PLOS ONE

RESEARCH ARTICLE

GraphMHC: Neoantigen prediction model applying the graph neural network to molecular structure

Hoyeon Jeong[1], Young-Rae Cho[2], Jungsoo Gim[3], Seung-Kuy Cha[4], Maengsup Kim[5], Dae Ryong Kang[1,6]*

1 Department of Biostatistics, Yonsei University, Wonju, Gangwon State, Republic of Korea, **2** Division of Software, Yonsei University Mirae Campus, Wonju, Gangwon State, Republic of Korea, **3** Department of Biomedical Science, Chosun University, Gwangju, Republic of Korea, **4** Department of Physiology, Yonsei University Wonju College of Medicine, Wonju, Gangwon State, Republic of Korea, **5** Research Center, Mustbio, Suwon-si, Gyeonggi-do, Republic of Korea, **6** Department of Precision Medicine, Yonsei University Wonju College of Medicine, Wonju, Gangwon State, Republic of Korea

* dr.kang@yonsei.ac.kr

Check for updates

OPEN ACCESS

Citation: Jeong H, Cho Y-R, Gim J, Cha S-K, Kim M, Kang DR (2024) GraphMHC: Neoantigen prediction model applying the graph neural network to molecular structure. PLoS ONE 19(3): e0291223. https://doi.org/10.1371/journal.pone.0291223

Editor: Kyle Ellrott, Oregon Health & Science University, UNITED STATES

Received: August 23, 2023

Accepted: March 5, 2024

Published: March 27, 2024

Copyright: © 2024 Jeong et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data used are publicly available: The MHC class I dataset, http://tools.iedb.org/static/main/binding_data_2013.zip; the MHC class II dataset, http://tools.iedb.org/static/download/classII_binding_data_Nov_16_2009.tar.gz; TCGA-SKCM data, http://firebrowse.org/; HLA types, https://static-content.springer.com/esm/art%3A10.1186%2Fs12920-020-0694-1/MediaObjects/12920_2020_694_MOESM3_ESM.xlsx; immunity scores, https://bioinformatics.mdanderson.org/estimate/tables/skin_cutaneous_melanoma_RNAseqV2.txt. The code of this

Abstract

Neoantigens are tumor-derived peptides and are biomarkers that can predict prognosis related to immune checkpoint inhibition by estimating their binding to major histocompatibility complex (MHC) proteins. Although deep neural networks have been primarily used for these prediction models, it is difficult to interpret the models reported thus far as accurately representing the interactions between biomolecules. In this study, we propose the GraphMHC model, which utilizes a graph neural network model applied to molecular structure to simulate the binding between MHC proteins and peptide sequences. Amino acid sequences sourced from the immune epitope database (IEDB) undergo conversion into molecular structures. Subsequently, atomic intrinsic informations and inter-atomic connections are extracted and structured as a graph representation. Stacked graph attention and convolution layers comprise the GraphMHC network which classifies bindings. The prediction results from the test set using the GraphMHC model showed a high performance with an area under the receiver operating characteristic curve of 92.2% (91.9-92.5%), surpassing a baseline model. Moreover, by applying the GraphMHC model to melanoma patient data from The Cancer Genome Atlas project, we found a borderline difference (0.061) in overall survival and a significant difference in stromal score between the high and low neoantigen load groups. This distinction was not present in the baseline model. This study presents the first feature-intrinsic method based on biochemical molecular structure for modeling the binding between MHC protein sequences and neoantigen candidate peptide sequences. This model can provide highly accurate responsibility information that can predict the prognosis of immune checkpoint inhibitors to cancer patients who want to apply it.

research model is published on the following GitHub page: https://github.com/recognizability/GraphMHC.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Cancer is the leading cause of death in a large number of countries worldwide [1] and the number one cause of death in South Korea [2]. Cancer mainly results from somatic mutations [3]. Cancer immunotherapy basically works by encouraging tumor cells to recognize the MHC protein non-self foreign, leading to activation of the cytotoxic T cell receptor (TCR) and CD8 coreceptor. However, the mechanism by which this occurs acts as an immune checkpoint, suppressing overactivation of the immune system, with proteins such as the programed cell death protein 1 (PD-1) and cytotoxic T lymphocyte antigen-4 (CTLA-4) the main actors in these pathways. Third-generation anticancer drugs block or inhibit this immune checkpoint [4]. Despite these efforts, only a small number of subjects respond well to immune treatment [5], and it is limited because of the high expense [6]. Neoantigens, or neoplasmic antigens, are tumor-specific antigenic determinants or epitopes that consist of 9-mer-long peptides that are cleaved by proteasome internal organelles and act as biomarkers predicting immune checkpoint inhibition [7, 8], which can be estimated by predicting the binding potential of major histocompatibility complex (MHC) proteins to candidate peptides associated with somatic mutations.

MHC proteins bind with peptides via non-covalent hydrogen bonds. The half maximal inhibitory concentration (IC_{50}) between MHC and isolated peptides can be experimentally determined [9], and the results of binding attempts can be found in the Immune Epitope Database (IEDB) [10].

MHC-peptide binding models based on deep neural networks dealing with amino acid sequences

Deep neural network models that utilize binding information between MHC and peptides from IEDB have been published. NetMHCpan-4.0 [11] and 4.1 [12] researches are representative of attempts to model the bond between MHC and peptide by constructing a deep neural network using known IEDB. According to NetMHCpan-1.0 [13], because the amino acids of MHC are long, only 34 amino acids corresponding to polymorphic residues with high mutation frequency are used, and they are used for input by concatenating with the peptide sequence. Although it is provided as a predictive value by default on the IEDB website and is widely used in clinical practice, there are limitations in terms of structure and accuracy for interaction modeling.

Several models using more advanced modern deep neural networks have also been reported. MHCAttnNet uses a recurrent neural network (RNN) for string processing [14], DeepNeo uses a convolutional neural network (CNN) [15], DeepImmuno used CNN, graph convolutional networks (GCN), which dealt with the relationships between amino acids, not between the constituent atoms [16].

Since these models are merely attempts to connect one-dimensional vectors or make them into two-dimensional matrices, they have limitations in simulating multidimensional interactions between biopolymers.

Graph neural networks for modeling molecular structures

Graph neural networks (GNN) [17], graph convolutional networks (GCN) [18], and graph attention networks (GAT) [19] are useful for graph classification as well as node classification. They are effective in expressing molecular structures and intrinsic attributes. To represent feature matrices and adjacency matrices derived from the graph structure, operations such as

convolution and attention are stacked, and sigmoid or softmax functions are applied for classification.

Unlike sequence-based neural networks or array-based neural networks, these models use not only features but also connection information between nodes to enhance accuracy by extracting more information [20, 21].

Graph neural networks that model SMILES-based interaction or affinity

Meanwhile, studies using the Simplified Molecular Input Line Entry System (SMILES) [22], an expression for molecular structure in the interaction between polymers in an adjacent academic field, have been reported as input to deep neural networks. Looking at each field, it corresponds to the protein-protein interaction (PPI), the drug-target affinity (DTA), and the drug-drug interaction (DDI). In terms of interaction methods, they can be grouped into two categories: concatenation [23–29] or coattention [30, 31]. These graph-graph interaction can be combination between embedded vectors extracted from each graph rather than direct combination between the nodes constituting the two graphs. There are limitations to model the interaction between all components of the graph.

