AA	0.555	NaN	0.737	0.751	NaN
	CHASMplus	0.721	0.788	0.725	0.783	NaN
	ParsSNP	0.537	0.749	0.74	0.764	0.792
	CanDrA	0.706	0.756	0.645	0.768	0.657
	TransFIC	0.613	0.74	0.592	0.729	0.626
	CHASM	0.691	0.688	0.578	0.645	NaN

Fig. 4 Results of five benchmark datasets. **A** Bar graph of ROC-AUC score. **B** Table of ROC-AUC score. For each benchmark dataset, tools were ranked based on the ROC-AUC score and represented on the color scale: white signifies a lower rank, while shades of red, culminating in dark red, represent higher ranks. In Network&AA, we could not acquire prediction scores of variants in IARC TP53 and CGC-recurrent. In CHASMplus and CHASM, variants in CGC-recurrent did not have prediction scores

caused by the variety and differences between the five benchmark datasets in terms of the label definition and the data source.

For the Kim et al. dataset, the models with graphs showed higher performance than conventional tools. Kim et al. dataset was derived from in vivo and evaluated for the impact of tumor growth on missense mutation in mice. It can be speculated that the consideration of molecular networks may be useful in predicting in vivo datasets, such as this data. The proposed model "Cancer pathway + Molecular interaction" correctly predicted AKT1 p.Q79K as driver and IDH1 p.P33S as passenger. AKT1 p.Q79K well known to be a hotspot mutation [19] and relate with oncogenicity [20]. Moreover, AKT1 p.Q79K has been shown to be involved in acquired BRAF inhibitor resistance in melanoma [21, 22].

IDH1 p.P33S has been shown to be neutral to oncogenesis because the gene expression patterns in cells with p.P33S mutation in IDH1 were similar to those of wildtype [11].

On the other hand, in the prediction of IARC TP53 and Ng et al. datasets, the prediction performances of the models with graphs were not increased. IARC TP53 and Ng et al. datasets were derived from in vitro. The criteria for driver and passenger were transactivation levels for TP53 targets in IARC TP53 dataset and cell viability in Ng et al. dataset, respectively; these criteria differed from those in the training dataset.

The prediction performance of our models for the Gene panel (OncoKB) dataset was not improved. This dataset uses OncoKB annotations as labels. Mutations annotated with "Oncogenic" and "Likely Oncogenic" were defined as positive, and mutations with other annotations were defined as negative. In the negative data, there were missense mutations not only annotated with "Likely Neutral" and "Inconclusive" but also "Unknown" in OncoKB. Therefore, there is the possibility that some missense mutations which were labeled as negative in the dataset may be positive. Among 38,925 mutations in the Gene panel (OncoKB) dataset, the proposed model "Cancer pathway + Molecular interaction" predicted 8876 mutations as the driver mutations, 8624 of which were annotated as "Unknown" in the provided Gene panel (OncoKB) dataset. Then, we confirmed the latest annotations for these 8624 mutations. As of March 20, 2023, in the OncoKB database, 12 mutations (RUNX1 p.D198N, PIK3CA p.D549N, KRAS p.D33E, PDGFRB p.N666K, ATM p.R3008C, MET p.H1094Y, PIK3CB p.A1048V, ERBB2 p.Q709L, BRAF p.S467L, BRAF p.N581I, MAP2K1 p.L177V, JAK3 p.R657W) were annotated as "Oncogenic" and 593 mutations as"Likely Oncogenic." This finding implies that our proposed method is expected to identify driver mutations from a lot of mutations of uncertain significance. In addition, of the 8876 mutations in the benchmark dataset that our model predicted as the driver mutations, 7966 mutations remained classified as "Unknown" according to OncoKB as of March 20, 2023. Some of these mutations have potential to be confirmed as driver mutations through further experimental validation.

In the prediction of CGC-recurrent, the models with graphs performed better than models without graphs. This dataset includes a variety of genes with mutations. In a dataset consisting of such a large number and variety of genes, the graph features of each molecule learned in the graph part may be useful to increase the prediction performance.