GraphMHC: A model that predicts neoantigens using a graph neural network using molecular structure

This study proposes the GraphMHC model, which uses data from the IEDB and the GNN model to illustrate the binding between MHC proteins and peptide sequences via the molecular structure. The determination of the feature extracted from combining an MHC protein with a peptide as a neoantigen is performed through multi-layered graph convolution when using this method. This model is a method of extracting feature vectors from both graphs, rather than connecting individual feature vectors extracted from each of the two graphs, so comprehensive and interactive feature extraction is possible. The GraphMHC model was validated against the baseline model [12], and applied to melanoma patient data from the Cancer Genome Atlas (TCGA) project. Data were divided into low and high neoantigen load groups for comparison of the clinical differences.

Materials and methods

In this section, the methods used to build and validate the neoantigen prediction model GraphMHC are presented.

GraphMHC: A neoantigen prediction model based on the GNN using MHC class I from the IEDB

The processes for building GraphMHC, which is a neoantigen prediction model based on GNNs, are presented in detail in this section. The overall pipeline can be seen in Fig 1.

The Immune Epitope Database (IEDB) [10], a binding affinity data set for peptides and HLA types, was used to construct the model for predicting neoantigens. The MHC class I dataset was used for neoantigen prediction. The use of the MHC class II dataset is described separately in the section describing extra-validation.

Referring to previous studies, a result with a binding affinity of 500 nM or less was determined to be a neoantigen [32–34].

Conversion from the molecular structure of amino acids to dataset of interatomic graph structures. In the IEDB dataset, 157,325 rows pertaining to humans were utilized. These rows were then transformed into amino acids that constitute the MHC protein,

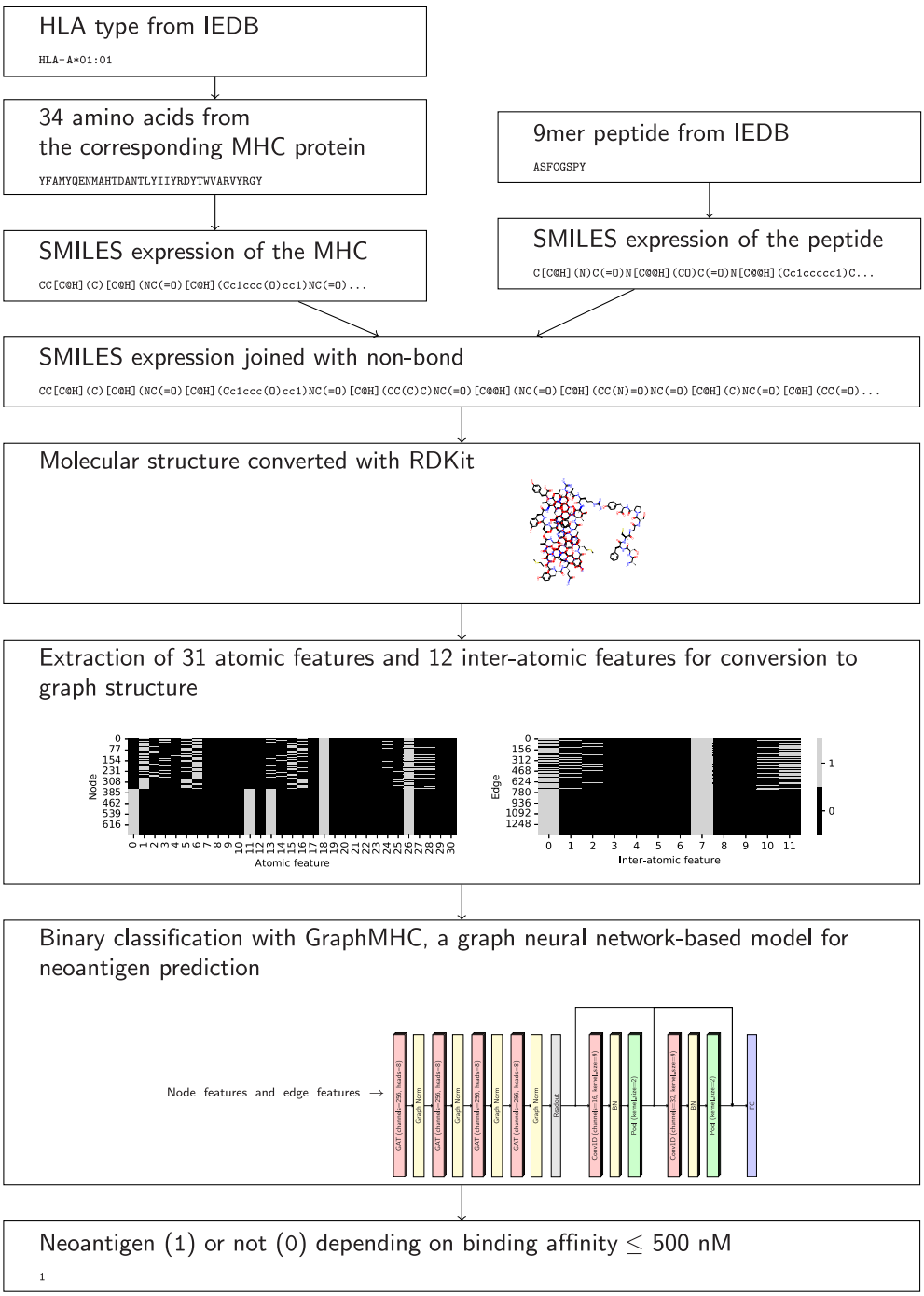

Fig 1. Overall pipeline for modeling neoantigen prediction. HLA types and 9-mer peptides from the IEDB are converted into molecular structures, and the GraphMHC model is used to classify probable neoantigens from the extracted atoms and interatomic features.

https://doi.org/10.1371/journal.pone.0291223.g001

employing the conversion data from NetMHCpan-4.1 [12]. Out of these, 157,084 were converted, excluding types for which conversion information was unavailable. Classification was based on binding affinity, with binding assumed for $IC_{50} \leq 500$ nM, and non-binding for $IC_{50} > 500$ nM. The dataset was split, with 80%, or 125 667, for training and 20%, or 31 417, for testing. The conversion process from database to dataset is as follows.

First, human leukocyte antigens (HLA) from the IEDB are converted to MHC amino acid sequences. Second, MHC and peptide sequences are converted using SMILES [22] structures using the RDKit 2022.03.2 library. The expressions used are given in S1 Table. Third, join the two SMILES strings with non-bond notation together (.). Fourth, convert to molecular structure using RDKit. Hydrogen atoms that were omitted from the symbol must be expressed at this point. The molecular structure is expressed in S2 Fig. Fifth, convert to graph structure using the RDKit library to encode vectors and matrices. The feature information of the graph representation is shown in Table 1. Each feature is encoded via one-hot encoding and constructed as a sparse matrix. The graph structure is expressed in S2 Fig using NetworkX 2.8.4 library with Kamada Kawai layout [35]. Characteristics of graph representations describing bound and unbounded data are shown in the S2 Table. Sixth, convert graph dataset using the PyTorch Geometric (PyG) [36] 2.1.0 library.

Architectural design of the GNN GraphMHC for graph classification. The layer-by-layer architecture of GraphMHC, the GNN model for MHC-peptide binding, is presented in a schematic diagram provided in Fig 2. In this model, graph attention [37] was used as graph convolution, where the attention factor is multiplied to assign importance to nodes. Another noteworthy element of this model is the stacking of conventional one-dimensional convolution layers after graph convolutions. This procedure enables the re-extraction of a given feature vector multiple times for classification purposes. Another highlight is that skip connections are connected to pass weights between these convolutional layers. This prevents vanishing of weight passing and contributes to more precise tuning. The stacking steps of the model are subdivided as follows.