Interpretation of results

We used SHAP (Shapley Additive exPlanations) [23] to evaluate the contribution of graph features. Figure 5 shows the driver prediction scores obtained from "Cancer pathway + Molecular interaction" and "No graph," and the contribution rate of graph features in "Cancer pathway + Molecular interaction" for the prediction of Kim et al. dataset. Here the graph feature importance rate was defined as the sum of SHAP values of graph features (32 features) divided by the sum of SHAP values of all features (123 features). The straight lines in Fig. 5 show the thresholds for the driver and passenger of each model. The upper left plots show missense mutations which "Cancer pathway + Molecular interaction" predicted to be positive and "No graph" predicted to be negative. Three missense mutations (AKT1 p.E267G, KRAS p.D33E, AKT1 p.R370C) indicated by the arrows have positive labels as the correct answers. These three mutations were correctly predicted by "Cancer pathway + Molecular interaction" and incorrectly predicted by "No graph." The graph feature importance rates for these variants show that the graph features had a significant impact on the predictions of these three mutations. In other words, by using the graph features, the model with the graph could have been predicted as positive for mutations correctly, which have been predicted as negative in models without the graph.

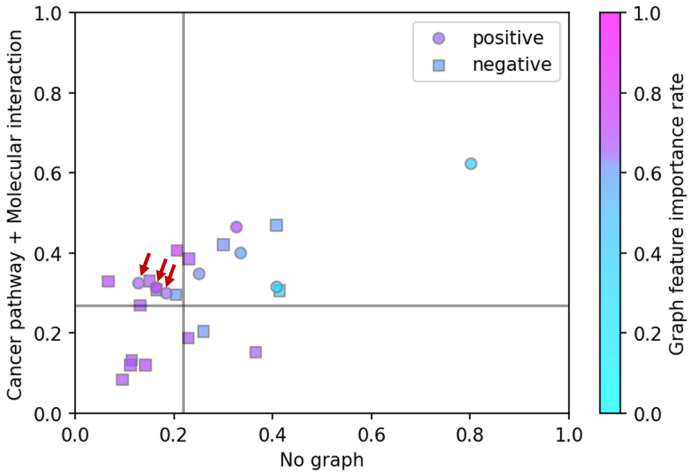


Fig. 5 Interpretation of the results for Kim et al. dataset. This figure shows the prediction scores for "Cancer pathway + Molecular interaction" and "No graph," and the graph feature importance rate for "Cancer pathway + Molecular interaction." The circle plots are positive variants and square plots are negative variants in this benchmark dataset. The straight lines are thresholds for positive and negative for "Cancer pathway + Molecular interaction" and "No graph" ("Cancer pathway + Mo-lecular interaction" = 0.270, "No graph" = 0.218). The plots with red arrows were predicted negative incorrectly for "No graph" but positive correctly for "Cancer pathway + Molecular interaction"

Upon a detailed examination of the contribution levels of individual features in the predictions when using "Cancer pathway + Molecular interaction," we observed that while numerous variant features rank highly, the contribution of graph features remains substantial (Additional file 1: Fig. S3A).

The graph part in the proposed method assigns random values to the initial graph features, and then graph features are trained to represent graph structure during the GNN training process. To ascertain whether the utility of graph features for prediction resulted from the GNN process, we also investigated the contribution of the initial graph features (Additional file 1: Fig. S3B). The results showed that the initial graph features did not contribute significantly to the predictions as much as variant features. These results suggest that the GNN training process is key to obtaining beneficial features for driver predictions.

Conclusion

In this study, we proposed Net-DMPred, which predicts driver mutation considering the entire architecture of molecular networks. The performance of the proposed method showed higher than or comparable to the conventional methods. This result shows that it is important to have information about molecular networks in predicting cancer driver mutation.

The prediction performances differed by the graph architecture, considering not only the local network, "Cancer pathway," but also the large-scale network, "Molecular interaction" improved the performance. This result indicates the importance of the construction of the graph appropriately.

We investigated the contribution of the graph to the prediction results using SHAP. It was confirmed that the graph features representing molecular networks contributed

to the prediction of cancer driver mutation. However, the proposed method has limitations. The graph part, which learns the graph representing the molecular network, and the prediction part, which learns whether the mutation is driver or passenger, are independent of each other. Thus, we cannot interpret the contribution of each node in the graph, that is, related molecules in driver prediction.

Although the development of genome sequencing technology has facilitated the detection of variants, a lot of variants of uncertain significance have been accumulated. In this study, we demonstrated that Net-DMPred was able to predict the driver mutations that were previously unannotated but recently have been shown to be involved in cancer as the data accumulated. This result suggests that Net-DMPred holds promise in identifying driver mutations from a large number of mutations whose association with cancer is not yet known.