First, stack four layers of graph attention [38] using PyTorch Geometric (PyG) [36] 2.1.0 library. Second, for graph classification, the readout layer is the mean value of the node feature vectors. Third, two layers of 1-dimensional convolution are stacked with the PyTorch [39] 1.11.0+cu113 framework. At this time, add a skip connection [40] between layers. Fourth, classify via sigmoid in fully-connected layer.

Model training and evaluation. ADAM [41] was used as the optimizer for learning. All hyperparameters of each layer in these networks are described in Fig 2, and the batch size is set to 64 and the number of epochs is set to 100. The reason why the number of attention heads was set to 8 is because if it is lower than this, there is a performance degradation, and if it is higher than this, there is a computational load. For tensor calculation during the deep neural

Table 1. Components of graph representation.

Graph representation	Components	Number of features
Node feature vectors with one-hot encoding	Atom symbol (H, C, N, O, S), hybridisation (SP3, SP2, SP, S, SP3D, SP3D2), degree of covalent bonding, number of bonded hydrogen atoms, chirality (CCW, CW or others), aromaticity, inclusion in rings, formal charge, and number of radical electrons	The number of features per node is 31
Edge indices (adjacency matrix)	Between the starting atom and the ending atom	
Edge feature vectors with one-hot encoding	covalent bond type (single, aromatic, double, triple), stereo type (any, cis, E, none, trans, Z), ring bond, conjugate bond	The number of features per edge is 12

https://doi.org/10.1371/journal.pone.0291223.t001

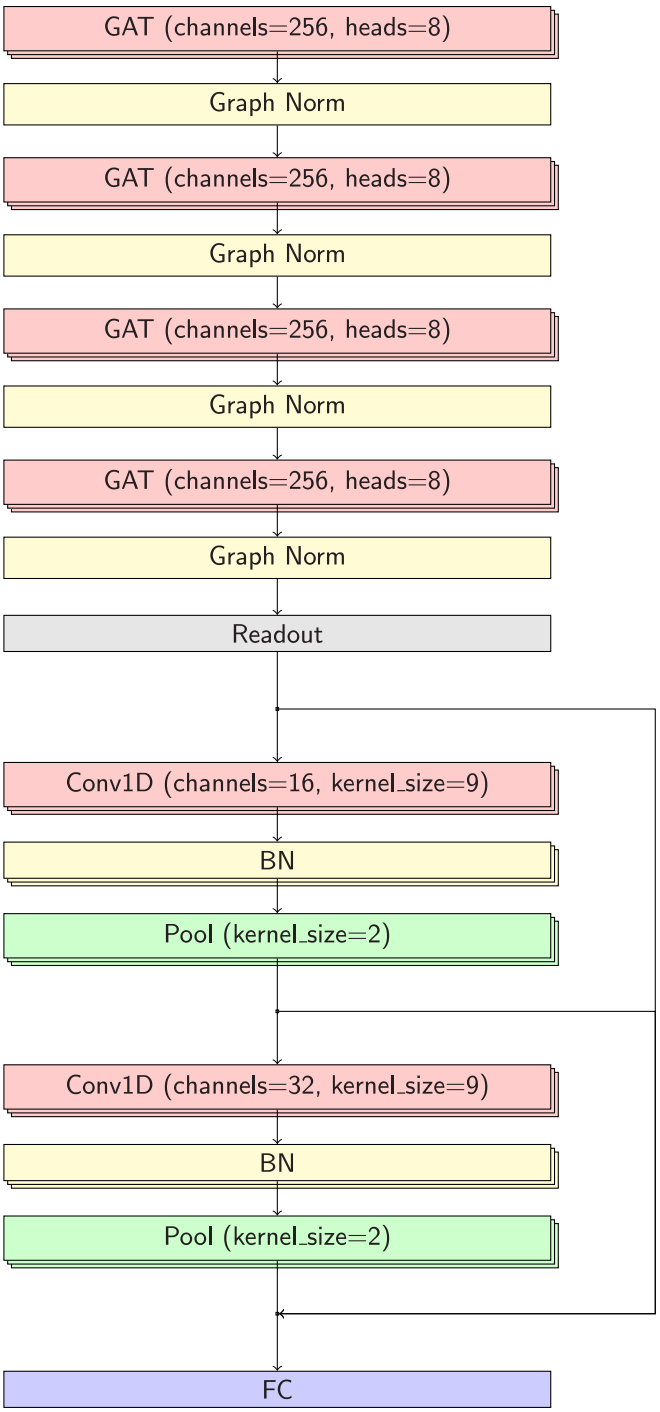

Fig 2. Architecture of the GraphMHC model, a GNN that predicts neoantigens via MHC-peptide binding information. Bypass arrows between Conv1D layers mean that skip connection. Dropout was set at 0.1 for all layers. GAT: graph attention layer, FC: fully-connected layer, Graph Norm: graph normalization, Conv1D: 1-dimensional convolution, BN, batch normalization, Pool: average pooling.

https://doi.org/10.1371/journal.pone.0291223.g002

network model training, massive parallel processing was performed using RTX 3090 24GB graphic processing units (GPU). Scikit-learn [42] 1.1.2 was used to evaluate the classification model, and the 95% AUC-ROC confidence interval was obtained through the DeLong method [43] in MedCalc [44] 20.106. 95% confidence intervals for other metrics were calculated directly.

For comparison with models comprising fewer layers, two cases were constructed; one of which involved subtracting two of four GNN layers, while the other involved the subtraction of two CNN layers.

Neoantigen load predictions on next generation sequencing (NGS) data from The Cancer Genome Atlas database

The constructed model was applied to next-generation sequencing (NGS) data from actual cancer patients. Data were sourced from The Cancer Genome Atlas (TCGA) [45]. Skin cutaneous melanoma (SKCM) data was used because of the effective response that this type of cancer has to immunotherapy [46]. Data were primarily downloaded from the Broad Institute Firehose website.

Sequence information for normal samples, which is not disclosed by TCGA-SKCM, is required for HLA typing and was thus obtained from previous studies [47]. The HLA types were reported to be Optitype [48] and were converted to 34-mer amino acids using NetMHCpan-4.1 [12].

The procedure for translating the mutation into a 9-mer peptide and identifying it as a neoantigen is as follows.

First, convert from mutation annotation format (MAF) to variant calling format (VCF) using maf2vcf [49]. Second, annotate as Variant Effect Predictor (VEP) [50]. At this time, the sequence information for the GRCh38 reference genome is used. Third, select only missense variants. Fourth, convert to amino acids with R customProDB [51]. The chromosome nomenclature must be consistent for this process. Fifth, only the region in which the actual mutation appears is truncated to a length of 9-mer [52]. Sixth, convert the MHC protein sequence and peptide sequence together to form a graph structure. Seventh, predict the binding affinities using GraphMHC.

Thus, data from a total 310 subjects were used for predicting neoantigen load by combining 9-mer peptides that are considered neoantigen candidates with HLA type information. For more details of the selection process, see S3 Fig Candidates are then predicted neoantigens or not using GraphMHC.

Comparison of groups according to high and low neoantigen load. Two groups were formed around the median, with high and low neoantigen loads. Comparison was performed using survival analysis and immunity score analysis. Clinical information on survival was also obtained from the Firehose website. This study uses overall survival as the analysis target, for which Lifelines 0.27.0 was used.

Estimates of stromal cell score, immune cell score, and tumor purity were obtained from the ESTIMATE website and utilized [53]. Data were calculated from the TCGA expression data. The stromal score refers to the number of stroma cells within tumor tissue [54, 55], and was used because stromal cells have been reported to be involved in tumor growth and disease progression. The immune score indicates the infiltration of immune cells into tumor tissues and is an immunological biomarker for prognosis prediction and therapeutic response [56]. The ESTIMATE score relates to tumor purity and is a combination of the stromal and the immune scores. Scipy 1.9.3 was used for the comparison test.

Extra-validation of the model using MHC class II from the IEDB

Although neoantigens are related to MHC class I and $CD8^+$, MHC class II and $CD4^+$ have also been reported complementary, with less variation between patients [57, 58]. Several model studies have investigated this concept [12, 59], and a model with the same architecture as the proposed model was trained using MHC class II data from the IEDB for extra-validation in this study. The same classification threshold of 500 nM was used. Data with reduced similarity were selected, and training and test sets comprising 85 708 and 21 427 data points, respectively, were used after pre-processing.

Baseline models for comparison

The model NetMHCpan-4.1 [12] was used as the baseline. Classification using IEDB data and comparison between groups using TCGA-SKCM data were applied equally. NetMHCIIpan-4.0 was used for extra-validation [12]. In order to convert the obtained binding affinity to a value between 0 and 1, studies [11, 60, 61] such as those involving NetMHCpan-4.0 used the expression $1 - \log(\cdot)/\log(50000)$, whereas an equation $1/(1 + \exp(\cdot))$ similar to the sigmoid was used for the same comparison in this study.

Results

In this section, the GraphMHC model, which is proposed for neoantigen prediction based on GNNs, is validated using intra-validation, inter-validation, clinical application, and extra-validation. The datasets and models used are clearly shown in Table 2. Matplotlib 3.5.3 was used to plot the ROC and PR curves, and Seaborn 0.12.0 was used to create violin plots.

Intra-validation: Combining graph convolution with convolution improves prediction accuracy

The comparison results obtained for the different models according to the layer configuration of the neural network are shown in the center columns of Table 3. The model consisting of 4-layer GNNs and 2-layer CNNs (GraphMHC) demonstrates the highest performance. A model consisting of a 2-layer GNN and a 2-layer CNN has higher performance than a model consisting of only a 4-layer GNN, because feature extraction is performed effectively in the CNN layers. Additionally, it was found that a 4-layer GNN contributed to performance improvement compared to a 4-layer GNN. This is due to the repeated aggregation and updates from neighboring nodes.

Table 2. Four methods used for GraphMHC model validation: Intra-validation, inter-validation, clinical application, and extra-validation.

Validation	Models	Dataset
Intra-validation according to the architecture of neural networks	4 GNNs and 2 CNNs (GraphMHC)	IEDB of MHC class I (Train set 125,667 and test set 31,417)
	2 GNNs and 2 CNNs	
	4 GNNs	
Inter-validation with the baseline model	GraphMHC	
	NetMHCpan-4.1	
Clinical applications of neoantigen prediction model	GraphMHC	TCGA-SKCM (310 subjects)
	NetMHCpan-4.1	
Extra-validation	GraphMHC	IEDB of MHC class II (Train set 85,708 and test set 21,427)
	NetMHCIIpan-4.0	

https://doi.org/10.1371/journal.pone.0291223.t002

Table 3. Intra-validation according to model architecture of neoantigen classification and inter-validation using the baseline model.

Metric	4 GNNs and 2 CNNs (GraphMHC)	2 GNNs and 2 CNNs	4 GNNs	NetMHCpan-4.1
AUC-ROC	**0.922** (0.919–0.925)	0.872 (0.869–0.876)	0.688 (0.683–0.693)	0.904 (0.900–0.907)
Sensitivity	**0.884** (0.881–0.888)	0.839 (0.834–0.843)	0.586 (0.580–0.591)	0.834 (0.830–0.838)
Specificity	0.810 (0.805–0.814)	0.746 (0.742–0.751)	0.690 (0.685–0.695)	**0.943** (0.941–0.946)
F_1-score	0.730 (0.725–0.734)	0.657 (0.651–0.662)	0.476 (0.471–0.482)	**0.836** (0.832–0.840)

Abbreviations: GNNs, graph neural networks; CNNs, convolutional neural networks; AUC-ROC, area under receiver operating characteristic curve. Values in parentheses indicate 95% confidence interval. Boldface typefaces represent the highest values among methods.

https://doi.org/10.1371/journal.pone.0291223.t003

Inter-validation: Graph-based model shows better sensitivity than string-based model

The results of comparing GraphMHC with the baseline model, NetMHCpan-4.1, in the right-most column of the Table 3 indicate that GraphMHC perform well in terms of AUC-ROC and sensitivity and badly for specificity and F_1-score. These results indicate a high rate for true positives and a low rate for false negatives, suggesting that it can be used as a meaningful indicator as a predictive biomarker for fatal diseases such as cancer.

Clinical applications: Groups divided by the graph-based model are clinically discriminated

The most common method of using the median as the threshold value for dividing groups was inherited in this study [62–64]. Examination of the 5-year survival in Fig 3A indicates that the differences between groups divided using GraphMHC are borderline (p=0.061). On the other hand, the results obtained using NetMHCpan-4.1 were not significant in any of the observation periods, aligning with the results of previous studies in which no significance has been reported (p=0.567) for 10-year survival under this method [64]. Comparison of the biomarker scores in two groups in Fig 3B show that significant differences were obtained for stromal scores when using GraphMHC, but not for other cases.

Extra-validation: GraphMHC model shows the best performance for most metrics in MHC class II data

The following are the evaluation results of the test set of 21 427 samples obtained from the MHC class II datasets in the IEDB. Comparison of GraphMHC and the corresponding baseline model NetMHCIIpan-4.0, applied to MHC class II data in Table 4, confirms GraphMHC showed higher performance in terms of AUC-ROC and sensitivity, as indicated by the results of inter-validation. Other scores showed similar or slightly lower values. The F_1-score was the same as the value derived using the baseline model, and the specificity obtained was low.

Discussion

Implications of the neoantigen prediction study using the GNN

This study is the first to predict binding by modeling the biochemical molecular structure using information that describes MHC protein sequences and peptide sequences as candidate neoantigens. It is noteworthy that feature extraction and binding modeling are possible only when the inherent structural information from the sequence data is used and no additional external information is included. In other words, no separate interaction mechanism is required; the structure and connection information between nodes and edges in the graph