Our prediction model has room for improvement in two aspects. In the graph part, the performances were affected by graph architecture, and our proposed model can improve depending on the design of the graph. In the prediction part, while we used features such as amino acid properties, sequence conservation, and protein structure, our proposed model can utilize these features and various additional features such as cancer type and the result of conventional prediction tools. Additionally, while the effectiveness of the graph features obtained through the GNN training process was confirmed in this study, it is challenging to interpret the specific implications of each feature within the proposed model's framework. Therefore, future considerations could involve expanding to an end-to-end framework that integrates the graph part.

Methods

Training dataset of gene mutation

The training dataset provided by CHASMplus consisted of 2051 driver mutations and 616,515 passenger mutations from the missense mutation dataset based on The Cancer Genome Atlas (TCGA) [24]. Positive data (driver mutations) were defined as mutations that occurred in genes listed in Cancer Genome Landscapes [25] and with a lower mutation frequency (less than 500 mutations) in samples. Negative data (passenger mutations) were defined as other mutations.

This dataset was imbalanced in two respects; driver mutations were in a limited number of genes, and there were many more passenger mutations than driver mutations. We performed the following sampling steps on the training dataset to resolve these imbalances. First, driver and passenger mutations were sampled randomly for each gene up to the median of the driver mutations per gene (=21) to prevent mutations from being in limited genes. As a result, 928 driver mutations and 310,450 passenger mutations were obtained. Next, because 928 passenger mutations out of 310,450 occurred in genes with driver mutations, they were selected and always included in the training dataset. This process prevented excessive dependence on gene features on the dataset in predicting driver and passenger mutations. Finally, passenger mutations of genes that did not have driver mutations were randomly sampled in order that four times as many as driver mutations. As a result, the training dataset contained 4640 mutations (928 driver mutations and 3712 passenger mutations).

Features of gene mutations

As the features of individual gene mutations, the proposed method uses 91 features. We obtained 88 features (Additional file 5: Supplementary Method) from SNVBox. It is a database of precomputed features for codons in the human exome, such as amino acid properties, sequence conservation, and protein structure. Other three features were obtained by running HotMAPS 1D. It is a method for estimating each gene's recurrently mutated genome region (hotspot regions). In this study, we used gene mutation data extracted from TCGA as input data. We ran this method with window sizes (hyperparameters) of 0, 5 and 10, following the precedent set in the previous study [3]. Then, we used the results (p-values) of the estimated regions as features of individual variants.

Molecular network dataset

A knowledge graph representing molecular networks was constructed from two databases. The first was from molecular interaction data from Pathway Commons v12. The second was from cancer signaling pathway data from PathwayMapper. These datasets describe molecules and their relationships in living organisms.

Pathway Commons integrates more than 20 public pathways and interactions databases and describes relationships between proteins and small chemical compounds with 13 types of binary relationships.

PathwayMapper is a web-based visualizing tool that includes various cancer-related pathways. Each cancer pathway contains information on regulatory relationships between molecules, such as activation and inhibition.

Construction of knowledge graph

The molecular network dataset can be represented as a knowledge graph with molecules as nodes and molecular relationships as edges, and the type of molecular relationship can be represented as an edge label.

In this study, we constructed three knowledge graphs using the molecular networks and compared each graph's difference in prediction performance. The first is the cancer pathway graph, "Cancer pathway," obtained from PathwayMapper. The second is the molecular interaction graph, "Molecular interaction," obtained from Pathway Commons. The third is the graph, "Cancer pathway + Molecular interaction," combining "Cancer pathway" and "Molecular interaction."

Net-DMPred model architecture

Graph part

A graph consists of a pair of nodes and edges $G = (V, E)$. An edge is represented $(u, r, v) \in \mathrm{E}$ using node u, v, and the relation r where $u, v \in V$, $r \in R$. R is a finite set of edge labels, E is a finite set of edges, and V is a finite set of nodes. In this study, background knowledge was molecular network, V was a set of molecules, and G was a set of molecular relationships.

In the graph part, the node vectors $z_u \in R^C$ is calculated by the operation of the embedding layer and the graph neural network (GINConv) [26]:

$$z_u = GNN(u, G) \tag{1}$$

The architecture of this graph neural network has a connection between the output of each layer and the concatenate block to enhance expression of graph neural network. This graph neural network is constructed using a graph isomorphism network (GIN) block. A GIN block is defined as follows:

$$z^{\ell+1} = \sigma\left(\sum_{r} W_{(r)} \cdot \sigma\left(GINConv\left(z^{\ell}, G_{(r)}\right)\right) + b_{(r)}\right) \tag{2}$$

where z^{ℓ} represents a node vector of ℓ-th layer, $r \in R$ is relation (edge label) in the graph G, and $G_{(r)}$ is defined as a subgraph that extracts the edges with relation r from the graph G.

The learning of the graph is pre-trained by link prediction. Link prediction is predicting the probability of the existence of an edge (link) between nodes. This pre-training learns node features based on the observed graph structure. The loss function is defined as follows:

$$L = -\ell_{+}(u, v) - \ell_{-}(u, v\prime) \tag{3}$$

where u, v is randomly sampled from E and v' is randomly sampled from V.

$$\ell_{+}(u, v) = \log\left(\sigma\left(u^{T} v\right)\right) \tag{4}$$

$$\ell_{-}(u, v) = \log\left(\sigma\left(-u^{T} v\right)\right) \tag{5}$$

When a link with a relation is predicted, weight matrix is used:

$$\ell_{+}(u, r, v) = \log\left(\sigma\left(u^{T} \mathbf{W}_{r} v\right)\right) \tag{6}$$

$$\ell_{-}(u, r, v) = \log\left(\sigma\left(-u^{T} \mathbf{W}_{r} v\right)\right) \tag{7}$$

where $\boldsymbol{W}$ is a diagonal matrix.

In this study, the dimension of the node vector was $C = 16$ and 32, and the continuous random values were initially assigned to each vector. We employed a learning rate of 0.0001, and the model after 50 epochs was used for the analysis.

Prediction part

In the prediction part, the driver mutation prediction is performed using the features of individual variants combined with graph node features learned by the graph part. We used the Random Forest model [27] for prediction.

The probability of driver mutation $\widehat{y}$ is calculated using graph features and variant features as follows:

$$\hat{y}_i = RF(\mathrm{X}_i, \{z_u : u \in V\}) \tag{8}$$

where X_i is an variant feature of mutation i of gene u, and z_u is a graph feature of gene u.

We performed Random Forest using the scikit-learn package (version 0.24.2). The hyperparameters of the model were tuned by three-fold grid search. Missing values were complemented by the mean.

Performance evaluation with the benchmark datasets

To compare our approach with existing methods, we used five benchmark datasets obtained from Tokheim and Karchin [3] (http://karchinlab.org/data/CHASMplus/Tokheim_Cell_Systems_2019.tar.gz, acquired on October 28, 2021); Kim et al., IARC TP53, Ng et al., Gene panel (OncoKB), and CGC-recurrent.

The hyperparameters of our models used in the evaluation were obtained from the model with the best performance in the five-fold cross-validation. To ensure robust evaluation, we performed this trial three times. Then, three trained models were used to predict each benchmark dataset, and the models were evaluated on the average performance. The prediction performance was evaluated by ROC-AUC and PR-AUC. The thresholds for positive and negative were defined as the average value of the Youden index [28] in the three trained models: "Cancer pathway + Molecular interaction" = 0.270, "Molecular interaction" = 0.281, "Cancer pathway" = 0.264, "No graph" = 0.218.

We compared the prediction performance of our approach with 26 preceding tools. There are six tools to predict driver mutations in cancer (CHASM, TransFIC, CanDrA, ParsSNP, CHASMplus, and Network&AA) and there are 20 tools to predict the effect of gene mutations on proteins, not specific to gene mutations in cancer (SIFT [29], MutPred [30], LRT [31], Polyphen2_HVAR [32], Polyphen2_HDIV [32], MutationAssessor [33], PROVEAN [34], VEST4 [35], FATHMM [36], CADD [37], MutationTaster [38], MetaSVM [39], DANN [40], REVEL [41], M-CAP [42], DEOGEN2 [43], MPC [44], ClinPred [45], LIST-S2 [46], and MVP [47]).

We obtained prediction scores of CHASM, TransFIC, CanDrA, and ParsSNP from Tokheim et al. [3], Network&AA from Ozturk et al. [5], and other prediction scores from the dbNSFP database [48, 49].