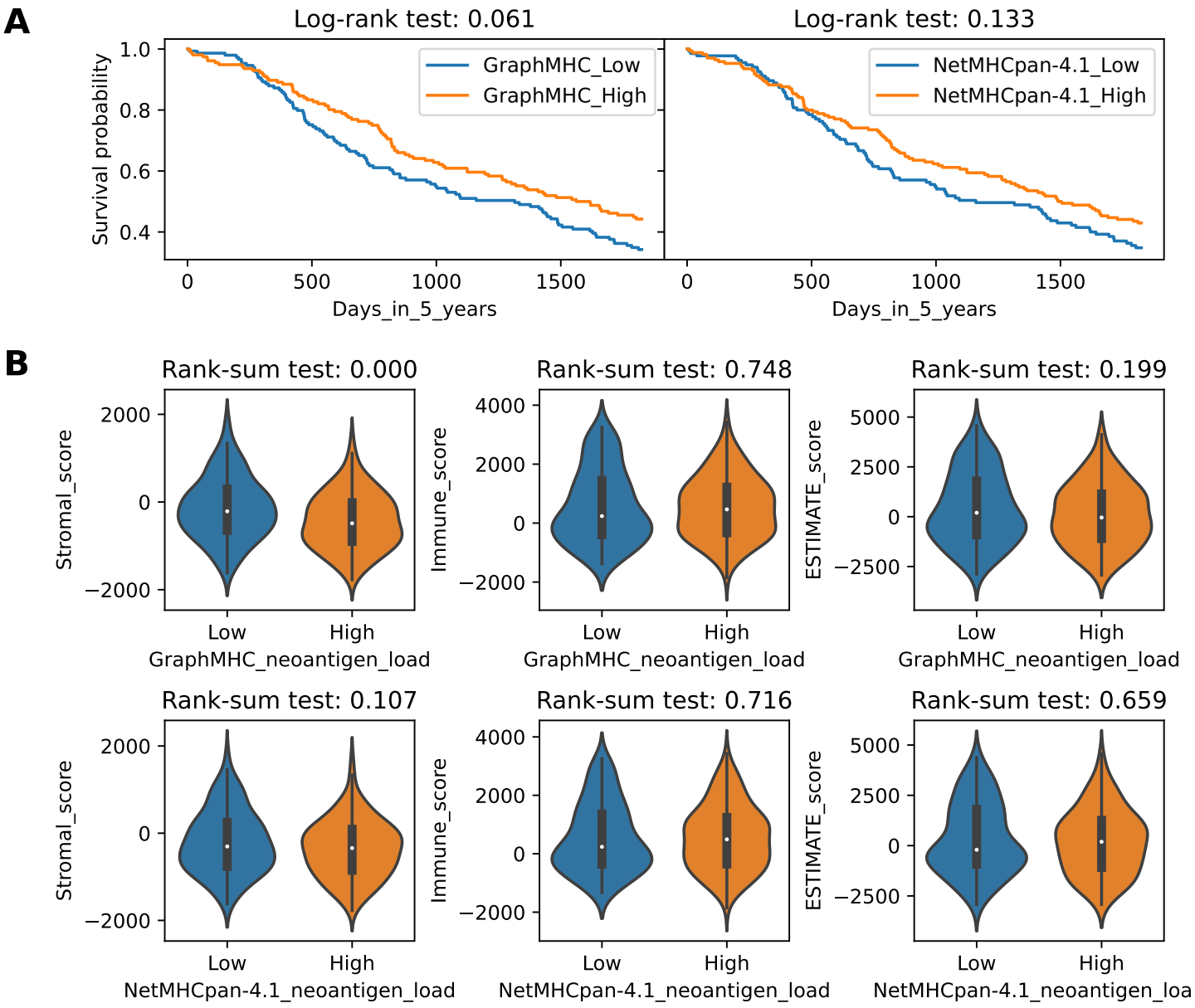

Fig 3. Comparison of high and low neoantigen load groups using TCGA-SKCM data. A Comparison of overall survival in high and low groups according to the median number of loaded neoantigens. The left is the result of grouping by GraphMHC while the right is obtained using NetMHCpan-4.1. Examination of the 5-year survival indicates that the differences between groups divided using GraphMHC are borderline. **B** Comparison of biomarker scores in high and low groups according to the median number of loaded neoantigens. The upper row is the result of grouping by GraphMHC while the lower row was obtained using NetMHCpan-4.1. Comparison of the stromal scores in two groups show that significant differences were obtained when using GraphMHC.

https://doi.org/10.1371/journal.pone.0291223.g003

structure itself are aggregated and extracted through layered automatic feature extraction, and the minute attractive and repulsive forces that occur between complex and diverse atoms are simulated when using this method. As a similar research case, a protein folding model using GNN with reduced parameters from AlphaFold 2 [65] has been reported to simulate the interaction between MHC and peptides [66], indicating the possibility of using GNN for the structural prediction of MHC-peptide binding.

Table 4. Extra-validation applied to MHC class II data.

Metric	GraphMHC	NetMHCIIpan-4.0
AUC-ROC	**0.874** (0.869–0.878)	0.834 (0.829–0.839)
Sensitivity	**0.813** (0.808–0.818)	0.788 (0.782–0.793)
Specificity	0.770 (0.764–0.776)	**0.798** (0.792–0.803)
F_1-score	**0.747** (0.741–0.753)	**0.747** (0.741–0.753)

Abbreviations: AUC-ROC, area under receiver operating characteristic curve. Values in parentheses indicate 95% confidence interval. Boldface typefaces represent the highest values obtained using different methods.

https://doi.org/10.1371/journal.pone.0291223.t004

In addition, the differences in the research results obtained in previous studies are not limited to the performance of the model. Application of the proposed model, GraphMHC, to clinical data suggested that it has better discrimination compared to those observed previously. These results suggest that the GraphMHC model can be used as a biomarker for the cancer immune response.

Potential application suggested from research results

This research model can be used in biological experiments or for medical prediction or prevention. One case in which binding affinity was evaluated using NetMHCpan in *in vitro* immunoprecipitation and liquid chromatography-tandem mass spectrometry experiments (LC-MS/MS) [67] and one in which the $CD8^+$ epitope was predicted and validated using NetMHC *in vivo* can be cited as close examples [68]. In terms of screening, it is possible that accurate information about responsibility can be obtained via sequencing for patients who are to receive immune checkpoint inhibitors, enabling patient-specific treatment. This model can also be used as a biomarker to divide groups and compare between two groups to predict response to cancer immunotherapy [15, 69]. Even more noteworthy is its possible use in preventing cancer by using it in the development of peptide vaccines that can activate T cells [70–72]. On the other hand, the results of this study could also be used to predict similar amino acid bindings such as the SARS-CoV-2 (COVID-19) antigen [16, 73]. Furthermore, the application of this model could be considered not only for amino acids, but also for modeling ligand or drug binding that can be represented by SMILES. From these various perspectives, this study can be considered a pioneering breakthrough in precision medicine research.

Limitations of the study

Despite the original suggestions and potential applications, this study also has some limitations. The information in the IEDB concerning the binding affinity between MHCs and peptides, including HLA-type polymorphisms is incomplete, although the experimental data are steadily accumulating. Inevitable uncertainties are also associated with the conversion and transformation of data via several different methods. In this regard, attempts to call variants more accurately using deep learning have been reported [74]. In addition, there is no guarantee that the conversion of the 34-mer amino acid sequence corresponding to the polymorphic residues in the HLA types referenced to by NetMHCpan is absolute. In terms of the model itself, the internal structure of the GNN model means that it occupies more memory than the string-based neural network model, indicating that a server-side service would be useful. Beyond the neoantigen load, several other points require consideration when expanding the research scope of immunotherapy. Even in cases where large numbers of neoantigens are loaded, the prognosis is often poor. Therefore, not only the foreignness of a tumor as an

antigenic mutation, but also the tumor sensitivity to mutations that are exported from the inside to the outside of a cancer cell should be considered and different machine learning methods applied [15, 75]. It is also necessary to consider TCR binding [76], for which other machine learning methods are being developed [77–79].

Research topics not included in this study

Since this study aimed to extract intrinsic information from a given amino acid sequence, the research methodology of using the extracted dataset with additional data was not included. This is because the methodologies used in studies reporting in this field are not as yet verifiable because of the small number or the artificial nature of datasets used. For example, one model that used molecular information about amino acids as well as binding affinity, Neopepsee [63], reached high prediction accuracy with an AUC of 0.981 by applying a support vector machine (SVM) model that combined the binding affinities from the IEDB with parameters such as immunogenicity, sequence similarity, and amino acid pairwise contact potentials [80]. The study was limited by the small sample size, with only 311 positive epitopes and 14,633 mutant negative peptides included, and the IMMA2 dataset used to calculate the amino acid pairwise contact potentials in this study was composed of only 558 immunogenic and 527 non-immunogenic peptide values [78], Another example, NetMHCpan-4.0, is a multilayer perceptron model that outputs binding affinities and eluted ligands from mass spectrometry, reached approximately 0.98 [11]. This study has a total of 85,217 entries, but the results are limited because the negative entries were artificially generated. Other models derived from it, such as MHCflurry [81] and DeepHLApan [82], have similar methods.

Further studies on precision oncology

In this study, only the genomic approach was presented using the graph neural network for cancer treatment, but a metabolic network approach and a gene regulatory network approach are also possible, and several such attempts have already been reported. These directions are outlined for further research on cancer-related networks.

In current precision oncology studies, the interaction effects within the tumor microenvironment, such as stromal cells and immune cells surrounding the tumor cells, are also being considered. In other words, in tumor-centric research, interaction between players will be treated as a time-dependent study in the future. While normal cells become proliferative through the cell cycle when there is a mitogenic growth signal, cancer cells overcome the antitumor defense by receiving oxygen from the blood vessel supply. This represents an important characteristic of early and midstage [83]. Tumor cells receive and symbiotically utilize glucose and lactic acid, so suppressing this supply can lead to the death of tumor cells [84]. If normalization of angiogenesis accompanies this, the effect of immunotherapy can be expected to increase [85, 86]. A metabolic network-based approach is useful for understanding the tumor microenvironment, and in a related study, a graph neural network was used to estimate the flux between cells [87].

After the study of induced pluripotent stem cells (iPS), which enabled dedifferentiation by regulating four genes [88, 89], an extended study revealed that two regulators, *BCL11A* and *HDAC1/2*, have been identified in the gene regulatory network for reprogramming cancer cells [88, 89]. The proposed model in this study only addressed the classification problem of graphs, but it can also be applied to link prediction [90] or community detection [91] problems in gene regulatory networks.

Conclusion

The GraphMHC model predicts neoantigens by converting MHC protein and peptide binding to graphic structure using the intrinsic features of the sequences themselves.

The GraphMHC model, which is based on the GNN, showed high accuracy with a low false negative rate for predicting neoantigens as compared to the baseline model. A significant difference was also observed when using the GraphMHC model to divide data into two groups for testing clinical discrimination. The GraphMHC model can thus be considered suitable for predicting the response prognosis of immune checkpoint inhibition.

Supporting information

S1 Table. SMILES representation of MHC and peptides.
(PDF)

S2 Table. Measurement statistics for all MHC-peptide graphs. Based on the binding affinity provided by the IEDB, non-binding is defined as $IC_{50} \leq 500$ nM and binding as $IC_{50} > 500$ nM. Statistics are expressed from the median (the 1st quartile—the 3rd quartile).
(PDF)

S1 Fig. Molecular structures of a MHC protein and a peptide, with upper part corresponding to the peptide and lower part corresponding to the MHC protein.
(EPS)

S2 Fig. Graph structure of a MHC protein and peptide. MHC protein and peptide chains are composed using disconnected graphs.
(EPS)

S3 Fig. Selection of subjects from TCGA-SKCM data.
(EPS)

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions.

Author Contributions

Conceptualization: Hoyeon Jeong.

Data curation: Hoyeon Jeong.

Formal analysis: Hoyeon Jeong.

Investigation: Hoyeon Jeong.

Methodology: Hoyeon Jeong.

Project administration: Hoyeon Jeong, Dae Ryong Kang.

Resources: Hoyeon Jeong.

Software: Hoyeon Jeong.

Supervision: Young-Rae Cho, Jungsoo Gim, Seung-Kuy Cha, Maengsup Kim, Dae Ryong Kang.

Validation: Hoyeon Jeong, Young-Rae Cho, Jungsoo Gim, Seung-Kuy Cha, Maengsup Kim, Dae Ryong Kang.

Visualization: Hoyeon Jeong.

Writing – original draft: Hoyeon Jeong.

Writing – review & editing: Hoyeon Jeong, Young-Rae Cho, Jungsoo Gim, Seung-Kuy Cha, Maengsup Kim, Dae Ryong Kang.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021; 71(3):209–249. PMID: 33538338
2. Jung KW, Won YJ, Hong S, Kong HJ, Lee ES. Prediction of cancer incidence and mortality in Korea, 2020. Cancer Research and Treatment: Official Journal of Korean Cancer Association. 2020; 52 (2):351. https://doi.org/10.4143/crt.2020.203 PMID: 32178488
3. Anand P, Kunnumakara AB, Sundaram C, Harikumar KB, Tharakan ST, Lai OS, et al. Cancer is a preventable disease that requires major lifestyle changes. Pharmaceutical research. 2008; 25(9):2097–2116. https://doi.org/10.1007/s11095-008-9661-9 PMID: 18626751
4. Sharma P, Allison JP. Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential. Cell. 2015; 161(2):205–214. https://doi.org/10.1016/j.cell.2015.03.030 PMID: 25860605
5. Nam J, Son S, Park KS, Zou W, Shea LD, Moon JJ. Cancer nanomedicine for combination cancer immunotherapy. Nature Reviews Materials. 2019; 4(6):398–414. https://doi.org/10.1038/s41578-019-0108-1
6. Ventola CL. Cancer immunotherapy, part 3: challenges and future trends. Pharmacy and Therapeutics. 2017; 42(8):514. PMID: 28781505
7. Jiang T, Shi T, Zhang H, Hu J, Song Y, Wei J, et al. Tumor neoantigens: from basic research to clinical applications. Journal of hematology & oncology. 2019; 12(1):1–13. https://doi.org/10.1186/s13045-019-0787-5 PMID: 31492199
8. Yi M, Qin S, Zhao W, Yu S, Chu Q, Wu K. The role of neoantigen in immune checkpoint blockade therapy. Experimental Hematology & Oncology. 2018; 7(1):1–11. https://doi.org/10.1186/s40164-018-0120-y PMID: 30473928
9. Sette A, Sidney J, del Guercio MF, Southwood S, Ruppert J, Dahlberg C, et al. Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. Molecular immunology. 1994; 31(11):813–822. https://doi.org/10.1016/0161-5890(94)90019-1 PMID: 8047072
10. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic acids research. 2015; 43(D1):D405–D412. https://doi.org/10.1093/nar/gku938 PMID: 25300482
11. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. The Journal of Immunology. 2017; 199(9):3360–3368. https://doi.org/10.4049/jimmunol.1700893 PMID: 28978689
12. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic acids research. 2020; 48(W1):W449–W454. https://doi.org/10.1093/nar/gkaa379 PMID: 32406916
13. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and-B locus protein of known sequence. PloS one. 2007; 2(8):e796. https://doi.org/10.1371/journal.pone.0000796 PMID: 17726526
14. Venkatesh G, Grover A, Srinivasaraghavan G, Rao S. MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model. Bioinformatics. 2020; 36 (Supplement_1):i399–i406. https://doi.org/10.1093/bioinformatics/btaa479 PMID: 32657386
15. Kim K, Kim HS, Kim JY, Jung H, Sun JM, Ahn JS, et al. Predicting clinical benefit of immunotherapy by antigenic or functional mutations affecting tumour immunogenicity. Nature communications. 2020; 11 (1):1–11.
16. Li G, Iyer B, Prasath VS, Ni Y, Salomonis N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. Briefings in bioinformatics. 2021; 22(6): bbab160. https://doi.org/10.1093/bib/bbab160 PMID: 34009266

17. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE transactions on neural networks. 2008; 20(1):61–80. https://doi.org/10.1109/TNN.2008.2005605 PMID: 19068426

18. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907. 2016;.

19. Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. stat. 2017; 1050:20.

20. Yi HC, You ZH, Huang DS, Kwoh CK. Graph representation learning in bioinformatics: trends, methods and applications. Briefings in Bioinformatics. 2022; 23(1):bbab340. https://doi.org/10.1093/bib/bbab340 PMID: 34471921

21. Ju W, Liu Z, Qin Y, Feng B, Wang C, Guo Z, et al. Few-shot molecular property prediction via Hierarchically Structured Learning on Relation Graphs. Neural Networks. 2023; 163:122–131. https://doi.org/10.1016/j.neunet.2023.03.034 PMID: 37037059

22. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences. 1988; 28(1):31–36. https://doi.org/10.1021/ci00057a005

23. Lin X. Deepgs: Deep representation learning of graphs and sequences for drug-target binding affinity prediction. arXiv preprint arXiv:200313902. 2020;.

24. Yang Z, Zhong W, Zhao L, Chen CYC. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. Chemical science. 2022; 13(3):816–833. https://doi.org/10.1039/d1sc05180f PMID: 35173947

25. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: Predicting drug–target binding affinity with graph neural networks. Bioinformatics. 2021; 37(8):1140–1147. https://doi.org/10.1093/bioinformatics/btaa921 PMID: 33119053

26. Jiang M, Li Z, Zhang S, Wang S, Wang X, Yuan Q, et al. Drug–target affinity prediction using graph neural network and contact maps. RSC Advances. 2020; 10(35):20701–20712. https://doi.org/10.1039/d0ra02297g PMID: 35517730

27. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics. 2018; 34(17):i821–i829. https://doi.org/10.1093/bioinformatics/bty593 PMID: 30423097

28. Nikolaienko T, Gurbych O, Druchok M. Complex machine learning model needs complex testing: Examining predictability of molecular binding affinity by a graph neural network. Journal of Computational Chemistry. 2022;. https://doi.org/10.1002/jcc.26831 PMID: 35201629

29. Joung JF, Han M, Hwang J, Jeong M, Choi DH, Park S. Deep Learning Optical Spectroscopy Based on Experimental Database: Potential Applications to Molecular Design. JACS Au. 2021; 1(4):427–438. https://doi.org/10.1021/jacsau.1c00035 PMID: 34467305

30. Nyamabo AK, Yu H, Shi JY. SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction. Briefings in Bioinformatics. 2021; 22(6):bbab133. https://doi.org/10.1093/bib/bbab133 PMID: 33951725

31. Deac A, Huang YH, Veličković P, Liò P, Tang J. Drug-drug adverse effect prediction with graph co-attention. arXiv preprint arXiv:190500534. 2019;.

32. Łuksza M, Riaz N, Makarov V, Balachandran VP, Hellmann MD, Solovyov A, et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. Nature. 2017; 551(7681):517–520. https://doi.org/10.1038/nature24473 PMID: 29132144

33. Vitiello A, Zanetti M. Neoantigen prediction and the need for validation. Nature biotechnology. 2017; 35(9):815–817. https://doi.org/10.1038/nbt.3932 PMID: 28898209

34. Richman LP, Vonderheide RH, Rech AJ. Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade. Cell systems. 2019; 9(4):375–382. https://doi.org/10.1016/j.cels.2019.08.009 PMID: 31606370

35. Kamada T, Kawai S, et al. An algorithm for drawing general undirected graphs. Information processing letters. 1989; 31(1):7–15. https://doi.org/10.1016/0020-0190(89)90102-6

36. Fey M, Lenssen JE. Fast Graph Representation Learning with PyTorch Geometric. In: ICLR Workshop on Representation Learning on Graphs and Manifolds; 2019.

37. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473. 2014;.

38. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint arXiv:171010903. 2017;.

39. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019; 32.

40. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

41. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011; 12:2825–2830.

43. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; p. 837–845. https://doi.org/10.2307/2531595 PMID: 3203132

44. Schoonjans F, Zalata A, Depuydt C, Comhaire F. MedCalc: a new computer program for medical statistics. Computer methods and programs in biomedicine. 1995; 48(3):257–262. https://doi.org/10.1016/0169-2607(95)01703-8 PMID: 8925653

45. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013; 45(10):1113–1120. https://doi.org/10.1038/ng.2764 PMID: 24071849

46. Eggermont AM, Spatz A, Robert C. Cutaneous melanoma. The Lancet. 2014; 383(9919):816–827. https://doi.org/10.1016/S0140-6736(13)60802-8 PMID: 24054424

47. Coelho ACM, Fonseca AL, Martins DL, Lins PB, da Cunha LM, de Souza SJ. neoANT-HILL: an integrated tool for identification of potential neoantigens. BMC Medical Genomics. 2020; 13(1):1–8. https://doi.org/10.1186/s12920-020-0694-1 PMID: 32087727

48. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014; 30(23):3310–3316. https://doi.org/10.1093/bioinformatics/btu548 PMID: 25143287

49. Park S, Won D, Kim DJ, Park SY, Lee ST. Genetic Alterations of Esophageal Squamous Cell Carcinoma in Korean Patients. 2021;.

50. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. Genome biology. 2016; 17(1):1–14. https://doi.org/10.1186/s13059-016-0974-4 PMID: 27268795

51. Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics. 2013; 29(24):3235–3237. https://doi.org/10.1093/bioinformatics/btt543 PMID: 24058055

52. Kim JH. Genome data analysis 2: NGS edition, cancer and disease genome. V. 2. Seoul, Republic of Korea: Panmun Education; 2020.

53. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications. 2013; 4(1):1–11. https://doi.org/10.1038/ncomms3612 PMID: 24113773

54. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. cell. 2011; 144(5):646–674. https://doi.org/10.1016/j.cell.2011.02.013 PMID: 21376230

55. Kalluri R, Zeisberg M. Fibroblasts in cancer. Nature reviews cancer. 2006; 6(5):392–401. https://doi.org/10.1038/nrc1877 PMID: 16572188

56. Galon J, Marincola FM, Thurin M, Trinchieri G, Fox BA, Gajewski TF, et al.. The immune score as a new possible approach for the classification of cancer; 2012.