Interpretation of results

We employed SHAP to interpret the contribution of each feature to the prediction results. SHAP is an approach to explain the prediction result, and the SHAP value expresses the contribution of each feature. A larger absolute SHAP value means a larger contribution of the feature to the prediction result. In this study, we evaluated the contribution of the graph features in predicting gene mutations in benchmark datasets using the SHAP values. SHAP Python package (version 0.39.0) was used to calculate the SHAP values.

Abbreviations

GNN	Graph neural network
ROC-AUC	Area under the receiver operating characteristic curve
PR-AUC	Area under the precision-recall curve
SHAP	Shapley additive explanations
TCGA	The Cancer Genome Atlas
GIN	Graph isomorphism network

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05507-6.

Additional file 1: Figure S1 The mean of ROC-AUC for each classifier when the graph node vector was 16 dimensions. **Figure S2** The mean of ROC-AUC for each classifier when the graph node vector was 32 dimensions. **Figure S3**. The top 30 contributed features in predicting Kim et al. dataset. The top 30 most contributed features are calculated by the average of the absolute SHAP values for each feature. Features prefixed with "graph_feature" are graph node features, with each number corresponding to a dimension of the node feature vectors. The remaining features, such as p-value_w5 and MGAEntropy, represent variant features. (A) The contributed features when using "Cancer pathway + Molecular interaction." (B) The contributed features when using the initial graph features.

Additional file 2: Table S1 "ROC-AUC for five benchmark datasets".

Additional file 3: Table S2 "PR-AUC for five benchmark datasets".

Additional file 4: Table S3 "Accuracy for five benchmark datasets (driver prediction tools)". **Table S4** "Precision for five benchmark datasets (driver prediction tools)". **Table S5** "Recall for five benchmark datasets (driver prediction tools)". **Table S6** "F1 for five benchmark datasets (driver prediction tools)".

Additional file 5. Supplementary Method. 88 features from SNVBox.

Acknowledgements
The analysis in this work were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Author contributions
MK, RK and YO conceived the ideas. RK and NH implemented NetDM-Pred. MK and NH performed the prediction and analyzed the experiment results. All authors contributed to the preparation of the manuscript.

Funding
This work was supported by MEXT as "Program for Promoting Researches on the Supercomputer Fugaku" (Application of Molecular DynamicsSimulation to Precision Medicine Using Big Data Integration System for Drug Discovery, JPMXP1020200201 and Simulation- and AI-driven next-generation medicine and drug discovery based on "Fugaku", JPMXP1020230120).

Availability of data and materials
The datasets and source codes of Net-DMPred are publicly available at the Github repository, https://github.com/clinfo/Net-DMPred. Project name: Net-DMPred. Project home page: https://github.com/clinfo/Net-DMPred. Operating system(s): Linux. Programming language: Python. Any restrictions to use by non-academics: No restrictions. The datasets described in this article can be freely and openly accessed at Pathway Commons (https://www.pathwaycommons.org/archives/PC2/v12/PathwayCommons12.All.hgnc.sif.gz), PathwayMapper (https://www.pathwaymapper.org/), CHASMplus training datasets (http://karchinlab.org/data/CHASMplus/formatted_training_list.txt.gz), five benchmark datasets (http://karchinlab.org/data/CHASMplus/Tokheim_Cell_Systems_2019.tar.gz), and SNVBox (http://karchinlab.org/data/CHASMplus/SNVBox_chasmplus.sql.gz). When running HotMAPS 1D (https://github.com/KarchinLab/probabilistic2020/tree/master), the input datasets were acquired from The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/), gene BED file (https://genome.ucsc.edu/cgi-bin/hgTables), and gene FASTA file (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 6 July 2023 Accepted: 2 October 2023
Published online: 10 October 2023

References

1. Saksena G, Mermel C, Getz G. Developing algorithms to discover novel cancer genes a look at the challenges and approaches. IEEE Signal Process Mag. 2012;29(1):89–97.
2. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009;69(16):6660–7.
3. Tokheim C, Karchin R. CHASMplus reveals the scope of somatic missense mutations driving human cancers. Cell Syst. 2019;9(1):9-23.e8.

4. Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K. CanDrA: cancer-specific driver missense mutation annotation with optimized features. PLOS ONE. 2013;8(10):e77945.
5. Ozturk K, Carter H. Predicting functional consequences of mutations using molecular interaction network features. Hum Genet. 2022;141(6):1195–210.
6. Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics. 2011;27(15):2147–8.
7. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, Masica DL, Karchin R. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer Res. 2016;76(13):3719–31.
8. Bahceci I, Dogrusoz U, La KC, Babur Ö, Gao J, Schultz N. PathwayMapper: a collaborative visual web editor for cancer pathways and genomic data. Bioinformatics. 2017;33(14):2238–40.
9. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al. Oncogenic signaling pathways in the cancer genome atlas. Cell. 2018;173(2):321-337.e10.
10. Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, Franz M, Siper MC, Cheung M, Wrana M, et al. Pathway commons 2019 update: integration, analysis and exploration of pathway data. Nucleic Acids Res. 2020;48(D1):D489–97.
11. Kim E, Ilic N, Shrestha Y, Zou L, Kamburov A, Zhu C, Yang X, Lubonja R, Tran N, Nguyen C, et al. Systematic functional interrogation of rare cancer variants identifies oncogenic alleles. Cancer Discov. 2016;6(7):714–26.
12. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum Mutat. 2007;28(6):622–9.
13. Ng PK, Li J, Jeong KJ, Shao S, Chen H, Tsang YH, Sengupta S, Wang Z, Bhavana VH, Tran R, et al. Systematic functional annotation of somatic mutations in cancer. Cancer Cell. 2018;33(3):450-462.e10.
14. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017;1:1–16.
15. Zehir A, Benayed R, Shah R, Syed A, Middha S, Kim H, Srinivasan P, Gao J, Chakravarty D, Devlin S, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017;23(6):703–13.
16. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45(D1):D777–83.
17. Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med. 2012;4(11):89.
18. Kumar RD, Swamidass SJ, Bose R. Unsupervised detection of cancer driver mutations with parsimony-guided learning. Nat Genet. 2016;48(10):1288–94.
19. Shrestha Bhattarai T, Shamu T, Gorelick AN, Chang MT, Chakravarty D, Gavrila EI, Donoghue MTA, Gao J, Patel S, Gao SP, et al. AKT mutant allele-specific activation dictates pharmacologic sensitivities. Nat Commun. 2022;13(1):2111.
20. Parikh C, Janakiraman V, Wu WI, Foo CK, Kljavin NM, Chaudhuri S, Stawiski E, Lee B, Lin J, Li H, et al. Disruption of PH-kinase domain interactions leads to oncogenic activation of AKT in human cancers. Proc Natl Acad Sci U S A. 2012;109(47):19368–73.
21. Shi H, Hugo W, Kong X, Hong A, Koya RC, Moriceau G, Chodon T, Guo R, Johnson DB, Dahlman KB, et al. Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. Cancer Discov. 2014;4(1):80–93.
22. Shi H, Hong A, Kong X, Koya RC, Song C, Moriceau G, Hugo W, Yu CC, Ng C, Chodon T, et al. A novel AKT1 mutant amplifies an adaptive melanoma response to BRAF inhibition. Cancer Discov. 2014;4(1):69–79.
23. Lundberg SM, Lee S-I: A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
24. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109–12.
25. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546–58.
26. Xu K, Hu W, Leskovec J, Jegelka S: How Powerful are graph neural networks? 2018: arXiv:1810.00826.
27. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
28. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J. 2005;47(4):458–72.
29. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.
30. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744–50.
31. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553–61.
32. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
33. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118.
34. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLOS ONE. 2012;7(10):e46688.
35. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14(Suppl 3):S3.
36. Shihab HA, Gough J, Cooper DN, Day IN, Gaunt TR. Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics. 2013;29(12):1504–10.
37. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
38. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11(4):361–2.

39. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125–37.
40. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761–3.
41. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, et al. REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877–85.
42. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581–6.
43. Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, Rooman M, Vranken W. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45(W1):W201–6.
44. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, Neale BM, Daly MJ. Regional missense constraint improves variant deleteriousness prediction. bioRxiv 2017;148353.
45. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. Am J Hum Genet. 2018;103(4):474–83.
46. Malhis N, Jacobson M, Jones SJM, Gsponer J. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. Nucleic Acids Res. 2020;48(W1):W154–61.
47. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, Guan Y, Shen Y. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun. 2021;12(1):510.
48. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32(8):894–9.
49. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12(1):103.

Publisher's Note