57. Pyke RM, Thompson WK, Salem RM, Font-Burgada J, Zanetti M, Carter H. Evolutionary pressure against MHC class II binding cancer mutations. Cell. 2018; 175(2):416–428. https://doi.org/10.1016/j.cell.2018.08.048

58. Sun Z, Chen F, Meng F, Wei J, Liu B. MHC class II restricted neoantigen: a promising target in tumor immunotherapy. Cancer letters. 2017; 392:17–25. https://doi.org/10.1016/j.canlet.2016.12.039 PMID: 28104443

59. Shao XM, Bhattacharya R, Huang J, Sivakumar I, Tokheim C, Zheng L, et al. High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggetsHigh-Throughput Prediction of Neoantigens with MHCnuggets. Cancer immunology research. 2020; 8(3):396–408. https://doi.org/10.1158/2326-6066.CIR-19-0464 PMID: 31871119

60. Zhang H, Lund O, Nielsen M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. Bioinformatics. 2009; 25(10):1293–1299. https://doi.org/10.1093/bioinformatics/btp137 PMID: 19297351

61. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2016; 32(4):511–517. https://doi.org/10.1093/bioinformatics/btv639 PMID: 26515819

62. McGranahan N, Furness AJ, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science. 2016; 351 (6280):1463–1469. https://doi.org/10.1126/science.aaf1490 PMID: 26940869

63. Kim S, Kim HS, Kim E, Lee M, Shin EC, Paik S. Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. Annals of Oncology. 2018; 29(4):1030–1036. https://doi.org/10.1093/annonc/mdy022 PMID: 29360924

64. Ghorani E, Rosenthal R, McGranahan N, Reading J, Lynch M, Peggs K, et al. Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma. Annals of oncology. 2018; 29(1):271–279. https://doi.org/10.1093/annonc/mdx687 PMID: 29361136

65. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844

66. Delaunay AP, Fu Y, Begue A, McHardy R, Djermani BA, Rooney M, et al. Peptide-MHC Structure Prediction With Mixed Residue and Atom Graph Neural Network. bioRxiv. 2022;.

67. Khan A, Shin JY, So MK, Na JH, Justesen S, Ansari AA, et al. Characterization of HLA-A* 33: 03 epitopes via immunoprecipitation and LC-MS/MS. Proteomics. 2022; 22(1-2):2100171. https://doi.org/10.1002/pmic.202100171 PMID: 34561969

68. Duarte A, Queiroz AT, Tosta R, Carvalho AM, Barbosa CH, Bellio M, et al. Prediction of CD8+ epitopes in Leishmania braziliensis proteins using EPIBOT: in silico search and in vivo validation. PLoS One. 2015; 10(4):e0124786. https://doi.org/10.1371/journal.pone.0124786 PMID: 25905908

69. Abbott CW, Boyle SM, Pyke RM, McDaniel LD, Levy E, Navarro FC, et al. Prediction of immunotherapy response in melanoma through combined modeling of neoantigen burden and immune-related resistance mechanisms. Clinical Cancer Research. 2021; 27(15):4265–4276. https://doi.org/10.1158/1078-0432.CCR-20-4314 PMID: 34341053

70. Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, Löwer M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature. 2017; 547(7662):222–226. https://doi.org/10.1038/nature23003 PMID: 28678784

71. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature. 2017; 547(7662):217–221. https://doi.org/10.1038/nature22991 PMID: 28678778

72. Zahm CD, Colluru VT, McNeel DG. Vaccination with high-affinity epitopes impairs antitumor efficacy by increasing PD-1 expression on CD8+ T cells. Cancer immunology research. 2017; 5(8):630–641. https://doi.org/10.1158/2326-6066.CIR-16-0374 PMID: 28634215

73. Prachar M, Justesen S, Steen-Jensen DB, Thorgrimsen S, Jurgons E, Winther O, et al. Identification and validation of 174 COVID-19 vaccine candidate epitopes reveals low performance of common epitope prediction tools. Scientific reports. 2020; 10(1):1–8. https://doi.org/10.1038/s41598-020-77466-4 PMID: 33235258

74. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nature biotechnology. 2018; 36(10):983–987. https://doi.org/10.1038/nbt.4235 PMID: 30247488

75. Blank CU, Haanen JB, Ribas A, Schumacher TN. The "cancer immunogram". Science. 2016; 352 (6286):658–660.

76. Fritsch EF, Rajasagi M, Ott PA, Brusic V, Hacohen N, Wu CJ. HLA-Binding Properties of Tumor Neoepitopes in HumansTumor Neoepitopes in Humans. Cancer immunology research. 2014; 2(6):522–529. https://doi.org/10.1158/2326-6066.CIR-13-0227 PMID: 24894089

77. Tung CW, Ho SY. POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. Bioinformatics. 2007; 23(8):942–949. https://doi.org/10.1093/bioinformatics/btm061 PMID: 17384427

78. Tung CW, Ziehm M, Kämper A, Kohlbacher O, Ho SY. POPISK: T-cell reactivity prediction using support vector machines and string kernels. BMC bioinformatics. 2011; 12(1):1–11. https://doi.org/10.1186/1471-2105-12-446 PMID: 22085524

79. Lu T, Zhang Z, Zhu J, Wang Y, Jiang P, Xiao X, et al. Deep learning-based prediction of the T cell receptor–antigen binding specificity. Nature Machine Intelligence. 2021; 3(10):864–875. https://doi.org/10.1038/s42256-021-00383-2 PMID: 36003885

80. Saethang T, Hirose O, Kimkong I, Tran VA, Dang XT, Nguyen LAT, et al. PAAQD: Predicting immunogenicity of MHC class I binding peptides using amino acid pairwise contact potentials and quantum topological molecular similarity descriptors. Journal of Immunological Methods. 2013; 387(1-2):293–302. https://doi.org/10.1016/j.jim.2012.09.016 PMID: 23058674

81. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. Cell systems. 2020; 11(1):42–48. https://doi.org/10.1016/j.cels.2020.06.010 PMID: 32711842

82. Wu J, Wang W, Zhang J, Zhou B, Zhao W, Su Z, et al. DeepHLApan: a deep learning approach for neoantigen prediction considering both HLA-peptide binding and immunogenicity. Frontiers in Immunology. 2019; p. 2559. https://doi.org/10.3389/fimmu.2019.02559 PMID: 31736974

83. Hanahan D, Weinberg RA. The hallmarks of cancer. cell. 2000; 100(1):57–70. https://doi.org/10.1016/S0092-8674(00)81683-9 PMID: 10647931

84. Sonveaux P, Végran F, Schroeder T, Wergin MC, Verrax J, Rabbani ZN, et al. Targeting lactate-fueled respiration selectively kills hypoxic tumor cells in mice. The Journal of clinical investigation. 2008; 118 (12):3930–3942. https://doi.org/10.1172/JCI36843 PMID: 19033663

85. Tian L, Goldstein A, Wang H, Ching Lo H, Sun Kim I, Welte T, et al. Mutual regulation of tumour vessel normalization and immunostimulatory reprogramming. Nature. 2017; 544(7649):250–254. https://doi.org/10.1038/nature21724 PMID: 28371798

86. Zheng X, Fang Z, Liu X, Deng S, Zhou P, Wang X, et al. Increased vessel perfusion predicts the efficacy of immune checkpoint blockade. The Journal of clinical investigation. 2018; 128(5):2104–2115. https://doi.org/10.1172/JCI96582 PMID: 29664018

87. Alghamdi N, Chang W, Dang P, Lu X, Wan C, Gampala S, et al. A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data. Genome research. 2021; 31(10):1867–1884. https://doi.org/10.1101/gr.271205.120 PMID: 34301623

88. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. cell. 2006; 126(4):663–676. https://doi.org/10.1016/j.cell.2006.07.024 PMID: 16904174

89. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. cell. 2007; 131(5):861–872. https://doi.org/10.1016/j.cell.2007.11.019 PMID: 18035408

90. Wang J, Ma A, Ma Q, Xu D, Joshi T. Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks. Computational and Structural Biotechnology Journal. 2020; 18:3335–3343. https://doi.org/10.1016/j.csbj.2020.10.022 PMID: 33294129

91. Sattar NS, Arifuzzaman S. Community detection using semi-supervised learning with graph convolutional network on GPUs. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE; 2020. p. 5237–5246.